NOTICE: THIS BOOK IS INTENDED TO INCLUDE THE MATHEMATICAL FOUNDATIONS & A COLLECTION OF EXERCISES NEEDED IN ECON 221 & ECON 222. THE TEXT & SOLUTIONS TO EXERCISES ARE STILL IN PROGRESS & IT IS POSSIBLE THAT EVERY TIME YOU VISIT THE 'TEACHING' PAGE, YOU'LL SEE A HOPEFULLY IMPROVED VERSION OF THE BOOK.

THIS VERSION WAS COMPILED ON **Wednesday 20th May, 2026** AT **14:50**.

SERKAN KARADEMIR & MUSTAFA ERAY YÜCEL

ECON 221 PROBABILITY AND STATISTICAL DISTRIBUTIONS &ECON 222 STATISTICAL ESTIMATION AND INFERENCE LECTURE NOTES

IHSAN DOGRAMACI BILKENT UNIVERSITY

lecture notes

ihsan dogramaci bilkent university

pdf: https://sites.google.com/view/erayyucel/probability-and-statistics

html: https://eray.bilkent.edu.tr/

this collection of notes is to assist the students in their study of econ 221 and econ 222. it is not meant to be a perfect substitute of in-class lectures. several issues of course conduct are described and explained in this document for better and accident-free communication, where it is the student's responsibility to "be informed" by thoroughly examining the content provided.

thanks are due to muzaffer akat for our common work at özyeğin university and tarik kara for our common work at ihsan dogramaci bilkent university, part of which is included here; arda kayra kolak and burcu yildiz for their assistance with the compilation of earlier in-class notes.

Compilation date and time: Wednesday 20th May, 2026at 14:50.

Contents

Essence of These Lecture Notes

This set of lecture notes was prepared to guide the students through ECON 221 & ECON 222 at Ihsan Dogramaci Bilkent University; yet, sticking to these notes alone may not result in the best possible outcomes in the absence of regular class attendance.

A variety of numerical or conceptual exercises are given in these lecture notes. To the extent possible, they appear after the topics they belong to. For more, Newbold, Carlson & Thorne, *Statistics for Business & Economics*, 8th edition, Pearson; McClave, Benson & Sincich, *Statistics for Business & Economics*, 13th edition, Pearson; Spiegel & Stephens, *Statistics*, 4th edition, Schaum's Outlines; Hill, Griffiths & Lim, *Principles of Econometrics*, 4th edition, John Wiley & Sons; Gujarati, *Basic Econometrics*, 4th edition, McGraw-Hill; and Brooks, *Introductory Econometrics for Finance*, 1st edition, Cambridge University Press may be useful.

Lecture hours are intended to deliver the theoretical knowledge as well as problem-solving skills. So, the subject matter is to be learned in the lectures; recitation hours are for practicing knowledge of probability theory and statistics, often on a computer.

Letter grades are assigned by the end of the semester considering the difficulty level of assessments, the statistical distribution of the class performance, and the particular student's individual performance in comparison to the first two. Typically, the average-performing student receives a grade around C. There is no compulsory or bonus-bearing attendance for any course activity other than the final exam. Failing to take the final exam returns an FX.

A good way to keep track of things is to utilize the instructor's office hours. The office hours are intended for regularly studying and lecture-attending students.

1 Describing data

As discussed in the opening lectures, an analytical and evidence-based approach to policymaking is a must for the modern & complex societies. As the famous engineer Edwards Deming put it, "Without data, you're just another person with an opinion." Data and its analysis are what distinguish well-designed policies from arbitrary and/or flawed ones. So, understanding the features & structure of data is an indispensable step of analysis.

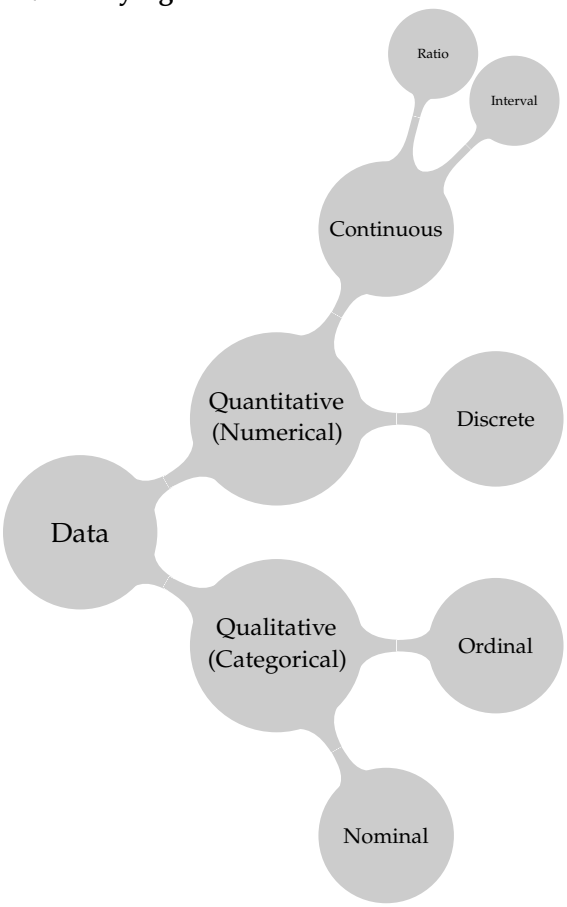
1.1 A taxonomy of data types

Practically every sort of empirical analysis begins with a need to describe a data set & its various elements. So, we begin our journey to learn probability theory & statistics from this simple yet crucial task. Recall from the class discussion that data may come in two main forms: qualitative and quantitative. While qualitative data qualifies 'things', quantitative data quantifies 'things', as the terms suggest. In that, qualitative data often have a categorical nature. If the values of a categorical variable are orderable (sortable) then this categrorical variable is called an 'ordinal' categorical variable. Otherwise, it is a 'nominal' categorical variable. While the responses in a satisfaction survey are ordinal (consider 1: least liked to 5: most liked), indicators of gender are nominal (F: female and M: male). Note that it is not always trivial to come up with a judgment: while we can treat age categories of 'young, middle-aged and old' for people or class/year categories of 'freshman, sophomore, junior and senior' for students as 'ordinal', another researcher may choose to treat them as nominal. All we need to reveal is our capability to sort a categorical variable with a clear understanding stripped of value-judgments. For instance, one cannot simply put one of the genders on top of others, regardless of the underlying way of thinking.

Quantitative data is by definition numerical. It can be either discrete 'as in the case of number of automobiles owned by households' (one cannot own fractional automobiles) or continuous 'as in the case of daily spending by households'. Household size, i.e., the number of

people forming the household is discrete, number of cities in a country is discrete, etc. The case of people's ages measured in years can be a little confusing: think about it.

Classifying data



An important point regarding the continuous data is the distinction between 'interval data' and 'ratio data'. A simple rule of thumb is: if there is an 'absolute zero' of the possible values of a data series, it is 'ratio data' & in the absence of an 'absolute zero' it is named 'interval data'. A trivial example of this is the temperature measurements using the Kelvin (K) versus Celsius scale ($^{\circ}C$). While the Kelvin scale has an absolute zero, i.e. $0K$, the Celsius scale does not. Freezing point of pure water (under certain conditions) is $0^{\circ}C$, yet this is not the lowest attainable temperature. Indeed, there are some $273.15^{\circ}C$ more to go down until that point & $-273.15^{\circ}C$ is defined as $0K$ and it is the lowest possible temperature in the Universe. While $200K$ is two times $100K$, $200^{\circ}C$ is not two times $100^{\circ}C$.

An easier example to understand the ratio data is the measurement of 'mass' (in kilograms, let's say). Mass has an absolute zero, which is '0 Kg' & a 20 Kg object is two times as heavy as a 10 Kg object (assuming

there is gravity).

1.2 *What is a "data set"?*

- A data set is composed of one or more data items (series, variables) for use in analysis (in our case statistical analysis)

- Each individual sub-item in a series is called a data value

- There is often a clear correspondence between the data values of different data items involved, controlled by a primary key (observation number, date or a combination of both)

It is possible and often necessary to convert one data series 'from numerical to categorical'. For instance, age data measured in 'years lived' can be expressed in terms of the qualifiers 'young, middle-aged & old'. Note that, this transformation results in some loss of information. Clearly, a numerical age series tell more about the people surveyed compared to simple categorization. Still, when properly made, a good categorization of numerical values may prove very useful in statistical (or in econometric) analysis.

Search & explore

Conversion 'from categorical to numerical' may not be so straightforward: come up with your cases/examples.

1.3 *Frequency*

In the Oxford English Dictionary, 'frequency' is defined as "the rate at which something occurs over a particular period of time or in a given sample". Our understanding covers the cases of 'being' in addition to 'occurring' or happening: Frequency is the numerical measure of 'how often something happens or how often some specific way of being is observed'. In that, as we can count car accidents in a certain hour, we can also count the people that survived a certain accident. So, we can count 'things' in time (we can call this temporal counting) and in space (we can call this spatial counting).

In a nutshell

In our learning and practice of the Probability theory and Statistics we will be 'counting the things', simply using our fingertips at the beginning, and using more sophisticated techniques then.

1.3.1 Frequency distribution

A frequency distribution is a tabular summary of how numerical values are distributed to classes in a data series.

First, determine the number of classes k, according to:

Number of observations	k
<50	$5-7$
$50-100$	$7-8$
$101-500$	$8-10$
$501-1,000$	$10-11$
$1,001-5,000$	$11-14$
>5,000	$14-20$

The table gives a rule of thumb and requires often your professional attention.

In a nutshell

A 'rule of thumb' is a broadly accurate guide or principle, based on practice rather than theory.

Second, determine the class width, w:

$$w = \frac{Maximum - Minimum}{k}$$

where *Maximum* is the 'largest observation' and *Minimum* is the 'smallest observation'. Always round the formula result up, to find w.

Third, construct the k classes; they are to be inclusive and non-overlapping.

Fourth, allocate your observations to classes and get the count of each class.

At the end, present your result as a table. What is obtained is a "frequency distribution table".

Consider the age data of 20 people (20 subjects) measured in years:

12	11	19	20	20
15	15	24	15	12
20	18	17	20	20
20	22	24	12	14

While summarizing the age data, it seems appropriate to use 5 classes following the rule of thumb given before. The *max* of our data series is 24 and the *min* is 11. Class width w, then, is calculated as:

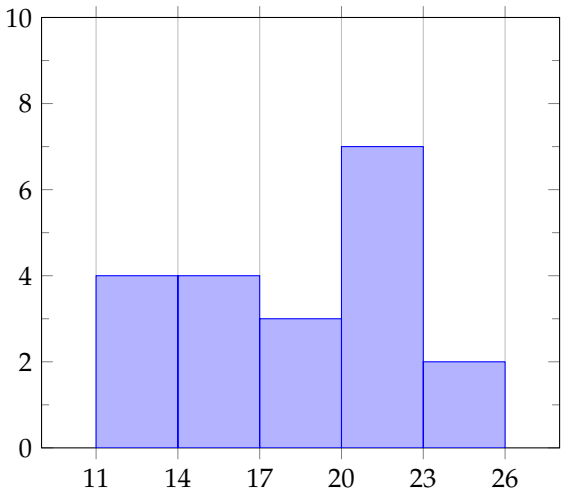
$$w = \frac{24 - 11}{5} = 2.6, \text{ always rounding up} \hat{\lceil} w = 3$$

Having calculated the class width, beginning from the *min* value (11 here) we establish our classes as $[11, 14)$, $[14, 17)$, $[17, 20)$, $[20, 23)$ and $[23, 26]$. Pay attention to openness and closedness of classes (intervals) on the left and on the right.

Once the classes are ready, we carefully count the data values falling into each interval and prepare the following table, a table that we call the 'frequency table'.

Class	*Frequency*
$[11, 14)$	4
$[14, 17)$	4
$[17, 20)$	3
$[20, 23)$	7
$[23, 26]$	2

The final step is to prepare (draw) the histogram of our data.

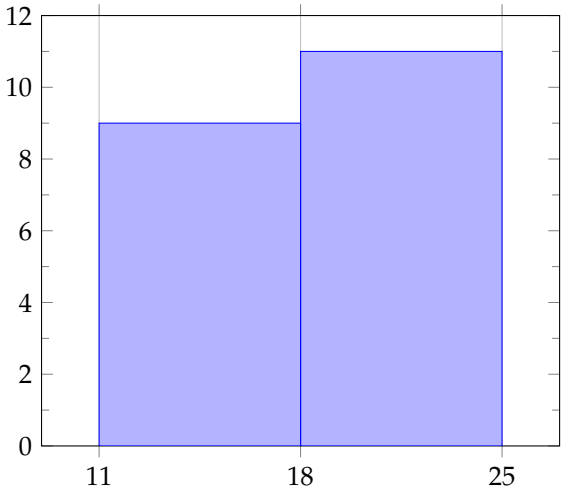


Consider another researcher who prefers arbitrarily to use 2 classes. In this case, the class width (w) will be:

$$w = \frac{24 - 11}{2} = 6.5, \text{ always rounding up} \hat{\lceil} w = 7$$

The classes will be $[11, 18)$ and $[18, 25)$, so our resulting frequency table will look like:

Class	*Frequency*
$[11, 18)$	9
$[18, 25]$	11



The final step is to prepare (draw) the histogram of our data again.

Which histogram (or frequency table) gives a better summary of the data? Avoid any confusions: the first histogram is the winner of the contest. It summarizes our data and conveys a tangible message. The second histogram, on the other hand, suffers from 'oversummarizing'. Here take our discussion to its limits and consider a third researcher who prefers to use 1 class only. Why would that be nonsense?

In a nutshell

In order to summarize numerical (quantitative) data we use 'frequency tables' and 'histograms'. The columns belonging to consecutive nonempty classes must touch each other while drawing a histogram

What about categorical (qualitative) data? Consider the following data series which consists category markings for 20 people (20 subjects), where Y, M and O stand for 'young', 'middle-aged' and 'old', respectively.

Y	Y	O	M	Y
O	Y	O	O	Y
M	M	Y	M	O
O	M	Y	Y	M

This time, forming a frequency table must be easier: we do not (indeed, we cannot) establish classes & simply count the frequency of each category:

Category	*Frequency*
Y	8
M	6
O	6

The final step is to prepare (draw) the bar chart of our data. It is more than trivial; do it yourself.

In a nutshell

In order to summarize categorical (qualitative) data we use 'frequency tables' and 'bar charts'. The bars will never touch each other while drawing a bar chart.

1.3.2 Relative frequency distribution

Once the counts in a frequency distribution table are divided by the total number of observations & expressed as "percentages" or as 'fractions between 0 and 1', the resulting table is called a "relative frequency distribution table". By construction, relative frequencies of all classes add up to 100% or 1.

1.3.3 Cumulative frequency distribution

Once the frequencies (counts) in a frequency distribution table are accumulated across classes, one row at a time and from the smallest to largest class, the resulting table is called a 'cumulative frequency distribution table'.

1.3.4 Relative cumulative frequency distribution

Once the relative frequencies in a relative frequency distribution table are accumulated across classes, one row at a time and from the smallest to largest class, the resulting table is called a 'relative cumulative frequency distribution table'.

In order to see the linkages between 'frequency', 'relative frequency', 'cumulative frequency' and 'cumulative relative frequency', examine the following table:

Class	Frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
$[10, 17)$	500	0.333	500	0.333
$[17, 24)$	250	0.167	750	0.500
$[24, 31)$	150	0.100	900	0.600
$[31, 38]$	600	0.400	**1500**	**1.000**
Total	**1500**	**1.000**	N.A.	N.A.

1.4 Representation of distributions

1.4.1 Histogram and relative frequency polygon

A histogram is a graph that consists of vertical bars constructed on a horizontal line on which intervals are marked for the variable being displayed.

- Horizontal intervals are the classes of a frequency or relative frequency distribution table
- Height of each bar is the frequency or relative frequency associated
  - Warning: not the cumulative figures
  - Warning: consecutive bars for non-empty classes are to touch each other, *i.e.*, no gaps

Histograms are traditionally used for continuous numerical data. When the midpoints of the top segment of each bar in a histogram are connected with line segments, what we obtain is called a frequency polygon. Note that a 'bar chart' resembles a histogram yet it differs in two main aspects: first, it is for categorical data & second, the bars in a bar chart are separated by a visible gap. Examples are provided in the upcoming exercises.

1.4.2 Symmetry and skewness

The shape of a distribution is said to be symmetric if the observations are balanced, or approximately evenly distributed, about its center. A distribution is skewed, or asymmetric, if the observations are not symmetrically distributed on either side of the center. A skewed-right distribution (sometimes called positively skewed) has a tail that extends farther to the right. A skewed-left distribution (sometimes called negatively skewed) has a tail that extends farther to the left.

1.4.3 *O-give*

An O-give, also called a cumulative line graph, is a line that connects points that are the cumulative percent of observations below the upper limit of each class (interval) in a cumulative frequency distribution. Even when not said so, an O-give is to present cumulative percentage figures. Beginning vertical value in an O-give is always 0 and ending vertical value is 1. Examples are provided in the upcoming exercises.

1.1 EXERCISES

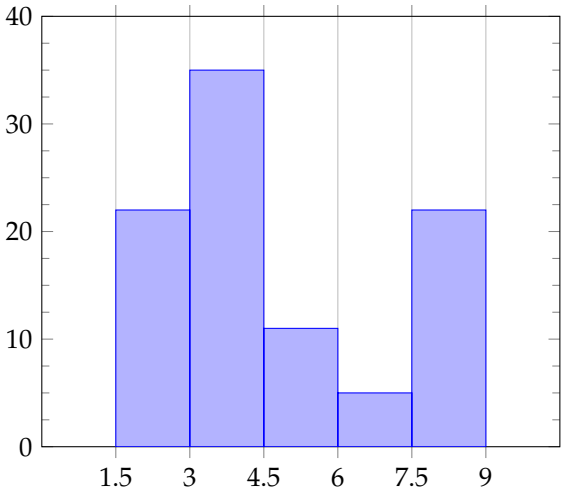
1. Fill the empty cells in following table:

Interval	Feq.	Rel. freq.(%)	Cum. freq.	Rel. cum. freq.(%)
$[0, 20]$	20	10		
$(20, 40]$			80	
$(40, 60]$	30			
$(60, 80]$				
$(80, 100]$		20		80
$(100, 120]$				

Solution: The complete table is as follows:

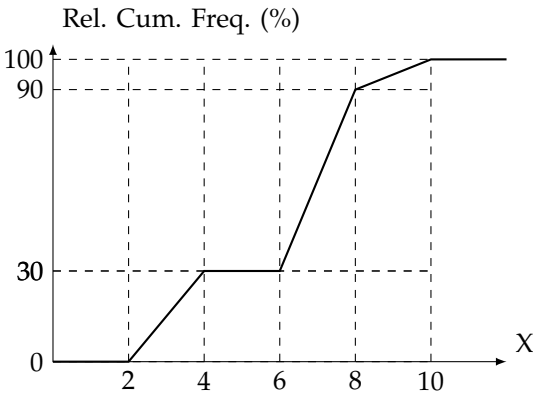
Interval	Freq.	Rel. freq.(%)	Cum. freq.	Rel. cum. freq.(%)
$[0, 20]$	20	10	20	10
$(20, 40]$	60	30	80	40
$(40, 60]$	30	15	110	55
$(60, 80]$	10	5	120	60
$(80, 100]$	40	20	160	80
$(100, 120]$	40	20	200	100

2. Consider the frequency histogram displayed below:



Draw the corresponding (relative frequency) o-give.
Solution: Prepare your Cartesian plane. The origin is $(0,0)$. Mark the following points on your graph space by paying attention to proportions: $(0,0)$, $(1.5,0.22)$, $(3.0,0.57)$, $(4.5,0.68)$, $(6.0,0.73)$, $(7.5,0.95)$, $(9.0,4.00)$. Then, connect these points with line segments from left to right. Once it is done, you will observe a properly drawn O-give. Make sure you have named the axes.

3. Consider the relative frequency o-give displayed below:



i. Draw the corresponding histogram.

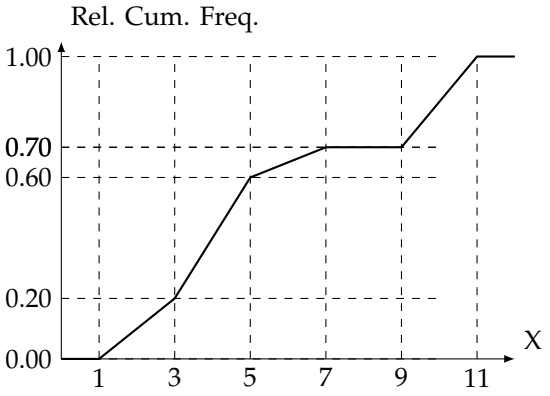
ii. What can you say about the percentage of observations that takes a value less than or equal to 6.5 (if you need to estimate it what would be a reasonable estimate)?

iii. What can you say about the percentage of observations that takes a value greater than or equal to 4.3 (if you need to estimate it what would be a reasonable estimate)?

Solution:

1. Consider the classes of $[0,2]$, $(2,4]$, $(4,6]$, $(6,8]$ and $(8,10]$. Taking simple differences, reveal that relative frequencies of these classes are 0, 0.3, 0, 0.6 and 0.1. Locate these numbers on cartesian plane to obtain the histogram. Recall that for two successive classes which are non-empty, the bars must touch each other.
2. Relative frequency of $(6,8]$ is 0.6. $(6.5-6)/(8-6) = 0.25$. So, approximately 0.25×0.6, i.e., 0.15 is the frequency (relative) of $(6,6.5]$. Since the relative frequency of $[0,6]$ is 0.3, the frequency of $[0,6.5]$ becomes 0.45. If the given o -give has been drawn properly, then we expect our estimate to be reliable.
3. Use the same approach. The solution should yield 0.70.

4. Consider the relative frequency o-give displayed below:



What can you say about the percentage of observations that takes a value between 4.5 and 9.5?
Solution: Use the same approach. $0.4/4 + 0.1 + 0 + 0.3/4$ yields 0.275.

5. A doctor's office staff studied the waiting times for patients who arrive at the office with a request for emergency service. The following data with waiting times in minutes were collected over a one-month period:

2, 5, 10, 12, 4, 4, 5, 17, 11, 8, 9, 8, 12, 21, 6, 8, 7, 13, 18, 117

i. Construct a histogram for this data set by including all observations.

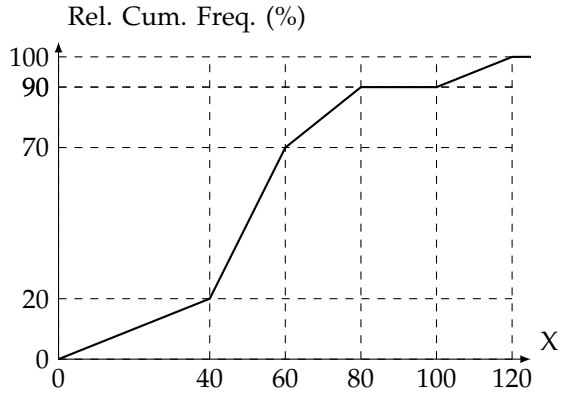
ii. Construct a histogram for this data set after excluding the value of 117. Note that you still have to show it on your histogram (but how?)

iii. Which histogram is more informative? Why?

Solution:

1. For ease in processing data sort/order the values as 2, 4, 4, 5, 5, 6, 7, 8, 8, 8, 9, 10, 11, 12, 12, 13, 17, 18, 21 and 117. The minimum is 2, the maximum is 117 and N is 20. 5 classes would work well here. Then, the class width becomes $(117-2)/5 = 23$, rounding up always, 24. This means that our classes will be $[2, 26]$, $(26, 50]$, $(50, 74]$, $(74, 98]$ and $(98, 122]$. When drawn (do it), this will turn out to be a funny histogram, as 19 observations will fall into the first class and only one observation (117) will fall into the last one.
2. When 117 is kept aside, the maximum becomes 21. Using 5 classes again, the class width becomes $(21-2)/5$, which is 3.8, rounding up always, 4. Our classes will be $[2, 6]$, $(6, 10]$, $(10, 14]$, $(14, 18]$ and $(18, 22]$. Respective frequencyies of these will be 6, 6, 4, 2 and 1. When drawn we will observe a neatly drawn histogram. (What about 117?)
3. The second histogram is more informative. It gives us more details, like the shape of the distribution.

6. Consider the relative frequency o-give displayed below:



Based on the information above estimate the median and the 3rd quartile (Q_3).
Solution: Hint: For the median, find the horizontal value at which the O-give has the value of 50. For Q_3, find the horizontal value at which the O-give has the value of 75.

7. In a data set, the frequency of the interval $(0, 10]$ is 0.10, frequency of the interval $(10, 20]$ is 0.20, frequency of the interval $(20, 30]$ is 0.30 and frequency of the interval $(30, 40]$ is 0.40. Construct the relative frequency O-give and calculate Q_3 for this data set.
Solution: Solving this must be straightforward now. Do it and discuss with classmates.

1.5 Measures of central tendency

Measures of central tendency or measures of concentration indicate 'where' on the real number line our data series is. The three terms connotate:

- Central tendency
- Concentration
- Location

As you'll see in the upcoming classes, the knowledge of this is critically important to make several statistical assessments.

1.5.1 Measures of central tendency: Mode

The "mode", whenever exists, is the most frequently occurring value in a data series.

- If the series has one mode, it is called "unimodal".
- If the series has two modes, it is called "bimodal".
- If the series has three modes, it is called "trimodal".
- If the series has more than two modes, it may simply be called "multimodal", use of the term "trimodal" is not that widespread in everyday professional use.

Note that, the mode is commonly used with (but not restricted to) categorical data.

Consider,

$$X : 1, 2, 3, 3, 3, 4, 4, 5, 6, 7$$

where $N = 10$. Among the values of X, the most frequent (mostly repeated) value is 3, so we say $Mode = 3$. If X included another 4 like:

$$X : 1, 2, 3, 3, 3, 4, 4, 4, 5, 6, 7$$

where $N = 11$, then we would say the Modes are 3 and 4.

1.5.2 Measures of central tendency: Mean (Arithmetic mean)

For $\{x_i\}_{i=1}^{N}$:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

is called the "population mean".

In addition, for $\{x_i\}_{i=1}^{n}$:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

is called the 'sample mean'

Considering,

$$X : 1, 2, 3, 3, 3, 4, 4, 4, 5, 6, 7$$

where $N = 11$, the mean is calculated as:

$$\begin{aligned}\mu &= \frac{\sum_{i=1}^{N} x_i}{N} \\ &= \frac{1+2+3+3+3+4+4+4+5+6+7}{11} \\ &= 3.81\end{aligned}$$

In another case for X, suppose the last value, i.e., 7, is replaced by 42; let's call this data series as X':

$$X : 1, 2, 3, 3, 3, 4, 4, 4, 5, 6, 42$$

where $N = 11$ again, the mean becomes:

$$\begin{aligned}\mu &= \frac{\sum_{i=1}^{N} x_i}{N} \\ &= \frac{1+2+3+3+3+4+4+4+5+6+42}{11} \\ &= 7\end{aligned}$$

As this example suggests, mean (μ) is sensitive to outliers/extreme values. However, this sensitivity does not imply that μ is a meaningless or a useless measure. On the contrary, it is a fundamental measure with many good statistical properties, as we will see in the upcoming chapters.

In a nutshell

Writing mathematics

Good mathematical writing involves:

- Using a relevant & consistent notation
- Flowing logically well the solution or proof steps
- Including verbal explanations & necessary definitions between steps
- Putting things gravitationally, i.e., top to bottom
- Keeping only the essentials, removing everything redundant, avoiding scratch work to remain

In a nutshell

Working with grouped data

In many situations, we the researchers are provided with a grouped summary of a data set, rather than the full data itself. Grouped data sets mostly come in the form of classes with their corresponding frequencies or relative frequencies, like in a histogram. Given population data of N observations grouped into K classes, with frequencies $f_1, f_2, \ldots, f_K$, if the midpoints of these classes are $m_1, m_2, \ldots, m_K$, then

$$\mu = \frac{\sum_{i=1}^{K} f_i m_i}{N}$$

can be written where

$$\sum_{i=1}^{K} f_i = N$$

Given sample data of n observations grouped into K classes, with frequencies $f_1, f_2, \ldots, f_K$, if the midpoints of these classes are $m_1, m_2, \ldots, m_K$, then

$$\bar{x} = \frac{\sum_{i=1}^{K} f_i m_i}{n}$$

can be written where

$$\sum_{i=1}^{K} f_i = n$$

Consider,

X	Frequency	Relative frequency
$[11, 14)$	4	$4/20$
$[14, 17)$	4	$4/20$
$[17, 20)$	3	$3/20$
$[20, 23)$	7	$7/20$
$[23, 26]$	2	$2/20$

Using the midpoints and frequencies of classes:

$$\begin{aligned}\mu &= \frac{\frac{11+14}{2} \cdot 4 + \frac{14+17}{2} \cdot 4 + \frac{17+20}{2} \cdot 3 + \frac{20+23}{2} \cdot 7 + \frac{23+26}{2} \cdot 2}{4+4+3+7+2} \\ &= 18.35\end{aligned}$$

is obtained.

Equivalently, one may use the relative frequencies of classes to calculate the same:

$$\mu = \frac{11+14}{2}\cdot\frac{4}{20} + \frac{14+17}{2}\cdot\frac{4}{20} + \frac{17+20}{2}\cdot\frac{3}{20} + \frac{20+23}{2}\cdot\frac{7}{20} + \frac{23+26}{2}\cdot\frac{2}{20}$$
$$= 18.35$$

1.5.3 *Measures of central tendency: Percentiles, Deciles and Quarties*

The "p-th" percentile in a data series is the smallest value which is greater than p% of observations. If there are N observations, we find

$$\frac{p}{100}(N+1)th$$

ordered position and read the observation at this position as the p-th percentile.

In a nutshell

Order (sort) the observations in ascending order first. Without ordering data, you'll not get correct results.

The "d-th" decile in a data series is the smallest value which is greater than d tenths of observations. The "q-th" quartile in a data series is the smallest value which is greater than q quarters of observations.

While we imagine we are slicing our ordered data set into 100 while finding percentiles, we slice it into 10 while finding deciles, and we slice it into 4 while finding quartiles.

By definition P_0, Q_0, D_0 are equal to the minimum observation and P_{100}, Q_4, D_{10} are equal to the maximum observation.

0th percentile	$\rightarrow$	0th decile	$\rightarrow$	0th quartile (Q_0)	$\rightarrow$	Minimum
10th percentile	$\rightarrow$	1st decile				
20th percentile	$\rightarrow$	2nd decile				
25th percentile	$\rightarrow$	$\rightarrow$	$\rightarrow$	1st quartile (Q_1)		
30th percentile	$\rightarrow$	3rd decile				
40th percentile	$\rightarrow$	4th decile				
50th percentile	$\rightarrow$	5th decile	$\rightarrow$	2nd quartile (Q_2)	$\rightarrow$	Median
60th percentile	$\rightarrow$	6th decile				
70th percentile	$\rightarrow$	7th decile				
75th percentile	$\rightarrow$	$\rightarrow$	$\rightarrow$	3rd quartile (Q_3)		
80th percentile	$\rightarrow$	8th decile				
90th percentile	$\rightarrow$	9th decile				
100th percentile	$\rightarrow$	10th decile	$\rightarrow$	4th quartile (Q_4)	$\rightarrow$	Maximum

1.5.4 Measures of central tendency: Median

Among the many percentiles, "Median" has a special place. "Median" in a data series is the smallest value which is greater than 50% of observations. It simply divides a data series into two equally-likely halves. Numerically, median is nothing but Q_2, P_{50}, D_5, which are all the same.

Consider now a variable X,

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th
1	1	2	3	5	8	13	21	34	55	89

To find, for instance, the 30th percentile of X we calculate:

$$\begin{aligned}\frac{p}{100}(N+1)th &= \frac{30}{100}(11+1)th \\ &= 3.6th\end{aligned}$$

Then, we find the observation value in the 3.6th position of the ordered data series. As seen here, there may not be such a physical position in data. As an approximation, we take the value of X in the 3rd position and add 0.6 times the difference between the value in 4th position and the value in 3rd position, i.e.,

$$\begin{aligned}P_{30} &= 2 + (3-2)\cdot 0.6 \\ &= 2.6\end{aligned}$$

is our 30th percentile.

To find the 80th percentile of X we calculate:

$$\begin{aligned}\frac{p}{100}(N+1)th &= \frac{80}{100}(11+1)th \\ &= 9.6th\end{aligned}$$

$$\begin{aligned}P_{80} &= 34 + (55-34)\cdot 0.6 \\ &= 46.6\end{aligned}$$

So, 46.6 is our 80th percentile.

To find the Median, i.e., the 50th percentile of X we calculate:

$$\begin{aligned}\frac{p}{100}(N+1)th &= \frac{50}{100}(11+1)th \\ &= 6th\end{aligned}$$

Without further calculations, 8 is our Median.

Consider another data series:

2	3	4	6	7
8	9	9	10	10
11	11	11	11	13
14	18	19	19	**19**
21	21	22	22	23
23	23	24	24	24
24	25	25	25	26
26	26	26	26	49

Solve yourself to see that the Median is the 20.5th value of this data series and it is equal to 20.

Before moving forward, consider finally:

Variable	1*st*	2*nd*	3*rd*	4*th*	**5th**	6*th*	7*th*	8*th*	9*th*	**Mean**
X	1	3	6	10	**15**	21	28	36	1000	**124.4**
X'	1	3	6	10	**15**	21	28	36	45	**18.3**

Did you notice anything?

In a nutshell

Unlike the mean (μ), the Median (Q_2) is not sensitive to outliers/extreme values. Equivalently we say, the Median is robust up to the presence of outliers/extreme values in a data series.

1.5.5 Where is my data? Five-number summary

When we give the five descriptive measures

$$minimum \leq Q_1 \leq median \leq Q_3 \leq maximum$$

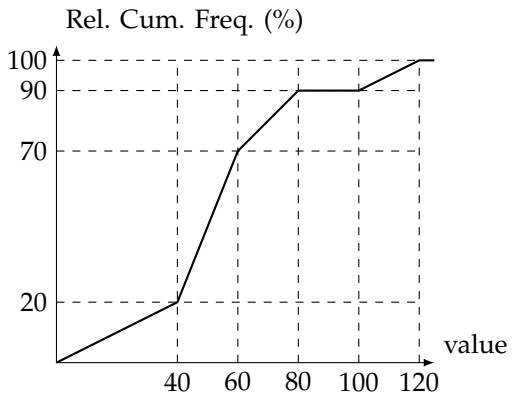
it is called a "five-number summary". This is a somehow ancient and still useful tool to summarize data sets.

For a variable X given as:

2	3	4	6	7
8	9	9	10	10
11	11	11	11	13
14	18	19	19	**19**
21	21	22	22	23
23	23	24	24	24
24	25	25	25	26
26	26	26	26	49

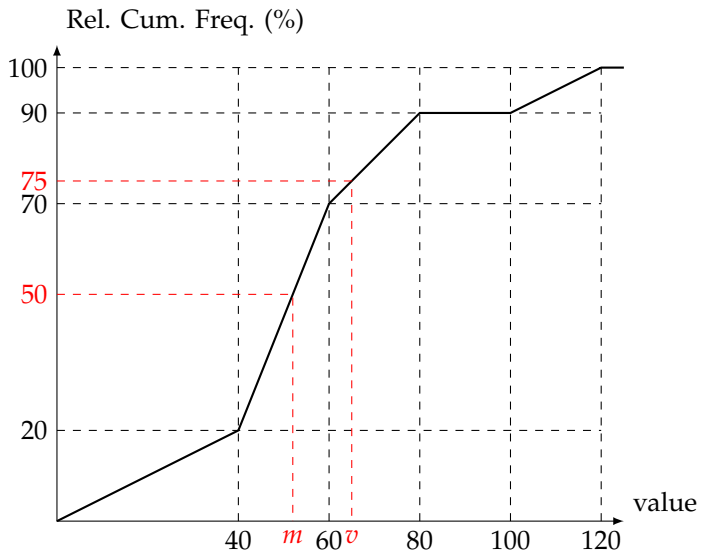
the five-number summary is $(2, 10.25, 20, 24, 49)$.

To make use of our knowledge gained up to this point, consider a data that is summarized with the following relative frequency o-give:



Based on the information above estimate the median, the mean, and the 3rd quartile.

In order to reach a good solution, note that the median is the 50th percentile and the 3rd quartile is the 75th percentile. Under the assumption that data is uniformly distributed over each class interval, a relative cumulative frequency o-give gives information about the percentage of observations that takes a value less than or equal to a given number, we can use the o-give to estimate the median and the 3rd quartile using the o-give. On the graph we mark the points that correspond to 50% and 75%.



From similarity of triangles we have:

$$\frac{m-40}{50-20} = \frac{60-40}{70-20}$$

which yields

$$m = 40 + 30\frac{20}{50} = 52$$

Similarly

$$\frac{v-60}{75-70} = \frac{80-60}{90-70}$$

yields

$$v = 60 + 5\frac{20}{20} = 65$$

We will use CM_l to denote the class mark of the lth class interval (the center of the lth interval). We will use RF_l to denote the relative frequency of the observations that takes values in the lth class interval.

The assumption of data being uniformly distributed over each class interval that the following formula yields a "reasonable" estimate for the mean:

$$RF_1CM_1 + RF_2CM_2 + RF_3CM_3 + RF_4CM_4 + RF_5CM_5$$

Thus our estimate for the mean is

$$\begin{aligned}
&(0.2-0)\frac{0+40}{2} + (0.7-0.2)\frac{40+60}{2} \\
&\quad + (0.9-0.7)\frac{60+80}{2} + (0.9-0.9)\frac{80+100}{2} \\
&\quad + (1.0-0.9)\frac{100+120}{2} \\
&= 54
\end{aligned}$$

1.6 Measures of dispersion

Measures of dispersion or measures of variation indicate how 'spread' on the real number line our data series is. The four terms connotate:

- Dispersion
- Variation
- Spread
- Fluctuation

Without properly assessing dispersion, the knowledge of location means only a little.

1.6.1 Measures of dispersion: Range

$$\text{Range} = Largestobs - Smallestobs$$

or

$$\text{Range} = Max - Min$$

Range measures the length of the interval on the real number line spanned by our data set.

1.6.2 Measures of dispersion: Interquartile range

The interquartile range (IQR) is defined as:

$$\text{IQR} = Q_3 - Q_1$$

IQR measures the length of the interval on the real number line spanned by the "central 50%" our data set.

1.6.3 Measures of dispersion: Box-Whisker plots

The five-number summary presented as a graph is called a Box-Whisker plot. Sometimes, near outliers and far outliers can also be added while constructing these plots.

- Outliers
- Near outliers
- Far outliers

1.6.4 *Measures of dispersion: Variance*

For $\{x_i\}_{i=1}^{N}$:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

is called the "population variance", and for $\{x_i\}_{i=1}^{n}$:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

is called the 'sample variance'

In a nutshell

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$

A practical way to calculate variance is to:

1. Calculate the mean of squares: $\frac{1}{N} \cdot \sum_{i=1}^{N} x_i^2$
2. Calculate the square of mean: $\mu^2 = (\frac{1}{N} \cdot \sum_{i=1}^{N} x_i)^2$
3. Subtract the second from the first

Consider:

1 3 6 10 15 21 28

For this series, $\mu = 12$ (calculate yourself) and the variance is calculated as follows:

$$\begin{aligned}
\sigma^2 &= \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \\
&= \frac{(1-12)^2 + (3-12)^2 + (6-12)^2 + (10-12)^2 + (15-12)^2 + (21-12)^2 + (28-12)^2}{7} \\
&= \frac{588}{7} \\
&= 84
\end{aligned}$$

A tabular approach may also be preferred:

i	x_i	$x_i - \mu$	$(x_i - \mu)^2$
1	1	-11	121
2	3	-9	81
3	6	-6	36
4	10	-2	4
5	15	3	9
6	21	9	81
7	28	16	256
			588
			$\sigma^2 = 588/7 = 84$

Finally, one may calculate the sum of squares as 1596 and the mean as 12, and calculates the variance σ^2 as $1596/7 - 12^2$, which is 84.

1.6.5 Measures of dispersion: Standard deviation

For $\{x_i\}_{i=1}^N$:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

is called the "population standart deviation", and for $\{x_i\}_{i=1}^n$:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

is called the 'sample variance'

Consider the following data series:

11	23	58	13	21	34
55	89	14	42	33	37
76	10	98	71	59	72
58	44	18	16	76	51
9	46	17	71	12	86
57	46	36	87	50	25
12	13	93	19	64	18
31	78	11	51	42	29

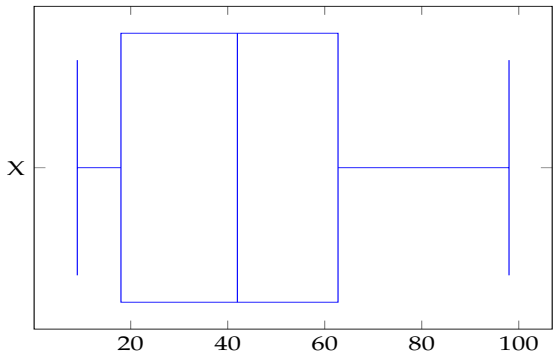
We are now asked to describe this data series, including its mean, five-number summary, range, interquartile range and variance. For ease in calculating the positional measures (quartiles here), it is a good practice to order the observations from the smallest to the largest, i.e., in ascending order:

9	10	11	11	12	12
13	13	14	16	17	18
18	19	21	23	25	29
31	33	34	36	37	42
42	44	46	46	50	51
51	55	57	58	58	59
64	71	71	72	76	76
78	86	87	89	93	98

The following are then found:

- $N = 48$
- $Minimum = 9$
- $Maximum = 98$
- Range $= Maximum - Minimum = 98 - 9 = 89$
- $\sum_{i=1}^{N} x_i = 2082$, so, $\mu = 2082/48 = 43.375$
- Q_2 is at the $(48+1) \cdot 0.5 = 24.5th$ position, $Q_2 = 42$
- Q_1 is at the $(48+1) \cdot 0.25 = 12.25th$ position, $Q_1 = 18$
- Q_3 is at the $(48+1) \cdot 0.75 = 36.75th$ position, $Q_3 = 62.75$
- $(9, 18, 42, 62.75, 98)$ is the five-number summary

For the same data series (call it X), the Box-Whisker plot looks like:



In a nutshell

Working with grouped data

Given population data of N observations grouped into K classes, with frequencies $f_1, f_2, \ldots, f_K$, if the midpoints of these classes are $m_1, m_2, \ldots, m_K$, then

$$\mu = \frac{\sum_{i=1}^{K} f_i m_i}{N}$$

$$\sum_{i=1}^{K} f_i = N$$

$$\sigma^2 = \frac{\sum_{i=1}^{K} f_i (m_i - \mu)^2}{N}$$

Given sample data of n observations grouped into K classes, with frequencies $f_1, f_2, \ldots, f_K$, if the midpoints of these classes are $m_1, m_2, \ldots, m_K$, then

$$\bar{x} = \frac{\sum_{i=1}^{K} f_i m_i}{n}$$

$$\sum_{i=1}^{K} f_i = n$$

$$s^2 = \frac{\sum_{i=1}^{K} f_i (m_i - \bar{x})^2}{n-1}$$

1.6.6 Measures of dispersion: Coefficient of variation

Population coefficient of variation:

$$\text{CV} = \frac{\sigma}{\mu} 100\% \text{ if } \mu \neq 0$$

Sample coefficient of variation:

$$\text{CV} = \frac{s}{\bar{x}} 100\% \text{ if } \bar{x} \neq 0$$

1.7 Measures of association for bivariate data

When we deal with one variable in our analysis, it is a case with "univariate" data. When we are concerned with patterns of change of two variables together, it is a case involving "bivariate" data. In these lecture notes,

$$\{x_i\}_{i=1}^{n_X} \text{ and } \{x_i\}_{i=1}^{n_Y}$$

indicate univariate data, but

$$\{(x_i, y_i)\}_{i=1}^{n}$$

indicate bivariate data.

Notice that, bivariate data come in "pairs", so one cannot change the correspondence between x's and y's.

1.7.1 *Measures of association for bivariate data: Covariance*

Covariance is a measure of the linear relationship between two variables. For $\{(x_i, y_i)\}_{i=1}^{N}$,

$$\text{Cov}(x, y) = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N}$$

For $\{(x_i, y_i)\}_{i=1}^{n}$,

$$\text{Cov}(x, y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

1.7.2 *Measures of association for bivariate data: Correlation*

The correlation for $\{(x_i, y_i)\}_{i=1}^{N}$ is given by

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

and for $\{(x_i, y_i)\}_{i=1}^{n}$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

When $|r| \geq \frac{2}{\sqrt{n}}$ we say the (linear) relationship is strong enough (or significant). Notice that it is always the case that

$$-1 \leq \rho_{xy} \leq 1$$
$$-1 \leq r_{xy} \leq 1$$

1.8 *Issues of unit and scale*

Despite not paid enough attention by economists and business administration people, almost every data series comes with a unit and scale. For instance, if my income is $TRY96,000$, the unit is TRY (international code for Turkish lira) and the scale is not explicitly said. If we write it as $TRY96K$, the unit is again TR and the scale is "thousands", so 96 means $96,000$ here.

Consider $\{(x_i, y_i)\}_{i=1}^{N}$ where x is the body weight in kilograms (kg) y is the height in centimeters (cm). Then,

Measure	Unit
μ_x	kg
μ_y	cm
σ_x^2	kg^2
σ_x	kg
σ_y^2	cm^2
σ_x	cm
$\text{CV}_x = \frac{\sigma_x}{\mu x}$	*Unitless*
$\text{CV}_y = \frac{\sigma y}{\mu x}$	*Unitless*
σ_{xy}	$kg.cm$
ρ_{xy}	*Unitless*

Similarly, quartiles, deciles and percentiles of a variable have the same unit as the variable. Range of a variable has the same unit as the variable. Interquartile range of a variable has the same unit as the variable. As a rule of thumb, linear operators do not alter the units.

Use of numerical scales is often a matter of practicality or convenience. Nobody likes to write 123,000,000,000,000 (except politicians) instead of writing 123 trillions or 123.10^{12}. One may need to learn two important practices of scaling numbers:

- Logarithmic scales
- Inverted scales

In this edition, these are left to readers as individual study.

1.2 EXERCISES

1. Consider a population with data values of:

5, 6, 3, 3, 6, 9, 10, 4, 10, 4
Compute the mean, range, standard deviation, median, and Q_1.
Solution: $\mu = 6$, Range $=$ max $-$ min $= 7$, $\sigma = 2.61$, $Q_2 = 5.5$ and $Q_1 = 3.75$.

2. Find the mean, median, mode(s), variance, range, 1st quartile, and the 80th percentile of the data given below:

9, 13, 6, 7, 8, 6, 6, 9, 13, 13
Solution: $\mu = 9$, $Q_2 = 8.5$, modes are 6 and 13, $\sigma^2 = 8$, Range $= 7$, $Q_1 = 6$ and $P_{80} = 13$.

3. A population has a range of R and it consists of two observations only. Calculate the variance of this data set.

Solution: x_1 and x_2 are the only two observations, and $x_2 - x_1 = R$ is given (suppose $x_1 < x_2$. Then $x_2 - \mu = R/2$ and $x_1 - \mu = -R/2$.

$$\begin{aligned}
\sigma^2 &= \frac{1}{2}\left[(x_1-\mu)^2 + (x_2-\mu)^2\right] \\
&= \frac{1}{2}\left[(-R/2)^2 + (R/2)^2\right] \\
&= \frac{1}{2}\cdot\frac{R^2}{2} \\
\sigma^2 &= R^2/4.
\end{aligned}$$

4. A researcher argues that median equals the simple average of the first and third quartiles. By giving a numerical example, show that this is incorrect.
Solution: Find/make up your own example.

5. Let a and b be any given real numbers. Let $x_1, x_2, \ldots, x_N$ and $y_1, y_2, \ldots, y_N$ be two data sets such that, for any $i = 1, 2, \ldots, N$, and $y_i = ax_i + b$.

i. What is the relation between the mean of the y-values and the mean of x-values?

ii. What is the relation between the variance of the y-values and the variance of x-values?

Solution: Needs some careful and patient elaboration.

1.

$$\begin{aligned}
\mu_y &= \frac{1}{N}\sum_{i=1}^{N} y_i \\
&= \frac{1}{N}\sum (ax_i + b) \\
&= \frac{1}{N}\sum ax_i + \frac{1}{N}\sum b \\
&= a\frac{1}{N}\sum x_i + \frac{1}{N}Nb \\
&= a\mu_x + b \\
\mu_y &= a\mu_x + b
\end{aligned}$$

2.

$$\begin{aligned}
\sigma_y^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mu_y\right)^2 \\
&= \frac{1}{N}\sum\left(ax_i + b - a\mu_x - b\right)^2 \\
&= \frac{1}{N}a^2\sum\left(x_i - \mu_x\right)^2 \\
&= a^2\frac{1}{N}\sum\left(x_i - \mu_x\right)^2 \\
&= a^2\sigma_x^2 \\
\sigma_y^2 &= a^2\sigma_x^2
\end{aligned}$$

6. Consider a bivariate data consisting of the 1st midterm and 2nd midterm grades of 216 students. It is known that the 1st midterm grade of each student is 8% less than his 2nd midterm grade. If the mean of the 2nd midterm grades is 64 and the variance is 9 what can you say about the correlation coefficient of this data?
Solution: Without any calculations we can say it is 1. Why?

7. If we remove a data point from a data series, variance decreases. True or false? Explain.
Solution: For a logical statement to be true, it must be true without any exceptions. Consider first $\{1, 5, 9\}$ and second $\{1, 9\}$. Which set of values has a larger variance? What is your conclusion?

8. When we multiply each point in a data series by the same factor, the variance increases. True or false? Explain.
Solution: Consider $y_i = kx_i, i = 1, 2, \ldots, N$. You have seen before that

$$\sigma_y^2 = k^2\sigma_x^2$$

Then, σ_y^2 is greater than σ_x^2 only when $k^2 > 1$, i.e., $|k| > 1$. So, the given statement is false (as we are able to find a counter example).

9. Below is the distribution of a variable X based on a sample of 40 observations. Compute the coefficient of variation.

X	Frequency
$10 - 14$	8
$15 - 19$	16
$20 - 24$	u
$25 - 29$	4
$30 - 34$	2

Solution: $\mu = 19$, $\sigma^2 = 28.5$ and $\sigma = 5.34$. So,

$$CV = \frac{\sigma}{\mu} = \frac{5.34}{19} = 0.28$$

10. Consider the two populations of bivariate data:

Population 1		Population 2	
x	y	x	y
2	2	2.9	3.8
6	3	-1	-4
10	4	-1.9	-5.8
4	2.5	4	6
-2	1	6	10

i. Find the covariance and correlation coefficient for each population.

ii. Plot the scatter plot for both populations.

iii. Standardize the x and y values in each population and plot the scatter plots for the standardized values.

Solution:

1. To find the covariance we first find the mean of x and y values for both populations:

$$\mu_{1,x} = \frac{1}{5}(2+6+10+4+(-2)) = 4$$
$$\mu_{1,y} = \frac{1}{5}(2+34+2.5+1) = 2.5$$
$$\mu_{2,x} = \frac{1}{5}(2.9+(-1)+(-1.9)+4+6) = 2$$
$$\mu_{2,y} = \frac{1}{5}(3.8+(-4)+(-5.8)+6+10) = 2$$

Thus

$$\text{Cov}_1 = \frac{1}{5}\sum_{i=1}^{5}(x_{1,i}-\mu_{1,x})(y_{1,i}-\mu_{1,y}) = 4$$
$$\text{Cov}_2 = \frac{1}{5}\sum_{i=1}^{5}(x_{2,i}-\mu_{2,x})(y_{2,i}-\mu_{2,y}) = 18.008 \approx 18$$

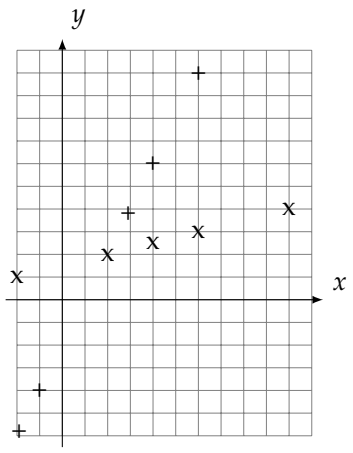
To find the correlation coefficient we first find the variances:

$$\sigma_{1,x} = \frac{1}{5}\sum_{i=1}^{5}(x_{1,i}-\mu_{1,x})^2 = 16$$
$$\sigma_{1,y} = \frac{1}{5}\sum_{i=1}^{5}(y_{1,i}-\mu_{1,y})^2 = 1$$
$$\sigma_{2,x} = \frac{1}{5}\sum_{i=1}^{5}(x_{2,i}-\mu_{2,x})^2 = 9.006 \approx 9$$
$$\sigma_{2,y} = \frac{1}{5}\sum_{i=1}^{5}(y_{2,i}-\mu_{2,y})^2 = 36.016 \approx 36$$

Thus

$$\rho_1 = \frac{\text{Cov}_1}{\sigma_{1,x}\sigma_{1,y}} = \frac{4}{4 \cdot 1} = 1$$
$$\rho_2 = \frac{\text{Cov}_2}{\sigma_{2,x}\sigma_{2,y}} = \frac{18}{3 \cdot 6} = 1$$

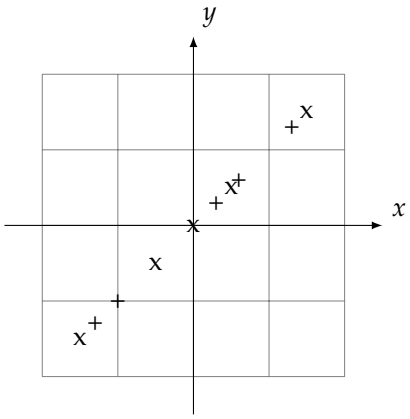
2. Population 1 is represented with a "x" and population 2 with a "+" in the scatter plot below:



3. Standardized values are obtained by subtracting the corresponding mean from each value and dividing the result by the standard deviation:

Population 1		Population 2	
x	y	x	y
-0.5	-0.5	0.3	0.3
0.5	0.5	-1	-1
1.5	1.5	-1.3	-1.3
0	0	0.6	0.6
-1.5	-1.5	1.3	1.3

The standardized values are plotted below:



Note that even though the original populations where on lines with different slope the standardized values are on a line with slope 1.

In a nutshell

Picking the appropriate statistic or visual

- Nominal data: Mode, Bar chart, Column chart
- Ordinal data: Median, Mode, Bar chart, Column chart
- Interval or Ratio data:
  - To describe center: Mean, Median, Mode, (Midrange), (Geometric Mean), (Midhinge), Histogram, Box plot
  - To describe variability: Range, IQR, Standard deviation, CV, (z-values), Histogram, Box plot
  - To describe shape: Mean vs Median, Skewness, (Kurtosis), Histogram, Box plot

1.9 *Chebyshev's theorem (Chebyshev's inequality)*

For any data set with a mean of μ and variance of σ^2, and any $k > 1$, at least

$$\left(1 - \frac{1}{k^2}\right) 100\%$$

of the observations will take a value in the interval $[\mu - k\sigma, \mu + k\sigma]$.

1.3 EXERCISES

1. Consider a population with a mean of 4 and variance of 36. Using Chebyshev's theorem find an interval that contains at least 70% of the observations.
Solution: $\mu = 4$ and $\sigma^2 = 36$, so $\sigma = 6$.

$$\left(1 - \frac{1}{k^2}\right) 100\% = 70\%$$
$$1 - \frac{1}{k^2} = 0.70$$
$$\frac{1}{k^2} = 0.30$$
$$k^2 = \frac{1}{0.30}$$
$$k = 1.83.$$

So, the requested interval is:

$$\begin{aligned}[\mu - k\sigma, \mu + k\sigma] &= [4 - 6(1.83), 4 + 6(1.83)] \\ &= [-6.95, 14.95]\end{aligned}$$

2. The monthly charges for credit card holders at a department store have a mean of \$250 and a standard deviation of \$100. Use Chebyshev's theorem to answer the following questions:

i. What can you say, for sure, about the percentage of card holders who will have monthly charges between \$100 and \$400?

ii. Provide a range for the credit card charges that will include at least 80% of all credit card customers.

Solution:

1. Use the same approach. For the interval $[100, 400]$, $k = 1.5$. Reveal why. Then, this interval contains at least

$$1 - \frac{1}{1.5^2} \cdot 100\% = (100 - 44)\% = 56\%$$

of all observations.

2. $[26.4, 473.6]$ is the answer.

3. In a stock exchange average return over a year turns out to be 1% with a standard deviation of 2%. Over the same year, average exam grade in a university is 60 *points* with a standard deviation of 24 *points*. Which one has a higher variability, the returns or the grades?
Solution: CV for returns is $2\%/1\% = 2$ and CV for grades is $24 points/60 points = 0.4$. So, returns have a higher variability.

4. When we replace a positive data value with its "additive inverse" in a data set, variance increases. Is this claim true or false? Either prove that it is true, or provide a counter example to show that it is false. Make sure you have used a formal mathematical notation.
Solution: Consider $\{5, -7\}$ and $\{-5, -7\}$. Which pair of values has higher variance? Then, come up with a conclusion.

1.10 Adding and multiplying terms over an index

If you've N numbers $x_1, x_2, \ldots, x_N$, the sum of these numbers, S, is:

$$S = x_1 + x_2 + \cdots + x_N$$

In expressing sums like this, we always write the first two terms, then three periods, then the last term. A shorter way to write S is:

$$S = \sum_{i=1}^{N} x_i$$

where i is the index of x, running from 1 to N.

Unless otherwise specified, i increases by 1 every time, from 1 to N. So, $S = \sum_{i=1}^{N} x_i$ is read as "sum of x_i, i from 1 to N". This means,

- We'll take $(i = 1)$, x_1 first,
- take $(i = 2)$: add x_2,
- take $(i = 3)$: add x_3
- $\vdots$
- and taking the next x_i each time, we'll take $(i = N)$ and add x_N the last.

For example,

$$S = \Sigma_{i=1}^{4} x_i = \sum_{i=1}^{4} x_i = \sum_{i=1}^{i=4} x_i = x_1 + x_2 + x_3 + x_4$$

If N is a number well-known in a problem:

$$S = \sum x_i$$

is a valid expression and it means "consider all x_i",

Notice that

$$\sum_{i=1}^{N} = \left(\sum_{i=1}^{N-1} \right) + x_N = \left(\sum_{i=1}^{N-2} \right) + x_{N-1} + x_N$$

and so forth.

In case we want to write

$$x_1 + x_3 + \cdots + x_{2k-1}$$

using our summation operator, we can write it as:

$$\sum_{i=1}^{k} x_{2i-1}$$

As you've seen, wisely using i solves many problems.

Consider:

$$S = 2 \cdot x_1 + 2 \cdot x_2 + \cdots + 2 \cdot x_N$$

which is equivalent to

$$\sum_{i=1}^{N} 2x_i$$

S, then, can be written as

$$2\sum_{i=1}^{N} x_i$$

So, if each number in the sequence $x_1, x_2, \ldots, x_N$ is multiplied by the same value which is not a function of i, this value can be taken out of the summation sign Σ.

Consider:

$$S = x_1 + y_1 + x_2 + y_2 + \cdots + x_N + y_N$$

Notice that, this is the same thing as:

$$\sum_{i=1}^{N} (x_i + y_i) = \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} y_i$$

Consider:

$$S = \sum_{i=1}^{N} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_N y_N$$

Notice that,

$$S = \sum_{i=1}^{N} x_i y_i \neq \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)$$

Sum of the products is not equal to product of the sums. Expand (write long) the expressions to see why not.

Consider:

$$S = x_1 + 2x_2 + 3x_3 + \cdots + Nx_N$$

How can we write this in short?

$$S = \sum_{i=1}^{N} i x_i$$

Notice that:

$$S = \sum_{i=1}^{N} ix_i \neq \left(\sum_{i=1}^{N} i\right)\left(\sum_{i=1}^{N} x_i\right)$$

Consider:

$$S = \sum_{i=1}^{N} (x_i\bar{y} + y_i\bar{x})$$

Since $\bar{x}$ and $\bar{y}$ are not indexed with i, the expression is equal to:

$$\bar{y}\sum_{i=1}^{N} x_i + \bar{x}\sum_{i=1}^{N} y_i$$

Consider:

$$\sum_{i=1}^{N} x_i^2,$$

Notice that

$$\sum_{i=1}^{N} x_i^2 \neq \left(\sum_{i=1}^{N} x_i\right)^2$$

Sum of the squares is not equal to square of the sum.

If you've N numbers $x_1, x_2, \ldots, x_N$, the product of these numbers, P, is:

$$P = x_1 \cdot x_2 \cdots x_N$$

In expressing products like this, we always write the first two terms, then three periods, then the last term. A shorter way to write P is:

$$P = \prod_{i=1}^{N} x_i$$

where i is the index of x, running from 1 to N.

Consider:

$$P = \prod_{i=1}^{N} i$$

What's this?

It is nothing but $P = \prod_{i=1}^{N} x_i$ with $x_i = i$. So P equals:

$$1 \cdot 2 \cdot 3 \cdots N = N!$$

Regarding our future purposes, an important property to remember is:

$$ln\prod_{i=1}^{N} x_i = \sum_{i=1}^{N} lnx_i$$

as we'll use while writing Likelihood functions in ECON 222.

Finally, consider:

$$\sum_{i=0}^{N} \binom{N}{i} x^{N-i}y^i$$

What's this? Expand it to see:

$$= \binom{N}{0} x^{N-0} y^0 + \binom{N}{1} x^{N-1} y^1 + \cdots + \binom{N}{N} x^{N-N} y^N$$

which is nothing but the binomial expansion

$$(x+y)^N$$

as we'll use while studying the Binomial and Poisson distributions in ECON 221.

1.4 EXERCISES

1. Consider the expression for the population variance:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

and simplify it until you see:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \mu^2$$

Solution:

$$\begin{aligned}
\sigma^2 &= \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \\
&= \frac{1}{N} \sum \left(x_i^2 - 2\mu x_i + \mu^2 \right) \\
&= \frac{1}{N} \sum x_i^2 - \frac{2\mu}{N} \sum x_i + \frac{1}{N} \sum \mu^2 \\
&= \frac{1}{N} \sum x_i^2 - 2\mu^2 + \frac{1}{N} N \mu^2 \\
\sigma^2 &= \frac{1}{N} \sum x_i^2 - \mu^2.
\end{aligned}$$

So, variance can be calculated by subtracting 'the square of the mean of observations' from 'the mean of squared observations'.

2 Probability basics

We refrain from giving a crystal-clear and simple definition of probability and provide you with three perspectives to obtain/come up with probabilities. Once we are done with this chapter, you'll understand the reason for our avoidance to give a definition and you'll be able to find/calculate probabilities.

2.1 Modeling a Random Experiment

A random experiment is a process leading to two or more possible and unsure outcomes. The process can be physical/mechanical or purely theoretical or thought-based.

The set S of all possible outcomes of a particular experiment is called the "sample space" for the experiment.

Then, an event is any collection of possible outcomes of an experiment, that is, any subset of S, including S itself.

Once we reduce the description of events into sets, we can enjoy (use) the knowledge of set theory to comprehend what probability is. So, we'll first go over a number of definitions about set operations and some useful properties. At the very beginning, a set is a collection of distinct objects (things). These things can be people, animals, kitchen utensils, countries or most usefully "numbers".

Consider the random experiment of tossing a fair coin: the coin will show one of its two faces, i.e., either a Head (H) or a Tail (T). These two are the basic outcomes of our random experiment. So, the sample space S can be written as $S = \{\text{Head}, \text{Tail}\}$ or as $S = \{H, T\}$. It is also possible to use 0 to denote a Head and 1 to denote a Tail and write the sample space as $S = \{0, 1\}$. Then, any of the following is an event in S, as they are all subsets of $S = \{\text{Head}, \text{Tail}\}$:

$$A = \{\text{Head}\}$$
$$B = \{\text{Tail}\}$$
$$C = \{\text{Head}, \text{Tail}\}$$
$$D = \{\}$$

In the random experiment of tossing two fair coins simultaneously and observing the joint outcome, each coin may yield a H or a T & we write our sample space as

$$S = \{(H,H),(H,T),(T,H),(T,T)\}$$

or

$$S = \{(c_1,c_2) \mid c_1,c_2 \in \{H,T\}\}$$

Since we observe the joint outcome of the random experiment (this is usually why we toss two coins simultaneously), we define a basic outcome as an ordered pair (c_1,c_2), where c_1 is the first coin outcome and c_2 is the second coin outcome. c_1 **and** c_2 **individually 'are not' basic outcomes.** For instance, H is not a basic outcome here, nor is T.

Consider the simultaneous toss of a fair coin and a fair die where we write a basic outcome as an ordered pair like (coin outcome, die outcome). As the coin can give us a face in $\{H,T\}$ and the die can give us a value in $\{1,2,3,4,5,6\}$, the sample space looks like:

$$S = \{(H,1),(H,2),(H,3),(H,4),(H,5),(H,6),(T,1),(T,2),(T,3),(T,4),(T,5),(T,6)\}$$

or equivalently as:

$$S = \{(x,y) \mid x \in \{H,T\}, y \in \{1,2,3,4,5,6\}\}$$

As in the previous example, H itself is not a basic outcome here, nor is 3; yet $(H,3)$ is. Since we write a basic outcome as an ordered pair like (coin outcome, die outcome), $(3,H)$ is not a basic outcome, either.

Then, we can define the event **A: Coin shows H** as:

$$A = \{(H,1),(H,2),(H,3),(H,4),(H,5),(H,6)\}$$

and the event **B: Die shows an even number** as:

$$B = \{(H,2),(H,4),(H,6),(T,2),(T,4),(T,6)\}$$

and the event **C: Coin shows H and die shows 4** as:

$$C = \{(H,4)\}$$

Given any two sets (events) A and B:

- The union of A and B, written $A \cup B$, is the set of elements that belong to **either** A or B or both:

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}$$

- The intersection of A and B, written $A \cap B$, is the set of elements that belong to **both** A and B:

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}$$

- The complement of A, written A^c, is the set of all elements that are not in A:

$$A^c = \{x \mid x \notin A\}$$

In a nutshell

Properties of Sets
For any three events A, B, and C defined on a sample space S:

- Commutativity
  - $A \cup B = B \cup A$
  - $A \cap B = B \cap A$
- Associativity
  - $A \cup (B \cup C) = (A \cup B) \cup C$
  - $A \cap (B \cap C) = (A \cap B) \cap C$
- Distributivity
  - $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
  - $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- De Morgan's laws
  - $(A \cup B)^c = A^c \cap B^c$
  - $(A \cap B)^c = A^c \cup B^c$

2.1.1 *Mutually exclusive events*

Two events A and B are mutually exclusive (or disjoint) if

$$A \cap B = \emptyset$$

The events $A_1, A_2, \ldots$ are pairwise mutually exclusive (or pairwise disjoint) if

$$A_i \cap A_j = \emptyset, \forall i \neq j$$

2.1.2 *Collectively exhaustive events*

Given the K events $E_1, E_2, \ldots, E_K$ in the sample space S, if

$$E_1 \cup E_2 \cup \cdots \cup E_K = S$$

these K events are collectively exhaustive events.

2.2 *Properties of a probability measure: Probability postulates*

S being the sample space, O_i the basic outcomes, and A an event, if $\mathrm{P}(A)$ is defined, it obeys the following:

- $A \subset S \implies 0 \leq \mathrm{P}(A) \leq 1$
- $\mathrm{P}(A) = \sum_A \mathrm{P}(O_i)$
- $\mathrm{P}(S) = 1$

where these three statements are called the probability postulates. Here, the first defines the boundaries of any probability measure, the second states that probability of any event is the sum of the probabilities of basic outcomes making the event, and the third ensures that the sample space has a probability of 1.

In a nutshell

When we are given

$A \subset S \implies 0 \leq \mathrm{P}(A) \leq 1$

$\mathrm{P}(A) = \sum_A \mathrm{P}(O_i)$

$\mathrm{P}(S) = 1$

we simply take them as granted universally. This is as we do to any scientific postulates anywhere. The simple reason for this is "postulates are the basis of our scientific study". Without them, we lack the suitable environment to work in. Putting it differently, we never try to and not required to prove the postulates, but we can establish and prove anything starting from them. In some books, probability postulates are called **probability axioms**.

2.1 EXERCISES

1. The sample space S for an experiment is $S = \{a, b, c\}$. Is it possible for a probability measure to have values

$$\mathrm{P}(\{a, b\}) = \tfrac{2}{3},\ \mathrm{P}(\{a, c\}) = \tfrac{1}{3},\ \mathrm{P}(\{b, c\}) = \tfrac{1}{3}$$

Why or why not?
Solution:

$$\begin{aligned}\mathrm{P}(\{a, b\}) = \tfrac{2}{3} \text{ implies } \mathrm{P}(\{a\}) + \mathrm{P}(\{b\}) = \tfrac{2}{3}\\ \mathrm{P}(\{a, c\}) = \tfrac{1}{3} \text{ implies } \mathrm{P}(\{a\}) + \mathrm{P}(\{c\}) = \tfrac{1}{3}\\ \mathrm{P}(\{b, c\}) = \tfrac{1}{3} \text{ implies } \mathrm{P}(\{b\}) + \mathrm{P}(\{c\}) = \tfrac{1}{3}\end{aligned}$$

Then,

$$2 \cdot \mathrm{P}(\{a\}) + 2 \cdot \mathrm{P}(\{b\}) + 2 \cdot \mathrm{P}(\{c\}) = \tfrac{2}{3} + \tfrac{1}{3} + \tfrac{1}{3}$$

and,

$$\mathrm{P}(\{a\}) + \mathrm{P}(\{b\}) + \mathrm{P}(\{c\}) = \tfrac{2}{3}$$

Since $S = \{a, b, c\}$, this sum cannot have a value other than 1.

2. Given an experiment such that $\mathrm{P}(A^c \cap B) = 0.1$, $\mathrm{P}(A \cap B^c) = 0.4$, and $\mathrm{P}\big((A \cap B)^c\big) = 0.6$.

i. Compute $\mathrm{P}(A)$

ii. Compute $\mathrm{P}(B)$

iii. Compute $\mathrm{P}(A \cup B)$

iv. Compute $\mathrm{P}(A^c \cup B)$

Solution:

1. Using a Venn diagram, see the answer is 0.8.
2. Using a Venn diagram, see the answer is 0.5.
3. Using a Venn diagram, see the answer is 0.9.
4. Using a Venn diagram, see the answer is 0.6.

3. Suppose events A and B are such that $\mathrm{P}(A) = \frac{2}{5}$, $\mathrm{P}(B) = \frac{2}{5}$, and $\mathrm{P}(A \cup B) = \frac{1}{2}$. Find $\mathrm{P}(A \cap B)$.
Solution: Solve on your own.

4. If $\mathrm{P}(A) = \frac{1}{3}$, $\mathrm{P}(A \cup B) = \frac{1}{2}$, and $\mathrm{P}(A \cap B) = \frac{1}{4}$. Find $\mathrm{P}(B)$.
Solution: Solve on your own.

5. Two mutually exclusive and collectively exhaustive events are complementary. True or false? Explain.
Solution: Solve on your own.

6. The events A, B, and C are such that $P(A) = 0.43$, $P(C) = 0.45$, $P(B \cap C) = 0.12$, $P(A \cup C) = 0.75$, and $P(A \cup (B \cap C)) = 0.45$. Find $P((A \cup B) \cap C)$.
Solution: $P(A) = 0.43$, $P(C) = 0.45$, $P(A \cup C) = 0.75$. Then,

$$P(A \cap C) = 0.43 + 0.45 - 0.75 = 0.13$$

$P(B \cap C) = 0.12$, $P(A \cup (B \cap C)) = 0.45$. Then,

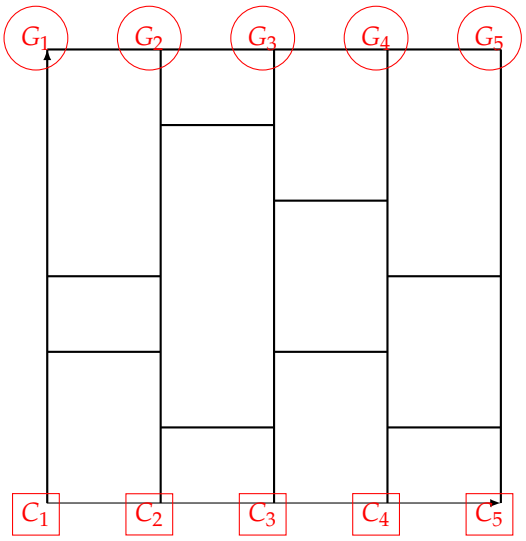
$$P(A \cap B \cap C) = 0.43 + 0.12 - 0.45 = 0.10$$

Finally,

$$\begin{aligned} P((A \cup B) \cap C) &= P((A \cap C) \cup (B \cap C)) \\ &= P(A \cap C) + P(B \cap C) - P(A \cap B \cap C) \\ &= 0.13 + 0.12 - 0.10 \\ &= 0.15 \end{aligned}$$

7. "I don't know half of you half as well as I should like; and I like less than half of you half as well as you deserve" says Bilbo Baggins in the Fellowship of the Ring by J.R.R. Tolkien. Try to present Bilbo's saying using sets.
Solution: Try on your own if you have time, and just for fun.

8. Consider the following children's game: each child, C_1, C_2, C_3, C_4, C_5, begins crawling from her/his spot and moves forward ("up" in the chart), at each turn she/he faces she/he has to take the turn until reaching one of the gifts G_1, G_2, G_3, G_4, G_5.



i. Defining an event as "a certain kid having reached a certain gift", write the sample space for this setup.

ii. Is there a random nature here?

Solution: A careful examination would reveal that there is no randomness in this setup. Everything is deterministic, i.e., each child reaches a different gift by design.

2.3 *Probability versus possibility*

A clear distinction is needed between the two terms before our scientific study of probabilities. While possibility is an expression of whether something may happen, probability is a numerical assessment of this happening. Putting it right, "calculation of a probability is based on counting the possibilities."

2.4 *Methods of assigning probability*

In our study of the probabilistic universe, we can mainly resort to three definitions of or approaches to probability, namely Classical probability, Relative frequency probability & Subjective probability. Keeping the order intact, these are also called the Classical view, the Frequentist view & the Subjective view.

2.4.1 *Classical probability*

Assuming that all outcomes in a sample space are equally likely to occur, classical probability is the proportion of times that an event will occur. So, the typical approach to problem is to count the members of the event A, then to count the members of the sample space S, and finally to divide the former by the latter:

- Number of members of A (cardinality of A):

$$n(A) = |A|$$

- Number of members of S (cardinality of S):

$$n(S) = |S|$$

- Probability of event A:

$$\mathrm{P}(A) = \frac{n(A)}{n(S)} = \frac{|A|}{|S|}$$

The essence of classical probability is that one can develop a probability from fundamental reasoning about the process. In practical terms, the classical statement of probability is based on the requirement that we are able to (or we can) count outcomes in the sample space. This also indicates that we have the full comprehension of the random experiment of interest.

2.4.2 *Relative frequency probability*

The relative frequency probability is the limit of the proportion of times that event A occurs in a large number of trials. Number of A outcomes:

- Number of A outcomes:

$$n_A$$

- Total number of outcomes (trials):

$$n$$

- Probability of event A:

$$\mathrm{P}(A) = \frac{n_A}{n}, n \to \infty$$

Note that tends to infinity points at a "sufficiently large number of replications" in everyday implementation of the approach. Think about what the role of data is in forming/stating relative frequency probabilities.

2.4.3 *Subjective probability*

Subjective probability expresses an individual's degree of belief about the chance that an event will occur. These subjective probabilities are used in certain management decision procedures. Think about what the role of experience is in forming/stating subjective probabilities.

Think about what the role of data is in forming/stating subjective probabilities. Is this something "judgmental"?

2.5 *Counting*

As said before, classical probability perspective is built upon the notion of counting. Due to the complexity of the problems at hand, simple counting by index finger as in our everyday lives is not the notion of counting we are talking about.

In a nutshell

A hypothetical shepherd boy's counting problem
Suppose the boy has a bag, empty at the beginning. In the morning, when emptying the barn, for each sheep going out he puts a little stone into his bag. When the barn gets empty, there are as many stones as the sheep. May be the boy cannot even explicitly write this as many as a number. In the evening, he removes one stone per sheep returning to barn. If there is no unmatched sheep or stones, he is alright. Otherwise, he may be in some trouble. What the shepherd boy facilitates here is called **one-to-one correspondence** and it has strong implications for our scientific/technical practices.

2.5.1 *Multiplication rule*

Going beyond the shepherd boy's problem, we fulfill practically all our counting needs in our study of probability and statistics using the **multiplication rule**, also called **product rule**.

Multiplication rule says, if there are K stages of a task and if we have x_i ways to complete each stage $i = 1, 2, \ldots, K$, then the whole task can be completed in

$$x_1 \cdot x_2 \cdots x_K$$

different ways. If we can go to library from faculty in 2 ways and can go to supermarket from library in 3 ways, then we can go to supermarket from faculty in $2 \cdot 3 = 6$ ways.

2.5.2 *Permutations*

The multiplication principle is useful, as we can find the number of different orderings of n distinct objects. So, in how many different ways can we order n distinct objects?

Let's show the n orders using n slots:

Slot 1	Slot 2	Slot 3	...	Slot n

We can place now n objects into Slot 1, then we can place only $(n-1)$ objects into Slot 2, then only $(n-2)$ into Slot 3, and so forth. Continuing till the end, the picture will look like:

Slot 1	Slot 2	Slot 3	...	Slot n
n	n-1	n-2		1

Now, using the multiplication rule, we have

$$n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1$$

different orderings. This is what we call $\mathrm{P}(n,n)$, which is read as **permutation-n-n**, and as you've seen is given by:

$$\mathrm{P}(n,n) = n!$$

So, 5 distinct things can be ordered in $5! = 120$ ways; 4 things in $4! = 24$ ways; 3 things in $3! = 6$ ways, and so forth. 1 thing can be ordered in $1! = 1$ way, and finally 0 things can also be ordered in $0! = 1$ way, however awkward it sounds.

Next, consider the n distinct objects again and answer: in how many different ways can I order n distinct objects in ranks of $r \leq n$? You'll now see the very same multiplication rule will happily help us to build and answer.

Consider the $r \leq n$ slots below and the allocation of objects:

Slot 1	Slot 2	Slot 3	...	Slot r
n	n-1	n-2	...	n-r+1

We can place now n objects into Slot 1, then I can place only $(n-1)$ objects into Slot 2, only $(n-2)$ into Slot 3, and so forth. Finally, $(n-r+1)$ objects into Slot r. So, I'll have

$$n \cdot (n-1) \cdot (n-2) \cdots (n-r+1)$$

different orderings.

This is what we call $(\mathrm{P}(n,r))$, which is read as **permutation-n-r**, and as you've seen it is given by:

$$\mathrm{P}(n,r) = n \cdot (n-1) \cdots (n-r+1)$$

But wait, this can be written in a simpler way as:

$$\mathrm{P}(n,r) = \frac{n!}{(n-r)!}$$

Notice that:

$$\mathrm{P}(n,n) = \frac{n!}{(n-n)!} = \frac{n!}{0!} = n!$$

2.5.3 *Circular Permutations*

Now, consider a slightly different problem: in how many different ways can we order n distinct objects in n ranks arranged on a circle? It seems again that we have $n!$ different orderings. Yet, this is incorrect! Incorrect, simply because we can start counting the orderings from any of the n positions (slots). This reveals every possible ordering is repeated n times. So, the correct count must be $n!$ **divided by** n, which is $(n-1)!$.

2.5.4 *Combinations*

Now, we move to another problem in which we are interested in different ways to select objects without paying any attention to ordering. Notice that, this problem is equivalent to finding the $r \leq n$ member subsets of a set of n members. Why not to use a thing we've derived earlier? Let's begin with permutations $\mathrm{P}(n,r)$. We know that

$$\mathrm{P}(n,r) = \frac{n!}{(n-r)!}$$

What if we can remove/discount the different number of orderings of the r objects from this formula? Yes, this must work, and we'll come up with the number of selections:

$$\frac{\mathrm{P}(n,r)}{\mathrm{P}(r,r)} = \frac{\frac{n!}{(n-r)!}}{r!}$$

This is called $\binom{n}{r}$, and is read as **combination-n-r**, and is given by:

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!}$$

As you'll see soon in your studies, this short list of formulas will do all the job when they're used wisely. A useful suggestion is: when confused with permutations and combinations, retreat back to multiplication rule, and study everything from scratch. After all, this principle was our point of departure. Another advice is: in every problem, ask yourself the question: **Am I ordering or selecting?**

2.5.5 *Pigeonhole principle (Dirichlet drawer principle)*

For any function $f : D \to R$, there exists i elements $d_1, d_2, \ldots, d_i \in D$,

$$i = \left\lceil \frac{|D|}{|R|} \right\rceil$$

such that

$$f(d_1) = f(d_2) = \ldots = f(d_i)$$

Intuitively, the Pigeonhole principle says, **if there are more pigeons than the holes, at least one hole will be occupied by more than one pigeon.** Keep this in mind as a life saving mathematical construct.

2.2 EXERCISES

1. In how many different ways can one order n distinct objects in $r \leq n$ ranks arranged on a circle?

Solution: Divide the question into two steps: (1) In how many different ways can I pick $r \leq n$ objects out of n? (2) In how many different ways can I order r different objects in r ranks arranged on a circle? Once you have the answers for both parts, use the multiplication rule.

2. Consider the experiment of tossing a fair coin until two heads or two tails appear in succession.

i. Describe the sample space.

ii. What is the probability that the experiment ends before the sixth toss?

Solution: This exercise is left as self-study.

3. A die is rolled and a coin is tossed. We observe the number of dots on the face of the die that turns up and the faces of the coin that turns up.

i. Write the sample space of this random experiment.

ii. Write using mathematical notation the event of observing a head.

iii. Write using mathematical notation the event of observing an odd number of dots on the die and a head on the coin.

Solution:

1. The die yields a number from $\{1,2,3,4,5,6\}$ and the coin yields a letter from $\{H,T\}$. So,

$$S = \{(d,c) \mid d \in \{1,2,3,4,5,6\}, c \in \{H,T\}\}$$

2.

$$A = \{(1,H),(2,H),(3,H),(4,H),(5,H),(6,H)\}$$

or

$$A = \{(d,c) \mid d \in \{1,2,3,4,5,6\}, c = H\}$$

3.

$$A = \{(1,H),(3,H),(5,H)\}$$

4. We first choose a number from the interval $[0,1]$ at random. Let x_1 be the number we choose. Then we choose a number from the interval $[x_1,1]$ at random.

i. Write the sample space of this random experiment of choosing two numbers as described above.

ii. Write the event which describes the situation where the second number is twice the first number.

Solution: This exercise is left as self-study.

5. A salesperson makes contact with two customers. Each contact results with a sale, a request for a return call later, or no sale. Hence the salesperson is involved in a random experiment.

i. Write the sample space of this experiment.

ii. Let A be the event that contact with customer 1 results with a sale and B be the event that contact with customer 2 results in a sale. List the elements of A and B.

iii. Let A and B be as above. Describe in plain English the event $A \cap B^c$.

Solution:

1. Denote "Sale" with 2, "Request for a return call" with 1, and "No sale" with 0. Let x_1 be the customer 1 outcome and x_2 be the customer 2 outcome. Then,

$$S = \{(x_1, x_2) \mid x_1, x_2 \in \{0, 1, 2\}\}$$

2.

$$A = \{(2,0), (2,1), (2,2)\}$$
$$B = \{(0,2), (1,2), (2,2)\}$$

3. $A \cap B^c$ is the event that call to customer 1 results in a sale **and** the call to customer 2 results in anything but a sale.

$$A = \{(2,0), (2,1), (2,2)\}$$
$$B^c = \{(0,0), (1,0), (2,0), (0,1), (1,1), (2,1)\}$$
$$A \cap B^c = \{(2,0), (2,1)\}$$

6. Let $S = \{(x, y) \mid 21 \leq x \leq 25, 21 \leq y \leq 25\}$.

$$\mathrm{P}(x, y) = \begin{cases} \frac{kx}{y} & x \leq y \\ \frac{ky}{x} & y \leq x \end{cases}$$

Determine the value of k.

Solution:

$$
\begin{aligned}
&k\left(\frac{21}{21}+\frac{21}{22}+\frac{21}{23}+\frac{21}{24}+\frac{21}{25}+\right.\\
&\frac{21}{22}+\frac{22}{22}+\frac{22}{23}+\frac{22}{24}+\frac{22}{25}+\\
&\frac{21}{23}+\frac{22}{23}+\frac{23}{23}+\frac{23}{24}+\frac{23}{25}+\\
&\frac{21}{24}+\frac{22}{24}+\frac{23}{24}+\frac{24}{24}+\frac{24}{25}+\\
&\left.\frac{21}{25}+\frac{22}{25}+\frac{23}{25}+\frac{24}{25}+\frac{25}{25}\right)=1\\
&k\left(\frac{21}{21}+\frac{64}{22}+\frac{109}{23}+\frac{156}{24}+\frac{205}{25}\right)=1\\
&k\left(1+\frac{32}{11}+\frac{109}{23}+\frac{13}{2}+\frac{41}{5}\right)=1\\
&k\left(\frac{2530+7360+11990+16445+20746}{2530}\right)=1\\
&k=\frac{2530}{59071}=0.04283
\end{aligned}
$$

7. Ali and Berna are taking a mathematics course. The course has only three grades: A, B, and C. The probability that Ali gets a B is 0.3. The probability that Berna gets a B is 0.4. The probability that neither gets an A but at least one gets a B is 0.1. What is the probability that at least one gets a B but neither gets a C?
Solution: Let (a, b) denote the grade outcomes of Ali and Berna as an ordered pair. So,

$$S = \{(a, b) \mid a, b \in \{A, B, C\}\}$$

or

$$S = \{(A,A),(A,B),(A,C),(B,A),(B,B),(B,C),(C,A),(C,B),(C,C)\}$$

It is given that

$$\mathrm{P}(\{(B,A),(B,B),(B,C)\}) = 0.3$$

and

$$\mathrm{P}(\{(A,B),(B,B),(C,B)\}) = 0.4$$

and

$$\mathrm{P}(\{(B,B),(B,C),(C,B)\}) = 0.1$$

Then,

$$\begin{aligned} \mathrm{P}(\{(B,A),(A,B),(B,B)\}) &= 0.3 + 0.4 - 0.1 \\ &= 0.6 \end{aligned}$$

8. An urn contains five red balls and four white balls. We select three balls at random, without replacement, from the urn.

i. What is the probability that the first ball we selected is a red ball?

ii. What is the probability that all three balls are of the same color?

iii. What is the probability that we have selected more red balls than white balls?

Solution: This exercise is left as self-study.

9. There are 4 women and 3 men working for a company. The boss selects 4 people at random for a holiday bonus. What is the probability that at least one of the bonus winners is a woman? Explain your steps clearly.
Solution: The number of ways the boss picks at least one woman is:

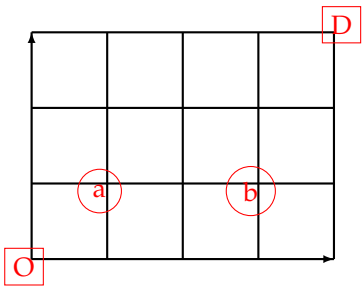
$$\begin{aligned} &= \binom{4}{1}\binom{3}{3} + \binom{4}{2}\binom{3}{2} + \binom{4}{3}\binom{3}{1} + \binom{4}{4}\binom{3}{0} \\ &= 4 \cdot 1 + 6 \cdot 3 + 4 \cdot 3 + 1 \cdot 1 \\ &= 4 + 18 + 12 + 1 \\ &= 35 \end{aligned}$$

The number of ways the boss picks 4 people among 7 is:

$$= \binom{7}{4} = 35$$

So, the requested probability is 1.
Alternatively: as there are only 3 men in the company, after selecting all of them, the boss *must* pick one woman. So, the requested probability is 1.

10. The following is a local map of streets:



You choose a random shortest path from the lower left to upper right corner. That is, at each intersection you either move north or east, unless you are at a boundary.

i. Write a sample space for this random experiment.

ii. What is the probability that your path will pass through a?

iii. What is the probability that your path will pass through b?

iv. What is the probability that your path will pass through a and b?

v. What is the probability that your path will pass through a or b?

Solution:
i. Finding the total count of shortest paths is enough:

$$\frac{(4+3)!}{4!3!} = \frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} = 35 = n(S)$$

ii. O to a: $2!/1!1! = 2$; a to D: $5!/3!2! = 10$; so, O to D through a: 20; $\mathrm{P}(\text{pass through } a) = 20/35$
iii. O to b: $4!/3!1! = 4$; b to D: $3!/1!2! = 3$; so, O to D through b: 12; $\mathrm{P}(\text{pass through } b) = 12/35$
iv. O to a: $2!/1!1! = 2$; a to b: $2!/2!0! = 1$; b to D: $3!/1!2! = 3$; so, O to D through a and b: 6; $\mathrm{P}(\text{pass through } a \text{ and } b) = 6/35$
v. $\mathrm{P}(\text{through } a \text{ or } b)$

$$\begin{aligned}
&= \mathrm{P}(\text{through } a) + \mathrm{P}(\text{through } b) \\
&\quad - \mathrm{P}(\text{through } a \text{ and } b) \\
&= \frac{20}{35} + \frac{12}{35} - \frac{6}{35} \\
&= \frac{26}{35}
\end{aligned}$$

11. From a group of five students, all with different CGPA's, we will select three at random. What is the probability that the student with the highest CGPA, *i.e.* the student who is ranked first with respect to CGPA's, is among the selected students?
Solution: Label the highest GPA student with H and the other four with four L's (L_1,L_2,L_3 and L_4). So, we are selecting 3 among (H, L_1, L_2, L_3, L_4). There are

$$\binom{5}{3} = \frac{5!}{3!2!} = 10$$

ways of making this selection.
One H can be selected in 1 way out of one H. Two L's can be selected in

$$\binom{4}{2} = \frac{4!}{2!2!} = 6$$

ways.
So, the probability that H will be among the selected students is:

$$\frac{6}{10} = \frac{3}{5} = 0.60$$

To verify, consider S that consists of (H, L_1, L_2), (H, L_1, L_3), (H, L_1, L_4), (H, L_2, L_3), (H, L_2, L_4), (H, L_3, L_4), (L_1, L_2, L_3), (L_1, L_2, L_4), (L_1, L_3, L_4), (L_2, L_3, L_4).

12. If we toss a fair coin n times, what is the probability of observing exactly k heads, $k \in \{0, 1, \ldots, n\}$?
Solution: $\mathrm{P}(k \text{ heads}) = \binom{n}{k} 0.5^k 0.5^{n-k} = \binom{n}{k} 0.5^n$

13. We have two urns such that, the first urn contains n_1 balls of which m_1 of them are white and the rest is black. The second urn contains n_2 balls, of which m_2 are white and the rest are black. A ball is chosen at random from the first urn. Without observing the color of the ball we transfer it to the second urn. After that a ball is drawn from the second urn. What is the probability that the ball (the ball drawn from the second urn) is white?
Solution: The answer is:

$$= \frac{m_1}{n_1} \cdot \frac{m_2 + 1}{n_2 + 1} + \frac{n_1 - m_1}{n_1} \cdot \frac{m_2}{n_2 + 1}.$$

14. A coin is tossed until a head is observed. Given that the probability of observing a head in any toss of the coin is p, what is the probability that the coin will be tossed $k \in N$ times?
Solution: As the coin is tossed until a H is observed and as we want to calculate the probability that the experiment will end at k-th toss, in the first $k - 1$ tosses no H is observed. So, the probability requested is

$$= (1 - p)^{k-1} \cdot p$$

where, $(1 - p)^{k-1}$ is the probability of no H in the first $k - 1$ tosses and p is the probability of a H in the k-th toss.

15. A manager has available a pool of 6 employees who could be assigned to a project-monitoring task. 3 of the employees are women and 3 are men. 2 of the men are brothers. The manager is to make the assignment at random so that each of the 6 employees is equally likely to be chosen. Let A be the event chosen employee is a man and B the event chosen employee is one of the brothers.

i. Calculate $\mathrm{P}(A)$

ii. Calculate $\mathrm{P}(B)$

iii. Calculate $\mathrm{P}(A \cap B)$

iv. Calculate $\mathrm{P}(A \cup B)$

Solution: This exercise is left as self-study.

16. An experimenter rolls a fair die twice and records the resulting numbers as x and y. Calculate the probability that x^y is less than or equal y^x. Show your sample space and event sets explicitly.
Solution: This exercise is left as self-study.

17. Two integers are randomly selected from the integers $1, 2, \ldots, 100$. Calculate the probability that their difference is **exactly seven**.
Solution: This exercise is left as self-study.

18. Seven distinct car accidents occurred in a week. What is the probability that they all occurred on the same day?
Solution: $\frac{7}{7^7}$

19. There are M people attending a party. What is the probability that **at least** 2 people among them have the same birthday (month and day)?
Solution: This question is reserved for in-class discussions. The answer is:

$$\begin{aligned} &= 1 - \frac{365 \cdot 364 \cdots (365 - M + 1)}{365^M} \\ &= 1 - \frac{365!}{(365 - M)! 365^M} \end{aligned}$$

for $M \leq 365$. If $M > 365$, the requested probability becomes 1.

20. Ten distinct passengers got into an elevator on the ground floor of a 20-story building. What is the probability that they will all get off at different floors?
Solution: This is just a re-worded version of the birthday question.

21. A project-based course has 20 students. Each student should choose a topic from a given list of topics. Students choose the topic at random from the list without knowing which topic their classmates choose. How many topics should there be in the list so that the probability that at least one pair of students choosing the same topic drops below 0.5?
Solution: This exercise is left as self-study.

22. Two integers are selected from $1, 2, \ldots, n-1$. What is the probability that their sum is larger than n?
Solution: This exercise is left as self-study.

23. A box has 10 balls numbered $1, 2, \ldots, 10$. A ball is picked at random and then a second ball is picked at random from the remaining nine balls. Find the probability that the numbers on the two selected balls differ by **two or more**.

Solution: This exercise is left as self-study.

24. A box has 10 balls numbered $1, 2, \ldots, 10$. Two balls are picked simultaneously, and randomly from the box. Find the probability that the numbers on the two selected balls differ by **two or more**.
Solution: This exercise is left as self-study.

25. A box has 10 balls, 6 of which are black, and 4 of which are white. Three balls are removed from the box, their color unnoted. Find the probability that a fourth ball removed from the box is white.
Solution: Solution requires a careful examination of the first three balls removed. The answer is:

$$\begin{aligned}
&= \frac{\binom{6}{3}\binom{4}{0}}{\binom{10}{3}} \cdot \frac{4}{7} + \frac{\binom{6}{2}\binom{4}{1}}{\binom{10}{3}} \cdot \frac{3}{7} + \frac{\binom{6}{1}\binom{4}{2}}{\binom{10}{3}} \cdot \frac{2}{7} + \frac{\binom{6}{0}\binom{4}{3}}{\binom{10}{3}} \cdot \frac{1}{7} \\
&= \frac{2}{5}.
\end{aligned}$$

26. A machine consists of 4 components linked in parallel, so that the machine fails only if all four components fail. Assume component failures are independent of each other. If the components have probabilities 0.1, 0.2, 0.3 & 0.4 of failing when the machine is turned on, what is the probability that the machine will function when turned on?
Solution: The answer is:

$$\begin{aligned}
&= 1 - (0.1)(0.2)(0.3)(0.4) \\
&= 1 - 0.0024 \\
&= 0.9976
\end{aligned}$$

27. If Sam and Peter are among n men who are arranged at random in a line, what is the probability that exactly k men stand between them?
Solution: Sam, Peter and the other $n-2$ men (totalling to n) can arrange in $n!$ different ways. k can be at least o (no others between Sam and Peter) and at most $n-2$ (all others are between Sam and Peter). Use K to denote the k-people sequence between S and P.
Consider the group $S-K-P$ when S is the first person in the line. This pattern can shift $(n-k-2)$ times until P becomes the last person in the line. So, the group $S-K-P$ can be located in a total of $1+(n-k-2) = n-k-1$ ways across the line.
S and P among themselves can be ordered in 2! ways.
$(n-2)$ people (those who are not S and P) can be ordered in $(n-2)!$ ways.
So, S and P and k people in between can be ordered in:

$$(n-k-1)\cdot 2!\cdot (n-2)!$$

ways.
So, the probability asked is:

$$\frac{2(n-k-1)(n-2)!}{n!}, 0 \leq k \leq n-2$$

which is equal to:

$$\frac{2(n-k-1)}{n(n-1)}, \quad 0 \leq k \leq n-2.$$

Some selected cases will be covered in class for further clarification.

28. A business person lives near a subway station. She has two offices: A and B. To go to A, she takes a train on the uptown side of the platform; to go to B, she takes a train on the downtown side of the platform. Since she is indifferent between the two offices, she simply takes the first train that comes along. In this way she lets chance determine whether she goes to A or to B. She reaches the subway platform at a random moment every day. A and B trains arrive at the station equally often, which is every 10 minutes. Yet for some obscure reason she finds herself spending most of her time at A: in fact, on the average she goes there nine out of every ten times. Can you think of a reason why the chance factor so heavily favors A? Explain in sufficient detail using a proper terminology and mathematical expressions when needed.
Solution: Consider the possibility that one of the trains is passing at 09:00, 09:10, 09:20, ... and the other at 09:09, 09:19, 09:29, ...

2.6 Conditional probability

Let A and B be two events. The conditional probability of event A given that B has occurred is denoted by $\mathrm{P}(A \mid B)$, and is defined as:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}, \textit{if } \mathrm{P}(B) > 0$$

Similarly,

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)}, \textit{if } \mathrm{P}(A) > 0$$

The information that "given that event B has occurred" is called the **prior information**. Availability of prior information restricts our sample space from S to B; so, when we study the probability of "A given B" we simply consider the part of A that is also in B.

Based on our new definition and the earlier knowledge of probability calculations:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\frac{n(A \cap B)}{n(S)}}{\frac{n(B)}{n(S)}} = \frac{n\,(A \cap B)}{n\,(B)}$$

Also notice & establish:

$$\mathrm{P}(A \mid B)\,\mathrm{P}(B) = \mathrm{P}(A \cap B) = \mathrm{P}(B \mid A)\,\mathrm{P}(A)$$

Let $E_1, E_2, \ldots, E_K$ be K mutually exclusive & collectively exhaustive events. This means:

- $E_i \cap E_j = \emptyset, \forall i \neq j$
- $E_1 \cup E_2 \cup \cdots \cup E_K = S$

Then, probability of any event A can be written as:

$$\begin{aligned}
\mathrm{P}(A) &= \mathrm{P}(A \cap S) \\
&= \mathrm{P}(A \cap (E_1 \cup E_2 \cup \cdots \cup E_K)) \\
&= \mathrm{P}((A \cap E_1) \cup (A \cap E_2) \cup \cdots \cup (A \cap E_K)) \\
&= \mathrm{P}(A \cap E_1) + \mathrm{P}(A \cap E_2) + \cdots + \mathrm{P}(A \cap E_K) \\
&= \mathrm{P}(A \mid E_1)\,\mathrm{P}(E_1) + \mathrm{P}(A \mid E_2)\,\mathrm{P}(E_2) + \cdots + \mathrm{P}(A \mid E_K)\,\mathrm{P}(E_K)
\end{aligned}$$

Think about why **mutual exclusion** is a key property while establishing this result. Think also about the essence of **collective exhaustion**.

2.7 *Bayes' Theorem*

Let A and B be two events. Then,

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A)\,\mathrm{P}(A)}{\mathrm{P}(B)}, \textit{if } \mathrm{P}(B) > 0$$

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(A \mid B)\,\mathrm{P}(B)}{\mathrm{P}(A)}, \textit{if } \mathrm{P}(A) > 0$$

Indeed, this is very intuitive. Bayes' theorem provides a way of revising conditional probabilities by using available information. It also provides a procedure for determining how probabilities should be adjusted in the light of additional information. If you're interested in machine learning, note that one of its sub-branches is built solely on this theorem.

Let $E_1, E_2, \ldots, E_K$ be K mutually exclusive & collectively exhaustive events & let A be another event of interest. Then,

$$\mathrm{P}(E_i \mid A) = \frac{\mathrm{P}(A \mid E_i)\,\mathrm{P}(E_i)}{\mathrm{P}(A)}$$

or, equivalently:

$$\mathrm{P}(E_i \mid A) = \frac{\mathrm{P}(A \mid E_i)\,\mathrm{P}(E_i)}{\mathrm{P}(A \mid E_1)\,\mathrm{P}(E_1) + \mathrm{P}(A \mid E_2)\,\mathrm{P}(E_2) + \cdots + \mathrm{P}(A \mid E_K)\,\mathrm{P}(E_K)}$$

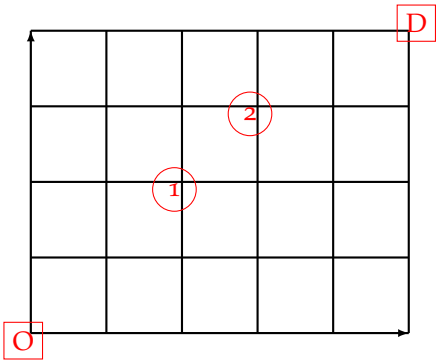
Regarding the denominator of the previous expression, remember that:

$$\begin{aligned}
&\mathrm{P}(A \mid E_1)\,\mathrm{P}(E_1) + \mathrm{P}(A \mid E_2)\,\mathrm{P}(E_2) + \cdots + \mathrm{P}(A \mid E_K)\,\mathrm{P}(E_K) \\
&= \mathrm{P}(A \cap E_1) + \mathrm{P}(A \cap E_2) + \cdots + \mathrm{P}(A \cap E_K) \\
&= \mathrm{P}((A \cap E_1) \cup (A \cap E_2) \cup \cdots \cup (A \cap E_K)) \\
&= \mathrm{P}(A \cap (E_1 \cup E_2 \cup \cdots \cup E_K)) \\
&= \mathrm{P}(A \cap S) \\
&= \mathrm{P}(A)
\end{aligned}$$

2.3 EXERCISES

1. Consider an epidemic where K out of every 100 people in the community is infected. Since the symptoms are not openly manifesting, a blood test is necessary to diagnose whether a given person is infected or not, and we know L out of every 100 people in the community has been tested in a perfectly random fashion. We also know that the test gives an erroneous result 1 out of every 20 times. The District Governor has made a press release today saying that the infection rate in the community is found to be 10% after some testing. **Given** this information, what is the rate of actual infections? How does your estimate depend on K and L? Answer using your knowledge of conditional probabilities and Bayes theorem.
   Solution: This exercise is left as self-study.

2. Below is the map of a city in which tourists are trying to go from the lower-left corner to upper-right corner via a shortest path, where line segments are the streets to be followed.



Supposing that a selected tourist knows very well what a shortest path is, what is the probability that his shortest path will pass through the point marked with (1) given that his shortest path passes through the point marked with (2)? In your solution, make sure you have clearly shown the sample space, the events of interest, cardinalities of the sets involved and calculations.
Solution: This exercise is left as self-study.

2.8 *Independence of events*

Let A and B be two events. If

$$\mathrm{P}(A) = \mathrm{P}(A \mid B)\,, \textit{when}\ \mathrm{P}(B) > 0$$

or

$$\mathrm{P}(B) = \mathrm{P}(B \mid A)\,, \textit{when}\ \mathrm{P}(A) > 0$$

then A and B are statistically independent events, and vice versa.

If A and B are independent events, then

$$\mathrm{P}(A) = \mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

and

$$\mathrm{P}(B) = \mathrm{P}(B \mid A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)}$$

both of which to yield:

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$$

Checking for either the first two expressions or the third one, we can test for the independence of two events.

The events $E_1, E_2, \ldots, E_K$ are mutually statistically independent if and only if

$$\mathrm{P}(E_1 \cap E_2 \cap \cdots \cap E_K) = \mathrm{P}(E_1)\,\mathrm{P}(E_2) \cdots \mathrm{P}(E_K)$$

In a nutshell

When two events are independent, **we say they are independent**; when they are not independent, **we say 'they are not independent'**, but **not say 'they are dependent'**. The reason for such a naming is that **independence** is a mathematically defined property where **dependence** is not.

2.4 EXERCISES

1. Let $S = \{1,2,3,4\}$ and assume each point has a probability of $\frac{1}{4}$. Set $A = \{1,2\}$, $B = \{1,3\}$, $C = \{1,4\}$. Show that the pairs of events A and B, A and C & B and C are independent.
Solution:

$$S = \{1,2,3,4\}$$
$$\mathrm{P}(\{1\}) = \mathrm{P}(\{2\}) = \mathrm{P}(\{3\}) = \mathrm{P}(\{4\}) = 1/4.$$

Consider $A = \{1,2\}$ and $B = \{1,3\}$.
check whether $\mathrm{P}(A \mid B) = \mathrm{P}(A)$.

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$
$$A \cap B = \{1\}, \mathrm{P}(A \cap B) = 1/4, \mathrm{P}(B) = 1/2$$
$$\text{So, } \mathrm{P}(A \mid B) = \frac{1/4}{1/2} = 1/2.$$
$$\mathrm{P}(A) = 1/2. \text{ So, } \mathrm{P}(A \mid B) = \mathrm{P}(A), \text{ indicating}$$

that events A and B are independent.
Checking for the pairs A and C and B and C are straightforward.
Lesson out of this exercise is not to confuse the concepts of independence and mutual exclusion.

2. Assume that, for a given family, any birth is equally likely to be a girl or boy. Given this, answer the following questions:

i. If the family has 3 kids, what is the probability that exactly one of them is a girl?

ii. If the family has 3 kids, what is the probability that at least one of them is a girl?

iii. If the family would like to have at least one girl with probability 0.95 (or more), at least how many kids should they plan on having?

Solution: Regarding 3 kids, all possible gender configurations can be listed as (B,B,B), (B,B,G), (B,G,B), (G,B,B), (B,G,G), (G,B,G), (G,G,B), (G,G,G), B denoting a boy and G denoting a girl.

1. In three out of these 8 configurations, the family has exactly one baby girl. So, the answer is $3/8$.

2. $7/8$

3. Suppose the family will have at least one baby girl among n kids. Then,

$$\begin{aligned} 1 - \frac{1}{2^n} &\geq 0.95 \\ 0.05 &\geq \frac{1}{2^n} \\ 2^n &\geq 20 \end{aligned}$$

is found; indicating that n must be at least 5.

3. Forty-four percent of the students at a given university are males. Ten percent of the students are engineering students. Two percent of the students are men in engineering. If a student is selected at random, find the conditional probability that:

i. the student is male given that the student is an engineering student

ii. the student is an engineering student given that the student is male

Solution: This exercise is left as self-study.

4. Two fair dice are rolled. What is the probability that at least one lands on 6 given that the dice land on different numbers?
Solution: This exercise is left as self-study.

5. Consider an urn containing 8 white and 4 red balls. Four balls are to be drawn with replacement. What is the conditional probability that the first and third balls drawn will be white given that the sample drawn contains exactly 3 white balls?
Solution: This exercise is left as self-study.

6. The probability that a new machine functions for over 30 months is 0.8, the probability that it functions for over 60 months is 0.4. If a new machine is still working after 30 months, what is the probability that its total life will exceed 60 months?
Solution: T denoting the duration of time during which the machine functions, we are given that $\mathrm{P}(T > 30) = 0.8$ and $\mathrm{P}(T > 60) = 0.4$.

$$\begin{aligned} \mathrm{P}(T > 60 \,|\, T > 30) &= \frac{\mathrm{P}(T > 60)}{\mathrm{P}(T > 30)} \\ &= \frac{0.4}{0.8} \\ &= 0.5 \end{aligned}$$

7. Let A, B, and C be three events such that A and B are independent, A and C are mutually exclusive, *i.e.* $A \cap C = \emptyset$, and $\mathrm{P}(A) = 0.4$, $\mathrm{P}(B) = 0.3$, $\mathrm{P}(C) = 0.2$, $\mathrm{P}(C \,|\, B) = 1/3$. Find:

i. $P(A \cup C)$

ii. $P(A \cup B)$

iii. $P(B \cup C)$

iv. $P(A \cup B \cup C)$

Solution: This exercise is left as self-study.

8. We have 3 boxes. Box labeled 1 contains 3 white and 5 black balls. Box labeled 2 contains 2 white and 6 black balls and box labeled 3 contains 4 white and 4 black balls. We first select a box at random, then choose a ball from that box at random.

i. What is the probability that the chosen ball is white.

ii. Given that the ball chosen is white, what is the probability that the ball was chosen from box 1.

iii. Are the events of choosing a white ball and choosing box 1 independent?

iv. Are the events of choosing a white ball and choosing box 2 independent?

Solution: This exercise is left as self-study.

9. A red die and a white die are rolled.

i. Are the event that the sum of the numbers is equal to 7 and the event the red die is 1 independent?

ii. Are the event that the sum of the numbers is equal to 7 and the event that at least one of the dice turned up 1 independent?

Solution: This exercise is left as self-study.

10. An Economics instructor would like to find out the percentage of students who cheat in a homework. Since students does not like to look like a cheater in the eye of the instructor, the instructor constructed the following procedure in order to eliminate the incentive to misreport:
The student is asked to roll a die and without showing the result to the instructor answer the question "Did you cheat in the homework?" as YES or NO as follows: If the number shown is 1*, then answer as NO, no matter what the true answer is. If the number shown is* 6*, then answer as YES, no matter what the true answer is. If the number shown is* 2*,* 3*,* 4 *or* 5*, then answer truthfully.*
After surveying a large group of students, using this method, the instructor calculated that 60% of the students answered "yes. What can you say about the percentage of students who cheated in the homework? That is, what is the probability that a randomly selected student has cheated?
Solution: This exercise is left as self-study.

11. In a simple experimental setup, the subjects (individuals) roll a fair die and toss a coin in a hidden chamber. If the number shown by die is 1, they report the coin toss as Head regardless of the actual outcome. If the number shown by die is 2 or 3, they report the coin toss as Tail regardless of the actual outcome. If the number shown by die is 4, 5 or 6, they report the coin toss result truthfully. After experimenting with a large number of subjects, we observe that 47% of the subjects reported Head. Based on this information, can we say that the coin is fair?
Solution: H: Head event T: Tail event D_1: Die shows 1 event D_{23}: Die shows 2 or 3 D_{456}: Die shows 4, 5 or 6 x: True probability of a Head. The question gives us:

$$\begin{aligned} \mathrm{P}(H) &= \mathrm{P}(H \mid D_1)\,\mathrm{P}(D_1) + \mathrm{P}(H \mid D_{23})\,\mathrm{P}(D_{23}) + \mathrm{P}(H \mid D_{456}) \\ 0.47 &= 1 \cdot \frac{1}{6} + 0 \cdot \frac{2}{6} + x \cdot \frac{3}{6} \\ 0.5x &= 0.47 - \frac{1}{6} \\ x &= 0.61 \end{aligned}$$

Since $0.61 \neq 0.5$, the coin is not fair.

12. Firm A is considering whether it should submit a bid for a new shopping center. In the past, Firm A's competitor, Firm B, has submitted bids 70% of the time. If Firm B does not bid, the probability that Firm A will get the job is 0.5. If Firm B does bid, the probability that Firm A will get the job is 0.25.

i. What is the probability that Firm A will get the job?

ii. If we are told that Firm A got the job what is the probability that Firm B did not bid?

Solution: This exercise is left as self-study.

13. 90% of students who have not properly prepared for an exam receive a low grade. 10% of students who prepared well also receive a low grade. We know that 3 out of every 4 students in a college do prepare well for their exams.

i. What is the probability that a randomly selected student in this university is a low-scorer?

ii. If we know a student is a low-scorer, what is the probability that she/he has not properly prepared for her/his exam?

Solution: This exercise is left as self-study.

14. We have 3 boxes. Box labeled 1 contains 3 white and 5 black balls. Box labeled 2 contains 2 white and 6 black balls and box labeled 3

contains 4 white and 4 black balls. We first select a box at random, then choose a ball from that box at random.

i. What is the probability that the chosen ball is white?

ii. Given that the ball chosen is white, what is the probability that the ball was chosen from box 1?

Solution: This exercise is left as self-study.

15. A certain disease is seen in 1 out of every M people in a society. Luckily, there is a diagnostic blood test to detect the disease which yields a wrong result in 1 out of every N applications.

i. If a person is said to have had the disease after the blood test, what is the probability that she/he actually is sick? Express your answer in terms of M and N.

ii. Under which configuration of M and N the test is more reliable?

Solution: This exercise is left as self-study.

16. Three students are given the same problem to solve. They work on the problem independently and student 1, 2, and 3 have probabilities 0.8, 0.7, and 0.6 of solving it, respectively.

i. What is the probability that none of them solves the problem?

ii. What is the probability that the problem will be solved (that is, by at least one of them)?

Solution: This exercise is left as self-study.

17. There is a 30% chance that it rains on any particular day. What is the probability that there is at least one rainy day within a 7-day period? Given that there is at least one rainy day, what is the probability that there are at least two rainy days?
Solution: This exercise is left as self-study.

18. A symmetric die is rolled 3 times. If it is known that face 1 appeared at least once, what is the probability that it appeared exactly once?
Solution: This exercise is left as self-study.

19. Suppose two identical and perfectly balanced coins are tossed once:

i. Find the conditional probability that both coins show a head given that the first shows a head.

ii. Find the conditional probability that both are heads given that at least one of them is a head.

Solution: This exercise is left as self-study.

20. Suppose that the population of a certain city is 40% male and 60% female. Suppose also that 50% of the males and 30% of the females smoke. Find the probability that a smoker is male.
Solution: This exercise is left as self-study.

21. Winner, Loser and Tailung are three cats. Winner jumps on the kitchen table 5 times in a typical day, Loser 10 times and Tailung 15 times. Each time Winner jumps on the table, there is a 50% probability that she will push something off to floor. For Loser and Tailung this probability is 30% and 10%, respectively. Suppose now, I've been watching TV in the living room armchair when I heard the noise of something fell from the kitchen table. What is the probability that Winner has done it?
Solution: This exercise is left as self-study.

2.9 Bivariate probabilities

Consider A_1, A_2, ..., A_K which are K mutually exclusive and collectively exhaustive events in a random experiment. Consider also B_1, B_2, ..., B_L which are L mutually exclusive and collectively exhaustive events in another random experiment. Now, consider the simultaneous happenings from the two random experiments; this is nothing but a joint view of the two random experiments & by definition another random experiment. In this joint random experiment, an event C_{ij} is defined as:

$$C_{ij} = A_i \cap B_j, i = 1, 2, \ldots K; j = 1, 2, \ldots, L$$

Of course,

$$\sum_i \sum_j \mathrm{P}(C_{ij}) = 1$$

As there are two variables here, namely A_i's and B_i's, this setup is called a bivariate probability setup. Considering all its outcomes, we'll have:

	B_1	B_2	...	B_L
A_1	$\mathrm{P}(A_1 \cap B_1)$	$\mathrm{P}(A_1 \cap B_2)$	...	$\mathrm{P}(A_1 \cap B_L)$
A_2	$\mathrm{P}(A_2 \cap B_1)$	$\mathrm{P}(A_2 \cap B_2)$	...	$\mathrm{P}(A_2 \cap B_L)$
...	...	...	...	...
A_K	$\mathrm{P}(A_K \cap B_1)$	$\mathrm{P}(A_K \cap B_2)$	...	$\mathrm{P}(A_K \cap B_L)$

As noted earlier, $\sum_i \sum_j \mathrm{P}(C_{ij}) = 1$; so, the probabilities given in the table add up to 1.

2.10 *Joint, marginal and conditional probabilities*

In the bivariate setup:

$$\mathrm{P}\left(A_i \cap B_j\right)$$

are the **joint probabilities**.

$$\mathrm{P}(A_i) = \sum_{j=1}^{L} \mathrm{P}\left(A_i \cap B_j\right), \forall i$$

$$\mathrm{P}\left(B_j\right) = \sum_{i=1}^{K} \mathrm{P}\left(A_i \cap B_j\right), \forall j$$

are the **marginal probabilities**.

$$\mathrm{P}\left(A_i \mid B_j\right) = \frac{\mathrm{P}\left(A_i \cap B_j\right)}{\mathrm{P}\left(B_j\right)}$$

and,

$$\mathrm{P}\left(B_j \mid A_i\right) = \frac{\mathrm{P}\left(A_i \cap B_j\right)}{\mathrm{P}(A_i)}$$

are the **conditional probabilities**, using our previous knowledge, and notations. Referring to our table of bivariate probabilities:

- Joint probabilities are the figures given in the table
- Column and row sums are the marginal probabilities
- To calculate the conditional probabilities for a column or row, we divide the joint probabilities in the column or row by the respective marginal probability of the column or row

2.11 *Independence of Events*

Let A and B be a pair of events, where A is broken into K mutually exclusive and collectively exhaustive events

$$A_1, A_2, \ldots, A_K$$

and B is broken into L mutually exclusive and collectively exhaustive events

$$B_1, B_2, \ldots, B_L$$

If every A_i is statistically independent of every B_j, then A and B are independent events.

2.5 EXERCISES

1. The following cross tabulation has been formed using a corporation's data:

(Frequency)		
	Promotion status last year	Promotion status last year
Gender	Promoted	Not promoted
Male	60	30
Female	70	105

i. Are gender events and promotion events independent?

ii. Are gender and promotion independent?

Solution: The following cross tabulation has been formed using a corporation's data:

(Frequency)		
	Promotion status last year	Promotion status last year
Gender	Promoted	Not promoted
Male	60	30
Female	70	105

1. Are gender events and promotion events independent?
2. Are gender and promotion independent?

3 *Random variables*

Earlier we entered the world of data and learned a rich collection of descriptive statistics, followed by developing a solid understanding of the probability theory. The knowledge of this chapter will now allow us to understand and practice the probability theory by means of the standard tools of calculus.

3.1 Random Variables

Formally, **a random variable is a function from a sample space S to the set of real numbers** $\mathbb{R}$. Note at the very beginning that we denote random variables with uppercase letters and their particular values with the corresponding lowercase letters. So, a random variable X can take a value x.

When we define the outcomes of a random experiment via a random variable, we can generalize the very structure of the experiment and get rid off our context-dependence.

After internalizing the knowledge of this chapter, we will be able to state and solve a long array of problems with more formalism. In the rest of this chapter, we will first study the concepts of 'Cumulative distribution function' (CDF) and 'Probability distribution function' (PDF). At the cost of a spoiler, we can say that CDF and PDF are the theoretical counterparts of **O-give** and **histogram**, respectively. Secondly, we will study the concepts of the **Expected value** and **Variance** along with their key properties. While doing so, we will create and refer to a number of *ad hoc* random variables. *Ad hoc* is a Latin phrase meaning literally '**to this**'. In English, it is used to describe 'something that has been formed or used for a special and immediate purpose, without previous planning'. In that, until we reach the section entitled 'Random variablest and distributions: Discrete probability laws', we will be creating, using and disposing several random variables that serve our specific scientific/ technical purposes.

On one hand, our discussion and use of those *ad hoc* random variables and distributions will prove quite useful to handle a long list of probabilistic or statistical cases/problems. On the other hand, staying

'*ad hoc*' is not good for a full-fledged practice of science, as our journey will reveal. As a matter of fact, a rich set of probability laws (Discrete probability laws and Continuous probability laws) will allow us to categorize, model and solve a variety of real-world statistical problems in a sound as well as practical fashion. Note that the use of the term '**Law**' may not be the best alternative available in scientific nomenclature; yet, it is part of the tradition. Those who are not comfortable with the use of the term 'Law' may replace it with the term '**Distribution**'. As an example, 'Uniform probability, law' and 'Uniform probability distribution' are simply the same thing as each other.

Now, we can proceed with our quest to learn things. Recall our repeatedly used random expertment of 'tossing a fair coin'. Head and Tail (or, H and T) being the two sides of a coin, we already know the following:

$$\begin{aligned} &S = \{H, T\} \\ &\mathrm{P}(\text{Coin shows a Head}) = \mathrm{P}(H) = 1/2 \\ &\mathrm{P}(\text{Coin shows a Tail}) = \mathrm{P}(T) = 1/2 \end{aligned}$$

Upon these, we are allowed to define and study everything that is relevant. Despite its simplicity, such an approach lacks one important feature: mathematical generalization. Indeed, the real-world hosts a bunch of random experiments with two basic outcomes; a student to pass or to fail an exam, a patient to survive or not survive a sickness, an asteroid to hit or not to hit our planet Earth, and so on. Notice that each of these experiments look like tossing a coin. Furthermore, if the probability of passing the exam, the probability to survive and the probability to hit the Earth are equal to $1/2$, these random experiments are 'identical' with tossing a coin, except for the details of naming. So, why not to suggest a random variable X along with its probability distribution to address all these random experiments?

Consider $X \in \{0, 1\}$ (or $x \in [0, 1]$) for which $\mathrm{P}(x = 0) = 1/2$ and $\mathrm{P}(x = 1) = 1/2$. This is nothing but a direct equivalent of the random experiment of tossing a coin, without referring to coin explicitly. Let us leave this discussion for a while to consider another random experiment.

Now consider (or recall from our in-class discussions) the random experiment of drawing a number from the interval $[1, 5]$ in a fully blind-folded fashion; so, there are infinitely many basic outcomes, which are the real numbers from 1 to 5 (as one cannot guarantee to pick intergers only, when blind-folded). With regard to this case, we already know the following:

$$S = \{x \mid x \in [1, 5], x \in \mathbb{R}\}$$

or simply,

$$S = [1, 5]$$

and

$$\mathrm{P}(\text{The number picked is } x) = 0, \forall x$$

The final statement should trivially follow from chapter 4 (and should not sound weird to your ears anymore).

Following a similar agenda, to what we used in the case of tossing a fair coin above, we can say that the random experiment of picking a real number from $[1, 5]$, from $[2, 6]$, from $[3, 7]$ or from $[1001, 1005]$ should not differ. You may confirm this expectation once you have measured the length (size) of each of these intervals as 4 .

Define now $Y \in [1, 5]$ and leave this discussion aside until we cover the following definitions. Each of these definitions is crucial for our subsequent study of probability theory and statistics. Combining/pairing with your in-class notes, use these definitions to come up with a holistic picture of the things (objects) involved.

3.2 *Cumulative distribution function: CDF*

The cumulative distribution function or **CDF** of a random variable X is denoted by $F_X(x)$ or $F(x)$, and is defined as:

$$F_X(x) = \mathrm{P}_X(X \leq x)$$

or

$$F(x) = \mathrm{P}(X \leq x)$$

In a nutshell

The function $F(x)$ is a CDF if and only if it satisfies the following conditions:

- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
- $F(x)$ is **non-decreasing** in x
- $F(x)$ is **right-continous**, *i.e.* $\lim_{x \to x_0^+} F(x) = F(x_0)\,, \forall x_0$

3.3 *Continuous and discrete random variables*

A random variable X is continuous if $F(x)$ is a continuous function of x. A random variable X is discrete if $F(x)$ is a step function of x.

Additionally, the random variables X and Y are identically distributed if for every set A,

$$\mathrm{P}(X \in A) = \mathrm{P}(Y \in A)$$

where this does not necessarily mean $X = Y$. If X and Y are identically distributed,

$$F_X(x) = F_Y(x)\,, \forall x$$

3.4 *Probability distribution functions*

3.4.1 *Probability distribution function: Case of discrete random variables*

The probability distribution function **PDF** of a discrete random variable X is given by:

$$f_X(x) = \mathrm{P}(X = x)\,, \forall x$$

or

$$f(x) = \mathrm{P}(X = x)\,, \forall x$$

For discrete random variables, probability distribution function is also called the probability mass function.

3.4.2 *Probability distribution function: Case of continuous random variables*

The probability distribution function **PDF** of a continuous random variable X is the function $f_X(x)$ that satisfies:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt, \forall x$$

For continuous random variables, probability distribution function is also called the probability density function.

In a nutshell

X has a distribution given by $f(x)$ is abbreviated by $X \sim f(x)$, where we read the symbol "$\sim$" as **is distributed as**.

In a nutshell

The function $f(x)$ is a PDF if and only if it satisfies the following conditions:

- $f(x) \geq 0, \forall x$
- $\sum_x f(x) = 1$, *(Discrete X)*
  $\int_{-\infty}^{\infty} f(x)\,dx = 1$, *(Continuous X)*

3.4.3 *Connection between CDF and PDF*

Using the Fundamental Theorem of Calculus, if $f(x)$ is continuous

$$\frac{d}{dx}F(x) = f(x)$$

The analogy with the discrete case is almost exact. We "add up" the point probabilities $f(x)$ to obtain interval probabilities $F(x)$. With a slight abuse of the notation:

$$\Delta F(x) = f(x)$$

In a nutshell

Numerical meaning of PDF: Values of $f_X(x)$ or $f(x)$ are the probabilities if X is discrete. However, when X is continuous, $f_X(x)$ or $f(x)$ values are **not** probabilities. They are, rather, likelihoods or density ordinates. So, notice that:

$$\mathrm{P}(X = x) = f_X(x)\,, \text{ (Discrete X)}$$

and

$$\mathrm{P}(X = x) = 0, \text{ (Continous X)}$$

Having been exposed to formal definitions of the functions and operators involved, now we will reconsider the random experiment of tossing a fair coin (discrete random variable case) and random experiment of picking a number from $[1, 5]$ (continuous random variable case), in that order.

Using our newly acquired knowledge, we can now define the following:

$$X \sim f(x)$$

$$f(x) = \begin{cases} 1/2, & x = 0 \\ 1/2, & x = 1 \end{cases}$$

$$F(x) = \begin{cases} 0, & x < 0 \\ 1/2, & 0 \leq x < 1 \\ 1, & x \leq 1 \end{cases}$$

Here, X is nothing but the random variable that describes the outcomes of the random experiment of tossing a fair coin.

Consider also:

$$Y \sim g(y)$$

$$g(y) = \begin{cases} 1/4, & 1 \leq y \leq 5 \\ 0, & \text{otherwise} \end{cases}$$

$$G(y) = \begin{cases} 0, & y < 1 \\ \frac{y-1}{4}, & 1 \leq y \leq 5 \\ 1 & 5 \leq y \end{cases}$$

You must have noticed that Y is the random variable that describes the outcomes of the random experiment of picking a number from $[1, 5]$.

Since $F(x)$ is a discrete function, X is a discrete random variable and since $G(y)$ is a continuous function, Y is a continuous random variable.

In a nutshell

A function $f(x)$ is a continuous function of x if:

$$\lim_{x \to x_0^-} f(x) = \lim_{x \to x_0^+} f(x) = f(x_0), \forall x_0$$

A function $f(x)$ is a left-continuous function of x if:

$$\lim_{x \to x_0^-} f(x) = f(x_0), \forall x_0$$

A function $f(x)$ is a right-continuous function of x if:

$$\lim_{x \to x_0^+} f(x) = f(x_0), \forall x_0$$

A function $f(x)$ which is both left-continuous and right-continuous in its domain is a continuous function.

In the cases of $X \sim f(x)$ and $Y \sim g(y)$ above, observe that:

$$\begin{aligned} f(-0.5) &= 0 \\ f(0.5) &= 0 \\ f(1.5) &= 0 \\ g(-0.5) &= 0 \\ g(0) &= 0 \\ g(1.5) &= 1/4 \\ g(2.5) &= 1/4 \\ g(5.1) &= 0 \\ g(10) &= 0 \end{aligned}$$

Also, notice that:

$$\mathrm{P}(X = 0) = f(0) = 1/2$$

and

$$\mathrm{P}(Y = 3) = 0$$

while

$$g(3) = 1/4$$

Your mind should be crystal clear in this distinction of probabilities and likelihoods for continuous random variables.

But, how we define/refer to probabilities and calculate them in the case of continuous random variables? The answer should be trivial to you: since the point probabilities are all zero for a continuous random variable, we can talk about the 'probabilities of intervals' only. Then, the following calculation for the random variable Y above is legitimate:

$$\begin{aligned} P(2 \leq Y \leq 4) &= \int_2^4 g(y)dy \\ &= \int_2^4 \frac{1}{4}dy \\ &= \frac{y}{4}\Big|_2^4 \\ &= \frac{4}{4} - \frac{2}{4} \\ &= \frac{2}{4} \\ &= \frac{1}{2} \end{aligned}$$

Alternatively,

$$\begin{aligned} P(2 \leq Y \leq 4) &= G(4) - G(2) \\ &= \frac{4-1}{4} - \frac{2-1}{4} \\ &= \frac{3}{4} - \frac{1}{4} \\ &= \frac{2}{4} \\ &= \frac{1}{2} \end{aligned}$$

yields the same solution. Now, give an effort to show these solutions on the graphs of $g(y)$ and $G(y)$.

3.5 *Expected Value*

The expected value or mean of a random variable X is:

$$E(X) = \sum_x xf(x), \textit{ if } X \textit{ is discrete}$$

and

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx, \textit{if } X \textit{ is continuous}$$

provided that the sum or integral exists.

In a nutshell

The expected value or mean of a random variable $g(X)$ is:

$$\mathrm{E}(g(X)) = \sum_x g(x) f(x), \textit{ if } X \textit{ is discrete}$$

and

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f(x)\, dx, \textit{if } X \textit{ is continous}$$

provided that the sum or integral exists.

In a nutshell

Let X be a random variable and a, b and c be constants. Then, for any $g_1(x)$ and $g_2(x)$ whose expectations exist:

- $\mathrm{E}(ag_1(x) + bg_2(x) + c) = a\,\mathrm{E}(g_1(x)) + b\,\mathrm{E}(g_2(x)) + c$
- If $g_1(x) \geq 0, \forall x$, then $\mathrm{E}(g_1(x)) \geq 0$
- If $g_1(x) \geq g_2(x), \forall x$, then $\mathrm{E}(g_1(x)) \geq \mathrm{E}(g_2(x))$
- If $a \leq g_1(x) \leq b, \forall x$, then $a \leq \mathrm{E}(g_1(x)) \leq b$

3.6 *Variance and standard deviation*

The variance of a random variable X is defined as:

$$\mathrm{Var}(X) = \mathrm{E}(X - \mathrm{E}(X))^2$$

For a discrete random variable X:

$$\mathrm{Var}(X) = \sum_x (x - \mathrm{E}(X))^2 f(x)$$

where

$$\mathrm{E}(X) = \sum_x x f(x)$$

For a continuous random variable X:

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mathrm{E}(X))^2 f(x)\, dx$$

where

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

The positive square root of $\text{Var}(X)$ is the standard deviation of X. If X is a random variable with finite variance, then for any constants a and b:

$$\text{Var}(aX+b) = a^2 \text{Var}(X)$$

In a nutshell

An alternative and easier formula for the variance is given by:

$$\text{Var}(X) = \text{E}\left(X^2\right) - (\text{E}(X))^2$$

The simple proof is as follows:

$$\begin{aligned}\text{Var}(X) = \text{E}(X-\text{E}(X))^2 &= \text{E}\left(X^2 - 2X\,\text{E}(X) + \left(\text{E}(X)^2\right)\right)\\ &= \text{E}\left(X^2\right) - 2\,\text{E}(X\,\text{E}(X)) + \text{E}(\text{E}(X))^2\\ &= \text{E}\left(X^2\right) - 2\,\text{E}(X)\,\text{E}(X) + (\text{E}(X))^2\\ &= \text{E}\left(X^2\right) - (\text{E}(X))^2\end{aligned}$$

Regarding the same X and Y defined above, we can now study/compute the Expected value and the Variance:

$$\begin{aligned}\text{E}(X) &= \sum_x x f(x)\\ &= 0\cdot\frac{1}{2} + 1\cdot\frac{1}{2}\\ &= \frac{1}{2}\end{aligned}$$

$$\begin{aligned}\text{Var}(X) &= \sum_x (x-\text{E}(X))^2 f(x)\\ &= \left(0-\frac{1}{2}\right)^2\cdot\frac{1}{2} + \left(1-\frac{1}{2}\right)^2\cdot\frac{1}{2}\\ &= \frac{1}{4}\cdot\frac{1}{2} + \frac{1}{4}\cdot\frac{1}{2}\\ &= \frac{1}{4}\end{aligned}$$

$$\begin{aligned}\text{E}\left(X^2\right) &= \sum_x x^2 f(x)\\ &= 0^2\cdot\frac{1}{2} + 1^2\cdot\frac{1}{2}\\ &= \frac{1}{2}\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}\left(X^2\right) - (\mathrm{E}(X))^2 \\ &= \frac{1}{2} - \left(\frac{1}{2}\right)^2 \\ &= \frac{1}{2} - \frac{1}{4} \\ &= \frac{1}{4}\end{aligned}$$

$$\begin{aligned}\mathrm{E}(Y) &= \int_{-\infty}^{\infty} y g(y) dy \\ &= \int_1^5 y \frac{1}{4} dy \\ &= \left.\frac{y^2}{8}\right|^5 \\ &= \frac{25}{8} - \frac{1}{8} \\ &= \frac{24}{8} \\ &= 3\end{aligned}$$

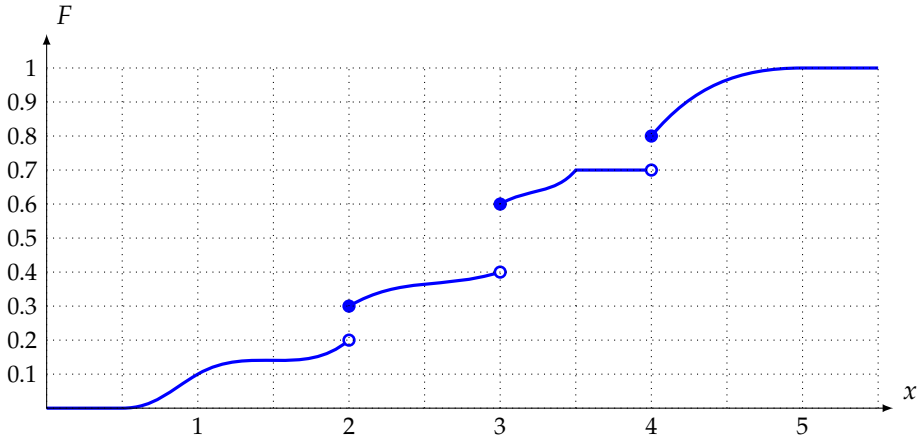
$$\begin{aligned}\mathrm{Var}(Y) &= \int_{-\infty}^{\infty} (y - \mathrm{E}(Y))^2 g(y) dy \\ &= \int_1^5 (y-3)^2 \frac{1}{4} dy \\ &= \left.\frac{(y-3)^3}{12}\right|_1^5 \\ &= \frac{8}{12} - \left(-\frac{8}{12}\right) \\ &= \frac{16}{12} \\ &= \frac{4}{3}\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left(Y^2\right) &= \int_{-\infty}^{\infty} y^2 g(y)dy \\
&= \int_1^5 y^2 \frac{1}{4} dy \\
&= \frac{y^3}{12}\Big|_1^5 \\
&= \frac{125}{12} - \frac{1}{12} \\
&= \frac{124}{12} \\
&= \frac{31}{3}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\,(Y) &= \mathrm{E}\left(Y^2\right) - (\mathrm{E}(Y))^2 \\
&= \frac{31}{3} - 3^2 \\
&= \frac{31-27}{3} \\
&= \frac{4}{3}
\end{aligned}$$

3.1 EXERCISES

1. Let X be a random variable with the following cumulative distribution function F:



Calculate the following probabilities:

i. $\mathrm{P}(X \leq 4)$

ii. $P(2 < X \leq 4)$

iii. $P(2 \leq X \leq 4)$

iv. $P(3.5 \leq X < 4)$

v. $P(X = 4)$

vi. $P(X > 3)$

Solution:

i. $P(x \leq 4) = 0.8$

ii. $P(2 < x \leq 4) = 0.5$

iii. $P(2 \leq x \leq 4) = 0.6$

iv. $P(3.5 \leq x < 4) = 0$

v. $P(x = 4) = 0.1$

vi. $P(x > 3) = 0.4$

2. Let X be a discrete random variable with the following PDF, f:

x	1	3	5	7	9
$f(x)$	0.4	0.1	0.2	0.2	0.1

i. $P(3 < X < 7)$

ii. $P(3 < X < 7 \,|\, X > 5)$

iii. Draw the graph of the CDF of X. Find the expected value of X.

Solution:

1. $P(3 < x < 7) = f(5) = 0.2$
2. $P(3 < x < 7 \,|\, x > 5) = 0$
3. This exercise is left as self-study.

$$\begin{aligned} E(X) = \sum_x xf(x) &= 1 \cdot 0.4 + 3 \cdot 0.1 + 5 \cdot 0.2 + 7.0.2 + 9 \cdot 0.1 \\ &= 0.4 + 0.3 + 1.0 + 1.4 + 0.9 \\ &= 4 \end{aligned}$$

3. Explain why each of the following is or is not a valid probability distribution for a discrete random variable X:

i.

x	0	1	2	3
$f(x)$	0.1	0.3	0.3	0.2

ii.

x	-2	-1	0
$f(x)$	0.25	0.50	0.25

iii.

x	4	9	20
$f(x)$	-0.3	0.4	0.3

iv.

x	2	3	5	6
$f(x)$	0.15	0.15	0.45	0.35

Solution:

1. $\sum_x f(x) = 0.1 + 0.3 + 0.3 + 0.2 = 0.9 < 1.0$ (not valid).
2. $f(x) \geq 0$ for all x values.

$$\sum_x f(x) = 0.25 + 0.50 + 0.25$$
$$= 1.0 = 1.0 \quad \text{(valid).}$$

3. $f(4) = -0.3 < 0$ (not valid).
4.

$$\sum_x f(x) = 0.15 + 0.15 + 0.45 + 0.35 > 1$$

(not valid).

4. The random variable X has the following discrete probability distribution:

x	1	3	5	7	9
$f(x)$	0.1	0.2	0.4	0.2	0.1

i. List the values x may assume.

ii. What value of x is the most probable?

iii. Graph the probability distribution.

iv. Find $\mathrm{P}(X = 7)$

v. Find $\mathrm{P}(X \geq 5)$

vi. Find $\mathrm{P}(X > 2)$

vii. Find $\mathrm{E}(X)$

Solution:

1. x can take any of the values from $\{1, 3, 5, 7, 9\}$.
2. $f(5)$ is greater than all other $f(x)$ valued; so, $x = 5$ is the most probable.
3. This exercise is left as self-study.
4. $\mathrm{P}(x = 7) = f(7) = 0.2$

5. $P(x \geq 5) = f(5) + f(7) + f(9) = 0.4 + 0.2 + 0.1 = 0.7$

6.

$$\begin{aligned} P(x > 2) &= f(3) + f(5) + f(7) + f(9) \\ &= 0.2 + 0.4 + 0.2 + 0.1 \\ &= 0.9 \end{aligned}$$

7. $E(X) = \sum_x xf(x) = 5$.

5. Consider the probability distributions,

x	0	1	2
$f(x)$	0.3	0.4	0.3

and

y	0	1	2
$f(y)$	0.1	0.8	0.1

i. Use your intuition to find the mean for each distribution.

ii. Which distribution appears to be more variable? Why?

Solution:

1. $f(x)$ is symmetric around $x = 1$. $f(y)$ is symmetric around $y = 1$. So, $E(X) = 1$ and $E(Y) = 1$.
2. X displays higher variation. Intuitively, "its values that are away from the expected value are more probable" compared to the case of Y.

6. Every morning, my mother gives me a random amount of money according to the following PDF, where X is the random variable that measures the amount of money:

x	20	30	40	50
$f(x)$	0.10	0.20	0.30	0.40

Right after that, my sister takes out of my pocket a random amount of money according to the following CDF, where Y is the random variable that measures the amount of money:

y	5	10	15
$F(y)$	0.30	0.70	1.00

Then I leave home and spend all my money before the day ends. Create a random variable W which shows the net amount of money before I leave home in the morning. Calculate $F(w)$ and present it in tabular format. Using these functions:

i. Calculate $\mathrm{E}(X)$

ii. Calculate $\mathrm{E}(Y)$

iii. Verify that $\mathrm{E}(W) = \mathrm{E}(X) - \mathrm{E}(Y)$

iv. Draw the graph of $f(w)$ and mark the value of $\mathrm{E}(W)$ on it

v. Calculate $\mathrm{Var}\,(W)$

Solution: i.

$$\begin{aligned}\mathrm{E}(x) &= \sum xf(x)\\ &= 20\cdot 0.10 + 30\cdot 0.20 + 40\cdot 0.30 + 50\cdot 0.40\\ &= 2 + 6 + 12 + 20\\ &= 40\end{aligned}$$

ii. First, we need to find $g(y)$:

$$\begin{aligned}g(y) &= \Delta G(y)\\ g(5) &= 0.30 \leftarrow 0.30\\ g(10) &= 0.40 \leftarrow 0.70 - 0.30\\ g(15) &= 0.30 \leftarrow 1.00 - 0.70\end{aligned}$$

$$\begin{aligned}\mathrm{E}(Y) = \sum yg(y) &= 5\cdot 0.30 + 10\cdot 0.40 + 15\cdot 0.30\\ &= 1.5 + 4 + 4.5\\ &= 10\end{aligned}$$

iii. First, we need to find the PDF of W, call it $h(w)$. Find the possible values of W and calculate the probability for each w. Those values are

$$\begin{aligned}&w \in \{5, 10, 15, 20, 25, 30, 35, 40, 45\}\\ &h(5) = f(20)\,g(15) = 0.10\cdot 0.30 = 0.03\\ &h(10) = f(20)\,g(10) = 0.10\cdot 0.40 = 0.04\\ &h(15) = f(20)\,g(5) + f(30)\,g(15) = 0.10\cdot 0.30 + 0.20\cdot 0.30 = 0.09\\ &h(20) = f(30)\,g(10) = 0.20\cdot 0.40 = 0.08\\ &h(25) = f(30)\,g(5) + f(40)\,g(15) = 0.20\cdot 0.30 + 0.30\cdot 0.30 = 0.15\\ &h(30) = f(40)\,g(10) = 0.30\cdot 0.40 = 0.12\\ &h(35) = f(40)\,g(5) + f(50)\,g(15) = 0.30\cdot 0.30 + 0.40\cdot 0.30 = 0.21\\ &h(40) = f(50)\,g(10) = 0.40\cdot 0.40 = 0.16\\ &h(45) = f(50)\,g(5) = 0.40\cdot 0.30 = 0.12\end{aligned}$$

Then,

$$\begin{aligned}
\mathrm{E}(w) =& \sum wh(w) \\
=& 5 \cdot 0.03 + 10 \cdot 0.04 + 15 \cdot 0.09 + 20 \cdot 0.08 \\
&+ 25 \cdot 0.15 + 30 \cdot 0.12 + 35 \cdot 0.21 + 40 \cdot 0.16 \\
&+ 45 \cdot 0.12 \\
=& 0.15 + 0.4 + 1.35 + 1.6 \\
&+ 3.75 + 3.6 + 7.35 + 6.4 \\
&+ 5.4 \\
=& 30
\end{aligned}$$

From the previous parts we know that $\mathrm{E}(X) = 40$ and $\mathrm{E}(Y) = 10$. In this part, we found $\mathrm{E}(W) = 30$. So, $\mathrm{E}(X) - \mathrm{E}(Y) = 40 - 10 = 30 = \mathrm{E}(W) \rightarrow$ verification done.
iv. Do on your own.
v. Calculate $\mathrm{Var}\,(W)$ as:

$$\begin{aligned}
\mathrm{Var}\,(W) &= \sum (w - \mathrm{E}(w))^2 h(w) \\
&= (5-30)^2 \cdot 0.03 + (10-30)^2 \cdot 0.04 \\
&+ (15-30)^2 \cdot 0.09 + (20-30)^2 \cdot 0.08 \\
&+ (25-30)^2 \cdot 0.15 + (30-30)^2 \cdot 0.12 \\
&+ (35-30)^2 \cdot 0.21 + (40-30)^2 \cdot 0.16 \\
&+ (45-30)^2 \cdot 0.12 \\
&= 115
\end{aligned}$$

As an alternative:

$$\begin{aligned}
\mathrm{E}\left(W^2\right) =& \sum w^2 h(w) \\
&= 25 \cdot 0.03 + 100 \cdot 0.04 + 225 \cdot 0.09 \\
&+ 400 \cdot 0.08 + 625 \cdot 0.15 + 900 \cdot 0.12 \\
&+ 1225 \cdot 0.21 + 1600 \cdot 0.16 + 2025 \cdot 0.12 \\
=& 1015
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\,(W) &= \mathrm{E}\left(W^2\right) - (\mathrm{E}(W))^2 \\
&= 1015 - 30^2 \\
&= 1015 - 900 \\
&= 115
\end{aligned}$$

(As a follow-up exercise: calculate $\text{Var}(X)$ and $\text{Var}(Y)$ on your own, and verify that $\text{Var}(W) = \text{Var}(X) + \text{Var}(Y)$).

7. Consider $X \sim f(x) = \frac{1}{4}, 4 \leq x \leq 8, \quad Y \sim g(y) = \frac{1}{3}, 0 \leq y \leq 3$ and another random variable W which is defined as $W = X - Y$. Calculate $\text{E}(X), \text{E}(Y), \text{E}(W)$, $\text{Var}(X), \text{Var}(Y), \text{Var}(W)$.
Solution:

$$\begin{aligned}\text{E}(X) = \int_{-\infty}^{\infty} x f(x)\, dx = \int_4^8 x \cdot \frac{1}{4} dx &= \frac{1}{4}\frac{x^2}{2}\Big|_4^8 \\ &= \frac{1}{8}(64 - 16) \\ &= 6\end{aligned}$$

$$\begin{aligned}\text{E}(Y) = \int_{-\infty}^{\infty} y g(y) dy = \int_0^3 y \frac{1}{3} dy &= \frac{1}{3}\frac{y^2}{2}\Big|_0^3 \\ &= \frac{1}{6}(9 - 0) \\ &= 3/2\end{aligned}$$

$$\begin{aligned}\text{E}\left(X^2\right) = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \int_4^8 x^2 \frac{1}{4} dx &= \frac{1}{4}\frac{x^3}{3}\Big|^8 \\ &= \frac{1}{12}(512 - 64) \\ &= 112/3\end{aligned}$$

$$\begin{aligned}\text{Var}(X) &= \text{E}\left(X^2\right) - (\text{E}(X))^2 \\ &= \frac{112}{3} - 36 \\ &= 4/3\end{aligned}$$

$$\begin{aligned}\text{E}\left(Y^2\right) = \int_{-\infty}^{\infty} y^2 g(y) dy = \int_0^3 y^2 \frac{1}{3} dy &= \frac{1}{3}\frac{y^3}{3}\Big|_0^3 \\ &= \frac{1}{9}(27 - 0) \\ &= 3\end{aligned}$$

$$\begin{aligned}\text{Var}(Y) &= \text{E}\left(Y^2\right) - (\text{E}(Y))^2 \\ &= 3 - \frac{9}{4} \\ &= 3/4\end{aligned}$$

Without finding $h(w)$, the following can be written:

$$\begin{aligned}W = X - Y \rightarrow \text{E}(W) &= \text{E}(X) - \text{E}(Y) \\ &= 6 - \frac{3}{2} \\ &= 9/2 \\ \rightarrow \text{Var}(W) &= \text{Var}(X) + \text{Var}(Y) \\ &= \frac{4}{3} - \frac{3}{4} \\ &= \frac{16-9}{12} \\ &= 7/12\end{aligned}$$

Another way to deal with W is:

$$\begin{aligned}&W \sim h(w), \quad h(w) = f(x)\, g(y) \\ &\text{E}(W) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} (x-y) f(x)\, g(y) dx dy \\ &= \int_{x=4}^{8} \int_{y=0}^{3} (x-y) \frac{1}{4} \frac{1}{3} dx dy \\ &= \frac{1}{12} \int_{x=4}^{8} \frac{-(x-y)^2}{2} \Big|_{y=0}^{3} dx \\ &= -\frac{1}{24} \int_{4}^{8} \left[(x-3)^2 - (x-0)^2\right] dx \\ &= -\frac{1}{24} \int_{4}^{8} (-6x+9) dx \\ &= -\frac{1}{24} \left(-6\frac{x^2}{2} + 9x\right) \Big|_{4}^{8} \\ &= -\frac{1}{24} \left(-3x^2 + 9x\right) \Big|_{4}^{8} \\ &= -\frac{1}{24} [(-3 \cdot 64 + 9 \cdot 8) - (-3 \cdot 16 + 9 \cdot 4)] \\ &= -\frac{1}{24} [-192 + 72 + 48 - 36] \\ &= -\frac{1}{24} (-108) \\ &= \frac{108}{24} = 9/2 \text{ (verifies } \text{E}(W) = \text{E}(X) - \text{E}(Y) \text{)}\end{aligned}$$

To calculate $\text{Var}(W)$ you do the following:

- Calculate $\text{E}\left(W^2\right)$

$$\text{E}\left(W^2\right) = \int_{x=4}^{8}\int_{y=0}^{3}(x-y)^2\frac{1}{4}\frac{1}{3}dxdy$$

- Then, find $\text{Var}(W)$

$$\text{Var}(W) = \text{E}\left(W^2\right) - (\text{E}(W))^2$$

If 'double integrals' were not in the curriculum of MATH 105 or MATH 106 and if you do not have a prior knowledge of it, you may safely skip this last part.

3.7 *Random variables and distributions: Discrete probability laws*

We consider here four (one being optional) discrete probability laws

- Bernoulli distribution
- Binomial distribution
- Poisson distribution
- Hypergeometric distribution
- Geometric distribution
- Negative Binomial distribution
- Discrete Uniform distribution

3.7.1 *Bernoulli distribution*

Bernoulli distribution is also called Bernoulli trial or Bernoulli process. Consider an experiment consists of 1 trial and let there be two possible outcomes, success and fail.

- Success $(x = 1)$ with probability of P
- Failure $(x = 0)$ with probability of $(1 - P)$

For $X \sim \text{Bernoulli}(P)$

$$\forall x \in \mathbb{R}, f(x) = \begin{cases} P & , x = 1 \\ 1 - P & , x = 0 \\ 0 & , otherwise \end{cases}$$

Despite its simplicity, Bernoulli distributton is a stunningly useful one, as a building block of some other distributions.

Observe below the PDF of $X \sim \text{Bernoulli}(0.80)$:

Expected value and Variance:

$$\begin{aligned} \mathrm{E}(X) = \sum_{x=0}^{1} x f(x) &= 0 \cdot (1-P) + 1 \cdot P \\ &= P \end{aligned}$$

$$\begin{aligned} \mathrm{E}\left(X^2\right) = \sum_{x=0}^{1} x^2 f(x) &= 0^2 (1-P) + 1^2 \cdot P \\ &= P \end{aligned}$$

$$\begin{aligned} \mathrm{Var}\,(X) &= \mathrm{E}\left(X^2\right) - (\mathrm{E}(X))^2 \\ &= P - P^2 \\ &= \mathrm{P}(1-P) \end{aligned}$$

In a nutshell

A practical hint for deriving $\text{Var}(X)$

As you have seen in the derivations of $\text{E}(X)$ and $\text{Var}(X)$ for the Bernoulli distribution, we first calculated the $\text{E}(X)$. This is an often seamless step. However, instead of attacking the $\text{Var}(X)$ directly, we preferred to calculate $\text{E}\left(X^2\right)$, which with the knowledge of $\text{E}(X)$ yields

$$\text{Var}(X) = \text{E}\left(X^2\right) - (\text{E}(X))^2$$

In the remaining derivations of this chapter, notice the use of

$$\text{E}(X(X-1))$$

to facilitate easier derivation / calculation of $\text{Var}(X)$. Especially when the PDF of X, ie., $f(x)$ involves combinations, $\text{E}(X(X-1))$ may be a life saver.

Notice that $\text{E}(X(X-1)) = \text{E}\left(X^2 - X\right) = \text{E}\left(X^2\right) - \text{E}(X)$

So, $\text{E}\left(X^2\right) = \text{E}(X(X-1)) + \text{E}(X)$, and:

$$\begin{aligned}\text{Var}(X) &= \text{E}\left(X^2\right) - (\text{E}(X))^2 \\ &= \text{E}(X(X-1)) + \text{E}(X) - (\text{E}(X))^2\end{aligned}$$

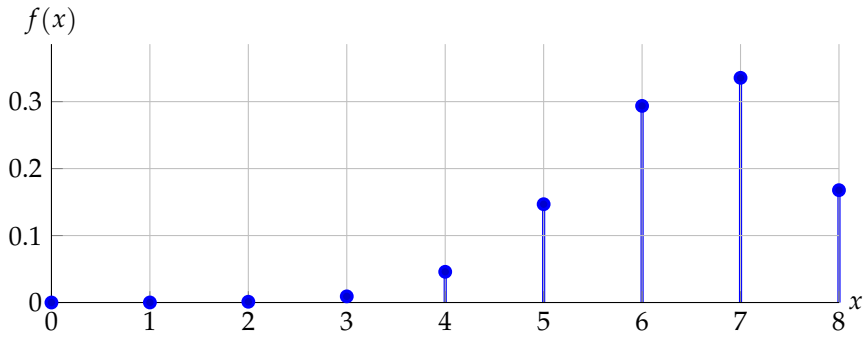
Examine/ practice this hint along this chapter.

3.7.2 *Binomial distribution*

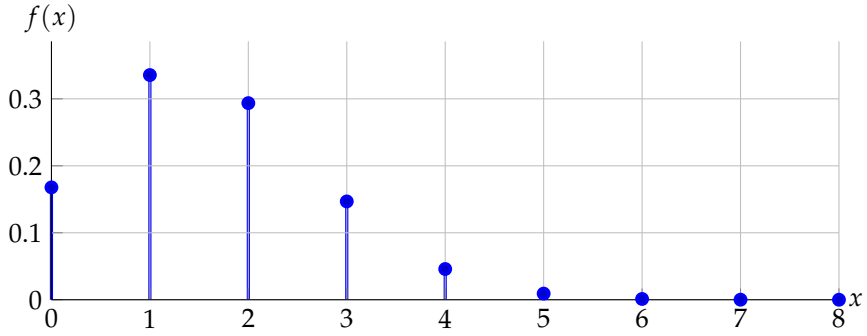
Consider an experiment which consists of n independent and identical Bernoulli trials;i.e, the probability of success (P) is the same across all the trials and a trial's outcome does not alter the outcomes of the subsequent trials. X being the number of successes in n trials, $X \sim \text{Binomial}(n, P)$, i.e., X has a Binomial distribution with parameters n and P:

$$\forall x \in \mathbb{R}, f(x) = \begin{cases} \binom{n}{x} P^x (1-P)^{n-x} & , x = 0, 1, 2 \ldots, n \\ 0 & , otherwise \end{cases}$$

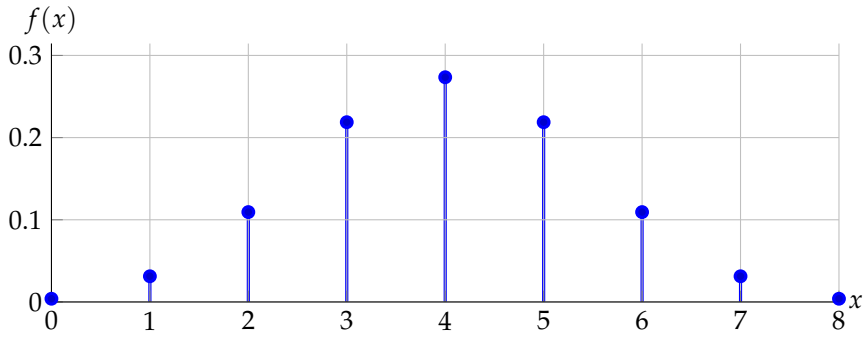
Observe below the PDF of $X \sim \text{Binomial}(8, 0.80)$:



Then the PDF of $X \sim \text{Binomial}(8, 0.20)$:



And finally the PDF of $X \sim \text{Binomial}(8, 0.50)$:



Having compared the PDFs of $\text{Binomial}(8, 0.80)$, $\text{Binomial}(8, 0.20)$, $\text{Binomial}(8, 0.50)$, can you identify the source of asymmetry of *Binomial* PDFs?

Expected value and Variance:

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} P^x (1-P)^{n-x} \\
&= nP \underbrace{\sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} P^{x-1} (1-P)^{n-x}}_{1} \\
&= nP
\end{aligned}$$

$$\mathrm{E}\left(X^2\right) = \sum_{x=0}^{n} x^2 \frac{n!}{x!(n-x)!} P^x (1-P)^{n-x}$$

is not practical to work with. So, consider:

$$\begin{aligned}
\mathrm{E}(X(X-1)) &= \sum_{x=0}^{n} x(x-1) \frac{n!}{x!(n-x)!} P^x (1-P)^{n-x} \\
&= n(n-1)P^2 \underbrace{\sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!} P^{x-2} (1-P)^{n-x}}_{1} \\
&= n(n-1)P^2
\end{aligned}$$

This means:

$$\begin{aligned}
\mathrm{E}\left(X^2\right) - \mathrm{E}(X) &= n(n-1)P^2 \\
\mathrm{E}\left(X^2\right) &= n(n-1)P^2 + nP
\end{aligned}$$

Then,

$$\begin{aligned}
\mathrm{Var}\,(X) &= \mathrm{E}\left(X^2\right) - (\mathrm{E}(X))^2 \\
&= n(n-1)P^2 + nP - (nP)^2 \\
&= n^2P^2 - nP^2 + nP - n^2P^2 \\
&= nP - nP^2 \\
&= nP(1-P)
\end{aligned}$$

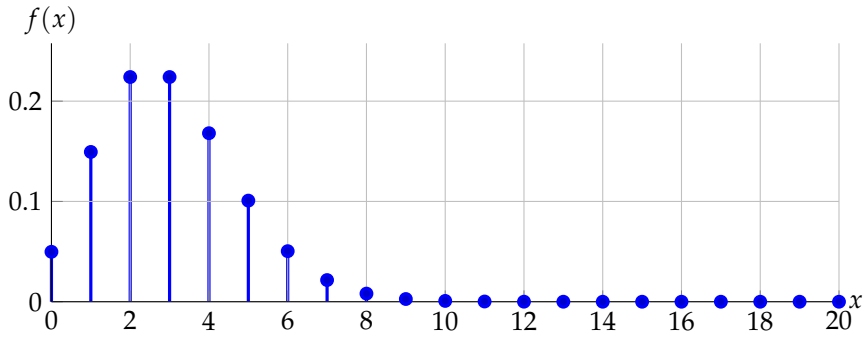
3.7.3 Poisson distribution

Consider an experiment which consists of counting the number of times a certain event occurs during a given unit of time or in a given area or volume. The probability that an event occurs in a given unit of

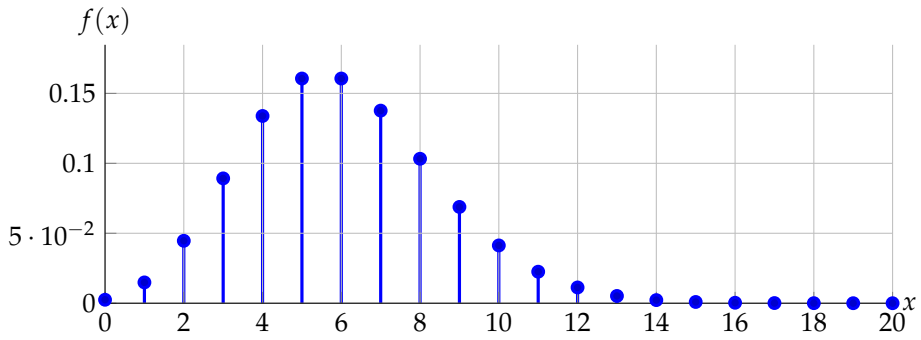
time, area or volume is the same for all units. The number of events that occur in one unit of time, area or volume is independent of the number that occur in any other mutually exclusive unit. The mean (or expected, or typical) number of events in each unit is denoted by λ. For $X \sim \text{Poisson}(\lambda)$:

$$\forall x \in \mathbb{R}, f(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & , x = 0, 1, 2 \dots \\ 0 & , otherwise \end{cases}$$

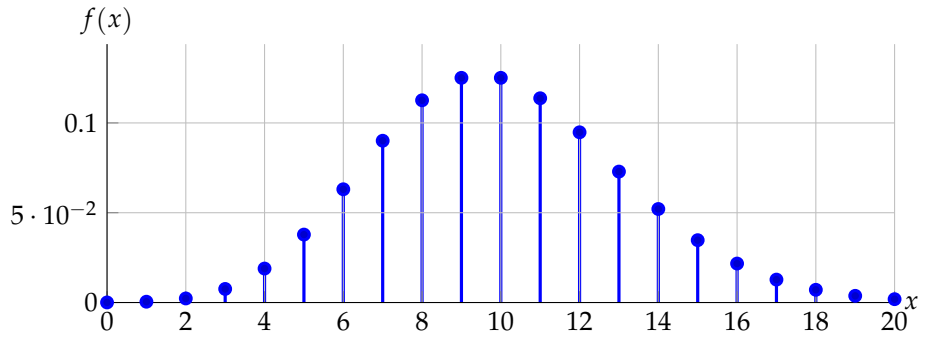
Recall that, e is called the **Euler's Number** where $e = 2.71828\dots$ Observe below the PDF of $X \sim \text{Poisson}(3)$:



Then the PDF of $X \sim \text{Poisson}(6)$:



And finally the PDF of $X \sim \text{Poisson}(10)$:



Is the last graph symmetric? Is it possible to have a Poisson(λ) PDF which is symmetric? Why?

Expected Value and Variance:

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} \\
&= \lambda e^{-\lambda} \underbrace{\sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}}_{e^{\lambda}} \\
&= \lambda
\end{aligned}$$

$$\mathrm{E}\left(X^2\right) = \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!}$$

is not useful againg. So, consider,

$$\begin{aligned}
\mathrm{E}(X(X-1)) &= \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda}\lambda^x}{x!} \\
&= e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!} \\
&= \lambda^2 e^{-\lambda} \underbrace{\sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!}}_{e^{\lambda}} \\
&= \lambda^2
\end{aligned}$$

This means,

$$\begin{aligned} \mathrm{E}\left(X^2\right) - \mathrm{E}(X) &= \lambda^2 \\ \mathrm{E}\left(X^2\right) &= \lambda^2 + \lambda \end{aligned}$$

Then,

$$\begin{aligned} \mathrm{Var}\,(X) &= \mathrm{E}\left(X^2\right) - (\mathrm{E}(X))^2 \\ &= \lambda^2 + \lambda - \lambda^2 \\ &= \lambda \end{aligned}$$

In a nutshell

Poisson approximation to Binomial distribution
Let X be the number of successes resulting from n independent trials, each with probability of success P. The distribution of the number of successes, X, is binomial, with mean nP. If the number of trials, n, is large and nP is of only moderate size (preferably $nP \leq 7$), this distribution can be approximated by the Poisson distribution with $\lambda = nP$. So,

$$f(x) = \frac{\mathrm{e}^{-nP}\,(nP)^x}{x!}, x = 0, 1, 2 \ldots$$

can safely be used to obtain a numerical result.

3.7.4 *How to derive Poisson distribution from Binomial distribution?*

Consider a *Binomial* (n, p) process with:

$$f(x) = \binom{n}{x} p^x\,(1-p)^{n-x}, x = 0, 1, 2, \ldots, n$$

Define

$$\lambda = np$$

So,

$$p = \frac{\lambda}{n}$$

Re-writting $f(x)$ as:

$$f(x) = \lim_{n\to\infty} \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

the derivation yields. Now,

$$
\begin{aligned}
f(x) &= \lim_{n\to\infty} \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \left(\frac{\lambda^x}{x!}\right) \lim_{n\to\infty} \underbrace{\left[\frac{n!}{(n-x)!}\left(\frac{1}{n^x}\right)\right]}_{A} \underbrace{\left[\left(1-\frac{\lambda}{n}\right)^n\right]}_{B} \underbrace{\left[\left(1-\frac{\lambda}{n}\right)^{-x}\right]}_{C}
\end{aligned}
$$

Consider now the parts A, B and C separately:

(A) The limit is 1; there are x terms linear in n in its numerator and x terms each of which is equal to n in its denominator

(B) The limit is $\mathrm{e}^{-\lambda}$; by definition of the Euler's number.

(C) The limit is 1 trivially.

Combining the limits:

$$f(x|\lambda) = \mathrm{P}(X = x \mid \lambda) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}$$

is found.

The intuition is as follows: When we consider a Binomial process in which a success occurs with an infinitesimal probability in every infinitesimal time period and when there are infinitely many time periods as such, what yields for a finite time period is nothing but the Poisson distribution. The derivation can be carried out in reference to space rather than time, if you wish.

3.7.5 Hypergeometric distribution

Consider an experiment which consists of randomly drawing n elements without replacement from a set of N elements, r of which are successes and $(N-r)$ of which are failures. X being the number of successes among n elements, $X \sim \mathit{Hypergeometric}\,(N, r, n)$:

$$\forall x \in \mathbb{R}, f(x) = \begin{cases} \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} & , x = \max\{0, n-(N-r)\}, \ldots, \min\{r, n\} \\ 0 & , \mathit{otherwise} \end{cases}$$

Expected Value and Variance:

$$\begin{aligned}\mathrm{E}(X) &= \sum_{x=0}^{n} x \frac{\left(\begin{array}{c} r \\ x \end{array}\right)\left(\begin{array}{c} N-r \\ n-x \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)} \\ &= \sum_{x=1}^{n} x \frac{\left(\begin{array}{c} r \\ x \end{array}\right)\left(\begin{array}{c} N-r \\ n-x \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)}\end{aligned}$$

$$\begin{aligned}x\left(\begin{array}{c} r \\ x \end{array}\right) &= x\frac{r!}{x!(r-x)!} \\ &= r\frac{(r-1)!}{(x-1)!(r-x)!} \\ &= r\left(\begin{array}{c} r-1 \\ x-1 \end{array}\right)\end{aligned}$$

$$\begin{aligned}\left(\begin{array}{c} N \\ n \end{array}\right) = \frac{N!}{n!(N-n)!} &= \frac{N(N-1)!}{n(n-1)!(N-n)!} \\ &= \frac{N}{n}\frac{(N-1)!}{(n-1)!(N-n)!} \\ &= \frac{N}{n}\left(\begin{array}{c} N-1 \\ n-1 \end{array}\right)\end{aligned}$$

$$\begin{aligned}&= \sum_{x=1}^{n} \frac{r\left(\begin{array}{c} r-1 \\ x-1 \end{array}\right)\left(\begin{array}{c} N-r \\ n-x \end{array}\right)}{\frac{N}{n}\left(\begin{array}{c} N-1 \\ n-1 \end{array}\right)} \\ &= \frac{nr}{N}\underbrace{\sum_{x=1}^{n} \frac{\left(\begin{array}{c} r-1 \\ x-1 \end{array}\right)\left(\begin{array}{c} N-r \\ n-x \end{array}\right)}{\left(\begin{array}{c} N-1 \\ n-1 \end{array}\right)}}_{1} \\ &= \frac{nr}{N}\end{aligned}$$

$$\begin{aligned}\mathrm{E}(X(X-1)) &= \sum_{x=0}^{n} x(x-1)\frac{\left(\begin{array}{c} r \\ x \end{array}\right)\left(\begin{array}{c} N-r \\ n-x \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)} \\ &= \sum_{x=2}^{n} x(x-1)\frac{\left(\begin{array}{c} r \\ x \end{array}\right)\left(\begin{array}{c} N-r \\ n-x \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)}\end{aligned}$$

Notice that:

$$\begin{aligned}x(x-1)\left(\begin{array}{c} r \\ x \end{array}\right) &= x(x-1)\frac{r!}{x!(r-x)!} \\ &= r(r-1)\frac{(r-2)!}{(x-2)!(r-x)!} \\ &= r(r-1)\left(\begin{array}{c} r-2 \\ x-2 \end{array}\right)\end{aligned}$$

and that:

$$\begin{aligned}\left(\begin{array}{c} N \\ n \end{array}\right) = \frac{N!}{n!(N-n)!} == \frac{N(N-1)(N-2)!}{n(n-1)(n-2)!(N-n)!} \\ = \frac{N(N-1)}{n(n-1)}\left(\begin{array}{c} N-2 \\ n-2 \end{array}\right)\end{aligned}$$

$$\begin{aligned}\mathrm{E}(X(X-1)) &= \sum_{x=2}^{n} \frac{r(r-1)\left(\begin{array}{c} r-2 \\ X-2 \end{array}\right)\left(\begin{array}{c} N-r \\ n-x \end{array}\right)}{\frac{N(N-1)}{n(n-1)}\left(\begin{array}{c} N-2 \\ n-2 \end{array}\right)} \\ &= \frac{nr}{N}\frac{(n-1)(r-1)}{N-1}\sum_{x=2}^{n}\frac{\left(\begin{array}{c} r-2 \\ X-2 \end{array}\right)\left(\begin{array}{c} N-r \\ n-x \end{array}\right)}{\left(\begin{array}{c} N-2 \\ n-2 \end{array}\right)}\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left(X^2 - X\right) &= \mathrm{E}\left(X^2\right) - \mathrm{E}(X) \\
&= \frac{nr}{N}\frac{(n-1)(r-1)}{N-1} \\
\mathrm{E}\left(X^2\right) &= \frac{nr}{N}\left(1 + \frac{(n-1)(r-1)}{N-1}\right)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\,(X) &= \mathrm{E}\left(X^2\right) - (\mathrm{E}(X))^2 \\
&= \frac{nr}{N}\left(1 + \frac{(n-1)(r-1)}{N-1}\right) - \left(\frac{nr}{N}\right)^2 \\
&= \frac{nr}{N}\left(1 + \frac{(n-1)(r-1)}{N-1} - \frac{nr}{N}\right) \\
&= \frac{nr}{N}\left(\frac{N(N-1) + N(n-1)(r-1) - (N-1)nr}{N(N-1)}\right) \\
&= \frac{nr}{N} \cdot \frac{(N-n)(N-r)}{N(N-1)}
\end{aligned}$$

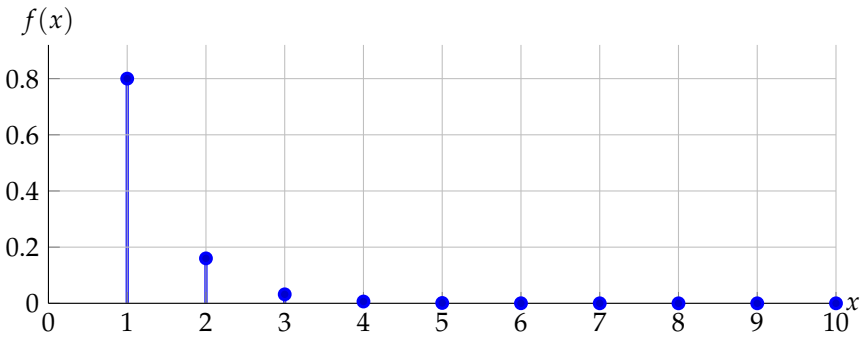
3.7.6 *Geometric distribution*

Consider an experiment which consists of a sequence of independent and identical Bernoulli trials; the expelriment ends when a (one) success is observed. X being the number of trials until one success, $X \sim$ Geometric(P), i.e., X has a geometric distribution with parameter P:

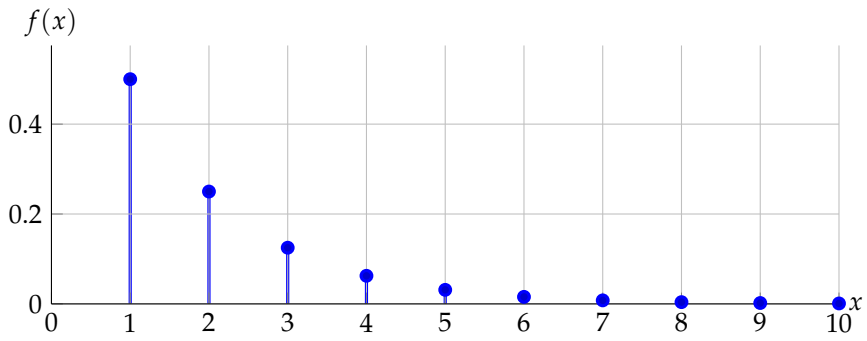
$$\forall x \in \mathbb{R}, f(x) = \begin{cases} (1-P)^{x-1}P & , x = 1, 2, \ldots \\ 0 & , otherwise \end{cases}$$

The construction of $f(x)$ is intuitive as the experiment will yield $x - 1$ failures before the 'one and only' success, which occurs at the end, by definition.

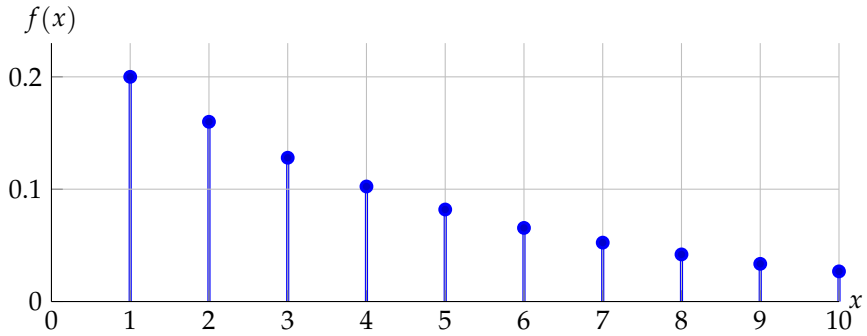
Observe below the PDF of $X \sim Geometric(0.80)$:



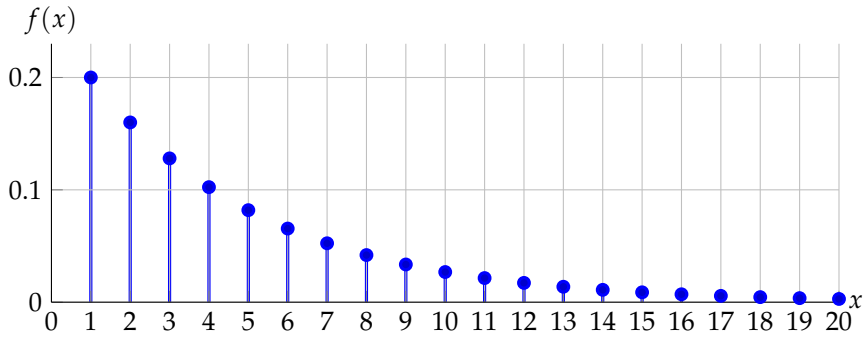
Observe below the PDF of $X \sim Geometric(0.50)$:



Observe below the PDF of $X \sim Geometric(0.20)$ for $1 \leq x \leq 10$



Observe below the PDF of $X \sim Geometric(0.20)$ for $1 \leq x \leq 20$



In a nutshell

Memoryless property of geometric distribution

Consider $X \sim Geometric(P)$, $f(x) = (1-P)^{x-1} \cdot P$. Suppose we know that $\mathrm{P}(x > h) = k$. What is the value of $\mathrm{P}(x > s+h \,|\, x > s)$?

$$\mathrm{P}(x > m) = (1-P)^m \quad \text{for any } m$$

So,

$$\begin{aligned}\mathrm{P}(x > s+h \,|\, x > s) &= \frac{(1-P)^{s+h}}{(1-P)^s} \\ &= (1-P)^h \\ &= \mathrm{P}(x > h) \\ &= k\end{aligned}$$

What does this result tell?

Expected Value and Variance:

$$\begin{aligned}\mathrm{E}(X) &= \sum_{x=1}^{\infty} x(1-P)^{x-1}P \\ &= 1\,\mathrm{P}(1-P)^0 + 2\,\mathrm{P}(1-P)^1 + 3\,\mathrm{P}(1-P)^2 + \cdots \\ (1-P)\,\mathrm{E}(X) &= 1\,\mathrm{P}(1-P) + 2\,\mathrm{P}(1-P)^2 + 3\,\mathrm{P}(1-P)^3 + \cdots \\ \mathrm{E}(X) - (1-P)\,\mathrm{E}(X) &= 1P + 1\,\mathrm{P}(1-P) + 1\,\mathrm{P}(1-P)^2 + \cdots \\ \mathrm{E}(X) + (P-1)\,\mathrm{E}(X) &= P + \mathrm{P}(1-P) + \mathrm{P}(1-P)^2 + \cdots \\ PE(X) &= \mathrm{P}\Big(1 + (1-P) + (1-P)^2 + \cdots\Big) \\ \mathrm{E}(X) &= \frac{1}{1-(1-P)} = \frac{1}{P}\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left(X^2\right) &= \sum_{x=1}^{\infty} x^2(1-P)^{x-1}P \\ &= P\sum_{x=1}^{\infty} x^2(1-P)^{x-1} \\ &= P\frac{2-P}{P^3} \\ &= \frac{2-P}{P^2}\end{aligned}$$

Note that, denoting $1 - P = q$

$$\sum_{x=1}^{\infty} x^2 q^{x-1} = \frac{1+q}{(1-q)^3} = \frac{1+1-P}{(1-1+P)^3} = \frac{2-P}{P^3}$$

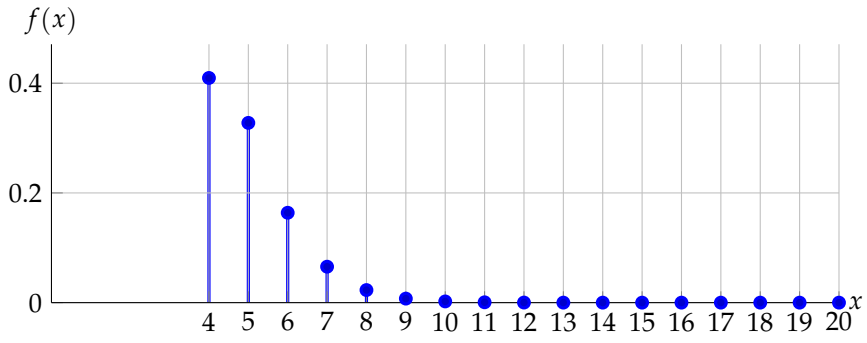
$$\begin{aligned}\text{Var}(X) &= \text{E}\left(X^2\right) - (\text{E}(X))^2 \\ &= \frac{2-P}{P^2} - (\frac{1}{P})^2 \\ &= \frac{1-P}{P^2}\end{aligned}$$

3.7.7 *Negative Binomial distribution*

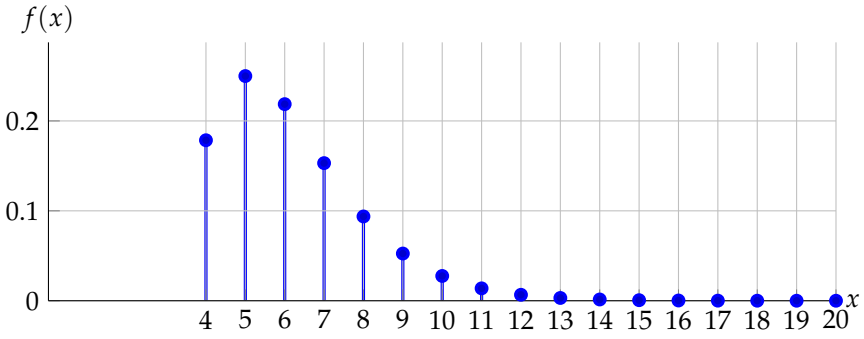
Consider an experiment which consists of a sequence of independent and identical Bernoulli trials; the expelriment ends when r successes are observed. X being the number of trials until r successes, $X \sim \text{Neg Bin}(r, P)$, i.e., X has a Negative Binomial distribution with parameters r and P:

$$\forall x \in \mathbb{R}, f(x) = \begin{cases} \binom{x-1}{r-1} P^r (1-P)^{x-r} & , x = r, r+1, \ldots \\ 0 & , otherwise \end{cases}$$

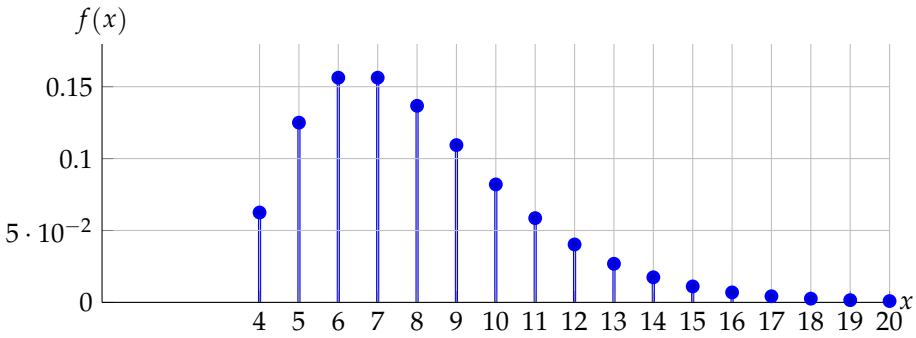
To develop an intuition of $f(x)$, notice that the last Bernoulli trial yields success, with a probability of P and the $x-1$ trials before that yield $r-1$ successes with a probability of $\binom{x-1}{r-1} P^{r-1}(1-P)^{x-r}$ according to a Binomial $(x-1, P)$ distribution, where the product of the two probabilities yield the Negative Binomial PDF. Observe below the PDF of $X \sim NegativeBinomial(4, 0.80)$:



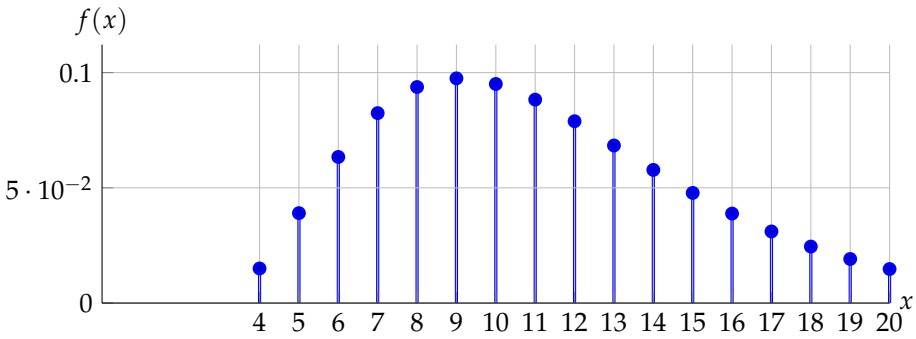
Observe below the PDF of $X \sim NegativeBinomial(4, 0.65)$:



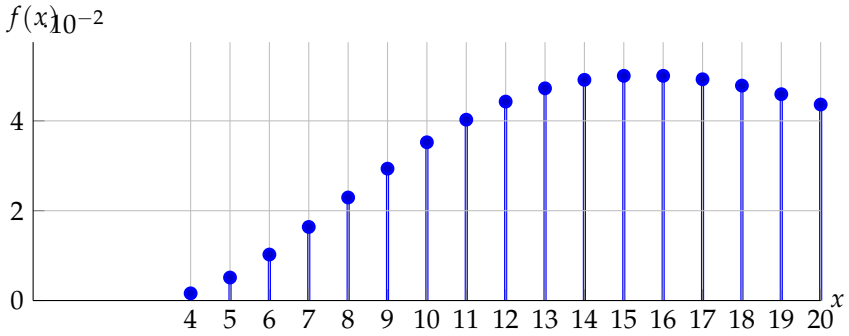
Observe below the PDF of $X \sim NegativeBinomial(4, 0.50)$:



Observe below the PDF of $X \sim NegativeBinomial(4, 0.35)$:



Observe below the PDF of $X \sim NegativeBinomial(4, 0.20)$:



In a nutshell

Among their many uses, one may consider the use of Geometric and Negative Binomial distributions in a Research and Development (R&D) environment. Suppose there is an R&D project consisting of a number of engineering trials. Though it is unrealistic, suppose also that each R&D trial is independent from others and all trials are identical. So, for simplicity of course, we assume that our R&D engineers are not learning across trials. Given these:

- We can assess the probability of x trials until a (one) success using a Geometric distribution
- We can assess the probability of x trials until r successes using a Negative Binomial distribution

Think: Is there a good use as such for budgeting purposes?

3.7.8 Discrete Uniform distribution

$$
\begin{aligned}
&X \sim \text{ Uniform } (a,b), a \leq x \leq b \in Z \\
&n = b - a + 1 \\
&f(x) = \frac{1}{n} \\
&\mathrm{E}(X) = \frac{a+b}{2} \\
&\text{Var}\,(X) = \frac{n^2-1}{12}
\end{aligned}
$$

$X \sim \text{Uniform}(a,b)$
Expected Value and Variance:

$$
\begin{aligned}
n =& b - a + 1 \\
f(x) =& \frac{1}{n}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}(X) &= \sum_{x=a}^{b} x \frac{1}{n} \\
&= \frac{na + \frac{(b-a)(b-a+1)}{2}}{n} \\
&= \frac{2na + (b-a)n}{2n} \\
&= \frac{2a + b - a}{2} \\
&= \frac{a+b}{2}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\left(X^2\right) &= \sum_{x=a}^{b} x^2 \frac{1}{n} \\
&= \frac{1}{n} \sum_{x=a}^{b} x^2 \\
&= \frac{1}{A} \frac{1}{6} (b-a+1)(2a^2 + 2ab - a + 2b^2 + b) \\
&= \frac{2a^2 + 2ab - a + 2b^2 + b}{6}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}\,(X) =& \mathrm{E}\left(X^2\right) - (\mathrm{E}(X))^2 \\
=& \frac{2a^2 + 2ab - a + 2b^2 + b}{6} - \left(\frac{a+b}{2}\right)^2 \\
=& \frac{2a^2 + 2ab - a + 2b^2 + b}{6} - \frac{a^2 + 2ab + b^2}{4} \\
=& \frac{4a^2 + 4ab - 2a + 4b^2 + 2b - 3a^2 - 6ab - 3b^2}{12} \\
=& \frac{a^2 - 2ab - 2a + 2b + b^2}{12} \\
=& \frac{n^2 - 1}{12}
\end{aligned}
$$

3.8 Random variables and distributions: Continuous probability laws

We consider here three continuous probability laws

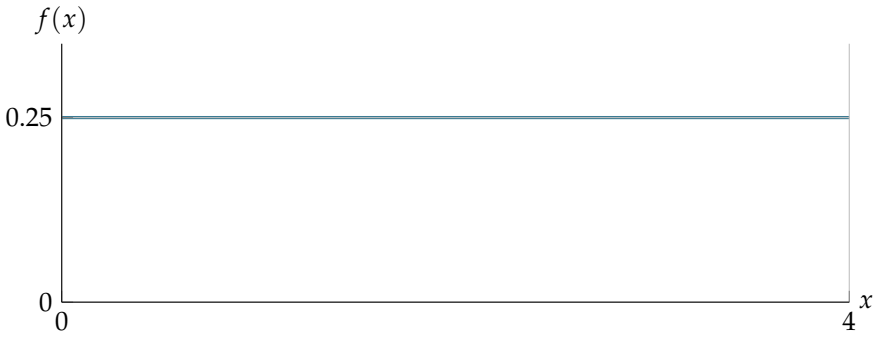
- Uniform distribution
- Triangular distribution
- Exponential distribution
- Normal distribution

3.8.1 Uniform distribution

$$X \sim Uniform\,(a, b)$$

$$\forall x \in \mathbb{R}, f(x) = \begin{cases} \frac{1}{b-a} & , a \leq x \leq b \\ 0 & , otherwise \end{cases}$$

The graph of the PDF of Uniform$(0, 4)$ looks like:



Expected Value and Variance:

$$\begin{aligned} \mathrm{E}(\mathrm{X}) = \int_2^b x \frac{1}{b-a} dx &= \frac{1}{b-a} \frac{x^2}{2} \Big|_a^b \\ &= \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{a+b}{2} \end{aligned}$$

$$\begin{aligned}\mathrm{E}\left(X^2\right) = \int_a^b x^2 \frac{1}{b-a} dx &= \frac{1}{b-a}\frac{x^3}{3}\Big|_a^b \\ &= \frac{b^3-a^3}{3(b-a)} \\ &= \frac{b^2+ab+a^2}{3}\end{aligned}$$

$$\begin{aligned}\mathrm{Var}\,(X) &= \mathrm{E}\left(X^2\right) - (\mathrm{E}(X))^2 \\ &= \frac{b^2+ab+a^2}{3} - \frac{(a+b)^2}{4} \\ &= \frac{4a^2+4ab+4b^2-3a^2-6ab-3b^2}{12} \\ &= \frac{a^2-2ab+b^2}{12} \\ &= \frac{(b-a)^2}{12}\end{aligned}$$

3.8.2 *Triangular distribution*

$X \sim \mathrm{Triangular}(a, b, c)$
a: Lower limit b: Mode c: Upper limit

$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)}, & a \le x \le b \\ \frac{2(c-x)}{(c-a)(c-b)}, & b \le x \le c \end{cases}$$

$$F(x) = \begin{cases} \frac{(x-a)^2}{(b-a)(c-a)}, & a \le x \le b \\ 1 - \frac{(c-x)^2}{(c-a)(c-b)}, & b \le x \le c \end{cases}$$

$$\mathrm{E}(X) = \frac{a+b+c}{3}$$

$$\mathrm{Var}\,(X) = \frac{a^2+b^2+c^2-ab-ac-bc}{18}$$

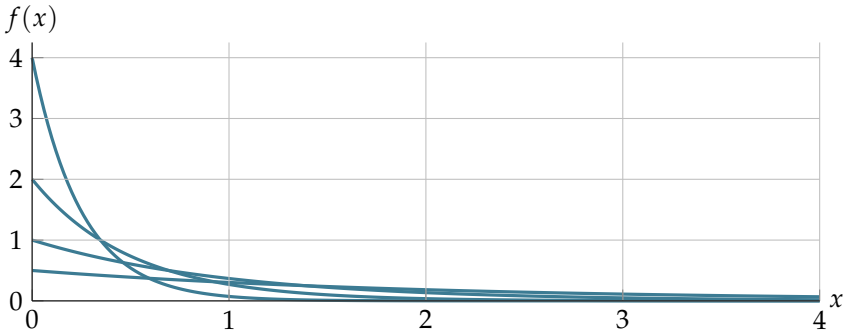
Triangular distribution is a practical model, mostly useful in business what-if analysis. A symmetric triangular is the sum of two identically distributed uniform variables.

3.8.3 *Exponential distribution*

$$X \sim Exponential\,(\lambda)$$

$$\forall x \in \mathbb{R}, f(x) = \begin{cases} \lambda e^{-\lambda x} & , x > 0 \\ 0 & , otherwise \end{cases}$$

Graphs of Exponential(0.5), Exponential(1.0), Exponential(2.0) and Exponential(4.0) PDFs can be seen in the following figure:



Expected Value and Variance:

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{\infty} x\lambda e^{-\lambda x} dx \\ &= \int_{0}^{\infty} x\lambda e^{-\lambda x} dx \\ &= \lambda \int_{0}^{\infty} \underbrace{x}_{u} \underbrace{e^{-\lambda x} dx}_{dv} \end{aligned}$$

$$\begin{aligned} u &= x \\ du &= dx \\ dv &= e^{-\lambda x} dx \\ v &= -\frac{1}{\lambda} e^{-\lambda x} \end{aligned}$$

$$\begin{aligned} \mathrm{E}(X) &= -\frac{x}{\lambda} e^{-\lambda x}\Big|_0^{\infty} + \frac{1}{\lambda} \int_0^{\infty} e^{-\lambda x} dx \\ &= 0 + \frac{1}{\lambda} \int_0^{\infty} e^{-\lambda x} dx \\ &= \frac{1}{\lambda} \left(-\frac{1}{\lambda} e^{-\lambda x}\Big|_0^{\infty} \right) \\ &= \frac{1}{\lambda^2} \\ \text{So, } \mathrm{E}(X) &= \lambda \frac{1}{\lambda^2} = \frac{1}{\lambda} \end{aligned}$$

$$\begin{aligned}\mathrm{E}\left(X^2\right) &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\ &= \lambda \int_0^\infty \underbrace{x^2}_{u} \underbrace{e^{-\lambda x dx}}_{dv}\end{aligned}$$

$$\begin{aligned}u &= x^2 \\ du &= 2x dx \\ dv &= e^{-\lambda x} dx \\ v &= -\frac{1}{\lambda} e^{-\lambda x}\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left(X^2\right) &= \cancel{\lambda}\left[-\frac{x^2}{\cancel{\lambda}} e^{-\lambda x}\Big|_0^\infty + \frac{2}{\cancel{\lambda}} \int_0^\infty x e^{-\lambda x} dx\right] \\ &= 2\int_0^\infty x e^{-\lambda x} dx \\ &= \frac{2}{\lambda^2}\end{aligned}$$

$$\begin{aligned}\mathrm{Var}\,(X) &= \mathrm{E}\left(X^2\right) - (\mathrm{E}(X))^2 \\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\ &= \frac{1}{\lambda^2}\end{aligned}$$

To gain some computational insight, consider the completion of a repetitive/routine task by an office employee. Suppose that every repetition of a task takes a random duration which is governed by an Exponential$(1/4)$ distribution. As $\lambda = 1/4$, one task is, on average, completed in 4 time units (let's say, days). Using this information, let's calculate the following:

1. What is the probability that a task will be completed in exactly 2 days?
   The answer here is 0, as time (X) here is a continuous random variable.

2. What is the probability that a task will be completed within 2 days?

$$X \sim \text{Exponential}(1/4)$$

$$
\begin{aligned}
f(x) &= \frac{1}{4}e^{-\frac{1}{4}x}, x > 0 \\
F(x) &= 1 - e^{-\frac{1}{4}x}, x > 0 \\
\mathrm{P}(x \leq 2) = F(2) &= 1 - e^{-\frac{1}{4}\cdot 2} \\
&= 0.3934
\end{aligned}
$$

3. What is the probability that a task will be completed within 4 days?

$$X \sim \text{Exponential}(1/4)$$

$$
\begin{aligned}
f(x) &= \frac{1}{4}e^{-\frac{1}{4}x}, x > 0 \\
F(x) &= 1 - e^{-\frac{1}{4}x}, x > 0 \\
\mathrm{P}(x \leq 4) = F(4) &= 1 - e^{-\frac{1}{4}\cdot 4} \\
&= 0.6321
\end{aligned}
$$

4. What is the probability that a task will be completed within 6 days?

$$X \sim \text{Exponential}(1/4)$$

$$
\begin{aligned}
f(x) &= \frac{1}{4}e^{-\frac{1}{4}x}, x > 0 \\
F(x) &= 1 - e^{-\frac{1}{4}x}, x > 0 \\
\mathrm{P}(x \leq 6) = F(6) &= 1 - e^{-\frac{1}{4}\cdot 6} \\
&= 0.7768
\end{aligned}
$$

5. What is the probability that a task will be completed between 2 and 6 days?

$$X \sim \text{Exponential}(1/4)$$

$$
\begin{aligned}
f(x) &= \frac{1}{4}e^{-\frac{1}{4}x}, x > 0 \\
F(x) &= 1 - e^{-\frac{1}{4}x}, x > 0 \\
\mathrm{P}(2 \leq x \leq 6) = F(6) - F(2) &= 0.7768 - 0.3934 \\
&= 0.3834
\end{aligned}
$$

In a nutshell

Memoryless (no memory) property of Exponential distribution
A watch repairer's repair times X follow an Exponential(λ) distribution (where λ is the typical/average number of repairs per unit time).

$$X \sim \text{Exponential}(\lambda)$$

$$f(x) = \lambda e^{-\lambda x}, x > 0$$
$$F(x) = 1 - e^{-\lambda x}, x > 0$$

At time 0, I left my watch for a repair and after waiting for T_1 time units I observed that my watch was not repaired. What is the probability that I will wait for an additional (extra) T_2 time units? A careful examination will reveal the followino: as T_1 units of time has already passed, total waiting time will be at least T_1, i.e., $x > T_1$ will hold. This is nothing but a condition imposed on my full sample space of $x > 0$. Based on this, the original question turns into "beginning at T_1, what is the probability that I will wait until $T_1 + T_2$?" In technical notation, the answer is:

$$\begin{aligned}
&\frac{\mathrm{P}(T_1 < x < T_1 + T_2)}{\mathrm{P}(T_1 < x)} \\
=&\frac{F(T_1 + T_2) - F(T_1)}{1 - F(T_1)} \\
=&\frac{1 - e^{-\lambda(T_1+T_2)} - \left(1 - e^{-\lambda T_1}\right)}{1 - \left(1 - e^{-\lambda T_1}\right)} \\
=&\frac{e^{-\lambda T_1} - e^{-\lambda(T_1+T_2)}}{e^{-\lambda T_1}} \\
=&e^{-\lambda T_1 + \lambda T_1} - e^{-\lambda T_1 - \lambda T_2 + \lambda T_1} \\
=&e^0 - e^{-\lambda T_2} \\
=&1 - e^{-\lambda T_2}
\end{aligned}$$

Now, we notice that the final expression is just equal to $F(T_2)$, i.e., $\mathrm{P}(x < T_2)$. T_1 simply drops out of the solution and "probability of waiting for an additional T_2 upon T_1" is equal to "probability of waiting until T_2 at time 0". Disappearance of T_1 (or its irrelevance) implies that the random process (random variable X here) does not remember its past. This is called the memoryless (no memory) property. Drawings and graphs will be covered in the lectures.

Simulation guide

Suppose we want to simulate and analyze the purchasing decisions of 1000 customers arriving at a store. In such a simulation, the very first step is to create/generate these 1000 customers. Under certain conditions (research for them), we can (and most of the time we should) assume that arrivals of customers follow a Poisson(λ) distribution; so, interarrival times of the same customers follow an Exponential(λ) distribution (which makes our lives quite easy).
The technique we use is called the "inverse transformation technique" and utilizes the inverse CDF, i.e., $F^{-1}(x)$. The steps are:

1. Generate a sequence of uniform $(0, 1)$ random values, call these values u.
2. Find the $F^{-1}(\cdot)$ expression, i.e., derive the inverse CDF.
3. Input u into $F^{-1}(\cdot)$ to obtain a sequence of x values. These x values are the randomly generated numbers obeying/following $F(\cdot)$.

In our case of generating Exponential(λ) interatrival times:

$$\mathrm{P}(X < x) = F(x) = u, 0 \leq u \leq 1.$$

So,

$$x = F^{-1}(u)$$

where

$$F^{-1}(u) = -\frac{\ln(1-u)}{\lambda}$$

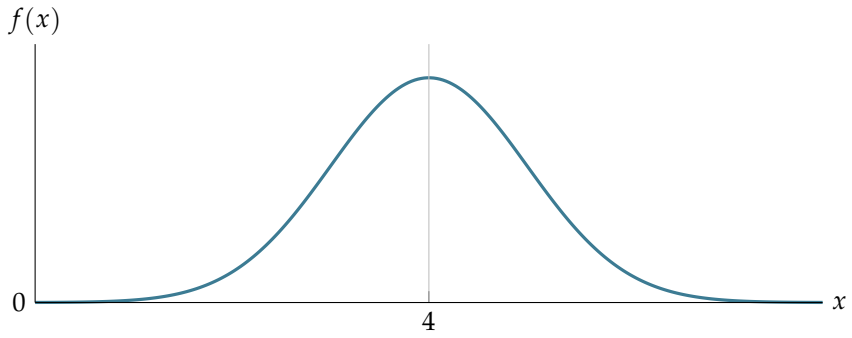
Accumulating the generated x (interarrival time) values, we can easily see the arrival times of our simulated customers. As the interarrival times have been randomly generated, arrival times are also random.
Think: What is the role and importance of Uniform$(0, 1)$ distribution here?
Think: Can you use this technique to generate random values that obey another statistical distribution?

3.8.4 Normal distribution

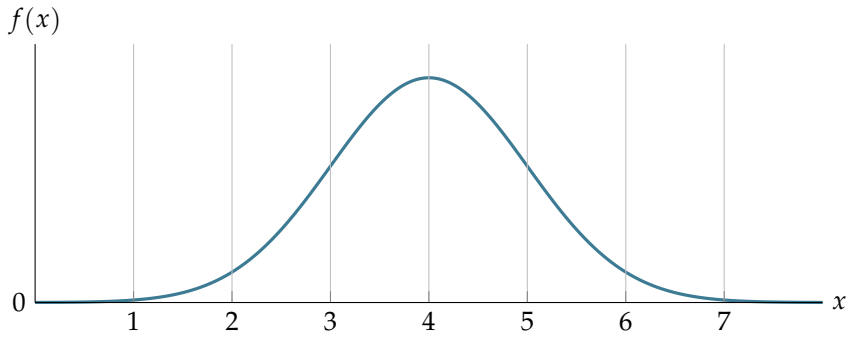
$$X \sim Normal\left(\mu, \sigma^2\right)$$

$$\forall x \in \mathbb{R}, f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, -\infty < x < \infty$$

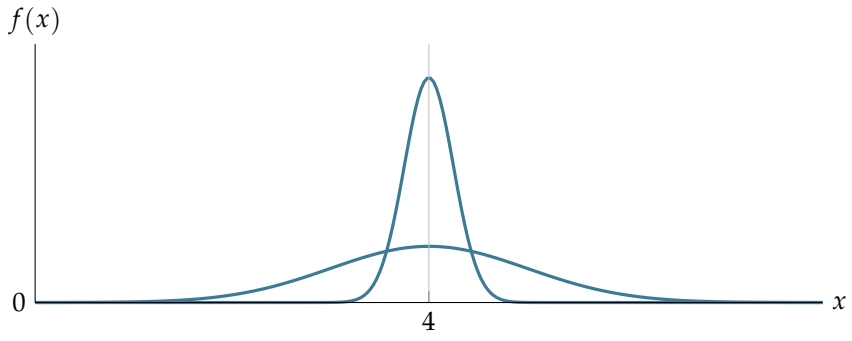
Below given the graph of Normal$(4, 1)$ PDF:



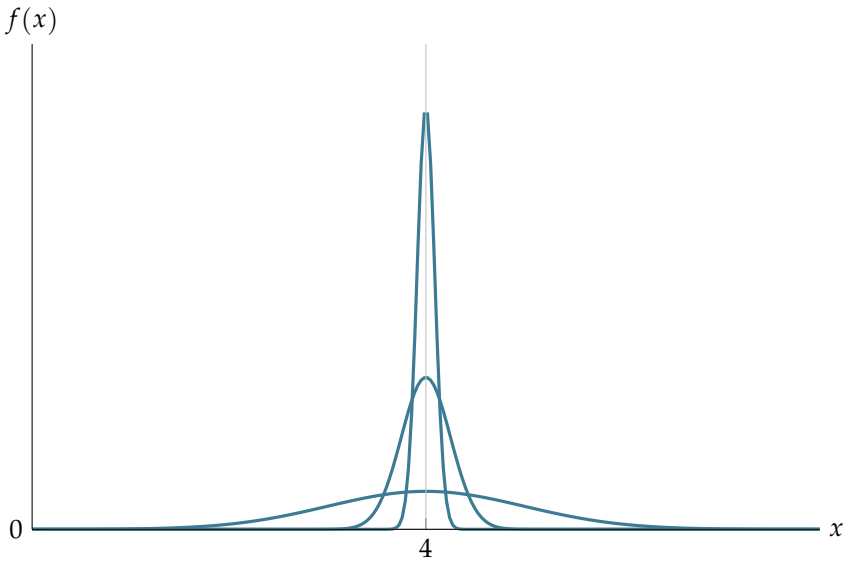
When we add the guidelines that show $\mu - 3\sigma$, $\mu - 2\sigma$, $\mu - \sigma$, $\mu + \sigma$, $\mu + 2\sigma$ and $\mu + 3\sigma$, the previous figure looks like:



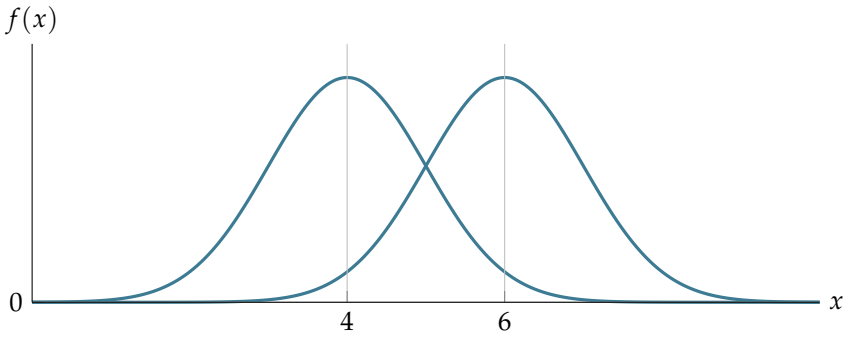
Displaying the PDFs of Normal$(4, 1)$ and Normal$(4, 0.25)$ together, we notice that the latter has a higher peak:

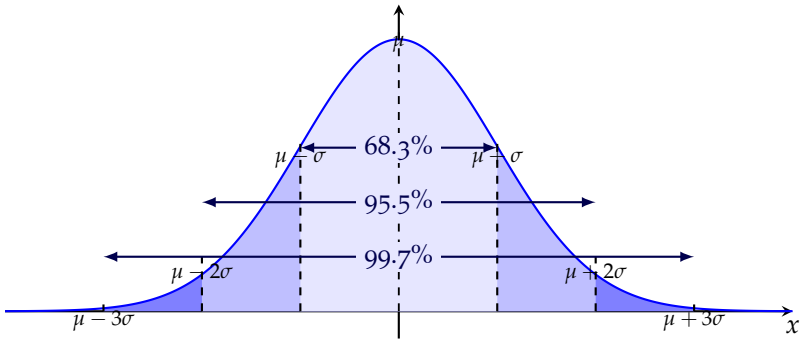


Displaying the PDFs of Normal$(4, 1)$, Normal$(4, 0.25)$ and Normal$(4, 0.09)$ together, we notice that the last has an even higher peak, the area under each PDF integrating to 1.



Keeping the variance σ^2 the same, a change in mean μ results in a shift of the PDF. Compare Normal$(4, 1)$ and Normal$(6, 1)$ below:





Expected Value and Variance:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$
\begin{aligned}
z =& \frac{x-\mu}{\sigma} \\
x =& \sigma z + \mu \\
dx =& \sigma dz
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}(X) =& \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (\sigma z + \mu)\, e^{-\frac{z^2}{2}} \sigma dz \\
=& \frac{1}{\sqrt{2\pi}\sigma^2} \int_{-\infty}^{\infty} \left(\sigma^2 z + \sigma\mu\right) e^{-\frac{z^2}{2}} dz \\
=& \underbrace{\frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} dz}_{0} + \frac{\mu}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz}_{\sqrt{2\pi}} \\
=& 0 + \frac{\mu}{\sqrt{2\pi}} \sqrt{2\pi} \\
=& \mu
\end{aligned}
$$

$$
\mathrm{E}\left(X^2\right) = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx
$$

$$
\begin{aligned}
z =& \frac{x-\mu}{\sigma} \\
x =& \sigma z + \mu \\
dx =& \sigma dz
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\left(X^2\right) =& \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (\sigma z + \mu)^2 e^{-\frac{z^2}{2}} \sigma dz \\
=& \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (\sigma^2 z^2 + 2\sigma\mu z + \mu^2) \sigma e^{-\frac{z^2}{2}} dz \\
=& \frac{\sigma^3}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz + \frac{2\sigma^2\mu}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} dz + \frac{\mu^2\sigma}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz \\
=& \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz + \frac{2\sigma\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} dz + \frac{\mu^2}{\sqrt{2\pi}} \sqrt{2\pi}
\end{aligned}
$$

$$\begin{aligned}\mathrm{E}\left(X^2\right) =& \frac{\sigma^2}{\sqrt{2\pi}}\sqrt{2\pi} + 0 + \mu^2 \\ =& \sigma^2 + \mu^2\end{aligned}$$

$$\begin{aligned}\mathrm{Var}\left(X\right) =& \mathrm{E}\left(X^2\right) - (\mathrm{E}(X))^2 \\ =& \sigma^2 + \mu^2 - \mu^2 \\ =& \sigma^2\end{aligned}$$

Consider $\int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz$. For $\alpha = \frac{1}{2}$:

$$\begin{aligned}\int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz =& \int_{-\infty}^{\infty} z^2 e^{-\alpha z^2} dz \\ =& -\int_{-\infty}^{\infty} \frac{d}{d\alpha} e^{-\alpha z^2} dz \\ =& -\frac{d}{d\alpha}\int_{-\infty}^{\infty} e^{-\alpha z^2} dz\end{aligned}$$

Set $\omega = \frac{z}{\sqrt{2\alpha}}$:

$$\begin{aligned}\int_{-\infty}^{\infty} e^{-\alpha z^2} dz =& \frac{1}{\sqrt{2\alpha}} \underbrace{\int_{-\infty}^{\infty} e^{-\frac{\omega^2}{2}} d\omega}_{\sqrt{2\pi}} \\ =& -\frac{d}{d\alpha}\sqrt{\frac{\pi}{\alpha}} \\ =& -\sqrt{\pi}\frac{d}{d\alpha}\alpha^{-1/2} \\ =& \frac{\sqrt{\pi}}{2}\alpha^{-3/2} \\ =& \frac{\sqrt{\pi}}{2}\left(\frac{1}{2}\right)^{-3/2} \\ =& \frac{\sqrt{\pi}}{2}2^{3/2} \\ =& \sqrt{2\pi}\end{aligned}$$

3.8.5 *Standard normal distribution*

$Z \sim Normal\,(0,1)$ has the standard normal distribution. If $X \sim Normal\left(\mu,\sigma^2\right)$, the random variable Z defined as:

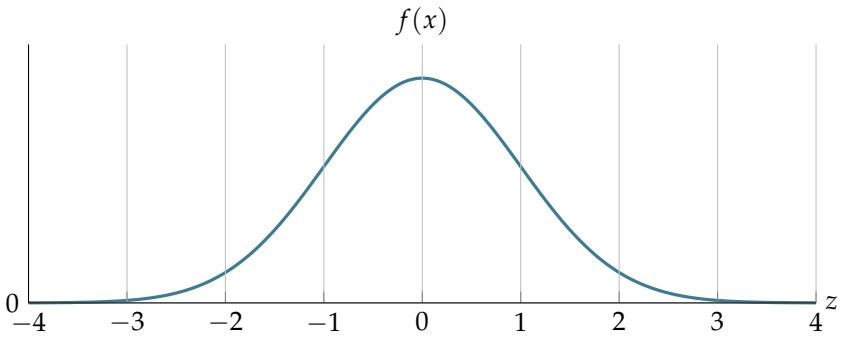
$$Z = \frac{X-\mu}{\sigma}$$

has a Normal$(0,1)$ distribution. A casual naming is **z-distribution**, and

$$f(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z^2}, -\infty < z < \infty$$

Recall that, e is called the **Euler's Number** where $\mathrm{e} = 2.71828\ldots$ and $\pi = 3.14159\ldots$

Notice/recall that the PDF of the Standard Normal (Z) random variable has a unique parametrization. Its PDF with the guidelines that show $\mu - 3\sigma = -3$, $\mu - 2\sigma = -2$, $\mu - \sigma = -1$, $\mu = 0$, $\mu + \sigma = 1$, $\mu + 2\sigma = 2$ and $\mu + 3\sigma = 3$ look like:



In a nutshell

Normal approximation to Binomial distribution
Let X be the number of successes resulting from n independent trials, each with probability of success P. The distribution of the number of successes, X, is binomial, with mean nP. If the number of trials, n, is large and $nP(1-P) > 5$, this distribution can be approximated by the Normal distribution with $\mu = nP$ and $\sigma^2 = nP(1-P)$. So,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, -\infty < x < \infty$$

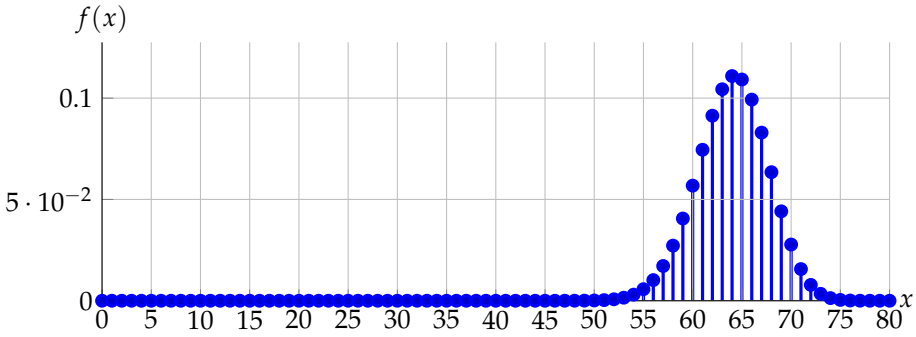
can safely be used to obtain a numerical result, where

$$Z = \frac{X - nP}{\sqrt{nP(1-P)}}$$

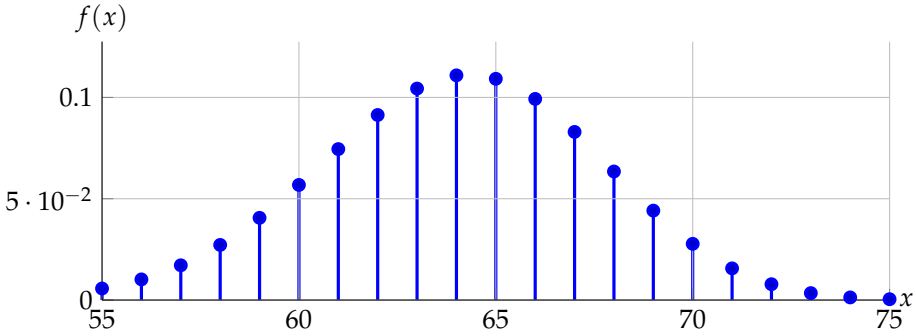
has a Standard Normal distribution.

To see how/why the Normal approximation to Binomial works, consider the PDF of $X \sim \text{Binomial}(80, 0.80)$ over the domains of $0, 1, \ldots, 80$ and $55, 56, \ldots, 75$ below:

PDF of $X \sim \text{Binomial}(80, 0.80)$ plotted over $0, 1, \ldots, 80$ looks like:



PDF of $X \sim \text{Binomial}(80, 0.80)$ plotted over $55, 56, \ldots, 75$ looks like:



Do you see the Normal-like behavior of $X \sim \text{Binomial}(80, 0.80)$ around its mean, i.e., $nP = 80 \cdot 0.80 = 64$? Can you obtain the same with $X \sim \text{Binomial}(8, 0.80)$? Why?

3.9 *Random variables and distributions: Moments of distributions [Optional material]*

For each integer k, the k^{th} moment of X is denoted as $\mu_K^{'}$ and is defined as:

$$\mu_K^{'} = \text{E}\left(\text{X}^k\right)$$

The k^{th} central moment of X is denoted as μ_k and is defined as:

$$\mu_K = \text{E}\left((X - \mu)^k\right)$$

Notice that $\mu = \mu_1^{'} = \text{E}(X)$ In addition to the mean (expected value) of a random variable, another important moment is the second central moment, as you've known as variance.

3.10 *Moment generating functions [Optional material]*

X being a random variable with CDF $F(x)$, the moment generating function (**MGF**) of X is denoted by $M_X(t)$ and is defined as:

$$M_X(t) = \mathrm{E}\left(\mathrm{e}^{tX}\right)$$

provided that the expected value exists for t in some neighborhood of zero. That is, there exists $h > 0$ such that for all $-h < t < h$, $\mathrm{E}\left(\mathrm{e}^{eX}\right)$ exists. Otherwise, the MGF is said not to exist. Explicitly,

$$M_X(t) = \int_{-\infty}^{\infty} \mathrm{e}^{tX} f(x)\, dx, \textit{continuous } X$$

or

$$M_X(t) = \sum_x \mathrm{e}^{tX} f(x)\, dx, \textit{discrete } X$$

3.10.1 *Moment generating functions for selected distributions [Optional material]*

Distribution	$M_X(t)$
Bernoulli(p)	$(1-p) + p\mathrm{e}^t$
Binomial(n,p)	$\left((1-p) + p\mathrm{e}^t\right)^n$
Poisson(λ)	$\mathrm{e}^{\lambda\left(\mathrm{e}^t - 1\right)}$
χ^2_n	$\left(\frac{1}{1-2t}\right)^{\frac{n}{2}},\ t < \frac{1}{2}$
Exponential(λ)	$\frac{1}{1-\frac{t}{\lambda}},\ t < \lambda$
F_{n_1,n_2}	*Does not exist*
Normal(μ,σ^2)	$\mathrm{e}^{\mu t + \frac{\sigma^2 t^2}{2}}$
t_n	*Does not exist*
Uniform(a,b)	$\frac{\mathrm{e}^{bt} - \mathrm{e}^{at}}{(b-a)t}$

If a random variable X has the MGF $M_X(t)$, then

$$\mathrm{E}(X^n) = \frac{d^n}{dt^n} M_X(t|_{t=0}$$

That is, the n^{th} moment of X is equal to the n^{th} derivative of $M_X(t)$ evaluated at $t = 0$. See after five years: convergence of MGF's.

3.2 EXERCISES

1. We roll a pair of fair dice. Let X be the random variable that assigns the minimum of the two numbers that turn up to each outcome.

i. Tabulate the probability density function and cumulative distribution function of X.

ii. If we know that one of the dice turned up a number less than or equal to 3, what is the probability that X takes a value greater than or equal to 2?

iii. If we know that one of the dice turned up a number less that or equal to 3, what is the probability that X takes a value equal to 3?

iv. Find the expected value of X.

v. Find the variance of X.

Solution:

1. PDF is tabulated as follows:

x	$f(x)$
1	$\frac{11}{36}$
2	$\frac{9}{36}$
3	$\frac{7}{36}$
4	$\frac{5}{36}$
5	$\frac{3}{36}$
6	$\frac{1}{36}$

2. This is a conditional probability question that you already are familiar with.

3. This is a conditional probability question that you already are familiar with.

4.

$$\begin{aligned}
\mathrm{E}(X) &= \sum_x x f(x) \\
&= 1 \cdot \frac{11}{36} + 2 \cdot \frac{9}{36} + 3 \cdot \frac{7}{36} + 4 \cdot \frac{5}{36} + 5 \cdot \frac{3}{36} + 6 \cdot \frac{1}{36} \\
&= \frac{11 + 18 + 21 + 20 + 15 + 6}{36} \\
&= \frac{91}{36} = 2.53
\end{aligned}$$

5.

$$\begin{aligned}
\mathrm{Var}\,(X) &= \sum_x (x - \mathrm{E}(X))^2 f(x) \\
&= (1 - 2.53)^2 \cdot \frac{11}{36} + \cdots + (6 - 2.53)^2 \cdot \frac{1}{36} \\
&= 1.97
\end{aligned}$$

2. Two balls are simultaneously chosen (*i.e.*., chosen without replacement) from an urn containing 3 white, 2 black, and 1 red balls. You are given 2TL for each white ball chosen, you have to pay 1TL for each black ball chosen, and you neither pay nor receive any money for a red ball that is chosen. For example if you have chosen 1 white and 1 black ball, you net winning is $2 + (-1) = 1$ TL. Let X be the random variable that gives your net winnings.

i. Construct a table that shows the possible values of X and the probabilities associated with each value, *i.e.*., tabulate the probability density (mass) function of X.

ii. Find the expected value of X.

Solution:

1. PDF is tabulated as follows:

x	$f(x)$
-2	$\frac{2}{30}$
-1	$\frac{4}{30}$
1	$\frac{12}{30}$
2	$\frac{6}{30}$
4	$\frac{6}{30}$

2.

$$\begin{aligned} \mathrm{E}(X) &= \sum_x x f(x) \\ &= (-2)\cdot\frac{2}{30} + (-1)\cdot\frac{4}{30} + 1\cdot\frac{12}{30} + 2\cdot\frac{6}{30} + 4\cdot\frac{6}{30} \\ &= \frac{-4-4+12+12+24}{30} \\ &= \frac{40}{30} = 1.33 \end{aligned}$$

3. A class in statistics has 20 students. In the first midterm 2 students scored 50, 10 scored 60, 1 scored 70, 5 scored 80, and 2 scored 100. Three students are selected at random without replacement. Let X be the median score of the three students.

i. Tabulate the probability density function of X.

ii. Find the probability of the median score being greater or equal to 80.

iii. Given that the median of the scores of the three students selected is greater than or equal to 70, what is the probability that their median is equal to 80?

iv. Find the expected value and variance of X.

Solution: Try on your own if you have time, and just for fun.

4. We have three coins such that when coin 1 is tossed the probability of observing a head is 0.4, when coin 2 is tossed the probability of observing a head is 0.7, and when coin 3 is tossed the probability of observing a head is 0.2. We first toss coin 1. If we observe a head we toss coin 2 otherwise we choose coin 1 or coin 3 at random and toss it.

i. What is the probability of observing a head on the second toss?

ii. Are the events of observing a head on the second toss and observing a head on the first toss independent?

Solution: This question is reserved for in-class discussions.

5. A fair die is rolled ten times. We are interested in the number of times 6 is obtained.

i. Given our interest, can we think of this experiment as a binomial experiment. If so describe each Bernoulli trial, *i.e.*. verbally describe the Bernoulli trial, state the outcome that you will call success and the probability of success in each trial.

ii. Let X be the random variable which assigns, to each outcome, the number of times 6 is obtained in the outcome. What is the distribution of X?

iii. With what probability will X take the value 1?

iv. With what probability will X take a value greater than or equal to 4?

Solution:

1. Success is observing a 6, failure is observing any of $\{1,2,3,4,5\}$. Since the die is fair, the probability of success is $1/6$. X being the random variable indicating the outcome of experiment,

$$f(x) = \begin{cases} 1/6, & x = 1 \\ 5/6, & x = 0 \end{cases}$$

that is, $X \sim$ Bernoulli $(1/6)$.

2. Considering the whole experiment, $X \sim$ Binomial$(10, 1/6)$.

$$f(x) = \begin{pmatrix} 10 \\ x \end{pmatrix} (1/6)^x (5/6)^{10-x}, x = 0, 1, 2, \ldots, 10.$$

3.

$$f(1) = \begin{pmatrix} 10 \\ 1 \end{pmatrix} (1/6)^1 (5/6)^9 = 0.3230$$

4.

$$\begin{aligned} \mathrm{P}(x \geq 4) &= 1 - f(0) - f(1) - f(2) - f(3) \\ &= 0.0697 \end{aligned}$$

6. It is known that 40% of all students of Economics are male. Independent observers note the gender of 12 random Economics students (a student's gender might be noted more than once) and we count the number of males observed.

i. What is the probability that **exactly** 2 of the observed students are male?

ii. What is the probability that the number of male students, in the group observed, is 5 **or less**?

iii. You have been told that **at least** 2 of the students that has been observed are female. What is the probability that the number of male students, in the observed group, is 5 **or less**?

Solution:

1. $X \sim \text{Binomial}(12, 0.40)$

$$\begin{aligned} f(2) &= \begin{pmatrix} 12 \\ 2 \end{pmatrix} 0.40^2 0.60^{10} \\ &= 0.0639 \end{aligned}$$

2. $x \sim \text{Binomial}(12, 0.40)$

$$\begin{aligned} \mathrm{P}(X \leq 5) &= f(0) + f(1) + f(2) + f(3) + f(4) + f(5) \\ &= F(5) \\ &= 0.6652 \end{aligned}$$

3. This question is reserved for in-class discussions.

7. Consider a game where a round of the game consists of rolling a fair die 10 times. Each time a 1 or 6 comes you win 1*TL*.

i. What is the probability that you will win 5 *TL* or less, if you played this game for one round?

ii. What is the probability that you will win exactly 5 *TL*, if you played this game for one round?

iii. You have learned that two of the rolls of the die resulted with a number different than 1 or 6, but you do not know what the result of the other rolls of the die was. What is the probability that you will win **more than** 5*TL*?

iv. What would your average (mean) winnings be if you played this game indefinitely?

Solution:

1. $X \sim \text{Binomial}(10, 1/3)$

$$\begin{aligned} \text{P}(X \leq 5) &= F(5) \\ &= 0.9234 \end{aligned}$$

2. $X \sim \text{Binomial}(10, 1/3)$

$$\begin{aligned} f(5) &= \begin{pmatrix} 10 \\ 5 \end{pmatrix} (1/3)^5 (2/3)^5 \\ &= 0.1366 \end{aligned}$$

3. This question is reserved for in-class discussions.
4. $X \sim \text{Binomial}(10, 1/3)$

$$\begin{aligned} \text{E}(X) &= nP = 10 \cdot \frac{1}{3} \\ &= 3.33 \end{aligned}$$

8. Based on past data, we know that, on average, 6 customers enter Coffee Break every 20 minutes.

i. What is the probability that **at least** 2 customers will enter Coffee Break during a given 20-minute time period?

ii. Define the probability of k customers entering Coffee Break in 20 minutes as a mathematical function. Describe what is what in your function clearly.

Solution:

1. $X \sim \text{Poisson}\,(6)$ $\text{P}(\text{At least 2 customers}) = 1 - \text{P}(\text{At most 1 customer})$ $= 1 - \frac{e^{-6}6^0}{0!} - \frac{e^{-6}6^1}{1!} = 0.9826$
2. $X \sim \text{Poisson}\,(6)$

$$f(x) = \frac{e^{-6}6^x}{x!}, x = 0, 1, 2, \ldots$$

X being the random variable that shows the number of customers arriving every 20 minutes. The rate of arriving customers is $\lambda = 6$. X is a Poisson random variable.

9. On an ordinary day, on average 3 white and 1 blue cars pass through a certain cross-section of a road every 5 minutes.

i. What is the probability that 6 white cars will pass in a 5-minute interval?

ii. What is the probability that 6 white cars will pass in a 10-minute interval?

iii. What is the probability that 3 cars (blue or white) will pass in a 5-minute interval?

Solution:

1. $W \sim \text{Poisson}(3)$

$$f(6) = \frac{e^{-3}3^6}{6!} = 0.0504$$

2. $W \sim \text{Poisson}(6)$

$$f(6) = \frac{e^{-6}6^6}{6!} = 0.1606$$

3.

$$X = W + B$$
$$W \sim \text{Poisson}(3)$$
$$B \sim \text{Poisson}(1)$$

As Poisson λ's are additive, $X \sim \text{Poisson}(4)$.

$$f(3) = \frac{e^{-4}4^3}{3!} = 0.1954$$

10. A Hypergeometric story: In a corporation, promotion decisions for employees are made by a committee of 5 people. The decision making procedure has the following steps:

1. Each of the 5 writes her vote (either 'Promote' or 'Not promote') on a piece of paper, folds the paper twice and casts the paper into a bowl.
2. Another person from outside the committee randomly picks 3 out of the 5 votes. (This is a step taken to anonymize votes).
3. The 3 picked papers are opened. If the employee gets 2 or 3 votes, then she is promoted. If she gets **no** votes or 1 vote, she is not promoted.

Consider Employee A for whom the chance of a promotion is P in the eyes of each committee member. That is, each committee member has a chance of P to promote Employee A. Also, preferences of committee members are independent from each other's. Is there a chance to be accidentally or unfairly promoted (or not promoted) in this kind of scheme?

Solution: The solution involves some steps:
First, a 'Promotion' vote being marked as Success, each committee member's vote is a Bernoulli trial:

$$X_i \sim \text{ Bernoulli } (P), i = 1, 2, 3, 4, 5$$

Then, total votes (total of successes) (Y) is a Binomial process:

$$Y = X_1 + X_2 + X_3 + X_4 + X_5$$
$$Y \sim \text{Binomial}(5, P), \quad y = 0, 1, 2, 3, 4, 5$$

Then, W being the number of 'Promotion' votes among the final 3, W has a Hypergometric distribution:
$W \sim Hypergeometric(5, Y, 5 - Y)$
So,

$$f(x_i) = \begin{cases} P, x_i = 1 \\ 1 - P, x_i = 0 \end{cases}$$
$$g(y) = \begin{pmatrix} 5 \\ y \end{pmatrix} P^y (1 - P)^{5-y}$$
$$h(w) = \frac{\begin{pmatrix} y \\ w \end{pmatrix} \begin{pmatrix} 5 - y \\ 3 - w \end{pmatrix}}{\begin{pmatrix} 5 \\ 3 \end{pmatrix}}$$

Now, your task is to find $g(y)$ for each value of y. Then you will calculate $h(w)$ for each different value of y. At the end, you will compare Employee A's chance to promote with and without the **Step 2&3** of the promotion procedure. Note that the result may be a little surprising.

11. Let X_1 be the random variable that gives the number of phone calls that you get between 1 PM and 2 PM. Let X_2 be the random variable that gives the number of phone calls that you get between 2PM and 4PM. Assume that X_1 is Poisson distributed with parameter 5 and X_2 is Poisson distributed with parameter 12. Let X be the random variable that gives the number of phone calls that you get between 1PM and 4PM. Find the PDF of X.
Solution: $X_1 \sim$ Poisson (5) and $X_2 \sim$ Poisson (12). $X = X_1 + X_2, X \sim$ Poisson (17). Make sure you have obtained this result by following the chapter's instructions.

12. Let X_1 be the random variable that gives the number of phone calls that you get between 1PM and 2PM. Let X_2 the random variable that gives the number of phone calls that your friend gets between 1PM and 2PM. Assume that X_1 is Poisson distributed with parameter λ_1 and X_2 is Poisson distributed with parameter λ_2. Find the distribution of $X = X_1 + X_2$, *i.e.*., the PDF of the random variable that gives the total number of phone calls that you and your friend receive between 1PM and 2PM.

Solution: The solution method is already arailable in the chapter.

13. Suppose that you buy 40 lottery tickets. Using the Poisson approximation find the probability of having **at least** 2 winning tickets, given that the probability of any ticket being a winning ticket is 0.02.
Solution: This is self-study for those who are interested. Not to appear in any examination.

14. Let X be a random variable that is uniformly distributed over $(-1, 3)$. Answer the following questions:

i. Find $P(X < 0)$

ii. Find $P\left(\frac{1}{2} < X < 1\right)$

iii. Find $P(X > 2)$

iv. What is the expected value and variance of X?

Solution:

1. $X \sim \text{Uniform}(-1, 3)$

$$f(x) = \frac{1}{3-(-1)} = \frac{1}{4}, -1 \leq x \leq 3$$
$$\text{P}(X < 0) = \int_{-1}^{0} \frac{1}{4} dx = \frac{x}{4}\Big|_{-1}^{0}$$
$$= \frac{0}{4} - \frac{-1}{4}$$
$$= 1/4$$

2.

$$\text{P}(1/2 < x < 1) = \int_{1/2}^{1} \frac{1}{4} dx$$
$$= \frac{x}{4}\Big|_{1/2}^{1}$$
$$= \frac{1}{4} - \frac{1}{8}$$
$$= 1/8$$

3.

$$\text{P}(X > 2) = \int_{2}^{3} \frac{1}{4} dx$$
$$= \frac{x}{4}\Big|_{2}^{3}$$
$$= \frac{3}{4} - \frac{2}{4}$$
$$= 1/4$$

4·

$$\begin{aligned} \mathrm{E}(X) &= \frac{-1+3}{2} \\ &= 1 \\ \mathrm{Var}\,(X) &= \frac{(3-(-1))^2}{12} \\ &= \frac{16}{12} \\ &= 4/3 \end{aligned}$$

15. A potato chips producer starts a promotion program in an effort to boost its sales. In that, gift tickets are placed in every 25 out of 100 chip bags in sale and the customers are required to collect **two** tickets to win a free soft drink. By the nature of such promotions, gift tickets are invisible from outside prior to purchase. In order to attain a probability of 90% at minimum to win a soft drink, how many bags of potato chips should an average customer buy? **Notes**:

- In case a manual solution for this problem is not feasible, you are required to provide a good mathematical formulation of the solution approach along with proper explanations.
- In a real life setting it is not likely to observe a high prize ratio like $\frac{25}{100}$. Instead, one may observe ratios even lower than $\frac{1}{100}$.

Solution: This is to be discussed in class only along with a computer demo.

16. Let X be a random variable with the following PDF:

x	-3	-1	0	1	2	3
$f(x)$	0.25	0.10	0.05	0.20	0.30	0.10

Define a new random variable Y as

$$Y = X^2 + 1$$

i. Find the PDF of Y

ii. Find the CDF of Y

iii. Find the expected value of Y and show that it is equal to $\sum \left(x^2+1\right) f_X(x)$, where $f(x)$ is the PDF of X

iv. Find the variance of Y

Solution: This question is reserved for in-class discussions.

17. Let X be a random variable normally distributed with expected value of 2 and variance of 9. Answer the following questions:

i. Find $P(X < 5.15)$

ii. Find $P(X < -1)$

iii. Find $P(X > 4)$

iv. Find $P(1.04 < X < 3.5)$

Solution:

1.

$$\begin{aligned} X &\sim \text{Normal}(2, 9) \\ P(x < 5.15) &= P\left(\frac{x-2}{3} < \frac{5.15-2}{3}\right) \\ &= P(Z < 1.05) \\ &= 0.85314 \end{aligned}$$

2.

$$\begin{aligned} P(X < -1) &= P\left(\frac{x-2}{3} < \frac{-1-2}{3}\right) \\ &= P(Z < -1) \\ &= P(Z > 1) \\ &= 1 - F(1) \\ &= 1 - 0.84134 \\ &= 0.15866 \end{aligned}$$

Reveal how symmetry property is used here.

3.

$$\begin{aligned} P(X > 4) &= P\left(\tfrac{x-2}{3} > \tfrac{4-2}{3}\right) \\ &= P(Z > 0.66) \\ &= 1 - F(0.66) \\ &= 1 - 0.74537 \\ &= 0.25463 \end{aligned}$$

4.

$$\begin{aligned} & P(1.04 < X < 3.5) \\ = \; & P\left(\tfrac{1.04-2}{3} < \tfrac{x-2}{3} < \tfrac{3.5-2}{3}\right) \\ = \; & P(-0.32 < z < 0.50) \\ = \; & F(0.50) + F(0.32) - 1 \\ = \; & 0.69146 + 0.62552 - 1 \\ = \; & 0.31698 \end{aligned}$$

Study this solution by drawing proper graphs of the PDF of the Standard normal distribution.

18. It is estimated that 45% of the freshmen entering a particular college will graduate from that college in four years.

i. For a random sample of 5 entering freshmen, what is the probability that exactly 3 will graduate in four years?

ii. For a random sample of 5 entering freshmen, what is the probability that a majority (more than half) will graduate in four years?

iii. 80 entering freshmen are chosen at random. Find the mean and variance of the number of these 80 that will graduate in four years.

Solution:

1. $X \sim \text{Binomial}(5, 0.45)$

$$f(3) = \begin{pmatrix} 5 \\ 3 \end{pmatrix} 0.45^3 0.55^2 = 0.2757$$

2. $X \sim \text{Binomial}(5, 0.45)$

$$f(3) + f(4) + f(5) = 0.4069$$

3. $X \sim \text{Binomial}(80, 0.45)$

$$\begin{aligned} \mathrm{E}(X) &= nP = 80 \cdot 0.45 = 36 \\ \mathrm{Var}\,(X) &= n\,\mathrm{P}(1-P) = 80 \cdot 0.45 \cdot 0.55 = 19.8 \end{aligned}$$

19. Bags of our packed by a particular machine have weights which are normally distributed with mean of $500gr$ and standard deviation of $20gr$.

i. What is the probability that a bag weights more than $515gr$ or less than $490gr$?

ii. If 2% of the bags are rejected for being underweight, what is the maximum weight for a bag to be rejected as underweight?

iii. Find an interval $[l, u]$ symmetric around the mean and the probability of the weight of a randomly selected bag being in the interval is 0.90.

Solution:

1. $X \sim \text{Normal}(500, 400)$

$$\begin{aligned} &\mathrm{P}(X < 490) + \mathrm{P}(X > 515) \\ =&\,\mathrm{P}\left(Z < \frac{490 - 500}{20}\right) + \mathrm{P}\left(Z > \frac{515 - 500}{20}\right) \\ =&\,\mathrm{P}(Z < -0.50) + \mathrm{P}(Z > 0.75) \end{aligned}$$

Use your z table, the answer is $0.30853 + 0.22662$, i.e., 0.53516.

2. This question is reserved for in-class discussions.

3. This question is reserved for in-class discussions.

20. X is distributed as $Bin\,(100, 0.04)$. Describe the steps to calculate $\mathrm{P}(X > k)$ for a given k by using a Poisson approximation.
Solution: This is self-study for thase who are interested. Not to appear in any examination.

21. We know that the number of vampires killed by Dean in a typical fight has a Poisson(5) distribution and the number of vampires killed by Sam in a typical fight has a Poisson(3) distribution. Show that the total number of vampires killed in a typical fight follows a Poisson(8) distribution.
Solution: The solution method is already available in the chapter.

22. X has a Uniform$(0, 100)$ distribution. Calculate $\mathrm{P}(33 < X < 67)$, $\mathrm{E}(X)$ and $\mathrm{Var}\,(X)$.
Solution: $X \sim \mathrm{Uniform}(0, 100)$

$$f(x) = \frac{1}{100-0}, 0 \leq x \leq 100$$
$$F(x) = \frac{x}{100}, 0 \leq x \leq 100$$

$$\begin{aligned}\mathrm{P}(33 < X < 67) &= F(67) - F(33)\\ &= \frac{67}{100} - \frac{33}{100}\\ &= \frac{34}{100}\end{aligned}$$

$$\mathrm{E}(X) = \frac{0+100}{2} = 50$$

$$\mathrm{Var}\,(X) = \frac{(100-0)^2}{12} = 833.33$$

23. Assume that the number of phone calls that you receive in a day is governed by a *Poisson* process. Answer the following questions assuming that on average you receive 3.4 phone calls in a day.

i. What is the probability that you will receive **at least**3 phone calls in a day?

ii. Given that you have already received a phone call, what is the probability that you will receive **at least**3 phone calls? [This part is a little annoying.]

Solution: This question is left as self-study.

24. The probability density function for a random variable X is defined as:

$$f(x) = \begin{cases} 0.5x & 0 < x < a \\ 0 & elsewhere \end{cases}$$

i. Find the value of a that makes $f(x)$ a well-defined probability function.

ii. Calculate $\mathrm{E}(X)$ and $\mathrm{Var}\,(X)$

Solution:

1. $\int_0^a 0.5x dx = 1 \;\; 0.5\frac{x^2}{2}\Big|_0^a = 1 \;\; a^2 - 0^2 = 4 \;\; a^2 = 4 \;\; a = 2$. When a equals 2, the given $f(x)$ becomes a well-defined probability (distribution) function.

2.

$$\begin{aligned} \mathrm{E}(X) &= \int_0^2 x 0.5x dx \\ &= 0.5 \int_0^2 x^2 dx \\ &= 0.5\frac{x^3}{3}\Big|_0^2 \\ &= 4/3 \\ \mathrm{Var}\,(X) &= \int_0^2 (x - 4/3)^2 0.5x dx \end{aligned}$$

Try on your own.

25. I roll a die repeatedly. In each roll, if the outcome is 3, my score increases by 1; nothing happens otherwise. Knowing that my score was initially zero, what are the expected value and variance of my score right after the 10000^{th} roll?
Solution: Reveal that this physical experiment generates a random variable X with $X \sim \mathrm{Binomial}(10000, 1/6)$. Then, $\mathrm{E}(X) = 10000 \cdot (1/6) = 1666.7$ and $\mathrm{Var}\,(X) = 10000 \cdot (1/6) \cdot (5/6) = 1388.9$

26. Suppose that you're in charge of marketing airline seats for a major carrier. Four days before the flight date you've 16 seats remaining on the aircraft. You know from the past experience that 80% of the people that purchase tickets in this time period will actually show up for the flight.

i. If you sell 20 tickets, what is the probability that you'll overbook the flight or have at least one empty seat?

ii. What if you sell 18 tickets?

Solution: This exercise is left as self-study.

27. A machine that produces stampings for automobile engines is malfunctioning and producing 10% defectives. The defective and non-defective stampings proceed from the machine in a random manner. If the next five stampings are tested, find the probability that three of them are defective.
Solution: This question is reserved for in-class discussions.

28. The variance of a *Poisson* random variable X is known to be 4. Calculate manually the probability that X takes a value of **at least**2. For ease, take e $= 3$.
Solution: $X \sim \text{Poisson}(\lambda)$, $\text{Var}(X) = 4$. Since for a Poisson random variable $X \sim \text{Poisson}(\lambda), \text{Var}(X) = \lambda$, $\lambda = 4$ here.

$$\begin{aligned}
\text{P}(x \geq 2) &= 1 - \text{P}(x \leq 1) \\
&= 1 - f(0) - f(1) \\
&= 1 - \frac{3^{-4}4^0}{0!} - \frac{3^{-4}4^1}{1!} \\
&= 1 - \frac{1}{81} - \frac{4}{81} \\
&= \frac{76}{81}
\end{aligned}$$

29. Median and the coefficient of variation for a random variable $X \sim N\left(\mu, \sigma^2\right)$ are given as 100 and 0.25, respectively. Given $F(-0.84) = 0.20$ for the standard normal distribution, calculate the 80^{th} pecentile of X.
Solution: Median of a Normal random variable is equal to μ, so $\mu = 100$.

$$\begin{aligned}
CV &= \frac{\sigma}{\mu} \\
0.25 &= \frac{\sigma}{100} \\
\sigma &= 25 \\
\sigma^2 &= 625
\end{aligned}$$

So, $X \sim$ Normal $(100, 625)$. $F(-0.84) = 0.20$ implies $F(0.84) = 1 - 0.20 = 0.80$ by the symmetry of the Standard normal distribution. Based on these, the $80th$ percentile of X is found as:

$$\begin{aligned}
\text{P}_{80}(X) &= 100 + 0.84 \cdot 25 \\
&= 121
\end{aligned}$$

Self-practice this solution by drawing the Normal and Standard normal PDFs.

30. Find the value of

$$\int_{-2}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

with proper explanations.
Solution: Notice that the question requires the calculation of $F(1) - F(-2)$ for the Standard normal random variable. It is nothing but the area under the standard normal PDF from -2 to 1. The answer is 0.8186.

31. Calculate $P \leq 8$ for $X \sim Bin\,(1000, 0.010)$ using the *Normal* approximation to *Binomial* distribution.
Solution: $X \sim Bin(1000, 0.010)$ has $\mathrm{E}(X) = 10$ and $\mathrm{Var}\,(X) = 9.9$. So, X can be approximated by $X \sim \mathrm{Normal}(10, 9.9)$. Calculation of $\mathrm{P}(X \leq 8)$ is then:

$$\begin{aligned} \mathrm{P}(X \leq 8) &= \mathrm{P}\left(\frac{X - 10}{\sqrt{9.9}} \leq \frac{8 - 10}{\sqrt{9.9}} \right) \\ &= \mathrm{P}(Z \leq -0.64) \\ &= 0.262 \end{aligned}$$

32. We make 100 independent observations from a normal population with mean 40 and standard deviation 20. Approximately, what is the probability that the mean of these observations will be less than or equal 37?
Solution: Indeed, this question is an early reference to sampling distributions. Each of the 100 observations has a Normal $(40, 400)$ distribution, name them as $X_1, X_2, \ldots, X_{100}$. Try to see $\bar{X} \sim \mathrm{Normal}(40, 4)$. Calculation of $\mathrm{P}(\bar{X} \leq 37)$ is then straightforward.

33. Suppose the PDF of a logistic random variable X is given by

$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2}, -\infty < x < \infty$$

Among the many of its parametrizations, this simple function is something you are familiar from your lab work during the semester.

i. Find $F(x)$ for the random variable X by performing the necessary calculus operations.

ii. Verify that the $F(x)$ you have found possesses the properties of a CDF

iii. Using the graph of $F(x)$ only, find the value of $\mathrm{E}(X)$

Solution:

1.

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} \frac{e^{-x}}{(1+e^{-x})^2} dx \\ &= \frac{1}{1+e^{-x}}\Big|_{-\infty}^{x} \\ F(x) &= \frac{1}{1+e^{-x}}, -\infty < x < \infty \end{aligned}$$

2. $F(-\infty) = 0$, $F(\infty) = 1$ and $F(\cdot)$ is nondecreasing. F is a proper CDF here.

3. Try it using WolframAlpha or GeoGebra.

34. The class grades after an exam has a normal distribution with a mean of 50 and a variance of 144. If a student is known to have a grade less than 70, what is the probability that she has received a grade between 40 and 60?
Solution: Except for the use of conditional probatrilities, this is a trivial question. Try on your own.

35. An experimenter tosses a coin (with $\mathrm{P}(Tail) = P$) until obtaining r successes (tails). What is the distribution of the number of tosses (X) to get r successes? Derive its PDF. Hint: $X = x$ can occur only if there are exactly $r-1$ successes in the first $x-1$ trials. When you notice the first $x-1$ trials have a Binomial structure, the rest is trivial.
Solution: $\mathrm{P}_r(x) = \mathrm{P}(\text{r-1 successes in the first x-1 trials}) \times \mathrm{P}(\text{a success at the x-th trial})$. The first term is nothing but the Binomial PDF & the second term is simply P. So,

$$\begin{aligned} \mathrm{P}_r(x) &= \binom{x-1}{r-1} P^{r-1}(1-P)^{x-1-(r-1)} \times P \\ &= \underbrace{\binom{x-1}{r-1} P^{r}(1-P)^{x-r}}, x = r, r+1, r+2, \ldots \end{aligned}$$

This is the PDF of the random variable described.

36. Consider a random variable x with $f(x) = \frac{k}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, 0 \leq x < \infty$. Notice the resemblance of this PDF of the standard Normal PDF. Though, domain of $f(x)$ spans the nonnegative real numbers. What should k be to make $f(x)$ a proper PDF? Plot $f(x)$ using your solution for k.
Solution: See the past exam questions for a solution.

3.11 *Random vectors [Optional material]*

Earlier we have studied the bivariate probabilities, yet we haven't described bivariate probabilities referring to random variables and distribution functions. This chapter provides a calculus-based treatment of the same topic in an attempt to complete our knowledge of the topic.

In our earlier study, we've discussed probability models and computation of probability for events involving one variable mostly. These were the univariate models. Now, we are diving into models that involve more than one random variable, called multivariate models. As we are talking about more than one random variables, they are best represented as an n-dimensional random vector. This random vector is a function from a sample space S into $\mathbb{R}^n$, *i.e.*. n-dimensional Euclidean space.

3.11.1 *Joint Distributions*

Let (X, Y) be a random vector. The function $f(x, y) : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x, y) = \mathrm{P}(X = x, Y = y)$$

is called the joint probability distribution function or **joint pdf** of (X, Y) if X and Y are discrete. We denote the function as $f_{X,Y}(x, y)$

If (X, y) is a continuous random vector, if for every $A \subset \mathbb{R}^2$

$$\mathrm{P}((X, Y) \in A) = \iint_A f(x, y)\, dxdy$$

$f(x, y) : \mathbb{R}^2 \to \mathbb{R}$ is called a joint probability density function or **joint PDF** of (X, Y).

Note that the following are to hold for properly defining **joint PDF**'s: Discrete case:

$$f_{X,Y}(x, y) \geq 0, \forall\, (x, y) \in \mathbb{R}^2$$

$$\sum_{(x,y)\in\mathbb{R}^2} f_{X,Y}(x, y) = \mathrm{P}\Big((X, Y) \in \mathbb{R}^2\Big) = 1$$

Continuous case:

$$f_{X,Y}(x, y) \geq 0, \forall\, (x, y) \in \mathbb{R}^2$$

$$\int_{*\infty}^{\infty} \int_{\infty}^{\infty} f(x, y)\, dxdy = 1$$

3.11.2 *Marginal distributions*

Let (X, Y) be a discrete bivariate random vector with joint PDF $f_X(x, y)$. Then, the marginal PDFs of X and Y are:

$$f_X(x) = \mathrm{P}(X = x) = \sum_{y \in \mathbb{R}} f_{X,Y}(x, y)$$

and

$$f_Y(Y) = \mathrm{P}(Y = y) = \sum_{x \in \mathbb{R}} f_{X,Y}(x, y)$$

Let (X, Y) be a continous bivariate random vector with joint PDF $f_X(x, y)$. Then, the marginal PDFs of X and Y are:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y), \infty < x < \infty$$

and

$$f_Y(Y) = \sum_{x \in \mathbb{R}} f_{X,Y}(x, y), \infty < y < \infty$$

3.11.3 *Conditional distributions*

Let (X, Y) be a discrete bivariate random vector with $f_{X,Y}(xy)$, $f_X(x)$, and $f_Y(y)$. Then,

$$f_{X|Y}(x|y) = \mathrm{P}(X = x \mid Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, f_Y(y) \neq 0$$

and

$$f_{Y|X}(y|x) = \mathrm{P}(Y = y \mid X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)}, f_X(x) \neq 0$$

Let (X, Y) be a continous bivariate random vector with $f_{X,Y}(xy)$, $f_X(x)$, and $f_Y(y)$. Then,

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, f_Y(y) \neq 0$$

and

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}, f_X(x) \neq 0$$

3.11.4 *Independence of random variables*

Let (X, Y) be a bivariate random vector with $f_{X,Y}(xy)$, $f_X(x)$. Then, X and Y are called independent random variables if, for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$

$$f_{X,Y}(xy) = f_X(x) f_Y(y)$$

If X and Y are independent

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = f_X(x)\,,$$

and

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = f_Y(y)\,,$$

Notice that, except for the minor changes in notation, these definitons are the same as before.

3.11.5 *Covariance and correlation*

The covariance of X and Y is the number defined by:

$$Cov\,(X,Y) = \sigma_{XY} = \mathrm{E}((X-\mu_X)\,(Y-\mu_Y))$$

The correlation of X and Y is the number defined by:

$$\rho\,(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

This value is also called the correlation coefficient. Note that, $-1 \leq \rho_{XY} \leq 1$.

For any random variables X and Y,

$$\begin{aligned}
\sigma_{XY} &= \mathrm{E}((X-\mu_X)\,(Y-\mu_Y)) \\
&= \mathrm{E}(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y) \\
&= \mathrm{E}(XY) - \mu_Y\,\mathrm{E}(X) - \mu_Y\,\mathrm{E}(Y) + \mu_X\mu_Y \\
&= \mathrm{E}(XY) - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y \\
&= \mathrm{E}(XY) + \mu_X\mu_Y
\end{aligned}$$

So,

$$\sigma_{XY} = \mathrm{E}(XY) - \mu_X\mu_Y$$

If X and Y are independent random variables, then

$$Cov\,(X,Y) = 0$$

and

$$\rho_{XY} = 0$$

Let X and Y be any two random variables, also let a and b are any two constants, then

$$\mathrm{Var}\,(aX+bY) = a^2\,\mathrm{Var}\,(X) + 2abCov\,(X,Y) + b^2\,\mathrm{Var}\,(Y)$$

or

$$\sigma^2_{aX+bY} = a^2\sigma^2_X + 2ab\sigma_{XY} + b^2\sigma^2_Y$$

If X and Y are independent random variables with moment generating functions $M_X(t)$ and $M_Y(t)$, then the moment generating function of $X+Y$ is given by:

$$M_{X+Y}(t) = M_X(t)\, M_Y(t)$$

3.3 EXERCISES

1. Fill the empty cells in the tables, which are tables of joint CDF and joint PDF of the random variables (X_1, X_2).

$f(x_1, x_2)$	0.0	0.5	1.0
1		0.1	
2			0.1
3			

$F(x_1, x_2)$	0.0	0.5	1.0
1	0.2		
2	0.5	0.65	0.75
3	0.5	0.8	

2. Let X and Y be two discrete random variables with joint density function $f(x,y) \in \mathbb{R}^2$ such that

$$f(x,y) = \begin{cases} cxy, & x \in \{1,2,3\} \text{ and } y \in \{1,2\}, \\ 0, & \text{otherwise.} \end{cases}$$

i. Find the value of c

ii. Find $\mathrm{P}(Y < X)$

iii. Find $\mathrm{P}(Y = X)$

iv. Find the PDF of X and PDF of Y

v. Find the conditional distribution of Y given $x = 1$

vi. Are the random variables X and Y independent?

vii. Find the expected value of X

viii. Find the variance of X

3. First we pick a number, at random, from the interval $(0,1)$, then we pick a number, at random, from the interval $(0, x_1)$. Let X_1 be the random variable that gives the value of the first number and X_2 the random variable of the second number. The distribution of X_1 is **uniform** over $(0,1)$ and the distribution of X_2 given that $x_1 = x_1$ is **uniform** over $(0, x_1)$

i. Find the joint PDF of (X_1, X_2).

ii. Find the PDF of X_2 (the second marginal distribution of (X_1, X_2)).

iii. Find the conditional distribution of X_1 given $X_2 = x_2$.

In a nutshell

In ECON 221 and ECON 222 we introduce the following statistical distributions:

- **ECON 221 Probability and Statistical Distributions:**
  - In full detail: Bernoulli, Binomial, Poisson, Discrete Uniform, Continuous Uniform, Exponential, Normal, Standard Normal
  - In exercises: Hypergeometric, Geometric, Negative Binomial, Triangular, Half Normal
- **ECON 222 Statistical Estimation and Inference:**
  - In full detail: t, χ^2, F

This is only a selection of essential distributions out of tens of them available. Learning how to acquire the knowledge of other distributions is a valuable asset.

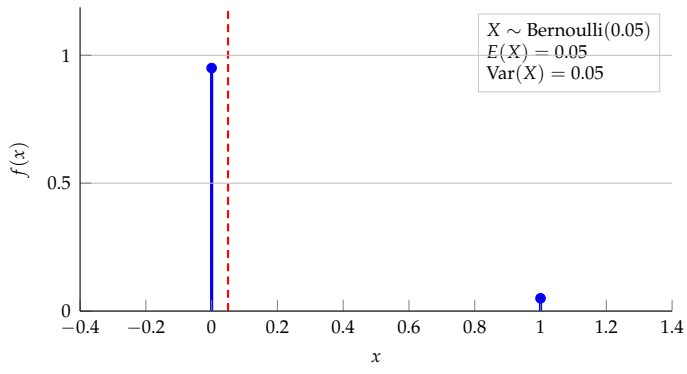
4 More on Distributions

This chapter carries the spreadsheet-style distribution demonstrations into the PDF. Each frame shows a probability graph together with a dynamic table of selected values of x, $f(x)$, and $F(x)$. The dashed red line marks the mean whenever it lies in the displayed range.

The animations play automatically in Adobe Acrobat Reader. Some PDF viewers may show only the first frame.

4.1 Discrete distributions

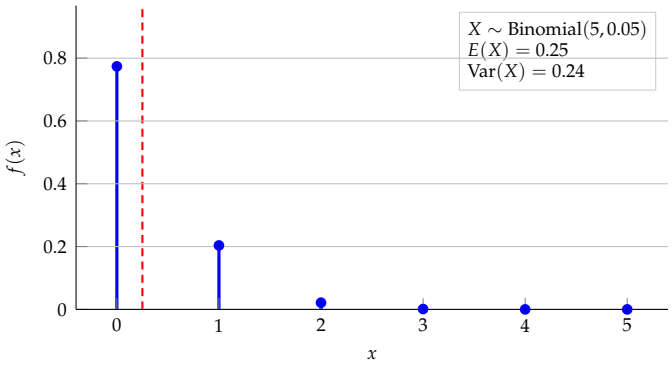
4.1.1 Bernoulli distribution



$X \sim \text{Bernoulli}(0.05)$

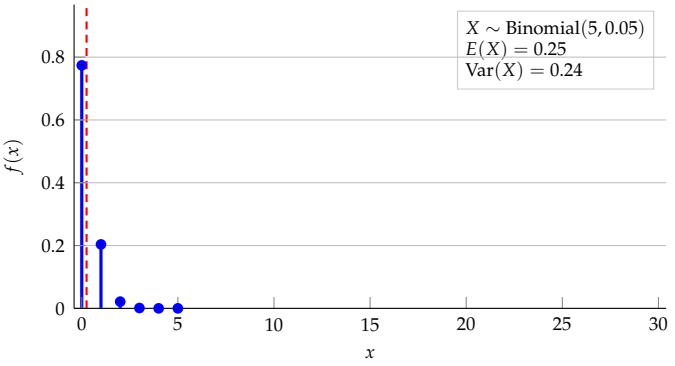
x	$f(x)$	$F(x)$
0	0.9500	0.9500
1	0.0500	1.0000

4.1.2 *Binomial distribution*



$X \sim \text{Binomial}(5, 0.05)$

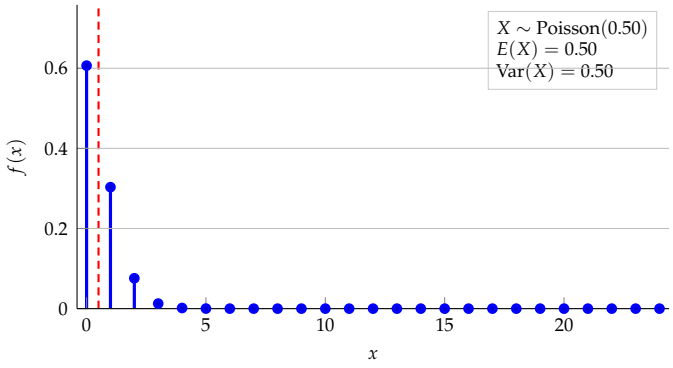
x	$f(x)$	$F(x)$
0	0.7738	0.7738
1	0.2036	0.9774
2	0.0214	0.9988
3	0.0011	1.0000
4	0.0000	1.0000
5	0.0000	1.0000



$X \sim \text{Binomial}(5, 0.05)$

x	$f(x)$	$F(x)$
0	0.7738	0.7738
1	0.2036	0.9774
2	0.0214	0.9988
3	0.0011	1.0000
4	0.0000	1.0000
5	0.0000	1.0000

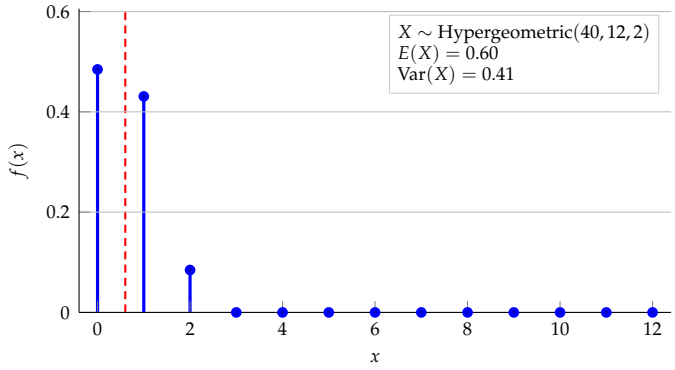
4.1.3 *Poisson distribution*



$X \sim \text{Poisson}(0.50)$

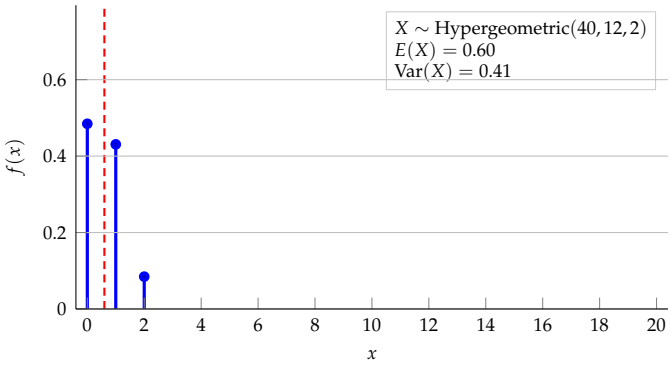
x	$f(x)$	$F(x)$
0	0.6065	0.6065
1	0.3033	0.9098
2	0.0758	0.9856
3	0.0126	0.9982
4	0.0016	0.9998
5	0.0002	1.0000
6	0.0000	1.0000
7	0.0000	1.0000
8	0.0000	1.0000
9	0.0000	1.0000
10	0.0000	1.0000
11	0.0000	1.0000
12	0.0000	1.0000

4.1.4 *Hypergeometric distribution*



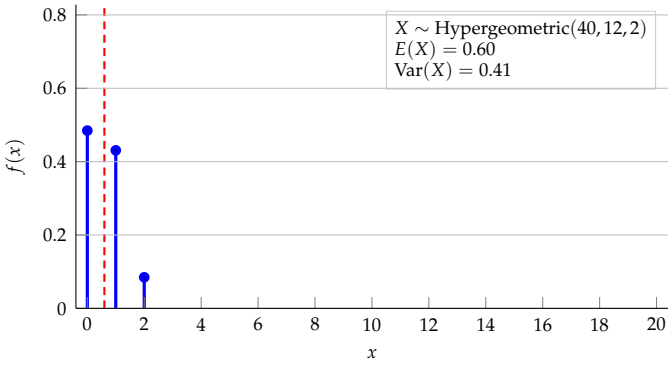
$X \sim \text{Hypergeometric}(40, 12, 2)$

x	$f(x)$	$F(x)$
0	0.4846	0.4846
1	0.4308	0.9154
2	0.0846	1.0000
3	0.0000	1.0000
4	0.0000	1.0000
5	0.0000	1.0000
6	0.0000	1.0000
7	0.0000	1.0000
8	0.0000	1.0000
9	0.0000	1.0000
10	0.0000	1.0000



$X \sim \text{Hypergeometric}(40, 12, 2)$

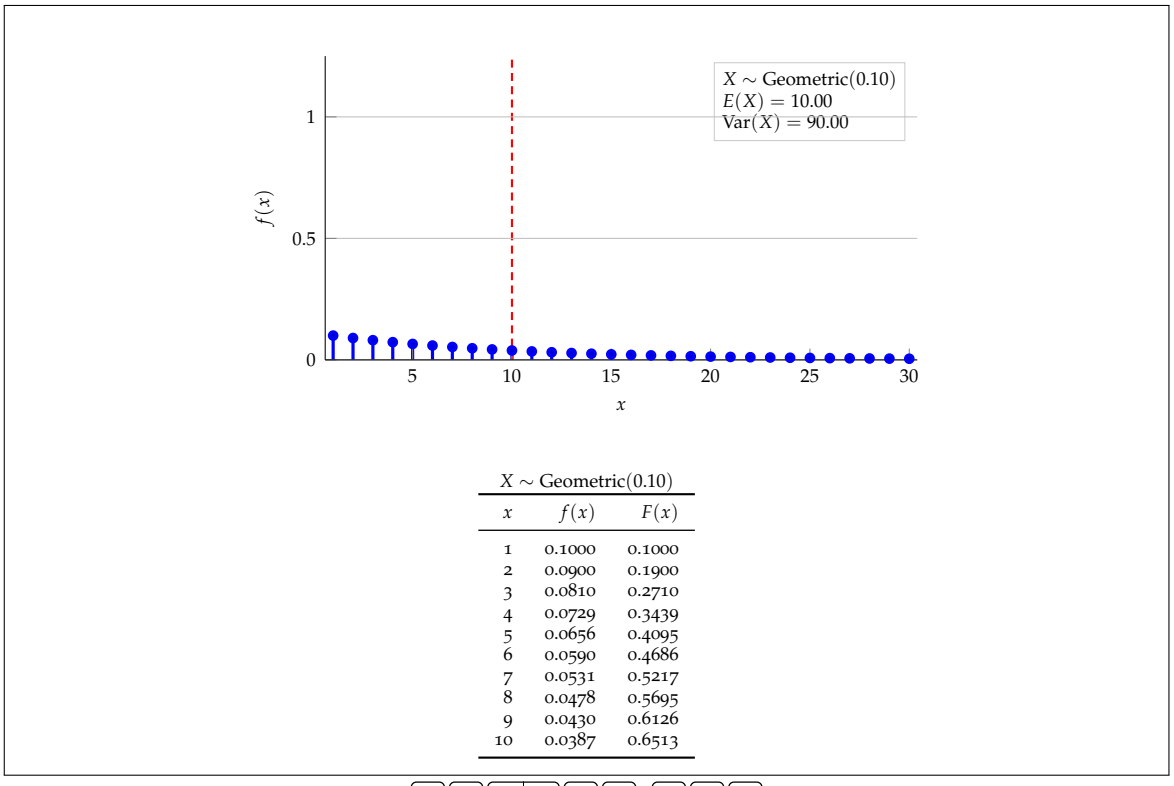
x	$f(x)$	$F(x)$
0	0.4846	0.4846
1	0.4308	0.9154
2	0.0846	1.0000



$X \sim \text{Hypergeometric}(40, 12, 2)$

x	$f(x)$	$F(x)$
0	0.4846	0.4846
1	0.4308	0.9154
2	0.0846	1.0000

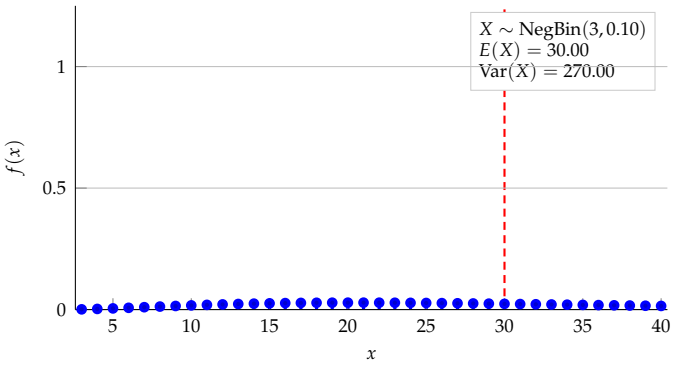
4.1.5 *Geometric distribution*



$X \sim \text{Geometric}(0.10)$

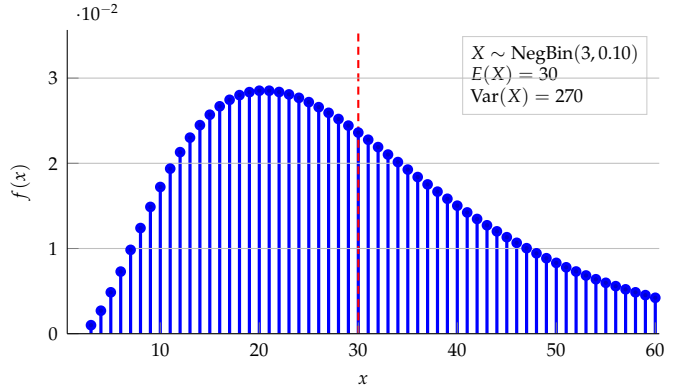
x	$f(x)$	$F(x)$
1	0.1000	0.1000
2	0.0900	0.1900
3	0.0810	0.2710
4	0.0729	0.3439
5	0.0656	0.4095
6	0.0590	0.4686
7	0.0531	0.5217
8	0.0478	0.5695
9	0.0430	0.6126
10	0.0387	0.6513

4.1.6 *Negative binomial distribution*



$X \sim \text{NegBin}(3, 0.10)$

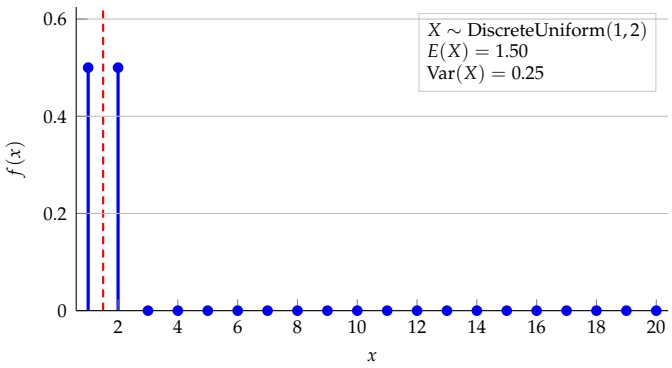
x	$f(x)$	$F(x)$
3	0.0010	0.0010
4	0.0027	0.0037
5	0.0049	0.0086
6	0.0073	0.0159
7	0.0098	0.0257
8	0.0124	0.0381
9	0.0149	0.0530
10	0.0172	0.0702
11	0.0194	0.0896
12	0.0213	0.1109



$X \sim \text{NegBin}(3, 0.10)$

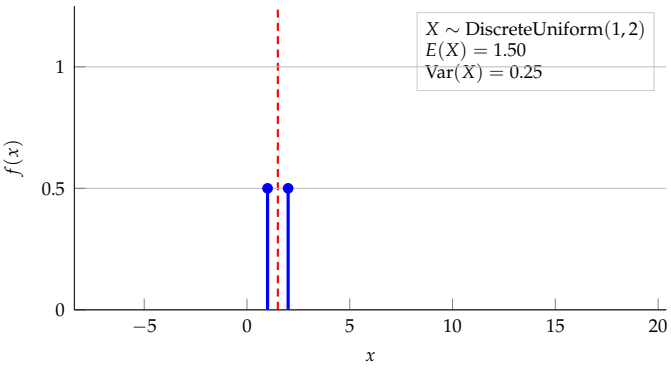
x	$f(x)$	$F(x)$
3	0.0010	0.0010
4	0.0027	0.0037
5	0.0049	0.0086
6	0.0073	0.0159
7	0.0098	0.0257
8	0.0124	0.0381
9	0.0149	0.0530
10	0.0172	0.0702

4.1.7 *Discrete uniform distribution*



$X \sim \text{DiscreteUniform}(1,2)$

x	$f(x)$	$F(x)$
1	0.5000	0.5000
2	0.5000	1.0000
3	0.0000	1.0000
4	0.0000	1.0000
5	0.0000	1.0000
6	0.0000	1.0000
7	0.0000	1.0000
8	0.0000	1.0000
9	0.0000	1.0000
10	0.0000	1.0000

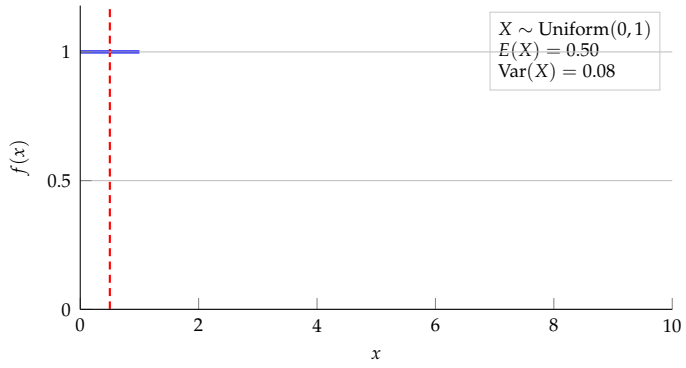


$X \sim \text{DiscreteUniform}(1,2)$

x	$f(x)$	$F(x)$
1	0.5000	0.5000
2	0.5000	1.0000

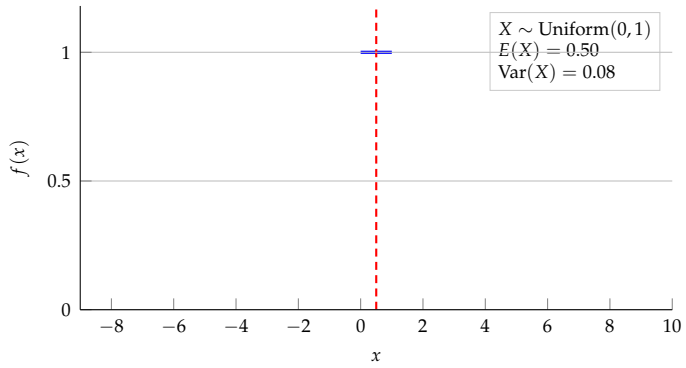
4.2 Continuous distributions

4.2.1 Uniform distribution



$X \sim \text{Uniform}(0,1)$

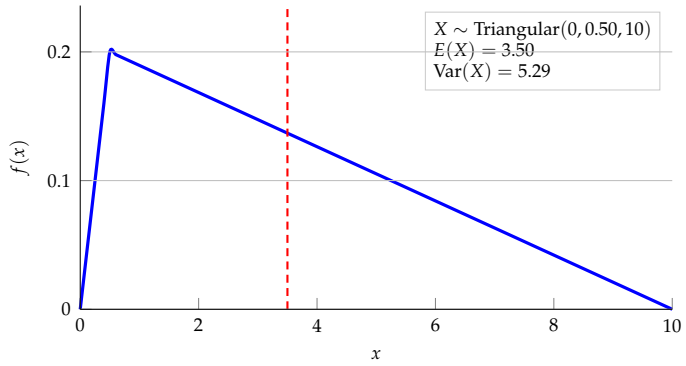
x	$f(x)$	$F(x)$
0	1.0000	0.0000
0.2500	1.0000	0.2500
0.5000	1.0000	0.5000
0.7500	1.0000	0.7500
1	1.0000	1.0000



$X \sim \text{Uniform}(0,1)$

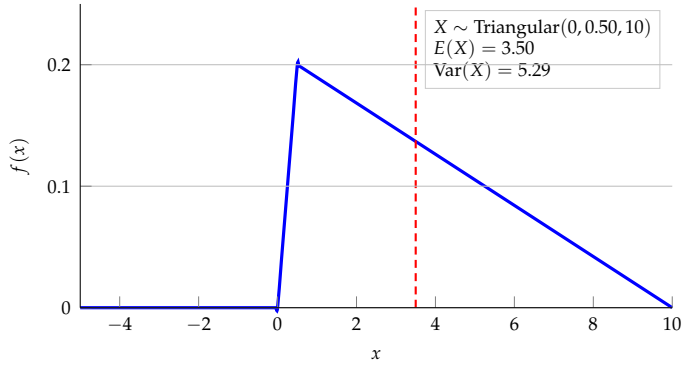
x	$f(x)$	$F(x)$
0	1.0000	0.0000
0.25	1.0000	0.2500
0.50	1.0000	0.5000
0.75	1.0000	0.7500
1	1.0000	1.0000

4.2.2 *Triangular distribution*



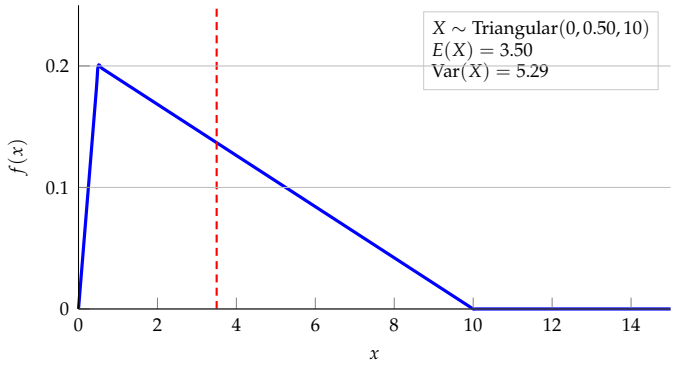
$X \sim \text{Triangular}(0, 0.50, 10)$

x	$f(x)$	$F(x)$
0	0.0000	0.0000
2	0.1684	0.3263
4	0.1263	0.6211
6	0.0842	0.8316
8	0.0421	0.9579
10	0.0000	1.0000



$X \sim \text{Triangular}(0, 0.50, 10)$

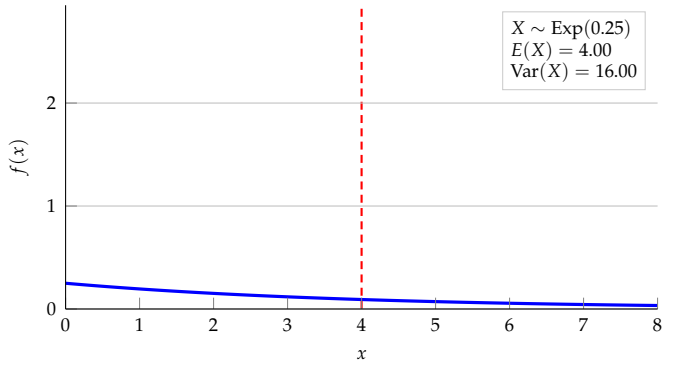
x	$f(x)$	$F(x)$
0	0.0000	0.0000
0.25	0.1000	0.0125
0.50	0.2000	0.0500
5.25	0.1000	0.7625
10	0.0000	1.0000



$X \sim \text{Triangular}(0, 0.50, 10)$

x	$f(x)$	$F(x)$
0	0.0000	0.0000
0.25	0.1000	0.0125
0.50	0.2000	0.0500
5.25	0.1000	0.7625
10	0.0000	1.0000

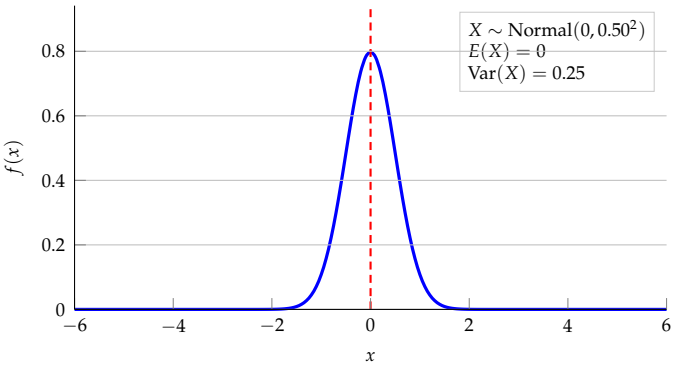
4.2.3 *Exponential distribution*



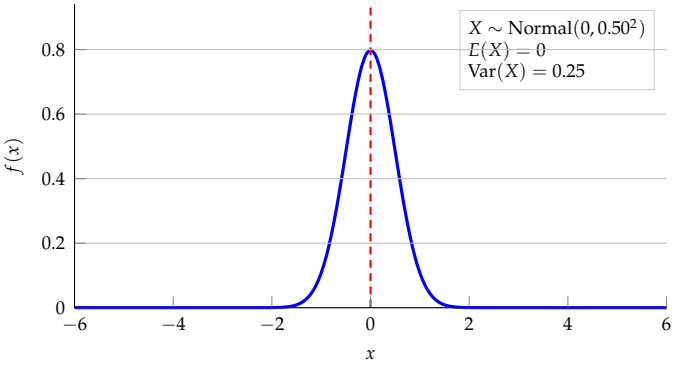
$X \sim \text{Exp}(0.25)$

x	$f(x)$	$F(x)$
0	0.2500	0.0000
0.5000	0.2206	0.1175
1	0.1947	0.2212
2	0.1516	0.3935
3	0.1181	0.5276
4	0.0920	0.6321
6	0.0558	0.7769

4.2.4 *Normal distribution*

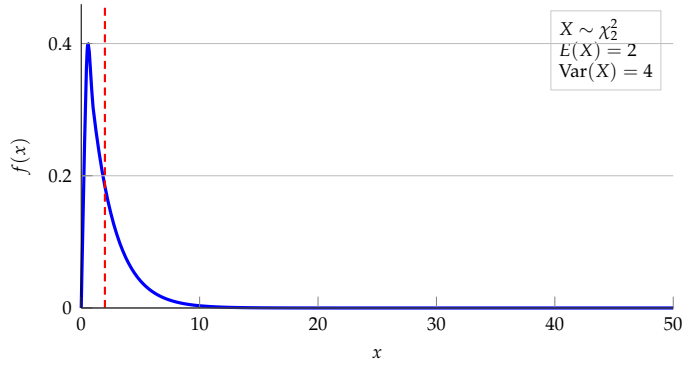


$X \sim \text{Normal}(0, 0.50^2)$		
x	$f(x)$	$F(x)$
-3	0.0000	0.0000
-2	0.0003	0.0000
-1	0.1080	0.0228
0	0.7979	0.5000
1	0.1080	0.9772
2	0.0003	1.0000
3	0.0000	1.0000



$X \sim \text{Normal}(0, 0.50^2)$		
x	$f(x)$	$F(x)$
-1.50	0.0089	0.0013
-1	0.1080	0.0228
-0.50	0.4839	0.1587
0	0.7979	0.5000
0.50	0.4839	0.8413
1	0.1080	0.9772
1.50	0.0089	0.9987

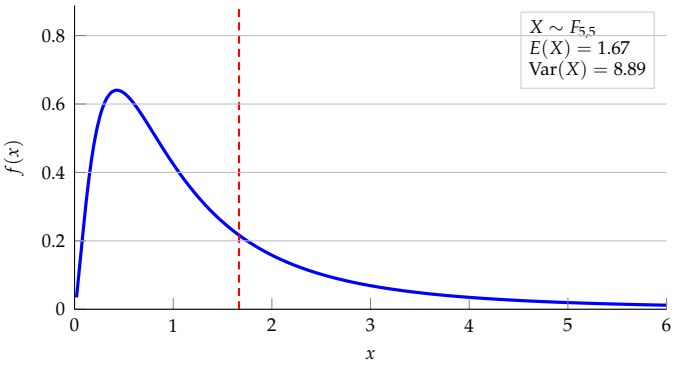
4.2.5 *Chi-square distribution*



$X \sim \chi^2_2$

x	$f(x)$	$F(x)$
0	0.0000	0.0000
1	0.3033	0.3935
2	0.1839	0.6321
5	0.0410	0.9179
10	0.0034	0.9933
20	0.0000	1.0000
30	0.0000	1.0000
50	0.0000	1.0000

4.2.6 *F distribution*



$X \sim F_{5,5}$

x	$f(x)$	$F(x)$
0.2500	0.5563	0.0772
0.5000	0.6323	0.2325
1	0.4244	0.5000
2	0.1581	0.7675
3	0.0689	0.8734
5	0.0195	0.9490

5 Sampling distributions

This chapter bridges our knowledge of the probability theory to statistical inference. Sampling distributions is the key to our understanding of how a small portion of a whole can represent the whole.

5.1 Chebyshev's theorem

For any random variable X with a finite expected value μ and finite variance σ^2,

$$\forall k > 1, \mathrm{P}(|X - \mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$$

or, alternatively, by setting $k = \frac{\epsilon}{\sigma}$,

$$\forall \epsilon > 0, \mathrm{P}(|X - \mu| \leq \epsilon) \geq 1 - \frac{1}{\frac{\epsilon^2}{\sigma^2}}$$

When σ^2 is known, one of the k and ϵ can be arbitrarily picked.

5.2 Law of large numbers theorem

Let $\{X_i\}_{i=1}^{N}$ be a sequence of identically and independently distributed random variables with a finite expected value μ. For each $n \leq N$, define the random variable $\bar{X}_n$ as:

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Then,

$$\forall \epsilon > 0, \lim_{n \to \infty} \mathrm{P}(|\bar{X}_n - \mu| \leq \epsilon) = 1$$

For sufficiently large n, the mean of independently and identically distrbuted (*i.i.d.*) n random variables will **almost surely** be **arbitrarily close** to the expected value of the individual random variables.

Noting that, $\mathrm{E}(\bar{X}_n) = \mu$ and $\mathrm{Var}\,(\bar{X}_n) = \frac{\sigma^2}{n}$ Then, $\forall n \leq N, \forall \epsilon > 0$,

$$\mathrm{P}(|\bar{X}_n - \mu| \leq \epsilon) \geq 1 - \frac{1}{\frac{\epsilon^2}{\sigma^2/n}}$$

as implied by the Chebyshev's theorem. As n approaches infinity, this expression reduces to:

$$\forall \epsilon > 0, \lim_{n \to \infty} \mathrm{P}(|\bar{X}_n - \mu| \leq \epsilon) = 1$$

which is known as the Law of large numbers and is true even when the variance of X_i is not finite.

5.3 Central limit theorem

Let $\{X_i\}_{i=1}^N$ be a sequence of identically and independently distributed random variables with a finite expected value μ, and a finite and positive variance σ^2. For each n, define the random variable $\bar{X}_n$ as:

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Let Z be a standard normal random variable. For any $z \in \mathbb{R}$, we have

$$\lim_{n \to \infty} \mathrm{P}\left(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z \right) = \mathrm{P}(Z \leq z)$$

Informally, CLT states that for **sufficiently large** n, the random variable

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

is **approximately** standard normal distributed regardless of the distribution of X_i and **exactly** standard normal distributed if X_i are normally distributed.

5.4 Distribution of sample means

Consider $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$ which is a random sample of n observations coming from a population with mean μ and variance σ^2. X_n being $\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$

$$\mathrm{E}(\bar{X}_n) = \mu$$

$$\mathrm{Var}\,(\bar{X}_n) = \frac{\sigma^2}{n}$$

If the population is distributed normally, then the distribution of the sample means is also normal. So,

$$Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

has a standard normal distribution.

5.4.1 Essence of sampling distributions

5.4.2 Distribution of sample proportions

Consider $\{x_i\}_{i=1}^{n} \subset \{x_i\}_{i=1}^{N}$ which is a random sample of n observations coming from a Bernoulli(p) population. X_n being $\bar{X}_n = \frac{X_1+X_2+\cdots+X_n}{n} = \hat{p}$

$$\begin{aligned} \mathrm{E}(\hat{p}) &= p \\ \mathrm{Var}\,(\hat{p}) &= \tfrac{p(1-p)}{n} \end{aligned}$$

If n is large,

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximately distributed as a standard normal.

5.4.3 Distribution of sample variances

Let s^2 denote the sample variance for a random sample of n observations from a population with a variance of σ^2.

$$\begin{aligned} \mathrm{E}\big(s^2\big) &= \sigma^2 \\ \mathrm{Var}\,\big(s^2\big) &= \tfrac{2\sigma^4}{n-1} \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum((x_i - \mu) - (\bar{x} - \mu))^2 \\ &= \sum(x_i - \mu)^2 - 2(\bar{x} - \mu)\sum(x_i - \mu) + \sum(\bar{x} - \mu)^2 \\ &= \sum(x_i - \mu)^2 - 2n(\bar{x} - \mu)^2 + n(\bar{x} - \mu)^2 \\ &= \sum(x_i - \mu)^2 - n(\bar{x} - \mu)^2 \end{aligned}$$

$$\begin{aligned} \mathrm{E}\Big(\sum(x_i - \bar{x})^2\Big) &= \mathrm{E}\Big(\sum(x_i - \mu)^2\Big) - n\,\mathrm{E}\Big((\bar{x} - \mu)^2\Big) \\ &= \sum \underbrace{\mathrm{E}\Big((x_i - \mu)^2\Big)}_{\sigma^2} - n\underbrace{\mathrm{E}\Big((\bar{x} - \mu)^2\Big)}_{\sigma^2/n} \\ &= n\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2 \end{aligned}$$

So,

$$\begin{aligned} \mathrm{E}\Big(s^2\Big) &= \mathrm{E}\Big(\frac{1}{n-1}\sum(x_i - \bar{x})^2\Big) \\ &= \frac{1}{n-1}\,\mathrm{E}\Big(\sum(x_i - \bar{x})^2\Big) \\ &= \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2 \end{aligned}$$

Given a random sample of n observations from a normally distributed population whose variance is σ^2, the sample variance s^2 has a χ^2 distribution with $(n-1)$ degrees of freedom?

$$\chi^2_{n-1} = \frac{(n-1)\,s^2}{\sigma^2}$$

is distributed as the **Chi-squared** (χ^2) **distribution** with $(n-1)$ degrees of freedom.

In a nutshell

When we talk about sampling distributions, notice that we are talking about the statistical properties of the n observations, rather than those of the N members of the population. By definition, properties of the sample relate to the properties of the population & the sample size n.

5.1 EXERCISES

1. We have two data sets consisting of identical values: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. We will use x_i denote the i^{th} value in one data set and y_j to denote the j_{th} value in the other data set. We construct a new data set by taking a number from each data set and finding their average, *i.e.*. the new data set consists of values of the form:

$$\frac{(x_i + y_j)}{2}$$

The constructed data set will consist of one 0, two 0.5's, three 1's, etc.

i. Find all values in the new data set and their corresponding frequencies

ii. Construct a graph that summarize your finding

iii. Find the mean and variance of the initial data set and the new data set

Solution: Solve yourself to explore the Central Limit Theorem.

2. We make 100 independent observations from a population with mean 40 and standard deviation 20. Approximately, what is the probability that the mean of these observations will be greater than 37?

Solution: Population $X \sim \cdot(40, 20^2)$. We take $n = 100$ observations and calculate $\bar{X}_{100}$.

$$\begin{aligned} \mathrm{E}(\bar{X}_{100}) &= \mu = 40 \\ \mathrm{Var}\,(\bar{X}_{100}) &= \frac{\sigma^2}{n} = \frac{400}{100} = 4 \\ \mathrm{P}(\bar{X}_{100} > 37) &= \mathrm{P}\left(\frac{\bar{X}_{100} - 40}{\sqrt{4}} > \frac{37 - 40}{\sqrt{4}} \right) \\ &= \mathrm{P}(z > -1.5) \\ &= 0.93319 \end{aligned}$$

3. The following table gives the relative frequency distribution of a population:

Value	Rel. Freq.
2	0.1
4	0.3
6	0.2
8	0.3
10	0.1

i. A number is selected from this population at random, what is the probability that the number selected is greater than or equal to 8?

ii. If we select two numbers at random (with replacement), what is the probability that the mean of these two numbers is less than or equal to 5?

iii. If 25 numbers are selected from the population at random (with replacement), what is the probability (approximately) that the mean these 25 numbers is less than 6.5?

Solution:

1. The only trick in this exercise is to begin with calculating $\mathrm{E}(X)$ and $\mathrm{Var}\,(X)$. Calculate and see $\mathrm{E}(X) = 6$ and $\mathrm{Var}\,(X) = 5.6$. There is no sampling as $n = 1$. Simply calculate $\mathrm{P}(X \geq 8)$.
2. Calculate $\mathrm{P}(\bar{X}_2 \leq 5)$.
3. Calculate $\mathrm{P}(\bar{X}_{25} < 6.5)$. Remember that $\mathrm{E}(\bar{X}_n) = \mu$ and $\mathrm{Var}\,(\bar{X}_n) = \sigma^2/n$.

4. We choose 36 numbers, with replacement, at random (*i.e.*., we take a random sample of size 36) from the interval $(0, 4)$. Let X be the random variable that assigns to each sample (outcome) the mean of the sample.

i. Find the expected value and variance of X

ii. Find (an approximate value for) the probability that the sample mean, *i.e.*. X, will be less than or equal to 2.3

Solution:

1. Since nothing further is instructed, assume that the population is Uniform$(0, 4)$. Then,

$$\mu = \frac{0+4}{2} = 2$$
$$\sigma^2 = \frac{(4-0)^2}{12} = \frac{16}{12}$$

X, here, is the mean of our 36 observations, i.e., $\bar{X}_{36}$ in our usual notation.

$$\mathrm{E}(X) = \mu = 2$$
$$\mathrm{Var}\,(X) = \frac{\sigma^2}{n} = \frac{16/12}{36} = \frac{1}{27}$$

2. Using the parameters in part (i), calculate $\mathrm{P}(X \leq 2.3)$, i.e., $\mathrm{P}(z \leq 1.56)$. The answer is 0.94062.

5. We choose 9 numbers from a normally distributed population of numbers. The mean of the population is unknown but the variance is know to be equal to 16. If μ denotes the mean of the population, then what is the probability that the mean of the 9 numbers that we choose will be in the interval $[\mu - 2, \mu + 2]$?
Solution: You do not need the value of μ in this exercise. The key to solution is that $\mathrm{Var}\,(\bar{X}_9) = 16/9$. So, performing the intermediate steps, the problem reduces to finding $\mathrm{P}(-1.5 \leq z \leq 1.5)$ and the answer is 0.86638.

6. In a certain university the CGPA's of students only takes the values 0, 1, 2, 3, 4. The distribution of CGPA's of students of this university is given below:

CGPA	Freq
0	5,000
1	10,000
2	20,000
3	5,000
4	10,000
Total	50,000

i. Let $\bar{X}_1$ denote the CGPA of a student who was chosen at random from the population of all students of this university. Tabulate the PDF of this random variable.

ii. Let $\bar{X}_2$ denote the average (mean) of the CGPA's of two randomly selected students from the population of all students of this university. What is the probability that the average CGPA of the two students is less than or equal to 1?

iii. Now we choose 36 students at random. What is the probability, approximately, that the average CGPA of these students ($\bar{X}_{36}$) is less than or equal to 2.3?

Solution: This exercise is left as self-study.

7. Consider a large population of which only 20% know basic concepts of statistics. We take a random sample of size 81 from this population and count the number of individuals, in the sample, who knows basic concepts of statistics. What is the probability that the sample will have between 15 and 18 (inclusive) individuals who know basic concepts of statistics?
Solution: This exercise is left as self-study.

8. A four sided fair die is rolled several times and we calculate the average (mean) of the values observed.

i. Let X_1 denote the random variable that gives the value observed when the die is rolled once. Find the expected value and variance of X_1.

ii. Let $\bar{X}_n$ denote the mean of the values observed when the die is rolled n times. What is the minimum number of times that the die should be rolled so that the mean (the value of $\bar{X}_n$) takes a value in the interval $[2.1, 2.9]$ with at least a probability of 0.9? Use *Chebyshev'sTheorem* to answer this problem.

iii. Using the *CentralLimitTheorem* and the value for n you found above, find an approximate value for the probability of $\bar{X}_n$ taking a value in the interval $[2.1, 2.9]$.

iv. Do the answers you found in item (ii) and item (iii) contradict each other? If there is a difference in what the two answers suggest, explain the reason for this.

Solution: i. When the die is rolled once, by definition we produce an outcome directly of the population (that is we don't do any sampling at all). So,

$$X_1 \sim f(x)\,, f(x) = 1/4, x = 1, 2, 3, 4 \quad (\text{Discrete}RV)$$

$$\begin{aligned} \mathrm{E}(X_1) &= 1 \cdot 1/4 + 2 \cdot 1/4 + 3 \cdot 1/4 + 4 \cdot 1/4 \\ &= 2.5 \end{aligned}$$

$$\begin{aligned}\mathrm{E}\left(X_1^2\right) &= 1 \cdot 1/4 + 4 \cdot 1/4 + 9 \cdot 1/4 + 16 \cdot 1/4 \\ &= 7.5\end{aligned}$$

$$\begin{aligned}\mathrm{Var}\left(X_1\right) &= 7.5 - 2.5^2 \\ &= 1.25\end{aligned}$$

So, $X_1 \sim \cdot(2.5, 1.25)$
ii. $\bar{X}_n$ is the RV denoting the sample mean, when We roll the die n times.

$$\begin{aligned}\mathrm{E}(\bar{X}_n) &= \mathrm{E}(X_1) = 2.5 \\ \mathrm{Var}\left(\bar{X}_n\right) &= \tfrac{\mathrm{Var}(X_1)}{n} = \tfrac{1.25}{n}\end{aligned}$$

Midpoint of the interval $[2.1, 2.9]$ is 2.5

$$\begin{aligned}\mathrm{E}(X_1) = \mathrm{E}(\bar{X}_n) &= 2.5 \\ 2.5 - 2.1 = 2.9 - 2.5 = 0.4 &\rightarrow \epsilon \\ \epsilon &= 0.4\end{aligned}$$

$$\mathrm{P}(|\bar{X}_n - 2.5| \leq 0.4) \geq 1 - \frac{1}{\frac{0.4^2}{\frac{1.25}{n}}} \leftarrow 0.9$$

$$\begin{aligned}1 - \frac{1}{\frac{0.4^2}{\frac{1.25}{n}}} &= 0.9 \\ \frac{1.25}{n} &= 0.1 \times 0.16\end{aligned}$$

$$\begin{aligned}n &= \frac{1.25}{0.016} \\ &= 78.125 \\ n &= 79 \leftarrow \text{Round } 78.125 \text{ up}\end{aligned}$$

Notice that in this part, we use the Chebyshev's theorem only. We get:

$$\mathrm{E}(\bar{X}_{79}) = 2.5$$

and

$$\text{Var}\left(\bar{X}_{79}\right) = \frac{1.25}{79} = 0.015823$$

iii.

$$\begin{aligned} \text{P}(2.1 \leq \bar{X}_{79} \leq 2.9) &= \text{P}\left(\frac{2.1-2.5}{\sqrt{0.015823}} \leq z \leq \frac{2.9-2.5}{\sqrt{0.015823}}\right) \\ &= \text{P}\left(\frac{-0.4}{0.1258} \leq z \leq \frac{0.4}{0.1258}\right) \\ &= 0.99852 \end{aligned}$$

Notice that in this part, we use the CLT only.
iv. No, they don't contradict each other: While Chebyshev's theorem sets a lowerbound of 0.90 here (in ii), (iii) gives the actual probability as 0.99852 which is larger than 0.90 (as it should be).
Though I am not a fan of such hints, here a useful hint may be underlined: When an interval is given in a question like this, always begin with observing (calculating) its midpoint. In this question, the midpoint was $(2.1 + 2.9)/2 = 2.5$, which is nothing but $\text{E}(X_1)$ and $\text{E}(\bar{X}_n)$. Once you have noticed this, it will be trivial to find ϵ to figure out the rest of the steps.
One additional point is:

$$\bar{X}_1 = \frac{X_1}{1}$$

So, sampling with $n = 1$ simply refers to population's distribution.

In a nutshell

If $Z_i, i = 1, 2, \ldots, m$ are all $N(0,1)$ random variables, then:

$$V = Z_1^2 + Z_2^2 + \ldots + Z_m^2$$

has a χ^2 distribution with m degrees of freedom.

$$V \sim \chi^2_{(m)}$$
$$\mathrm{E}(V) = m$$
$$\mathrm{Var}(V) = 2m$$

Recall that your Z_i's measure nothing but deviation from mean (here 0) in terms of standard deviations (here 1). So, when we sum the squares of Z_i's, we are calculating sum of squares of the deviations of a random variable from its mean. This should be something related to variance. Indeed, χ^2 turns out to be the sampling distribution of variance. Putting in another way, χ^2 distribution is the assessment benchmark for variability.
Consider a single $Z \sim N(0,1)$ random variable and,

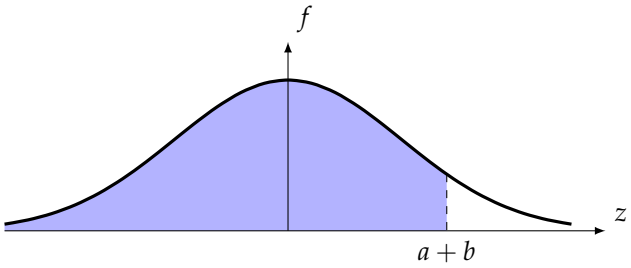
$$V = Z^2$$

and observe this is nothing but a χ^2 distribution with 1 degree of freedom.

$$V \sim \chi^2_{(1)}$$
$$\mathrm{E}(V) = 1$$
$$\mathrm{Var}(V) = 2$$

Ztandard Normal Diztribution

A cell which is at the intersection of the row labeled with a and column labeled with b gives the probability $P(Z \leq a + b)$.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5754
0.20	0.5793	0.5832	0.5871	0.5909	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.60	0.7258	0.7291	0.7324	0.7357	0.7389	0.7421	0.7454	0.7486	0.7518	0.7549
0.70	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7793	0.7823	0.7852
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	0.8849	0.8869	0.8888	0.8906	0.8925	0.8943	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.20	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.30	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.40	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936

2.50	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.60	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.70	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.80	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.90	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.00	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.10	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.20	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.30	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.40	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

In a nutshell

Probability theory → Statistics conceptual mapping:

ECON 221 → ECON 222

Probability theory → Statistics

$X \rightarrow$ Data

PDF, $f \rightarrow$ Relative frequency distribution (Histogram)

CDF, $F \rightarrow$ Relative cumulative frequency distribution (O-give)

$\mathrm{E}(X) \rightarrow \mu \leftarrow \bar{x}$

$\mathrm{Var}\,(X) \rightarrow \sigma^2 \leftarrow s^2$

6 Point estimators

6.1 Point estimation

Based on our earlier studies and practice, we now move toward the study of estimating population parameters. Let's use θ to denote, generally speaking, our parameter of interest. Very first of all, we don't know the value of θ. That's why we are estimating it. All we know/have is a sequence of values which are coming out of the population. Formally, $\{x_i\}_{i=1}^{N}$ is the population and $\{x_i\}_{i=1}^{n}$ is our sample. Naturally the second one is a subset of the first. As we'll discuss in detail later, our sample is better to be a random and large enough sample. So, our problem is to come up with a formula to find θ; using a function notation

$$\hat{\theta} = \hat{\theta}\left(\{x_i\}_{i=1}^{n}\right)$$

In the expression $\hat{\theta}\left(\{x_i\}_{i=1}^{n}\right)$, the $\hat{\theta}(.)$ is our **estimator**, *i.e.*. our formula to find a value for the unknown θ. The $\hat{\theta}$ on the left hand side is called an **estimate** of θ. **Estimation** is the name of the thing we are doing in here.

So, in the task of **estimation**, we use an **estimator** to obtain an **estimate**.

As we are seeking for a single value as an estimate here, what we do fall into the category of **point estimation**.

Notice that θ is unknown. $\{x_i\}_{i=1}^{n}$ is our data set, at hand and is known, possibly we collect or gather. $\hat{\theta}$ as a function of $\{x_i\}_{i=1}^{n}$ is an estimator, possibly we derive. $\hat{\theta}$ as a numerical result is an estimate, possibly we calculate. The task to find a value $\hat{\theta}$ for θ is estimation, possibly we perform.

Note that, while studying the properties of $\theta(.)$, we prefer writing it as

$$\hat{\theta}\left(\{X_i\}_{i=1}^{n}\right)$$

rather than

$$\hat{\theta}\left(\{x_i\}_{i=1}^{n}\right)$$

as this practice allows us to refer to statistical properties of X_i.

For our point estimators, we seek to attain three important properties which are **Unbiasedness**, **Consistency** and **Efficiency**.

6.1.1 *Unbiasedness*

$$\mathrm{E}\left(\hat{\theta}\right) = \theta$$

Unbiasedness is an individual property. If $\hat{\theta}$ is an unbiased estimator, repeating the task of estimation so several times, using a different random sample of n observations each time, we come up with θ, on average. That is $\mathrm{E}\left(\hat{\theta}\left(\{X_i\}_{i=1}^{n}\right)\right) = \theta$

6.1.2 *Efficiency*

Efficiency is a comparative property. $\hat{\theta}_1$ and $\hat{\theta}_2$ being two estimators of θ, if

$$\mathrm{Var}\left(\hat{\theta}_1\right) < \mathrm{Var}\left(\hat{\theta}_2\right)$$

then $\hat{\theta}_1$ is a **more efficient** estimator than $\hat{\theta}_2$. Notice that $\mathrm{Var}\left(\hat{\theta}_1\left(\{X_i^n\}\right)\right) < \mathrm{Var}\left(\hat{\theta}_2\left(\{X_i^n\}\right)\right)$ inequality must hold at all times, *i.e.*. for all n.

6.1.3 *Consistency*

$$\lim_{n\to\infty} \mathrm{Var}\left(\hat{\theta}\right) = 0$$

Consistency is an individual property. Consider $\hat{\theta} = \hat{\theta}\left(\{X_i\}_{i=1}^{n}\right)$. If $\hat{\theta}$ is a consistent estimator, repeating the task of estimation with larger samples reduces the error of estimation. That is,

$$\mathrm{Var}\left(\hat{\theta}\left(\{X_i\}_{i=1}^{n}\right)\right) \xrightarrow{n\to\infty} 0$$

6.1 EXERCISES

1. Consider a population that is uniformly distributed over the interval $[0, \beta]$ where β is an unknown positive number that we want to estimate. Three alternative estimators are given to serve our purpose as:

$$\hat{\beta}_1 = \max\{X_1, X_2, \ldots, X_n\} \quad \hat{\beta}_1 = \left(\tfrac{(n+1)}{n}\right)\max\{X_1, X_2, \ldots, X_n\}$$
$$\hat{\beta}_3 = 2\left(\tfrac{X_1, X_2, \ldots, X_n}{n}\right)$$

Estimator	Expected value	Variance
$\hat{\beta}_1$	$\frac{n}{n+1}\beta$	$\frac{n\beta^2}{(n+2)(n+1)^2}$
$\hat{\beta}_2$	β	$\frac{\beta^2}{(n+2)n}$
$\hat{\beta}_3$	β	$\frac{\beta^2}{3n}$

If you had to choose one of these estimators, which one would you pick? Why?

Solution: Among the three rival estimators, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$; $\hat{\beta}_2$ and $\hat{\beta}_3$ are unbiased as:

$$\mathrm{E}(\hat{\beta}_2) = \beta$$
$$\mathrm{E}(\hat{\beta}_3) = \beta$$

Both $\hat{\beta}_2$ and $\hat{\beta}_3$ are consistent as:

$$\lim_{n\to\infty} \mathrm{Var}\left(\hat{\beta}_2\right) = \lim_{n\to\infty} \frac{\beta^2}{(n+2)n} = 0$$

and

$$\lim_{n\to\infty} \mathrm{Var}\left(\hat{\beta}_3\right) = \lim_{n\to\infty} \frac{\beta^2}{3n} = 0$$

For $n > 1, \mathrm{Var}\left(\hat{\beta}_2\right) < \mathrm{Var}\left(\hat{\beta}_3\right)$. So, $\hat{\beta}_2$ is more efficient than $\hat{\beta}_3$. So, we choose $\hat{\beta}_2$.

2. Consider a population that is uniformly distributed over and interval $[0\beta]$ where β is an unknown positive number that we would like to estimate. We take a sample of size 3 from this population and use the **sample maximum** as an estimate for the population maximum.

i. What is the PDF of this estiamtor, *i.e.*, what is the sampling distribution of the sample maximum?

ii. What is the expected value of this estimator?

iii. Is this estimator an unbiased estimator for β?

iv. Suggest one more estimator for β which is unbiased.

Solution:

1. $\hat{\beta}$ is defined as:

$$\hat{\beta} = \max\{x_1, x_2, x_3\}$$

Where x_1, x_2 and x_3 are the sample observations. Representing each observation x_i as a random variable X_i:

$$\hat{\beta} = \max\{X_1, X_2, X_3\}$$

is attained, which is to be used in the subsequent computations. *CDF* of $\hat{\beta}$, $F_{\hat{\beta}}(t)$ is defined as a function of t as follows:

$$F_{\hat{\beta}}(t) = \mathrm{P}\left(\hat{\beta} \le t\right) = \mathrm{P}(\max\{X_1, X_2, X_3\} \le t)\,.$$

Assuming that X_1, X_2 and X_3 are independent random variables,

$$\begin{aligned}
&\mathrm{P}(\max\{X_1, X_2, X_3\} \le t)\\
=&\mathrm{P}(X_1 \le t \text{ and } X_2 \le t \text{ and } X_3 \le t)\\
=&\mathrm{P}(X_1 \le t) \cdot \mathrm{P}(X_2 \le t) \cdot \mathrm{P}(X_3 \le t)
\end{aligned}$$

can be written. Since X_1, X_2 and X_3 come from a Uniform$(0, \beta)$ distribution:

$$\begin{aligned} \mathrm{P}(X_1 \leq t) &= \frac{t}{\beta}, \\ \mathrm{P}(X_2 \leq t) &= \frac{t}{\beta}, \\ \mathrm{P}(X_3 \leq t) &= \frac{t}{\beta} \end{aligned}$$

Then, $F_{\hat{\beta}}(t) = \frac{t^3}{\beta^3}$ is obtained and it further yields:

$$f_{\hat{\beta}}(t) = \frac{dF_{\hat{\beta}}(t)}{dt} = \frac{3t^2}{\beta^3}.$$

2. $\mathrm{E}(\hat{\beta}) = \int_{-\infty}^{\infty} t f_{\hat{\beta}}(t)\, dt = \int_0^{\beta} t \frac{3t^2}{\beta^3} dt$

$$\begin{aligned} &= \frac{3}{4} \frac{t^4}{\beta^3} \Big|_0^3 \\ &= \frac{3}{4}\beta \end{aligned}$$

3. As $\mathrm{E}(\hat{\beta}) = \frac{3}{4}\beta < \beta$, $\hat{\beta}$ is not an unbiased estimator of β.
4. $\tilde{\beta} = \frac{4}{3} \max\{x_1, x_2, x_3\}$ is an unbiased estimator of β. Check why/how.

3. In a pasta factory, due to a miscalibration of the machinery, all spaghetti sticks produced had different lengths on a given day, then randomly packed and shipped to different locations and consumers. We wonder the length of the longest spaghetti stick produced that day. Offer a good statistical estimator to address this problem and discuss/evaluate its properties.
Solution: This exercise is left as self-study.

4. German tank problem
The problem is named after its use by Allied forces in World War II to estimate the monthly rate of German tank production from limited data. The approach enjoys the manufacturing practice of assigning and attaching ascending sequences of serial numbers to tank components, with some tanks being captured in battle by Allied forces. N being the total number of tanks produced, m being the highest serial number observed and k being the number of tanks captured, the estimator for N is given by:

$$\hat{N} = m + \frac{m}{k} - 1$$

Examine and explain each term in the formula.

Solution: The purpose in the German tank problem is to estimate the population size N. Supposing that German tank manufacturers number their tanks in an intuitive fashion, estimation of population size and estimation of population maximum are similar problems. m being the highest (largest) serial number among the captured tanks and k being the number of tanks captured (i.e., sample size), the estimator for N is given by:

$$\hat{N} = m + \frac{m}{k} - 1$$

In the formula, the first term (m) reflects the intuition that "sample maximum" is a good estimator of population maximum. However, when k is so small, it is not likely to observe higher serial numbers in the sample (among the captured tanks. To account for this, the m/k term is added to the first one. This term, for small k, yields a substantial addition to $\hat{N}$. For large k, the addition to $\hat{N}$ is limited. At the extreme, if $k = N$, then m is also equal to N. Such a configuration yields:

$$\hat{N} = N + \frac{N}{N} - 1 = N$$

indicating that 'when we capture all tanks, we, by definition know the total number of tanks.

Previously, we've studied the properties of point estimators in a thorough manner. However, we didn't devote any efforts to derivation of the estimators. In this chapter, we do it. This chapter introduces three techniques to yield point estimators of distributional parameters of populations or data generating processes. We'll consider three techniques (estimation criteria, estimation rules, or principles):

- Least squares (*LS*) technique
- Maximum likelihood (*ML*) technique
- Method of moments (*MM*) technique

As promised at the beginning of Chapter 8, we are now taking into consideration the problem of estimating distributional parameters. Our task here is to deriving the relevant mathematical formulas to generate numerical estimates. Recall that these formulas are called **estimators** as we've studied in Chapter 8.

6.2 *Least squares technique: LS*

Consider $\{x_i\}_{i=1}^{n} \subset \{x_i\}_{i=1}^{N} \sim \text{Normal}(\mu, \sigma^2)$ where we want to estimate unknown μ using our data. If we define a function S like:

$$S\left(\hat{\mu} | \{x_i\}_{i=1}^{n}\right) = \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

we can easily see it as a loss (or punishment) function. Solving:

$$\min_{\hat{\mu}} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

we find our estimator.

$$\begin{aligned}
\frac{dS(.)}{d\hat{\mu}} = \sum_{i=1}^{n} 2(x_i - \hat{\mu})(-1) = 0 \text{ (F.O.C.)} \\
-2\sum_{i=1}^{n} (x_i - \hat{\mu}) = 0 \\
\sum_{i=1}^{n} (x_i - \hat{\mu}) = 0 \\
\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \hat{\mu} \\
n\hat{\mu} = \sum_{i=1}^{n} x_i
\end{aligned}$$

So,

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

is found. By minimizing the loss function, we've obtained the *LS* estimator for μ, which is $\hat{\mu}$. Notice that, minimization of S here is equivalent to minimization of the unknown variance.

6.3 Maximum likelihood technique: ML

Consider $\{x_i\}_{i=1}^{n} \subset \{x_i\}_{i=1}^{N} \sim \text{Poisson}(\lambda)$ where we want to estimate unknown λ using our data. Let's denote the estimator as $\hat{\lambda}$. So, the likelihood of any x_i in our data set is:

$$f(x_i) = \frac{e^{-\hat{\lambda}} \hat{\lambda}^{x_i}}{x_i!}$$

Then, the likelihood of the whole data set becomes:

$$L(\hat{\lambda} | \{x_i\}_{i=1}^{n}) = \prod_{i=1}^{n} \frac{e^{-\hat{\lambda}} \hat{\lambda}^{x_i}}{x_i!}$$

which is to be maximized to find $\hat{\lambda}$.

For computational ease, we take the natural logarithm of the likelihood function and call it **log-likelihood function**, denoted $LogL$.

$$\begin{aligned}
\text{LogL}(\hat{\lambda} | \{x_i\}_{i=1}^{n}) &= \sum_{i=1}^{n} \left(\ln e^{-\hat{\lambda}} + \ln \hat{\lambda}^{x_i} - \ln x_i! \right) \\
&= \sum_{i=1}^{n} (-\hat{\lambda}) + \sum_{i=1}^{n} (x_i \ln \hat{\lambda}) - \sum_{i=1}^{n} (\ln x_i!)
\end{aligned}$$

Thus,

$$\text{LogL}(.) = -n\hat{\lambda} + \ln\hat{\lambda}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\ln x_i!$$

Solving,

$$\max_{\hat{\lambda}}\left(n\hat{\lambda} + \ln\hat{\lambda}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\ln x_i!\right)$$

$$\frac{d\,\text{LogL}(.)}{d\hat{\lambda}} = -n + \frac{1}{\hat{\lambda}}\sum_{i=1}^{n} x_i = 0 \text{ (F.O.C.)}$$

$$\frac{1}{\hat{\lambda}}\sum_{i=1}^{n} = n$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$$

is found. By maximizing the likelihood function, we've obtained the *ML* estimator for λ, which is $\hat{\lambda}$.

6.4 *Method of moments technique: MM*

Consider $\{x_i\}_{i=1}^{n} \subset \{x_i\}_{i=1}^{N} \sim \text{Normal}(\mu, \sigma^2)$ where we want to estimate unknown μ and σ^2 using our data. Remember that the theoretical moments for the population are:

$$\mu_1^{'} = \mu$$
$$\mu_2^{'} = \sigma^2 + \mu^2$$

Numerical values of each are all unknown. Yet, we know the values of data moments as:

$$M_1 = \frac{\sum_{i=1}^{n} x_i}{n}$$
$$M_2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$$

Now, all we need to do is to match population moments with data moments:

$$\mu_1^{'} = M_1 \to \mu = \frac{\sum_{i=1}^{n} x_i}{n}$$
$$\mu_2^{'} = M_2 \to \sigma^2 + \mu^2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$$

As we have a system of two equations with two unknowns, if it exists, the solution is:

$$\hat{\mu} = M_1$$
$$\hat{\sigma^2} = M_2 - M_1^2$$

$\hat{\mu}$ and $\hat{\sigma^2}$ are the *MM* estimators of μ and σ^2, respectively.

6.2 EXERCISES

1. Derive the **Least Squares** estimator for the mean of a population out of which you are given a data set of $\{x_i\}_{i=1}^n$. Analyze the unbiasedness and consistency of the estimator you've derived.
   Solution: Follow the replicate the solution given in the chapter. It simply yields:
   $$\hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$
   Writing our estimator $\hat{\mu}$ as a random variable $\bar{X}_n$, it is easy to compute $\mathrm{E}(\bar{X}_n)$ and $\mathrm{Var}(\bar{X}_n)$:
   $$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$
   $$\mathrm{E}(\bar{X}_n) = \mu$$
   $$\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$
   So, $\hat{\mu}$ is an unbiased and consistent estimator.

2. We want to estimate λ for a *Poisson* population. However, we have only data on interarrival times of occurrences rather than the count of occurrences per unit time. Denoting this data set as $\{x_i\}_{i=1}^n$, try to estimate λ using the **Maximum Likelihood** technique.
   Solution: Since *Poisson* and *Exponential* are sister distributions (sharing the same parameter λ as we have maintained in these Lecture Notes), we can use the interarrival time data $\{x_i\}_{i=1}^n$ via Maximum Likelihood technique to estimate λ.
   Consider $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N \sim \text{Exponential}(\lambda)$ where we want to estimate unknown λ using our data. Let's denote the estimator as $\hat{\lambda}$. So, the likelihood of any x_i in our data set is:
   $$f(x_i) = \hat{\lambda}\mathrm{e}^{-\hat{\lambda}x_i}$$
   Then, the likelihood of the whole data set becomes:
   $$L(\hat{\lambda}|\,\{x_i\}_{i=1}^n) = \prod_{i=1}^n \hat{\lambda}\mathrm{e}^{-\hat{\lambda}x_i}$$
   which is to be maximized to find $\hat{\lambda}$. Then, the log-likelihood function, *LogL* becomes:
   $$\mathrm{LogL}(\hat{\lambda}|\,\{x_i\}_{i=1}^n) = \sum_{i=1}^n (\ln\hat{\lambda} - \hat{\lambda}x_i)$$
   $$\mathrm{LogL}(.) = n\ln\hat{\lambda} - \hat{\lambda}\sum_{i=1}^n x_i$$

Solving,

$$\max_{\hat{\lambda}} \left(n \ln \hat{\lambda} - \hat{\lambda} \sum_{i=1}^{n} x_i \right)$$

$$\frac{d \operatorname{LogL}(.)}{d\hat{\lambda}} = \frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} x_i = 0 \text{ (F.O.C.)}$$

$$\frac{n}{\hat{\lambda}} = \sum_{i=1}^{n} x_i$$

$$\hat{\lambda} = \frac{1}{\frac{\sum_{i=1}^{n} x_i}{n}}$$

3. We believe the monthly book expenditures by students of a college has a **normal distribution**, yet we don't know μ and σ^2. Luckily, we have the following data on book expenditures:

5, 15, 20, 40, 40, 45, 50, 50, 50, 55

Using the **Methods of Moments**, estimate the population mean and variance.

Solution: The theoretical moments are:

$$\mu_1^{'} = \mu$$
$$\mu_2^{'} = \sigma^2 + \mu^2$$

We calculate the data moments for 5, 15, 20, 40, 40, 45, 50, 50, 50, 55 as:

$$M_1 = \frac{\sum_{i=1}^{n} x_i}{n} = 37$$
$$M_2 = \frac{\sum_{i=1}^{n} x_i^2}{n} = 1640$$

Matching the population (theoretical) moments with data moments:

$$\mu_1^{'} = M_1 \rightarrow \mu = 37$$
$$\mu_2^{'} = M_2 \rightarrow \sigma^2 + \mu^2 = 1640$$

Finally, the solution is:

$$\hat{\mu} = M_1 = 37$$
$$\hat{\sigma^2} = M_2 - M_1^2 = 1640 - 37^2 = 271$$

4. Given a sample data set $\{x_i\}_{i=1}^{n}$ out of a Uniform(a, b) population, estimate a and b using:

i. The *Least Squares* technique

ii. The *Maximum Likelihood* technique

iii. The *Method of Moments* technique

In which cases you are successful? Why?
Solution: This exercise is left as self-study.

5. We know that a population of values (X) has a $Funny\,(a,b)$ distribution. We also know that PDF of the $Funny\,(a,b)$ distribution is defined properly as:

$$f(x) = \binom{a}{x} b^x\,(1-b)^{a-x}\,, x = 0,1,2,\ldots,a; a \in \mathbb{Z}^+$$

First reveal and write the properties of b. Then, supposing a random sample of $\{x_i\}_{i=1}^n$ is available, derive the *Method of Moments* estimator of distributional parameters a and b. Explain whether you require any specific features for your sample so that you will be able to obtain numerically consistent estimates.
Solution: This exercise is left as self-study.

7 Confidence intervals

7.1 Confidence interval estimation: One population

At the surface, the problem of interval estimation seems to be a straightforward one. Despite this true from a computational viewpoint, not true from a more philosophical angle. We'll be discussing these issues in our lectures. The problem of interval estimation is to find a real number interval to contain an unknown population parameter of interest with a given or chosen value of probability.

In a nutshell

Interval estimation of μ: Construction of the problem

$$\begin{aligned} \mathrm{P}(L \leq \mu \leq U) &= 1-\alpha \\ \mathrm{P}(\mu - K \leq \mu \leq \mu + K) &= 1-\alpha \end{aligned}$$

where $L = \mu - K$ and $U = \mu + K$.

$$\begin{aligned} \mathrm{P}(-\mu - K \leq -\mu \leq -\mu + K) &= 1-\alpha \\ \mathrm{P}(\bar{x}_n - \mu - K \leq \bar{x}_n - \mu \leq \bar{x}_n - \mu + K) &= 1-\alpha \\ \mathrm{P}\left(\underbrace{\frac{\bar{x}_n - \mu - K}{\sigma/\sqrt{n}}}_{-z_c\ (1)} \leq \underbrace{\frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}}}_{z} \leq \underbrace{\frac{\bar{x}_n - \mu + K}{\sigma/\sqrt{n}}}_{z_c\ (2)} \right) &= 1-\alpha \end{aligned}$$

Considering (1) and (2) simultaneously $\rightarrow K = z_c \frac{\sigma}{\sqrt{n}}$. Then,

$$\begin{aligned} \mathrm{P}\left(\underbrace{\mu}_{\text{Subs.}\bar{x}} - z_c \frac{\sigma}{\sqrt{n}} \leq \mu \leq \underbrace{\mu}_{\text{Subs.}\bar{x}} + z_c \frac{\sigma}{\sqrt{n}} \right) &= 1-\alpha \\ \mathrm{P}\left(\bar{x} - z_c \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_c \frac{\sigma}{\sqrt{n}} \right) &= 1-\alpha \end{aligned}$$

Often enough, but not always, a confidence interval is symmetric around a mean. We'll formalize our discussion after our introductory exercise. Before that exercise, note that: **A confidence interval estimator for a population parameter is a rule for determining based on sample information, an interval that is likely to include the parameter. The corresponding estimate is called a confidence interval estimate.**

7.1.1 *Confidence interval estimation*
Mean of a normal population
Case: Known population variance

Consider $\{x_i\}_{i=1}^n$ a random sample of n observations from a normal population Normal(μ, σ^2). If the sample mean is $\bar{x}$, then a $(1-\alpha)\,100\%$ confidence interval for μ with known σ^2 is:

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Here,

$$\text{ME} = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

is called the margin or error (or sampling error),

$$w = 2ME$$

is called the width.

$$\text{UCL} = \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

is called **the upper confidence limit**, and

$$\text{LCL} = \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

is called **the lower confidence limit**.

Note that, definitions of ME, w, UCL and LCL will not be repeated in the other cases to save some space. Think about the ways to reduce the margin of error. Is everything under your control?

7.1 EXERCISES

1. A researcher wants to estimate a 95% confidence interval for the mean wage rate of workers in Ankara, for which the population variance is known to be 1000000 . She uses a sample of 64 workers and measures their mean wage rate as 7000. Calculate/estimate the confidence interval requested.
Solution: The population variance is known; so, the relevant distribution iz z & the 95% confidence interval for μ is:

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \rightarrow 7000 \pm 1.96 \frac{1000}{\sqrt{64}}$$

$$\rightarrow [6755.005, 7244.995]$$

7.1.2 *Confidence interval estimation*
Mean of a normal population
Case: Unknown population variance

Consider $\{x_i\}_{i=1}^{n}$ a random sample of n observations from a normal population Normal(μ, σ^2). If the sample mean is $\bar{x}$, then a $(1-\alpha)\,100\%$ confidence interval for μ with unknown σ^2 is:

$$\bar{x} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

where,

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

is the sample standard deviation. Go over the description of t-distribution in Chapter 10.

7.2 EXERCISES

1. A researcher wants to estimate a 95% confidence interval for the mean wage rate of workers in Ankara, for which the population variance is unknown. She uses a sample of 64 workers and measures their mean wage rate as 7000 , 'sample variance' being calculated as 640000. Calculate/estimate the confidence interval requested.
   Solution: The population variance is unknown; so, the relevant distribution iz t, the degrees of freedom is 63 & the 95% confidence interval for μ is:

$$\bar{x} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \rightarrow 7000 \pm 1.998 \frac{800}{\sqrt{64}}$$

$$\rightarrow [6800.166, 7199.834]$$

7.1.3 *Confidence interval estimation*
Population proportion

Consider $\{x_i\}_{i=1}^{n}$, a random sample of n observations from a Bernoulli(P) population. Notice that each x_i is either 1 $(success)$ or 0 $(failure)$. $\bar{x}$in this case is nothing but the **observed** proportion of successed, denoted as $\hat{p}$. Then, if $n\hat{p}(1-\hat{p})$, a $(1-\alpha)$ 100%confidence interval for p is:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

7.3 EXERCISES

1. A political candidate wants to know her nationwide support rate. Among a sample of 64 people, we know 35 support the political candidate. Calculate/estimate a 95% confidence interval for the candidate's nation-wide support rate.
   Solution: The relevant distribution is z.

$$\hat{p} = \frac{35}{64} = 0.547$$

The 95% confidence interval for P is:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} \rightarrow 0.547 \pm 1.96 \sqrt{\frac{0.547\,(1-0.547)}{64}}$$

$$\rightarrow [0.425, 0.669]$$

7.1.4 *Confidence interval estimation*
Variance of a normal population

Consider $\{x_i\}_{i=1}^{n}$, a random sample of n observations from a normal population Normal(μ, σ^2). If the observed sample variance is s^2, then a $(1-\alpha)\,100\%$ confidence interval for σ^2 is:

$$\left[\frac{(n-1)\,s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)\,s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right]$$

where

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

is the sample variance. Go over the description of χ^2-distribution in Chapter 10.

7.4 EXERCISES

1. A process engineer is concerned with the variation of temperature in an industrial furnace. She collects a random sample of temperatures as:

975 1075 1050 900
1000 950 1025 1050
975°C

Calculate/estimate a 95% confidence interval for the (population) variance of temperatures in this furnace.

Solution: s^2 for the 9 temperature readings is 3125, the relevant distribution is χ^2 and the degrees of freedom is 8. The 95% confidence interval for σ^2 is:

$$\left[\frac{(n-1)\,s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)\,s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right] \rightarrow \left[\frac{(9-1)\,3125}{17.535}, \frac{(9-1)\,3125}{2.180}\right]$$

$$\rightarrow [1425.757, 11469.306]$$

7.2 *Finite populations and correction*

When $n << N$, our procedures work seamlessly. However, when n is considerably high, *i.e.*

$$n > \frac{1}{20}N$$

we need to use a factor of

$$\frac{N-n}{N-1}$$

to correct the relevant variances involved. This factor is called the **finite population correction** fpc factor. Observe that fpc $= 1$ for $n = 1$ and fpc $= 0$ for $n = N$ in a very intuitive manner.

7.3 *Sample size determination*

Mean of a normally distributed population, known population variance:

$$n = \frac{z^2_{1-\frac{\alpha}{2}}\sigma^2}{\text{ME}^2}$$

Population proportion:

$$n = \frac{0.25z^2_{1-\frac{\alpha}{2}}}{\text{ME}^2}$$

7.5 EXERCISES

1. A random sample, of size 9, from a normally distributed population, with variance 9, yielded as sample mean of 7 and a sample variance of 4.

i. Construct a 90% confidence interval for the population mean of the population that the sample is taken from.

ii. What is the probability of a random sample of size 9 yielding a sample variance of 4 or less, given that the population variance is 9?

iii. What is the minimum sample size required if we would like the 90% confidence interval to be at most of length 2?

Solution: This question is under maintenance.

2. A random sample 100 consumers where asked if they made their purchasing decisions based on price or based on quality. 64% of the consumers in the sample stated that they mainly base their buying decisions on price. Based on this information, construct a 95% confidence interval for the percentage of consumers in the population who base their buying decision on price.

Solution: $n = 100$, $\hat{p} = 0.64$ are given. A 95% confidence interval for P is:

$$\begin{aligned}
&\hat{p} \mp z_c \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\
\rightarrow &0.64 \mp 1.96 \sqrt{\frac{0.64 \cdot 0.36}{100}} \\
\rightarrow &0.64 \mp 1.96 \cdot 0.048 \\
\rightarrow &0.64 \mp 0.0941.
\end{aligned}$$

So, $\mathrm{P}(P \in [0.5459, 0.7341]) = 0.95$.

3. Two statisticians, using the same sample data reported the following different confidence intervals for the population mean: $[3, 5]$ and $[2, 6]$. Given that they used the same sample and based their confidence interval on same estimators (the sample mean and sample variance), what is the source of the difference in the confidence interval?
Solution: This exercise is left as self-study.

4. A researcher has a strange habit of using a sample size of $\sqrt{N}$ where N is the size of the population of interest. Under which value of N does the researcher need to use a correction factor for the standard deviation of the sampling distribution while estimating a CI for μ?
Solution: This exercise is left as self-study.

7.4 Confidence interval estimation: Two populations

In scientific research, we often need to compare selected parameters of two populations, rather than only comparing to a single population parameter to a given value. Despite the problem gets slightly complicated, essence of the problem is unchanged. So, while considering the confidence interval estimation and hypothesis testing problems involving two populations, we'll first maintain a mechanical approach in what follows. Through the following pages, notice

- That we use a more shorthand notation
- That we use two samples every time:

$$\begin{aligned}
\{x_i\}_{i=1}^{n_x} &\subset \{x_i\}_{i=1}^{N_x} \\
\{y_i\}_{i=1}^{n_y} &\subset \{y_i\}_{i=1}^{N_y}
\end{aligned}$$

  where n_x and n_y are their respective sample sizes.

7.4.1 *Confidence interval estimation*

Difference between two normal population means

Case: Dependent (matched) samples

Let $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ be two matched samples. We can then creat:

$$\{d_i\}_{i=1}^n where d_i = x_i - y_i, \forall i$$

Then, a $(1-\alpha)\,100\%$ confidence interval for $\mu_d = \mu_x - \mu_y$ is:

$$\bar{d} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}$$

where

$$s_d = \sqrt{s_d^2} = \sqrt{\frac{\sum_{i=1}^n \left(d_i - \bar{d}\right)^2}{n-1}}$$

and,

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n}$$

7.6 EXERCISES

1. A company is about to release a new drug to assist weight loss, and we are in charge of assessing how effective the drug is. We pick a random sample of 8 people with the following pre-drug body weights:

$$90, 95, 105, 95, 110, 85, 100, 90$$

After using the drug for the designated test duration, the post-drug body weights are measured as:

$$85, 80, 110, 90, 110, 80, 95, 90$$

Calculate/estimate a 95% confidence interval for the pre-drug minus post-drug difference of mean body weights. Is the drug effective?

Solution: The difference series (pre-drug minus post-drug) is:

$$+5, +15, -5, +5, 0, +5, +5, 0$$

The relevant distribution is t, the degrees of freedom is 7 and the 95% confidence interval for $\mu_x - \mu_y$ is:

$$\bar{d} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}} \rightarrow 3.750 \pm 2.365 \frac{5.824}{\sqrt{8}}$$

$$\rightarrow [-1.120, 8.620]$$

7.4.2 *Confidence interval estimation*
Difference between two normal population means
Case: Independent samples & Known population variances

Let,

$$\{x_i\}_{i=1}^{n_x} \subset \{x_i\}_{i=1}^{N_x} \sim \text{Normal}(\mu_x, \sigma_x^2)$$
$$\{y_i\}_{i=1}^{n_y} \subset \{y_i\}_{i=1}^{N_y} \sim \text{Normal}(\mu_y, \sigma_y^2)$$

where σ_x^2 and σ_y^2 are known. Then a $(1-\alpha)\,100\%$ confidence interval for $\mu_x - \mu_y$ is:

$$\bar{x} - \bar{y} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

7.7 EXERCISES

1. A researcher wants to compare the mean wages of workers in Ankara and Istanbul. She has the following data and information: Mean wage rate of 49 workers from Ankara is 6000. Mean wage rate of 81 workers from Istanbul is 7000. Population variance of wages in Ankara and Istanbul are known to be 640000 and 810000, respectively. Calculate/ estimate a 95% confidence interval for the difference of (population) means of wages in Ankara and Istanbul.
Solution: The relevant distribution is z and the 95% confidence interval for $\mu_x - \mu_y$ is:

$$\bar{x} - \bar{y} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \to 6000 - 7000 \pm 1.96 \sqrt{\frac{640000}{49} + \frac{810000}{81}}$$
$$\to [-1297.639, -702.361]$$

7.4.3 *Confidence interval estimation*
Difference between two normal population means
Case: Independent samples & Unknown yet equal population variances

Let,

$$\{x_i\}_{i=1}^{n_x} \subset \{x_i\}_{i=1}^{N_x} \sim \text{Normal}(\mu_x, \sigma_x^2)$$
$$\{y_i\}_{i=1}^{n_y} \subset \{y_i\}_{i=1}^{N_y} \sim \text{Normal}(\mu_y, \sigma_y^2)$$

where σ_x^2 and σ_y^2 are unkown but assumed to be equal. Then a $(1-\alpha)\,100\%$ confidence interval for $\mu_x - \mu_y$ is:

$$\bar{x} - \bar{y} \pm t_{n_x+n_y-2,1-\frac{\alpha}{2}} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$$

In our formulation:

$$s_p^2 = \frac{(n_x - 1)\, s_x^2 + (n_y - 1)\, s_y^2}{n_x + n_y - 2}$$

and

$$s_x^2 = \frac{\sum_{i=1}^{n_x} (x_i - \bar{x})^2}{n_x - 1}$$
$$s_y^2 = \frac{\sum_{i=1}^{n_y} (y_i - \bar{y})^2}{n_y - 1}$$

7.8 EXERCISES

1. A researcher wants to compare the mean wages of workers in Ankara and Istanbul. She has the following data and information: Mean wage rate of 49 workers from Ankara is 6000. Mean wage rate of 81 workers from Istanbul is 7000. Population variances of wages in Ankara and Istanbul are unknown but they are assumed to be equal. Sample variance of wages in Ankara and Istanbul are calculated as 490000 and 640000, respectively. Calculate/ estimate a 95% confidence interval for the difference of (population) means of wages in Ankara and Istanbul.

   Solution:

$$s_p^2 = \frac{(n_x - 1)\, s_x^2 + (n_y - 1)\, s_y^2}{n_x + n_y - 2}$$
$$s_p^2 = \frac{(49 - 1)\, 490000 + (81 - 1)\, 640000}{49 + 81 - 2} \rightarrow s_p^2 = 583750$$

   The relevant distribution is t, the degrees of freedom is 128 and the 95% confidence interval for $\mu_x - \mu_y$ is:

$$\bar{x} - \bar{y} \pm t_{n_x+n_y-2,1-\frac{\alpha}{2}} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$$
$$\rightarrow 6000 - 7000 \pm 1.979 \sqrt{\frac{583750}{49} + \frac{583750}{81}}$$
$$\rightarrow [-1273.601, -726.399]$$

7.4.4 *Confidence interval estimation*

Difference between two normal population means

Case: Independent samples & Unknown and unequal population variances

$$\{x_i\}_{i=1}^{n_x} \subset \{x_i\}_{i=1}^{N_x} \sim \text{Normal}(\mu_x, \sigma_x^2)$$
$$\{y_i\}_{i=1}^{n_y} \subset \{y_i\}_{i=1}^{N_y} \sim \text{Normal}(\mu_y, \sigma_y^2)$$

where σ_x^2 and σ_y^2 are unknown and assumed not to be equal. Then a $(1-\alpha)\,100\%$ confidence interval for $\mu_x - \mu_y$ is:

$$\bar{x} - \bar{y} \pm t_{v,1-\frac{\alpha}{2}} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

In our formulation:

$$v = \frac{\left(\left(\frac{s_x^2}{n_x}\right) + \left(\frac{s_y^2}{n_y}\right)\right)^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y - 1}}$$

Notice that, if $n_x = n_y = n$

$$v = \left(1 + \frac{2}{\frac{s_x^2}{s_y^2} + \frac{s_y^2}{s_x^2}}\right)(n-1)$$

7.9 EXERCISES

1. A researcher wants to compare the mean wages of workers in Ankara and Istanbul. She has the following data and information: Mean wage rate of 49 workers from Ankara is 6000. Mean wage rate of 81 workers from Istanbul is 7000. Population variances of wages in Ankara and Istanbul are unknown and they are assumed to be unequal. Sample variance of wages in Ankara and Istanbul are calculated as 490000 and 640000, respectively. Calculate/ estimate a 95% confidence interval for the difference of (population) means of wages in Ankara and Istanbul.

Solution: The relevant distribution is t and the degrees of freedom is ν:

$$v = \frac{\left(\left(\frac{s_x^2}{n_x}\right) + \left(\frac{s_y^2}{n_y}\right)\right)^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y - 1}}$$

$$\rightarrow v = \frac{\left(\left(\frac{490000}{49}\right) + \left(\frac{640000}{81}\right)\right)^2}{\frac{\left(\frac{490000}{49}\right)^2}{49-1} + \frac{\left(\frac{640000}{81}\right)^2}{81-1}} \rightarrow 112$$

The 95% confidence interval for $\mu_x - \mu_y$ is:

$$\bar{x} - \bar{y} \pm t_{v,1-\frac{\alpha}{2}} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \rightarrow 6000 - 7000 \pm 1.982 \sqrt{\frac{490000}{49} + \frac{640000}{81}}$$

$$\rightarrow [-1265.125, -734.875]$$

7.4.5 *Confidence interval estimation*
Difference between two population proportions

Let,

$$\{x_i\}_{i=1}^{n_x} \subset \{x_i\}_{i=1}^{N_x} \sim \text{Bernoulli}(P_x)$$
$$\{y_i\}_{i=1}^{n_y} \subset \{y_i\}_{i=1}^{N_y} \sim \text{Bernoulli}(P_y)$$

Then, a $(1-\alpha)\,100\%$ confidence interval for $p_x - p_y$ is:

$$\hat{p}_x - \hat{p}_y \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_x\left(1-\hat{p}_x\right)}{n_x} + \frac{\hat{p}_y\left(1-\hat{p}_y\right)}{n_y}}$$

7.10 EXERCISES

1. A political candidate wonders how her support rate in Ankara and Istanbul compares. We know that among 64 people from Ankara 35 supports the candidate and among 81 people from Istanbul 45 supports the candidate. Calculate estimate a 95% confidence interval for the difference of population support rates in Ankara and Istanbul.
   Solution: $\hat{p}_x = 35/64 = 0.547$ and $\hat{p}_y = 45/81 = 0.556$ and the 95% confidence interval for $P_x - P_y$ is:

$$\hat{p}_x - \hat{p}_y \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_x\left(1-\hat{p}_x\right)}{n_x} + \frac{\hat{p}_y\left(1-\hat{p}_y\right)}{n_y}}$$

$$\rightarrow 0.547 - 0.556 \pm 1.96 \sqrt{\frac{0.547\left(1-0.547\right)}{64} + \frac{0.556\left(1-0.556\right)}{81}}$$

$$\rightarrow [-0.172, 0.154]$$

7.11 EXERCISES

1. Journal A of city X reports a 90% CI for the population mean income in city $X\,(\mu_x)$ as $[3400, 6400]$ and Journal B in city $Y\,(\mu_y)$ as $[3800, 6800]$. Each journal notes that the population is distibuted normally with a **known** variance. The journals report the odds for the approval of Mr. Doe, a political candidate, in their respective cities of X and Y as $750/500$ and $80/60$ where the fractions are the (sample count of approvals/sample count of disapprovals).

i. Estimate a 95% CI for μ_x

ii. Estimate a 99% CI for $\mu_x - \mu_y$

iii. Test: $H_0 : \mu_x - \mu_y = 0$ againts $H_1 : \mu_x - \mu_y < 0$ at $\alpha = 0.05$

iv. Estimate a 90% CI for the popularity (share of approvals) of Mr. Doe in city X (p_x)

v. Estimate a 95% CI for $p_x - p_y$

Interpret your result clearly in each case.

Solution: To come up with solutions to parts (i) to (v), we need to find/ calculate $\bar{x}, \bar{y}, \hat{p}_x, \hat{p}_y, n_x$ and n_y from the given information. In that,

$$\bar{x} = \frac{3400 + 6400}{2} = 4900$$
$$\bar{y} = \frac{3800 + 6800}{2} = 5300$$
$$\frac{\sigma_x}{\sqrt{n_x}} = \frac{4900 - 3400}{1.65} = 909.1$$
$$\frac{\sigma_y}{\sqrt{n_y}} = \frac{5300 - 3800}{1.65} = 909.1$$

These are sufficient to solve parts (i), (ii) and (iii). To solve (iv) and (v), we need the following:

$$\hat{p}_x = \frac{750}{750 + 500} = 0.60, n_x = 1250$$
$$\hat{p}_y = \frac{80}{80 + 60} = 0.57, \quad n_y = 140$$

Notice that, n_x and n_y are not to be necessarily and explicitly known while solving parts (i), (ii) and (iii).

2. A researcher wants to test whether three populations' means are equal to each other; *i.e.* whether (A) $H_0 : \mu_1 = \mu_2 = \mu_3$. She picks α as 0.10 and collects data from each population. She, then, tests separately (one at a time):

(B) $H_0 : \mu_1 = \mu_2$ (C) $H_0 : \mu_1 = \mu_3$ (D) $H_0 : \mu_2 = \mu_3$

against their two-sided alternatives. She fails to reject H_0 every time at $\alpha = 0.10$. So, she concludes: *Failure to reject H_0 in all B, C and D at $\alpha = 0.10$ is equivalent to failure to reject H_0 in A at $\alpha = 0.10$; so, all three means are equal with a confidence of* 90%.

Explain why her conclusion is wrong.

Solution: This question requires the execution of (1) obtaining the confidence levels for the tests A, B, and C, (2) noticing that these confidence levels are nothing but simple probabilities, (3) multiplying the individual confidence levels to obtain the joint

confidence level, (4) subtracting the joint confidence level from 1 to find the joint significance level (call it α'). In that,

$$\alpha' = 1 - (1 - \alpha)^3$$

and for $\alpha = 0.10, \alpha'$ becomes 0.271, so 'straightforwardly joining/merging the conclusions of separate tests of hypotheses' is not allowed in out professional practice.

In a nutshell

A t random variable with m degrees of freedom, denoted $t_{(m)}$ is found by:

$$t = \frac{Z}{\sqrt{\frac{\chi^2_{(m)}}{m}}} \sim t_{(m)}$$

if the numerator and denominator are independent random variables. Here, consider the meaning of:

$$\frac{\chi^2_{(m)}}{m}$$

Previously we said χ^2 should be something related to variance. Based on the definition of $\chi^2_{(m)}$, do you think the fraction above is the variance of **something**?

8 Hypothesis testing

8.1 Hypothesis testing: One population

A sizable volume of scientific efforts involve questions on a given population parameter. While we estimate/calculate an interval to contain an unknown population parameter with a probability of $(1-\alpha)\,100\%$ under the heading of Confidence interval estimation, here in Hypothesis testing, we question the viability of a given value as the value of our unknown population parameter.

So, what we do is to check for the validity of a claim about an unknown, in formal terms.

A statistical **hypothesis** is a statement about the numerical value of a parameter. The **null hypothesis**, denoted H_0, represents the hypothesis that is assumed to be true unless the data provide convincing counter evidence. This usually represents the **status quo** or some claim about the parameter that the researcher states.

The **alternative hypothesis**, denoted H_1, represents the hypothesis that will be maintained only if the data provide convincing evidence for its truth.

The **test statistic** is a sample statistic, computed from information provided in the sample, that the researcher uses to decide between the **null** and **alternative** hypotheses.

A **Type I error** occurs if the researcher rejects the null hypothesis in favor of the alternative hypothesis when, in fact, the null hypothesis is true. The probability of Type I error is denoted as α.

The **rejection region** of a statistical test is the set of possible values of the test statistic for which the researcher will reject the null hypothesis in favor of the alternative.

A **Type II error** occurs if the researcher fails to reject the null hypothesis when, in fact, the null hypothesis is false. The probability of Type II error is denoted as β.

In a nutshell

A step-by-step description of Hypothesis testing:

1. Establish H_0 and H_1
2. Determine test statistic (test score) and its statistical distribution
3. Set α
4. Experiment/collect data and calculate test statistic
5. If the value of the test statistic falls in the rejection region conclude **Reject** H_0, otherwise conclude **Fail to reject** H_0

In a nutshell

Two deadly sins of Hypothesis testing:

1. Never say **Accept** H_0 instead of **Fail to reject** H_0. When H_1 is not supported enough by the data, say:
   - Fail to reject H_0
   - Insufficient evidence to reject H_0
2. Never include H_1 in your concluding statement. Don't try to be creative when writing/speaking technical statments as such. Keep your appetite for the interpretation and discussion of findings.

Some perspective:
Note that confidence intervals are the back side of the coin for hypothesis tests. In testing $H_0 : \mu = 2$ versus $H_1 : \mu \neq 2$, if $\mu = 2$ lies in a 95% confidence interval, then we cannot reject $Ho : \mu = 2$ at $\alpha = 0.05$; otherwise we reject H_0. This highlights the fact that any value of μ that lies in this 95% confidence interval (assuming it was our null hypothesis) cannot be rejected at the 5% significance level by this sample. This is why we do not say 'accept H_0 ', but we rather say 'do not reject H_0'.

8.1 EXERCISES

1. For each of the situations below write the null and alternative hypotheses, corresponding to the test, in plain English; then write the null and alternative hypotheses using only mathematical symbols; then state what the symbols you used above represents:

i. We would like to test if the income share held by the highest earning 20% is less than 46%.

ii. We would like to test if the average income of males is greater than the average income of females.

iii. It is claimed that, among the people who drinks at least 2 liters of water every day, the percentage of those with a kidney problem is less than 5%. We suspect the truth of this statement and would like to test it.

iv. We would like to test if a coin is fair.

v. We would like to test if a coin is not a fair coin.

Solution: This exercise is left as self-study.

8.1.1 *Hypothesis testing*
Mean of a normal population
Case: Known population variance

Consider:

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$

and $\{x_i\}_{i=1}^n$, a random sample of n observations from a normal population Normal(μ, σ^2) with known σ^2.
At the statistical significance level α:

- H_0 is rejected if
$$\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha}$$
- and we fail to reject H_0 otherwise.

8.2 EXERCISES

1. A researcher wonders if the mean wage rate of workers in Ankara is greater than 6500. The population variance is known to be 1000000. The researcher measures the mean wage rate of a sample of 64 workers as 7000. Conduct and conclude the relevant hypothesis testate the significance level of 5%.
Solution:

$$H_0 : \mu \leq 6500$$
$$H_1 : \mu > 6500$$

The relevant distribution is z.
Since:

$$\frac{7000-6500}{\frac{1000}{\sqrt{64}}} = 4.00 > 1.65$$

we reject H_0 at $\alpha = 0.05$.

For:

$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

At the statistical significance level α:

- H_0 is rejected if

$$\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} < z_\alpha$$

- and we fail to reject H_0 otherwise.

8.3 EXERCISES

1. A researcher wonders if the mean wage rate of workers in Ankara is less than 7500. The population variance is known to be 1000000. The researcher measures the mean wage rate of a sample of 64 workers as 7000. Conduct and conclude the relevant hypothesis testate the significance level of 5%.
Solution:

$$H_0 : \mu \geq 7500$$
$$H_1 : \mu < 7500$$

The relevant distribution is z.
Since:

$$\frac{7000-7500}{\frac{1000}{\sqrt{64}}} = -4.00 < -1.65$$

we reject H_0 at $\alpha = 0.05$.

For:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

At the statistical significance level α:

- H_0 is rejected if

$$\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}} \text{ or } \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\frac{\alpha}{2}}$$

- and we fail to reject H_0 otherwise.

8.4 EXERCISES

1. A researcher wonders if the mean wage rate of workers in Ankara is different than 7500. The population variance is known to be 1000000. The researcher measures the mean wage rate of a sample of 64 workers as 7000. Conduct and conclude the relevant hypothesis testate the significance level of 5%.
Solution:

$$H_0 : \mu = 7500$$
$$H_1 : \mu \neq 7500$$

The relevant distribution is z.
Since:

$$\frac{7000 - 7500}{\frac{1000}{\sqrt{64}}} = -4.00 \text{ is outside of } [-1.96, 1.96]$$

we reject H_0 at $\alpha = 0.05$.

8.1.2 *Hypothesis testing*
Mean of a normal population
Case: Unknown population variance

Consider:

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$

and $\{x_i\}_{i=1}^n$, a random sample of n observations from a normal population Normal(μ, σ^2) where σ^2 is unkown.
At the statistical significance level α:

- H_0 is rejected if

$$\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} > t_{n-1,1-\alpha}$$

- and we fail to reject H_0 otherwise.

8.5 EXERCISES

1. A researcher wonders if the mean wage rate of workers in Ankara is greater than 6500 , for which the population variance is unknown.

The researcher measures the mean wage rate of a sample of 64 workers as 7000 and the 'sample variance' as 640000. Conduct and conclude the relevant hypothesis test at the significance level of 5%.

Solution:

$$H_0 : \mu \leq 6500$$
$$H_1 : \mu > 6500$$

The relevant distribution is t and the degrees of freedom is 63.
Since:

$$\frac{7000 - 6500}{\frac{800}{\sqrt{64}}} = 5.00 > 1.669$$

we reject H_0 at $\alpha = 0.05$.

For:

$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

At the statistical significance level α:

- H_0 is rejected if
$$\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} < t_{n-1,\alpha}$$
- and we fail to reject H_0 otherwise.

8.6 EXERCISES

1. A researcher wonders if the mean wage rate of workers in Ankara is less than 7500 , for which the population variance is unknown. The researcher measures the mean wage rate of a sample of 64 workers as 7000 and the 'sample variance' as 640000. Conduct and conclude the relevant hypothesis test at the significance level of 5%.
Solution:

$$H_0 : \mu \geq 7500$$
$$H_1 : \mu < 7500$$

The relevant distribution is t and the degrees of freedom is 63.
Since:

$$\frac{7000 - 7500}{\frac{800}{\sqrt{64}}} = -5.00 < -1.669$$

we reject H_0 at $\alpha = 0.05$.

For:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

At the statistical significance level α:

- H_0 is rejected if

$$\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} < t_{n-1,\frac{\alpha}{2}} \text{ or } \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} > t_{n-1,1-\frac{\alpha}{2}}$$

- and we fail to reject H_0 otherwise.

8.7 EXERCISES

1. A researcher wonders if the mean wage rate of workers in Ankara is different than 7500 , for which the population variance is unknown. The researcher measures the mean wage rate of a sample of 64 workers as 7000 and the 'sample variance' as 640000. Conduct and conclude the relevant hypothesis test at the significance level of 5%.
Solution:

$$H_0 : \mu = 7500$$
$$H_1 : \mu \neq 7500$$

The relevant distribution is t and the degrees of freedom is 63.
Since:

$$\frac{7000 - 7500}{\frac{800}{\sqrt{64}}} = -5.00 \text{ is outside of } [-1.998, 1.998]$$

we reject H_0 at $\alpha = 0.05$.

8.1.3 Hypothesis testing
Population proportion

Consider:

$$H_0 : P \leq P_0$$
$$H_1 : P > P_0$$

and $\{x_i\}_{i=1}^n$, a random sample of n observations from a Bernoulli(P) population.
At the statistical significance level α:

- H_0 is rejected if

$$\frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} > z_{1-\alpha}$$

- and we fail to reject H_0 otherwise.

8.8 EXERCISES

1. A political candidate wonders if her nationwide support rate exceeds 50%. Among a sample of 64 people, we know 35 support the political candidate. Conduct and conclude the relevant hypothesis test at the significance level of 5%.
   Solution:

$$H_0 : P \leq 0.50$$
$$H_1 : P > 0.50$$

   The relevant distribution is z.
   Since:

$$\frac{0.547 - 0.500}{\sqrt{\frac{0.500(1-0.500)}{64}}} = 0.750 \text{ is not greater than } 1.65$$

   we fail to reject H_0 at $\alpha = 0.05$.

   For:

$$H_0 : P \geq P_0$$
$$H_1 : P < P_0$$

At the statistical significance level α:

- H_0 is rejected if

$$\frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} < z_\alpha$$

- and we fail to reject H_0 otherwise.

8.9 EXERCISES

1. A political candidate wonders if her nationwide support rate falls short of 50%. Among a sample of 64 people, we know 30 support the political candidate. Conduct and conclude the relevant hypothesis test at the significance level of 5%.
   Solution:

$$H_0 : P \geq 0.50$$
$$H_1 : P < 0.50$$

The relevant distribution is z.
Since:

$$\frac{0.469 - 0.500}{\sqrt{\frac{0.500(1-0.500)}{64}}} = -0.500 \text{ is not less than } -1.65$$

we fail to reject H_0 at $\alpha = 0.05$.

For:

$$H_0 : P = P_0$$
$$H_1 : P \neq P_0$$

At the statistical significance level α:

- H_0 is rejected if

$$\frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} < z_{\frac{\alpha}{2}} \text{ or } \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} > z_{1-\frac{\alpha}{2}}$$

- and we fail to reject H_0 otherwise.

8.10 EXERCISES

1. A political candidate wonders if her nationwide support rate is different than 50%. Among a sample of 64 people, we know 35 support the political candidate. Conduct and conclude the relevant hypothesis test at the significance level of 5%.
Solution:

$$H_0 : P = 0.50$$
$$H_1 : P \neq 0.50$$

The relevant distribution is z.
Since:

$$\frac{0.469 - 0.500}{\sqrt{\frac{0.500(1-0.500)}{64}}} = -0.500 \text{ is not outside } [-1.96, 1.96]$$

we fail to reject H_0 at $\alpha = 0.05$.

8.1.4 Hypothesis testing
Variance of a normal population

Consider:

$$H_0 : \sigma^2 \leq \sigma_0^2$$
$$H_1 : \sigma^2 > \sigma_0^2$$

and $\{x_i\}_{i=1}^{n}$, a random sample of n observations from a normal population Normal(μ, σ^2).
At the statistical significance level α:

- H_0 is rejected if
$$\frac{(n-1)\,s^2}{\sigma_0^2} > \chi^2_{n-1,1-\alpha}$$
- and we fail to reject H_0 otherwise.

8.11 EXERCISES

1. A process engineer is concerned with the variation of - temperature in an industrial furnace and wonders if it exceeds 1500. She collects a random sample of temperatures as:

975	1075	1050	900
1000	950	1025	1050
975°C			

Conduct and conclude the relevant hypothesis test at the significance level of 5%.
Solution:

$$H_0 : \sigma^2 \leq 1500$$
$$H_1 : \sigma^2 > 1500$$

The relevant distribution is χ^2 and the degrees of freedom is 8.
Since:
$$\frac{(9-1)\,3125}{1500} = 16.667 > 15.507$$
we reject H_0 at $\alpha = 0.05$.

For:

$$H_0 : \sigma^2 \geq \sigma_0^2$$
$$H_1 : \sigma^2 < \sigma_0^2$$

At the statistical significance level α:

- H_0 is rejected if
$$\frac{(n-1)\,s^2}{\sigma_0^2} < \chi^2_{n-1,\alpha}$$
- and we fail to reject H_0 otherwise.

8.12 EXERCISES

1. A process engineer is concerned with the variation of - temperature in an industrial furnace and wonders if it is less than 2500. She collects a random sample of temperatures as:

975	1075	1050	900
1000	950	1025	1050
975°C			

Conduct and conclude the relevant hypothesis test at the significance level of 5%.
Solution:

$$H_0 : \sigma^2 \geq 2500$$
$$H_1 : \sigma^2 < 2500$$

The relevant distribution is χ^2 and the degrees of freedom is 8.
Since:

$$\frac{(9-1)\,3125}{2500} = 10.000 \text{ is not less than } 2.733$$

we fail to reject H_0 at $\alpha = 0.05$.

For:

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 \neq \sigma_0^2$$

At the statistical significance level α:

- H_0 is rejected if

$$\frac{(n-1)\,s^2}{\sigma_0^2} < \chi^2_{n-1,\frac{\alpha}{2}} \text{ or } \frac{(n-1)\,s^2}{\sigma_0^2} > \chi^2_{n-1,1-\frac{\alpha}{2}}$$

- and we fail to reject H_0 otherwise.

8.13 EXERCISES

1. A process engineer is concerned with the variation of - temperature in an industrial furnace and wonders if it is different than 2000. She collects a random sample of temperatures as:

975	1075	1050	900
1000	950	1025	1050
975°C			

Conduct and conclude the relevant hypothesis test at the significance level of 5%.
Solution:

$$H_0 : \sigma^2 = 2000$$
$$H_1 : \sigma^2 \neq 2000$$

The relevant distribution is χ^2 and the degrees of freedom is 8. Since:

$$\frac{(9-1)\,3125}{2000} = 12.500 \text{ is not outside } [2.180, 17.535]$$

we fail to reject H_0 at $\alpha = 0.05$.

8.14 EXERCISES

1. A manufacturer of automobile batteries claim that at least 80% of the batteries that it produces will last 36 months. A consumers' advocate group wants to evaluate this longevity claim and selects a random sample of 28 batteries to test. The following data indicate the length of time (in months) that each of these batteries lasted (*i. e.*, performed properly before failure):

42.3, 39.6, 25.0, 56.2, 37.2, 47.4, 57.5, 39.3, 39.2, 47.0, 47.4, 39.7, 57.3, 51.8, 31.6, 45.1, 40.8, 42.4, 38.9, 42.9, 34.1, 49.0, 41.5, 60.1, 34.6, 50.4, 30.7, 44.1

Now, we would like to test, at a significance level of 0.05, if there is a significant evidence that less than 80% of the batteries will last at least 36 months? Conduct and conclude the test.
Solution: The critical element of solution is that what we are testing here is not the mean product life, rather it is the proportion of items that last at least 36 months. So, begin by counting the product lifetimes (among the given 28 measurements), calculate $\hat{p}$ and proceed straightforwardly with the rest. 8402 This exercise is left as self-study.

2. Right after the poll stations are closed at 17:00, a political candidate receives the information that out of the 50 people interviewed her approval "count" is 24. As a statistics lover, she immediately tests the null hypothesis that her population approval rate is less than or equal 0.50 against its respective alternative, at the 5% level of statistical significance. What is the conclusion of this test? Suppose in every consecutive 15 minutes, number of people interviewed increases by 5 and approval count increases by 4. Find the earliest time, *HH:MM*, that she can declare her victory based on her tests of hypotheses. Note that a formal statistical/algebraic solution is expected with proper terminology and notation.
Solution: This exercise is left as self-study.

8.2 Hypothesis testing: Two populations

In this part, you are more than welcome to transfer your earlier, indeed recently acquired, knowledge to understand things better. Except for one case or two, the material remains fairly intact compared to the ones in confidence intervals for two populations.

8.2.1 Hypothesis testing
Difference between two normal population means
Case: Dependent (matched) samples

Consider:

$$H_0 : \mu_x - \mu_y \leq 0$$
$$H_1 : \mu_x - \mu_y > 0$$

Let $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, be two matched samples.
At the statistical significance level α:

- H_0 is rejected if
$$\frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} > t_{n-1,1-\alpha}$$
- and we fail to reject H_0 otherwise.

For:

$$H_0 : \mu_x - \mu_y \geq 0$$
$$H_1 : \mu_x - \mu_y < 0$$

At the statistical significance level α:

- H_0 is rejected if
$$\frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} < t_{n-1,\alpha}$$
- and we fail to reject H_0 otherwise.

For:

$$H_0 : \mu_x - \mu_y = 0$$
$$H_1 : \mu_x - \mu_y \neq 0$$

At the statistical significance level α:

- H_0 is rejected if
$$\frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} < t_{n-1,\frac{\alpha}{2}} \text{ or } \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} > t_{n-1,1-\frac{\alpha}{2}}$$
- and we fail to reject H_0 otherwise.

8.15 EXERCISES

1. A company is about to release a new drug to assist weight loss, and we are in charge of assessing how effective the drug is. We pick a random sample of 8 people with the following pre-drug body weights:

$$90, 95, 105, 95, 110, 85, 100, 90$$

After using the drug for the designated test duration, the post-drug body weights are measured as:

$$85, 80, 110, 90, 110, 80, 95, 90$$

Conduct and conclude a hypothesis test at the significance level of 5% to assess if 'pre-drug minus post-drug difference of mean body weights is positive'.
Solution:

$$H_0 : \mu_x - \mu_y \leq 0$$
$$H_1 : \mu_x - \mu_y > 0$$

The difference series (pre-drug minus post-drug) is:

$$+5, +15, -5, +5, 0, +5, +5, 0$$

The relevant distribution is t and the degrees of freedom is 7.
Since:

$$\frac{3.75}{\frac{5.825}{\sqrt{8}}} = 1.821 \text{ is not greater than } > 1.895$$

we fail to reject H_0 at $\alpha = 0.05$.

8.2.2 *Hypothesis testing*
Difference between two normal population means
Case: Independent samples & Known population variances

Consider:

$$H_0 : \mu_x - \mu_y \leq 0$$
$$H_1 : \mu_x - \mu_y > 0$$

Let,

$$\{x_i\}_{i=1}^{n_x} \subset \{x_i\}_{i=1}^{N_x} \sim \text{Normal}(\mu_x, \sigma_x^2)$$
$$\{y_i\}_{i=1}^{n_y} \subset \{y_i\}_{i=1}^{N_y} \sim \text{Normal}(\mu_y, \sigma_y^2)$$

where σ_x^2 and σ_y^2 are known.
At the statistical significance level α:

- H_0 is rejected if
$$\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} > z_{1-\alpha}$$
- and we fail to reject H_0 otherwise.

For:

$$H_0 : \mu_x - \mu_y \geq 0$$
$$H_1 : \mu_x - \mu_y < 0$$

At the statistical significance level α:

- H_0 is rejected if
$$\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} < z_{\alpha}$$
- and we fail to reject H_0 otherwise.

For:

$$H_0 : \mu_x - \mu_y = 0$$
$$H_1 : \mu_x - \mu_y \neq 0$$

At the statistical significance level α:

- H_0 is rejected if
$$\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} < z_{\frac{\alpha}{2}} \text{ or } \frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} > z_{1-\frac{\alpha}{2}}$$
- and we fail to reject H_0 otherwise.

8.16 EXERCISES

1. A researcher wonders if the mean wage of workers in Ankara falls short of that in Istanbul. She has the following data and information: Mean wage rate of 49 workers from Ankara is 6000. Mean wage rate of 81 workers from Istanbul is 7000. Population variance of wages in Ankara and Istanbul are known to be 640000 and 810000, respectively. Conduct and conclude a hypothesis test at the significance level of 5% to assess if mean wage rate in Ankara is less than the mean wage rate in Istanbul.
Solution:

$$H_0 : \mu_x - \mu_y \geq 0$$
$$H_1 : \mu_x - \mu_y < 0$$

The relevant distribution is z.
Since:

$$\frac{6000 - 7000}{\sqrt{\frac{640000}{49} + \frac{810000}{81}}} = -6.585 < -1.65$$

we reject H_0 at $\alpha = 0.05$.

8.2.3 *Hypothesis testing*

Difference between two normal population means
Case: Independent samples & Unknown yet equal population variances

Consider:

$$H_0 : \mu_x - \mu_y \leq 0$$
$$H_1 : \mu_x - \mu_y > 0$$

Let,

$$\{x_i\}_{i=1}^{n_x} \subset \{x_i\}_{i=1}^{N_x} \sim \text{Normal}(\mu_x, \sigma_x^2)$$
$$\{y_i\}_{i=1}^{n_y} \subset \{y_i\}_{i=1}^{N_y} \sim \text{Normal}(\mu_y, \sigma_y^2)$$

where σ_x^2 and σ_y^2 are unkown but assumed to be equal.
At the statistical significance level α:

- H_0 is rejected if

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}} > t_{n_x+n_y-2,1-\alpha}$$

- and we fail to reject H_0 otherwise.

In our formulation:

$$s_p^2 = \frac{(n_x - 1)\, s_x^2 + (n_y - 1)\, s_y^2}{n_x + n_y - 2}$$

and,

$$s_x^2 = \frac{\sum_{i=1}^{n_x} (x_i - \bar{x})^2}{n_x - 1}$$
$$s_y^2 = \frac{\sum_{i=1}^{n_y} (y_i - \bar{y})^2}{n_y - 1}$$

For:

$$H_0 : \mu_x - \mu_y \geq 0$$
$$H_1 : \mu_x - \mu_y < 0$$

At the statistical significance level α:

- H_0 is rejected if

$$\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}} < t_{n_x+n_y-2,\alpha}$$

- and we fail to reject H_0 otherwise.

For:

$$H_0 : \mu_x - \mu_y = 0$$
$$H_1 : \mu_x - \mu_y \neq 0$$

At the statistical significance level α:

- H_0 is rejected if

$$\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}} < t_{n_x+n_y-2,\alpha/2} \text{ or } \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}} > t_{n_x+n_y-2,1-\alpha/2}$$

- and we fail to reject H_0 otherwise.

8.17 EXERCISES

1. A researcher wonders if the mean wage of workers in Ankara falls short of that in Istanbul. She has the following data and information: Mean wage rate of 49 workers from Ankara is 6000. Mean wage rate of 81 workers from Istanbul is 7000. Population variances of wages in Ankara and Istanbul are unknown but they are assumed to be equal. Sample variance of wages in Ankara and Istanbul are calculated as 490000 and 640000, respectively. Conduct and conclude a hypothesis test at the significance level of 5% to assess if mean wage rate in Ankara is less than the mean wage rate in Istanbul.

Solution:

$$H_0 : \mu_x - \mu_y \leq 0$$
$$H_1 : \mu_x - \mu_y > 0$$

$$s_p^2 = \frac{(49-1)\,490000 + (81-1)\,640000}{49+81-2}$$

The relevant distribution is t and the degrees of freedom is 128. Since:

$$\frac{6000-7000}{\sqrt{\frac{583750}{49}+\frac{583750}{81}}} = -7.232 < -1.657$$

we reject H_0 at $\alpha = 0.05$.

8.2.4 *Hypothesis testing*

Difference between two normal population means

Case: Independent samples & Unknown and unequal population variances

Consider:

$$H_0 : \mu_x - \mu_y \leq 0$$
$$H_1 : \mu_x - \mu_y > 0$$

Let,

$$\{x_i\}_{i=1}^{n_x} \subset \{x_i\}_{i=1}^{N_x} \sim \text{Normal}(\mu_x, \sigma_x^2)$$
$$\{y_i\}_{i=1}^{n_y} \subset \{y_i\}_{i=1}^{N_y} \sim \text{Normal}(\mu_y, \sigma_y^2)$$

where σ_x^2 and σ_y^2 are unkown and assumed not to be equal.
At the statistical significance level α:

- H_0 is rejected if
$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} > t_{\nu, 1-\alpha}$$
- and we fail to reject H_0 otherwise.

In our formulation:

$$\nu = \frac{\left(\left(\frac{s_x^2}{n_x}\right) + \left(\frac{s_y^2}{n_y}\right)\right)^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y - 1}}$$

Notice that, if $n_x = n_y = n$

$$\nu = \left(1 + \frac{2}{\frac{s_x^2}{s_y^2} + \frac{s_y^2}{s_x^2}}\right)(n - 1)$$

For:

$$H_0 : \mu_x - \mu_y \geq 0$$
$$H_1 : \mu_x - \mu_y < 0$$

At the statistical significance level α:

- H_0 is rejected if
$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} < t_{\nu, \alpha}$$
- and we fail to reject H_0 otherwise.

For:

$$H_0 : \mu_x - \mu_y = 0$$
$$H_1 : \mu_x - \mu_y \neq 0$$

At the statistical significance level α:

- H_0 is rejected if

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} < t_{\nu,\frac{\alpha}{2}} \text{ or } \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} > t_{\nu,1-\frac{\alpha}{2}}$$

- and we fail to reject H_0 otherwise.

8.18 EXERCISES

1. A researcher wonders if the mean wage of workers in Ankara falls short of that in Istanbul. She has the following data and information: Mean wage rate of 49 workers from Ankara is 6000. Mean wage rate of 81 workers from Istanbul is 7000. Population variances of wages in Ankara and Istanbul are unknown and they are assumed to be unequal. Sample variance of wages in Ankara and Istanbul are calculated as 490000 and 640000, respectively. Conduct and conclude a hypothesis test at the significance level of 5% to assess if mean wage rate in Ankara is less than the mean wage rate in Istanbul.
Solution:

$$H_0 : \mu_x - \mu_y \leq 0$$
$$H_1 : \mu_x - \mu_y > 0$$

The relevant distribution is t and the degrees of freedom is ν:

$$\nu = \frac{\left(\left(\frac{490000}{49}\right) + \left(\frac{640000}{81}\right)\right)^2}{\frac{\left(\frac{490000}{49}\right)^2}{49-1} + \frac{\left(\frac{640000}{81}\right)^2}{81-1}} \to 112$$

Since:

$$\frac{6000 - 7000}{\sqrt{\frac{490000}{49} + \frac{640000}{81}}} = -7.474 < -1.659$$

we reject H_0 at $\alpha = 0.05$.

8.2.5 *Hypothesis testing*
Difference between two population proportions

Consider:

$$H_0 : P_x - P_y \leq 0$$
$$H_1 : P_x - P_y > 0$$

Let

$$\{x_i\}_{i=1}^{n_x} \subset \{x_i\}_{i=1}^{n_X} \sim \text{Bernoulli}(P_x)$$
$$\{y_i\}_{i=1}^{n_y} \subset \{y_i\}_{i=1}^{n_Y} \sim \text{Bernoulli}(P_y)$$

At the statistical significance level α:

- H_0 is rejected if

$$\frac{\hat{p}_x - \hat{p}_y}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} > z_{1-\alpha}$$

- and we fail to reject H_0 otherwise.

In our formulation:

$$\hat{p}_0 = \frac{n_x\hat{p}_x + n_y\hat{p}_y}{n_x + n_y}$$

For:

$$H_0 : P_x - P_y \geq 0$$
$$H_1 : P_x - P_y < 0$$

At the statistical significance level α:

- H_0 is rejected if

$$\frac{\hat{p}_x - \hat{p}_y}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} < z_{\alpha}$$

- and we fail to reject H_0 otherwise.

For:

$$H_0 : P_x - P_y = 0$$
$$H_1 : P_x - P_y \neq 0$$

At the statistical significance level α:

- H_0 is rejected if

$$\frac{\hat{p}_x - \hat{p}_y}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} < z_{\frac{\alpha}{2}} \text{ or } \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} > z_{1-\frac{\alpha}{2}}$$

- and we fail to reject H_0 otherwise.

8.19 EXERCISES

1. A political candidate wonders if her support rate in Ankara exceeds that in Istanbul. We know that among 64 people from Ankara 35 supports the candidate and among 81 people from Istanbul 45 supports the candidate. Conduct and conclude the relevant hypothesis test at the significance level of 5%.
Solution:

$$H_0 : P_x - P_y \leq 0$$
$$H_1 : P_x - P_y > 0$$

$$\hat{p}_0 = \frac{64 \cdot 0.547 + 81 \cdot 0.556}{64 + 81} \to 0.552$$

The relevant distribution is z.
Since:

$$\frac{0.547 - 0.556}{\sqrt{\frac{0.552(1-0.552)}{64} + \frac{0.552(1-0.552)}{81}}} \text{ is not greater than } 1.65$$

we fail to reject H_0 at $\alpha = 0.05$.

8.2.6 *Hypothesis testing*
Equality of variances of two normal populations

Consider:

$$H_0 : \sigma_x^2 \leq \sigma_y^2$$
$$H_1 : \sigma_x^2 > \sigma_y^2$$

Let

$$\{x_i\}_{i=1}^{n_x} \subset \{x_i\}_{i=1}^{N_x} \sim \text{Normal}(\mu_x, \sigma_x^2)$$
$$\{y_i\}_{i=1}^{n_y} \subset \{y_i\}_{i=1}^{N_y} \sim \text{Normal}(\mu_y, \sigma_y^2)$$

At the statistical significance level α:

- H_0 is rejected if
$$\frac{s_x^2}{s_y^2} > F_{n_x-1, n_y-1, 1-\alpha}$$
- and we fail to reject H_0 otherwise.

Go over the description of F-distribution in Chapter 10.
In our formulation:

$$s_x^2 = \frac{\sum_{i=1}^{n_x} (x_i - \bar{x})^2}{n_x - 1}$$

$$s_y^2 = \frac{\sum_{i=1}^{n_y} (y_i - \bar{y})^2}{n_y - 1}$$

For:

$$H_0 : \sigma_x^2 \geq \sigma_y^2$$
$$H_1 : \sigma_x^2 < \sigma_y^2$$

At the statistical significance level α:

- H_0 is rejected if
$$\frac{s_x^2}{s_y^2} < F_{n_x-1,n_y-1,\alpha}$$
- and we fail to reject H_0 otherwise.

For:

$$H_0 : \sigma_x^2 = \sigma_y^2$$
$$H_1 : \sigma_x^2 \neq \sigma_y^2$$

At the statistical significance level α:

- H_0 is rejected if
$$\frac{s_x^2}{s_y^2} < F_{n_x-1,n_y-1,\alpha/2} \text{ or } \frac{s_x^2}{s_y^2} > F_{n_x-1,n_y-1,1-\alpha/2}$$
- and we fail to reject H_0 otherwise.

8.20 EXERCISES

1. A process engineer wonders if the temperature variation in Furnace X exceeds that in Furnace Y. Sample variance of temperatures in Furnace X is 1600 on the basis of 10 temperature readings and sample variance of temperatures in Furnace Y is 1100 on the basis of 8 temperature readings. Conduct and conclude the relevant hypothesis test at the significance level of 5%.
Solution:

$$H_0 : \sigma_x^2 \leq \sigma_y^2$$
$$H_1 : \sigma_x^2 > \sigma_y^2$$

The relevant distribution is F with a numerator degrees of freedom of 9 and a denominator degrees of freedom of 7.
Since:

$$\frac{1600}{1100} = 1.455 \text{ is not greater than } 3.677$$

we fail to reject H_0 at $\alpha = 0.05$.

8.21 EXERCISES

1. Consider the hypotheses regarding two normal populations X and Y:

$$H_0 : \sigma_x^2 \leq \sigma_y^2 \; H_1 : \sigma_x^2 > \sigma_y^2$$

Sample values for X and Y are given as follows:

X:	2	8	5	4	3	7	9	6
Y:	26	24	23	25	22	27		

Conduct and conclude the test at $\alpha = 0.05$. Clearly state the test statistic, the distribution of test statistic and critical value(s). Find the necessary critical values from the end of your textbook or from Internet sources.

Solution: This exercise is left as self-study.

2. Consider two populations X and Y for which a researcher has estimated the following confidence intervals given that $\bar{x} = 150$ and $\bar{y} = 250$.

$$\mathrm{P}(\mu_x \in [100, \infty)) = 0.90$$

$$\mathrm{P}\big(\mu_y \in [-\infty, 400)\big) = 0.95$$

In her research report, the researcher noted that **she used an** Normal$(0, 1)$ **distribution in her calculations**. Based on these, calculate a 90% confidence interval for $\mu_x - \mu_y$

Solution: This exercise requires some little portion of creative thinking. As the researcher has wed the standard normal distribution in her calculations, this means σ_x^2 and σ_y^2 are both known (or given). As the given confidence intervals 'for μ_x and μ_y are one-sided, the critical values are -1.29 and 1.65, respectively. So,

$$\frac{150-100}{1.29} = \frac{\sigma_x}{\sqrt{n_x}}$$
$$\frac{400-250}{1.65} = \frac{\sigma_y}{\sqrt{n_y}}$$

Once these are known, estimation of a 90% C.I. for $\mu_x - \mu_y$ is straightforward.

8.3 p-value

p-value is defined as the **tail probability** of a test statistic. While conducting hypothesis tests manually, *i.e.* with a pencil on paper, use of a *p-value* is not essential, since calculation of *p-value* already requires a calculated test statistic. In some cases, we may need to do so, though. A *p-value* is especially practical when we do our analysis on a computer using a dedicated software. The rule is simple:

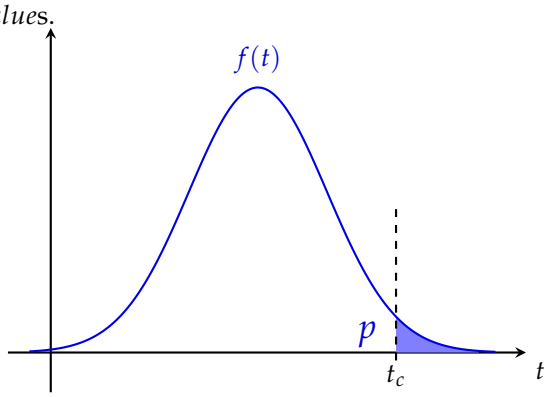
$$H_0 \text{ is rejected if } p - value < \alpha$$

In a nutshell

To calculate the *p-value* of a test statistic:

1. Look at H_1 to see the rejection side & calculate the test statistic
2. Calculate the tail probability according to the distribution of the test statistic. If the rejection side is the right tail, calculate the right tail probability. If the rejection side is the left tail, calculate the left tail probability. If the rejection side is both the right and left tails, calculate the tail probability at either side; multiply the result by 2

A course-related/pedagogical warning about the p-value is that, my expectation (from students) is to see the proper use of 'test score vs critical value' comparisons in concluding hypothesis tests rather than $p - value$ vs α' comparisons, unless otherwise stated. In your future/professional practice you will have full freedom to enjoy $p - values$.



8.4 *Type I and Type II errors and the Power of a hypothesis test*

Despite *Power* is not a difficult concept to grasp intuitively, its mathematics is often confusing to students. Patiently go over the following:

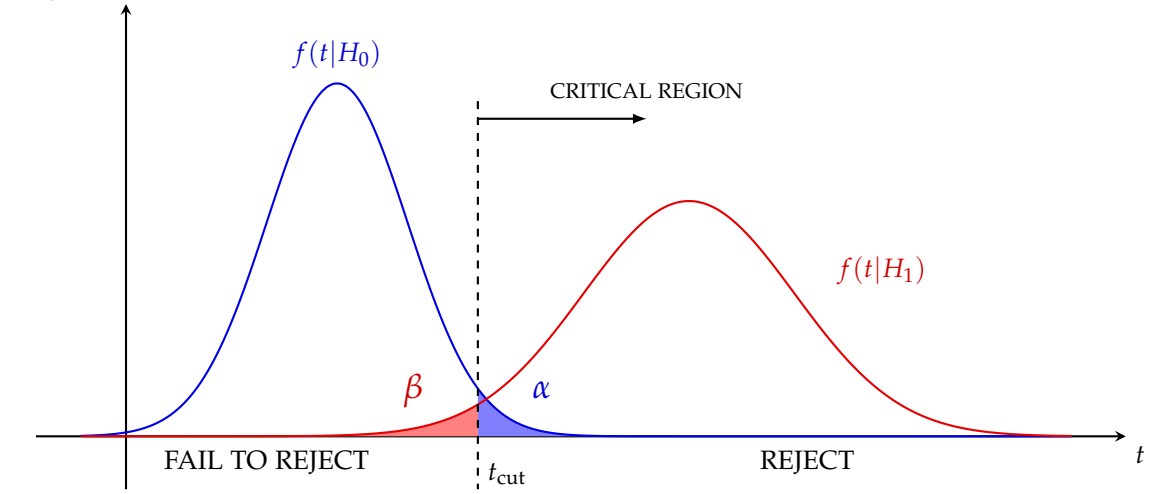
As you may recall, a Type I error occurs if the researcher rejects the null hypothesis in favor of the alternative hypothesis when, in fact, the null hypothesis is true. The probability of Type I error is denoted as α. A Type II error, on the other hand, occurs if the researcher fails to reject the null hypothesis when, in fact, the null hypothesis is false. The probability of Type II error is denoted as β. So,

$$\begin{aligned} \text{P(Reject } H_0 \mid H_0 \text{ true)} &= \alpha \\ \text{P(Fail to reject } H_0 \mid H_0 \text{ true)} &= 1 - \alpha \\ \text{P(Fail to reject } H_0 \mid H_0 \text{ false)} &= \beta \end{aligned}$$

$$\begin{aligned} \text{Power} = 1 - \beta &= \text{P(Fail to reject } H_0 \mid H_0 \text{ false)} \\ &= \text{P(Reject } H_0 \mid H_0 \text{ false)} \end{aligned}$$

So, in plain language, *Power* is the ability of a test to avoid a false null hypothesis.

(As a caution, note that there is no requirement of any sort like $\alpha + \beta = 1$)



In a nutshell

To calculate the *Power* of a test:

1. Determine the rejection condition and area in terms of the null distribution
2. Express the rejection condition in terms of the parameter concerned in the your hypotheses
3. Determine the alternative distribution by picking a value for the parameter concerned, in line with the alternative hypothesis
4. Calculate the probability of the rejection area according to the alternative distribution
5. Now, you have the *Power*.

As you may pick infinitely many alternative values for your parameter of interest, there is a multiplicity of values for *Power*. So, *Power*, is indeed a function. We often write it as a function of the difference between the alternative and hypothesized parameter values.

In a nutshell

Ceteris Paribus in each case:

- If $\mu_1 - \mu_0$ increases, *Power* increases
- If α decreases, *Power* decreases
- If σ^2 increases, *Power* decreases
- If n increases, *Power* increases
- Fun fact: *Power* is 0.50 at the critical value

As a closing remark, drawing graphs (rather than calculating) may be very useful to understand the *Type II error* as well as *Power*.

In a nutshell

The classical theory of hypothesis testing, known as the **Neyman-Pearson theory**, fixes $\alpha = \text{P(Type I error)} \leq$ a constant and minimizes β or maximizes $1 - \beta$. $(1 - \beta)$ is known as the Power of the test under the alternative hypothesis.

The Neyman-Pearson lemma: If C is a critical region of size α and k is a constant such that $(L_0/L_1) \leq k$ inside C and $(L_0/L_1) \geq k$ outside C, then C is a most powerful critical region of size α for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Here,

- Likelihood has to be completely specified under the null and alternative hypotheses. So, this lemma applies only to testing a simple hypothesis versus another simple hypothesis.
- L_0 is the likelihood under H_0 and L_1 is the corresponding likelihood function under H_1.

8.22 EXERCISES

1. In a two-sided (two-tailed) hypothesis test, the test statistic was calculated as 0.18. We know that the distribution of the test statistic (call this A distribution) has the triangular shaped union of the line segments $[AB]$ and $[BC]$, given $A(0.00, 0.00)$, $B(2.00, 0.50)$ and

$C(4.00, 0.00)$. Conclude the test at $\alpha = 0.005$ by calculating and using p-values only. In your answer, clearly define what p-value is.
Solution: This exercise is left as self-study.

2. Referring to $H_0 : \mu = 0$ against $H_1 : \mu > 0$ and using proper drawings of the relevant distributions, demonstrate that

i. Power of a hypothesis test gets higher as the sample size gets larger

ii. Power of a hypothesis test gets higher as population variance gets smaller

Make sure your drawings are clear and well-explained.
Solution: This exercise is left as self-study.

3. Consider a large box which contains many white (W) and black (B) balls. We have forgotten the percentage of white balls in the box, but remember that it is either $\frac{1}{3}$ or $\frac{2}{3}$. Even though we do not know the percentage of white balls in the box we strongly believe that it is $\frac{1}{3}$ (but still believe that it might be $\frac{2}{3}$). Hence we decide to test if the percentage of white balls in the box is $\frac{2}{3}$. For this purpose we draw 20 balls at random with replacement and note their color.

i. What are the hypotheses of this test?

ii. If we decide to use the number of white balls as our test statistic, what is the distribution of the test statistic?

iii. What is the decision rule?

iv. If the sample you observed was:

$$W\ W\ W\ B\ B\ W\ B\ W\ B\ W\ B\ B\ W\ B\ W\ B\ W\ W\ W\ W$$

what would your conclusion be?

v. What is the *p-value* corresponding to the above sample?

vi. What is the probability of a *Type I error* and the probability of a *Type II error*?

Solution: This exercise is left as self-study.

4. In the investigation of the average performance of produced kettles, a quality control engineer examines 49 kettles and measures the mean time to heat 1 liter of water from $25 \circ C$ to $100 \circ C$ as $75 seconds$. Knowing that this had a historical variance of $100 seconds^2$, he wants to test at $\alpha = 0.05$ whether the population mean time is equal to $60 seconds$ or not, as the producer's advertisements say "1 liter in 1 minute". Help him to correct the mistakes in his statistical test report.

Statistical test report
Engineer's name: Dumb Dumber Jr.
Date of test: 05 January 2019
Hypotheses:

$$H_0 = 60$$
$$H_1 < 60 \quad \textit{(He has a problem here ...)}$$

Since historical variance is known, I'll use t-distribution in my assessment. *(... and here)* The test statistic is:

$$t = \frac{\bar{x} - \mu_0}{s} \quad \textit{(... and here)}$$

and its value is:

$$t = \frac{75 - 60}{100} = 0.15$$

Lower critical t-value is:

$$t_{48,0.05} = -1.684$$

Since $0.15 > -1.684$, I reject $H_0 = 60$ *(...and here)* Finally, the mean time to boil water for our heaters is less than 60 seconds with some 95% confidence, as promised in the advertisements. *(... and of course, here.)*

Solution: The hypotheses involved should be written as:

$$H_0 : \mu = 60$$
$$H_1 : \mu > 60$$

As the historical variance of temperatures is known to be 100, the researcher should use:

$$z = \frac{75 - 60}{10} = 1.5$$

where the upper-critical z-value is 1.65 in this one-sided test. Since $1.5 < 1.65$, we fail to reject H_0. The mean time to boil water is not longer than 60 seconds, as promised in the advertisements.

5. A researcher investigates whether two different teaching methods yield similar impacts on learning of students. After Method 1 is used in Section 1 and Method 2 is used in Section 2 of the same course, the same final exam is given to both sections. Then the researcher forms a 95% confidence interval as $[9.82, 17.30]$ for the difference of exam grades (Section 1 grade minus Section 2 grade).

Can you analyze whether there is a difference of 15*points* between the grades of two sections?
Solution: This exercise is left as self-study.

6. The choice of confidence level $(1-\alpha)$ for statistical practices depend on the scientific/technical discipline. Referring to an economist/financial analyst (performing portfolio analysis), a computer scientist (designing and coding national payment systems), an international relations specialist (trying to avoid nuclear conflicts) and a physicist working for the CERN (searching for a very rare subatomic particle), explain how the confidence level must be chosen.
Solution: This exercise is left as self-study.

7. We have the following information:

- Researcher A tests $H_1 : \sigma^2 > a$ against $H_0 : \sigma^2 \leq a$ at $\alpha = 0.05$ and she uses in her report the critical value of c_1 to conduct and conclude the test, using a sample of size n_1.
- Researcher B tests $H_1 : \sigma^2 \neq b$ against $H_0 : \sigma^2 = b$ at $\alpha = 0.10$ and she uses in her report the critical values of d_1 and d_2, where $d_1 < d_2$, to conduct and conclude the test, using a sample of size n_2.
- Researcher C tries to test $H_1 : \sigma_X^2 > \sigma_Y^2$ against $H_0 : \sigma_X^2 \leq \sigma_Y^2$ at $\alpha = 0.05$, using n_2 observations of X and n_1 observations of Y. Unfortunately, he only has his own data of X and Y as well as the research reports of Researcher A and Researcher B, but he does not have a computer or any statistical tables.

Help him to find the critical value needed.
Solution: This exercise is left as self-study.

In a nutshell

A t random variable with m degrees of freedom, denoted $t_{(m)}$ is found by:

$$t = \frac{Z}{\sqrt{\frac{\chi^2_{(m)}}{m}}} \sim t_{(m)}$$

if the numerator and denominator are independent random variables. Here, consider the meaning of:

$$\frac{\chi^2_{(m)}}{m}$$

Previously we said χ^2 should be something related to variance. Based on the definition of $\chi^2_{(m)}$, do you think the fraction above is the variance of **something**? Reveal what this something is.

In a nutshell

$V_1 \sim \chi^2_{m_1}$ and $V_2 \sim \chi^2_{m_2}$ being two *chi-squared* random variables:

$$F = \frac{\frac{V_1}{m_1}}{\frac{V_2}{m_2}} \sim F_{(m_1,m_2)}$$

Can we say that F random variable handles the difference between two variances? Recall:

$$t_{(m)} = \frac{Z}{\sqrt{\frac{\chi^2_{(m)}}{m}}}$$

Now, square both sides of the expression:

$$\begin{aligned} t^2_{(m)} &= \frac{Z^2}{\frac{\chi^2_{(m)}}{m}} \\ &\to \frac{\chi^2_1}{\frac{\chi^2_m}{m}} \\ &\to \frac{\frac{\chi^2_1}{1}}{\frac{\chi^2_m}{m}} \\ &\sim F_{1,m} \end{aligned}$$

$F_{1,m}$ has a numerator degrees of freedom of 1 and denominator degrees of freedom of m.

In a nutshell

Just to joyfully remember:

- Sum of *Bernoulli*s is *Binomial*
- *Binomial* with a single trial is *Bernoulli*
- *Binomial* with an infinitely many trials and infinitesimal probability of success is *Poisson*
- Sum of *Poisson*s is another *Poisson*
- *Binomial* with a variance greater than 5 and infinite trials is *Normal*
- *Normal* with a zero mean and variance of unity is *Standard Normal*
- Sum of squares of *Standard Normal*s is *Chi-Squared*
- *Standard Normal* divided by the square root of a "*Chi-Squared* corrected for its degrees of freedom" is *t*
- Ratio of two *Chi-Squared*s each corrected for its degrees of freedom is *F*
- Square of a *t* is *F*

9 *Linear regression analysis*

In the previous chapters that served the whole ECON 221 and about a half of ECON 222, we studied the fundamentals of Probability theory and the key theory and toolset of Statistical inference. Remember that we focused solely on understanding statistical distributions and estimating the distributional parameters. In your future scientific, technical, professional practice, this body of knowledge will be quite fruitful.

Now, we are ready to study the theoretical background and applied dimensions of the 'curve-fitting' problem. To this end, in this chapter, we will consider the Linear Regression Models. Notice that what we will do here accounts for the first half of a traditionally designed 'Introductory Econometrics' course.

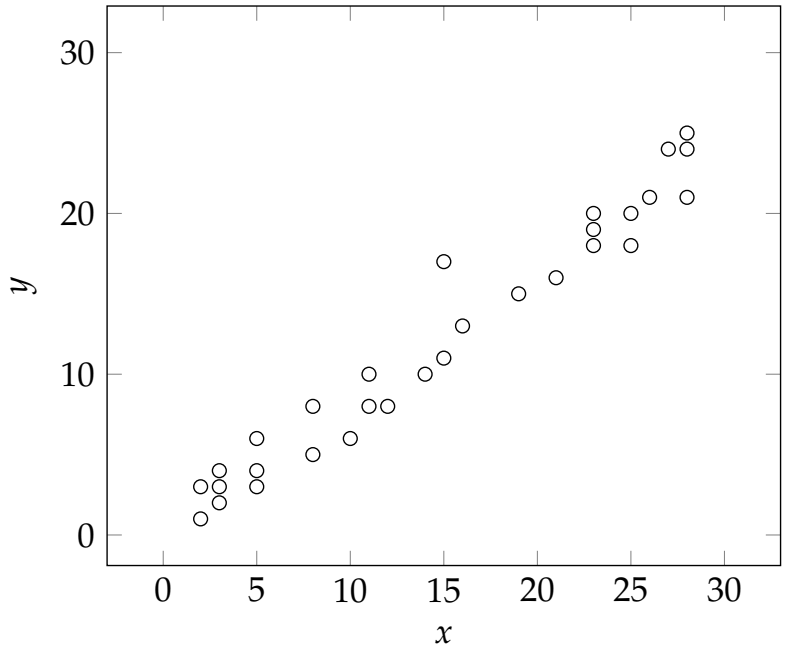
The term regression was coined by Francis Galton to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average, which is also known as regression toward the mean. For Galton, regression had only this biological meaning. His work was later extended by Udny Yule and Karl Pearson, and later by Fisher (in a way to come closer to Gauss's 1821 formulation of the problem). Once you have researched it, you will enjoy the history of this certain line of research.

As being the main pillar of it, the 'regression analysis' takes us to the rich analytical world of Econometrics. The literal meaning of the term econometrics (econo+metrics) is 'measurement in economics'. Econometrics is 'the branch of economics concerned with the use of statistical methods in describing and quantifying economic systems' (Oxford Dictionary). From a broader perspective, econometrics is a shared sub-field of Statistics (hence of Mathematics) and Economics. In that, our tools in Econometrics are those tools in Statistics as shaped and augmented by our knowledge of Economics. (One of the founders of the Econometrics Society, another pioneer of the field, Ragnar Frisch is credited with coining the term 'econometrics'.)

Renowned academic Badi Baltagi says "An econometrician has to be a competent mathematician and statistician who is an economist

by training. Fundamental knowledge of mathematics, statistics and economic theory are a necessary prerequisite for this field".

Our starting point is a scientific urge to find/formulate, measure and test the relationship between, say, two variables y and x. These variables may belong to natural sciences, social sciences or even to humanities; this is not something to mind. What matters often is that the linkage between our variables may not be (mostly is not) a perfect relationship like $y = mx + n$ (we prefer indeed a notation like $y = \beta_0 + \beta_1 x$, where n is β_0 and m is β_1). We rather observe there are deviations from a perfect relationship, as seen earlier in ECON 221. In that, actual y values are connected to actual x values through a relationship like $y = \beta_0 + \beta_1 x + e$ where e stands for a sequence of statistical errors (disturbances).



The error sequence e may stem from the random actions/choices of humans, unexpected shocks to socio-economic systems, misspecification of models, improper choices of mathematical functional forms or imprecision of the data. Note that, this picture is not specific to social sciences: in the natural science experiments there is a multiplicity of sources of uncertainty (hence of statistical errors or disturbances).

Goals of econometrics, as we understand it, will be (1) to find the relation between variables y and x, encapsulated in (β_0, β_1), (2) to validate and quantify theory and (3) forecasting.

Purpose of modeling and Simplicity

Deferring a detailed discussion of it to class gatherings, we will say here that 'a model is a downsized yet realistic representation of reality'. An immediate analogy from architecture would be useful: on an architectural model of a building we see things 'only as needed'. While we may not see the doorknobs (depending on the scale) on a model, we see the proportionality of distances clearly. After all, the purpose of the model is to give a broad yet accurate idea of/about things.

A similar idea applies in the other disciplines. In business models we do not see every tiny detail of the workplace or the manufacturing environment. In economic models we tend not to include all potential explanatory variables at once. We just try to remain 'accurate enough'.

Using our models we can present our scientific grasp of the nature or universe or the society. Once the model is well-parametrized and quantified, we can develop forecasts of the future, or we can (depending on the type of our model) develop counter factual and/or scenario analyses. Presenting a scientific view of ours and forecasting the future (while) are fairly pragmatic ends, a third use of a scientific model helps testing, validating/ invalidating theories, which calls for a more than pragmatic spirit. Regardless of the purposes cited, though, a model (any model) should display: a certain level of simplicity. Before proceeding, recall Albert Einstein saying "Everything should be made as simple as possible, but no simpler."

In our practice of statistical/ econometric modeling, the 'principle of parsimony' guides us. Equipped with a rich toolset of formal statistical tests and her judgmental skills, a good researcher tries to come up with an "as simple as possible but no simpler" model. Common sense says essentials will be included in while all the inessentials will be omitted from a model. Bad news is every researcher's practice has a couple bumps as to improving a sense of such in practice. Good news is honest and hard work pays back.

In Philosophy (and Science) there are several 'razors' to shave away the redundancies in models (or in scientific explanations). Here we will maintain the Occam's razor (or Ockham's razor) attributed to william of Ockham, an English philosopher of the 13th-14th centuries. Occam's razor is a principle of parsimony stating that among the explanations addressing the same thing, the simplest is to be picked! (William of Baskerville of the Name of the Rose by Umberto Eco is a tribute to William of Ockham) (Arthur Conan Doyle's Sherlock Holmes once utters "When you have eliminated the impossible, whatever remains, however improbable, must be the truth")

Occam's razor reads in Latin as "*pluralitas non est ponenda sine necessitate*" which translates into English as "plurality should not be posited

without necessity". The principle, so, calls for parsimony in 'deductive thinking'.

Despite what we do in applied statistics/econometrics is not purely (maybe not at all) deductive thinking, we rather try to reach an inference to the best explanation via a formal sequence of estimations/tests/calculations. In this practice, Occam's razor sheds some good light for us to see things clearly.

In the world of project development you may hear the same principle as an acronym of 'KISS'. Referring to a model, KISS reads as 'Keep It Small and Simple' or sometimes as 'Keep It Simple, Stupid'. (Search yourself for its relevance to the US Navy)

In the remainder of this chapter, we will study/learn the theory of elementary econometrics, a rich enough toolset pertaining to it along with a selection of applied problems.

9.1 EXERCISES

1. Refer to our in-class discussions to explain/discuss the following:

i. Occam's razor

ii. Principle of parsimony

iii. 'Keep It Small and Simple', i.e., KISS

iv. Purpose of modeling

v. Come up with a synthesis of the terms/phrases referred to above.

Solution: Left as self-exercise.

9.1 Overview of linear models

Overview of linear models

The specific meaning of linearity here is 'the linearity of a model in terms of (with respect to) its parameters. In that:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

is a linear model. So is

$$y = \beta_0 + \beta_1 x_1^2 + \beta_2 x_2^3 + e$$

However,

$$y = \beta_0 + \beta_1^2 x_1 + \beta_1 \beta_3 x_2 + \beta_3 x_3 + e$$

is not considered to be a linear model. Neither is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_1 \beta_2 x_3 + e.$$

In your future practice, you will be able to settle this issue in a crystal clear fashion.

Why do we resort to linear models? This is a very legitimate question once we observe a number of relationships in the nature and in societal life are, indeed, nonlinear (not linear). A straightfor ward answer reads as 'linear models are easy to use'. So, simplicity matters. Simplicity brings practicality to researchers, they are easy to compute, to interpret and to communicate. More importantly, as noted earlier, our linear regression models are linear with respect to their parameters while the independent variables of our models can be of any nonlinear form. All in all, one can establish/ form models that are nonlinear in their variables' using 'models that are linear in their parameters'. The good thing about models that are linear in parameters is that such a structure allows us to use the tools of linear algebra effectively in our computations.

Our curious nature often forces us to include many explanatory variables in a model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

However, a minimalist design is also possible:

$$y = \beta_0 + \beta_1 x$$

Even this may be a good enough model (think when):

$$y = \beta_0$$

The process of inference begins with the specification of an economic model. Then a statistical model describes the sampling process that we visualize was used to produce the sample data. See the structure below:

Economic model:

$$y = \beta_0 + \beta_1 x$$

Statistical model:

$$y = \beta_0 + \beta_1 x + e$$

The random error term (e) serves three main purposes:

1. e captures the combined effect of all other influences other than x. These other effects are assumed to be unobservable, otherwise they would be included in the model.

2. e captures any approximation error that arises because of the linear functional form

3. e captures any element of random behavior present in each individual observation.

See the structures below:
Case 1: Unconditional model of mean
Economic model:

$$y = \beta_0$$

Statistical model:

$$y = \beta_0 + e$$

Case 2: Simple Linear model
Economic model:

$$y = \beta_0 + \beta_1 x$$

Statistical model:

$$y = \beta_0 + \beta_1 x + e$$

Case 3: Multiple Linear model
Economic model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Statistical model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

9.2 *Transformations and functional forms*

Transformations and functional forms In economics and finance, like in other quantitative disciplines, we attribute a great deal of importance to measuring the impact of a change in one variable on one another. Considering $y = f(x)$ as a relationship between the variables y (dependent) and x (independent), the derivative $dy/dx = f'(x)$ describes that impact. When we consider $y = f(x_1, x_2, \ldots, x_k)$, the impact of an independent variable x_i on the dependent variable y is better described by the partial derivative $\partial y/\partial x_i$. Having formed and estimated a proper statistical/ econometric model, then, a researcher gains a good grasp of issues embedded in the research problem at hand.

Note that, as economists and finance specialists, we like to learn about a special class of impact measurements, namely the elasticities. Recall from your introductory economics classes that 'elasticity of y with respect to x is the percentage change in y against a one percentage point change in x '. in formal terms:

$$n_{y,x} = \frac{\%\Delta y}{\%\Delta x} = \frac{\frac{\Delta y}{y}}{\frac{\Delta x}{x}} = \frac{\Delta y}{\Delta x} \cdot \frac{x}{y}$$

So, as long as we can estimate $\Delta y/\Delta x$, we can come up with an estimate of $\eta_{y,x}$ by substituting appropriate values of x and y into x/y. We will see several examples as we progress through this chapter, where we will see estimating an elasticity is possible under a wide array of functional forms of $f(\cdot)$ in the expression $y = f(x)$.

One of the functional forms, i.e., the Log-Log form, yields elasticities directly as:

$$\eta_{y,x} = \frac{\Delta \ln y}{\Delta \ln x}.$$

We will discuss this topic further in our classes.

In a nutshell

Functional form: Linear

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Nonlinear form:

$$None$$

Impact at margin:

$$\frac{dy}{dx} = \beta_1$$

Elasticity:

$$\frac{dy}{dx} \cdot \frac{x}{y} = \beta_1 \frac{x}{y}$$

In a nutshell

Functional form: Reciprocal

$$y_i = \beta_0 + \beta_1 \frac{1}{x_j} + e_i$$

Nonlinear form:

$$None$$

Impact at margin:

$$\frac{dy}{dx} = -\beta_1 \frac{1}{x^2}$$

Elasticity:

$$\frac{dy}{dx} \cdot \frac{x}{y} = -\beta_1 \cdot \frac{1}{xy}$$

In a nutshell

Functional form: log − log

$$\ln y_i = \beta_0 + \beta_1 \ln x_i + \ell_i$$

Nonlinear form:

$$y_i = \alpha x_i^{\beta_1} e^{e_i}$$

Impact at margin:

$$\frac{dy}{dx} = \beta_1 \frac{y}{x}$$

Elasticity:

$$\frac{dy}{dx} \cdot \frac{x}{y} = \beta_1$$

In a nutshell

Functional form: Log-Linear (exponential)

$$\ln y_i = \beta_0 + \beta_1 x_i + e_i$$

Nonlinear form:

$$y_i = e^{\beta_0 + \beta_1 x_i + e_i}$$

Impact at margin:

$$\beta_1 y = \frac{dy}{dx}$$

Elasticity:

$$\frac{dy}{dx} \cdot \frac{x}{y} = \beta_1 x$$

In a nutshell

Functional form: Linear log(semilog)

$$y_i = \beta_0 + \beta_1 \ln x_i + e_i$$

Nonlinear form:

$$e^{y_i} = e^{\beta_0 + e_i} x_i^{\beta_1}$$

Impact at margin:

$$\frac{dy}{dx} = \beta_1 \frac{1}{x}$$

Elasticity:

$$\frac{dy}{dx} \cdot \frac{x}{y} = \beta_1 \frac{1}{y}$$

In a nutshell

Functional form: Log-Inverse

$$\ln y_i = \beta_0 - \beta_1 \frac{1}{x_i^i} + e_i$$

Nonlinear form:

$$y_i = e^{\beta_0 - \beta_1 \frac{1}{x_i} + e_i}$$

Impact at margin:

$$\frac{dy}{dx} = \beta_1 \frac{y}{x^2}$$

Elasticity:

$$\frac{dy}{dx} \frac{x}{y} = \beta_1 \frac{1}{x}$$

9.3 *Our approach to teaching/learning*

In the remainder of this chapter, we will maintain an approach which may slightly differ from the approaches of others. Sticking to this approach would facilitate better learning. Our approach folds out as:

- Depiction of an unconditional model of mean and the mechanics of estimation (without inference)
- Depiction of a Simple Linear Regression model and the mechanics of estimation (without inference)

- Depiction of a Multiple Linear Regression model and the mechanics of estimation (without inference)
- Goodness of fit of a model measured via R^2
- Handling statistical uncertainty: calculation of variances and covariances associated with a Multiple Linear Regression model
- Statistical inference

Having a pitstop here, the sequence of topics above will provide us with a solid understanding of the mechanical workings of our linear regression universe.

Once we have learned these, we will move to:

- Ideal econometric conditions: Gauss-Markov assumptions

In many, maybe all, books Gauss-Markov assumptions are covered before other things. Though, our approach maintains a different pedagogical perspective. In that, we take into consideration the Gauss-Markov assumptions, which are crucial in econometric theory and practice, upon a clear view of the working environment. After that we will move to:

- Model specification
- Regression analysis at work

Note that the above order of topics require us to stick to it without interruption or gaps for successful learning.

An artificial data set:

In our subsequent discussions we will be referring to the following data set frequently. While we can show a data set as an actual set (with proper mathematical notation) like:

$$\begin{array}{llllll} \text{A}=\{(2,1), & (2,3), & (3,2), & (3,3), & (3,4), & (5,3), \\ (5,4), & (5,6), & (8,5), & (8,8), & (10,6), & (11,8), \\ (11,10), & (12,8), & (14,10), & (15,11), & (15,17), & (16,13), \\ (19,15), & (21,16), & (23,18), & (23,19), & (23,20), & (25,18), \\ (25,20), & (26,21), & (27,24), & (28,21), & (28,24), & (28,25)\} \end{array}$$

it may be more practical to use a tabular listing of the data. A tabular structure improves visibility and exposition:

Observation i	x_i	y_i	Observation i	x_i	y_i
1	2	1	16	15	11
2	2	3	17	15	17
3	3	2	18	16	13
4	3	3	19	19	15
5	3	4	20	21	16
6	5	3	21	23	18
7	5	4	22	23	19
8	5	6	23	23	20
9	8	5	24	25	18
10	8	8	25	25	20
11	10	6	26	26	21
12	11	8	27	27	24
13	11	10	28	28	21
14	12	8	29	28	24
15	14	10	30	28	25

9.4 *Building and estimating an Unconditional Model of Mean: A model which is a non-model*

Consider a variable y that is modeled as:

$$y = \beta_0 + e$$

If we have a sample $y_1, y_2, \ldots, y_n$, this relationship can also be written as

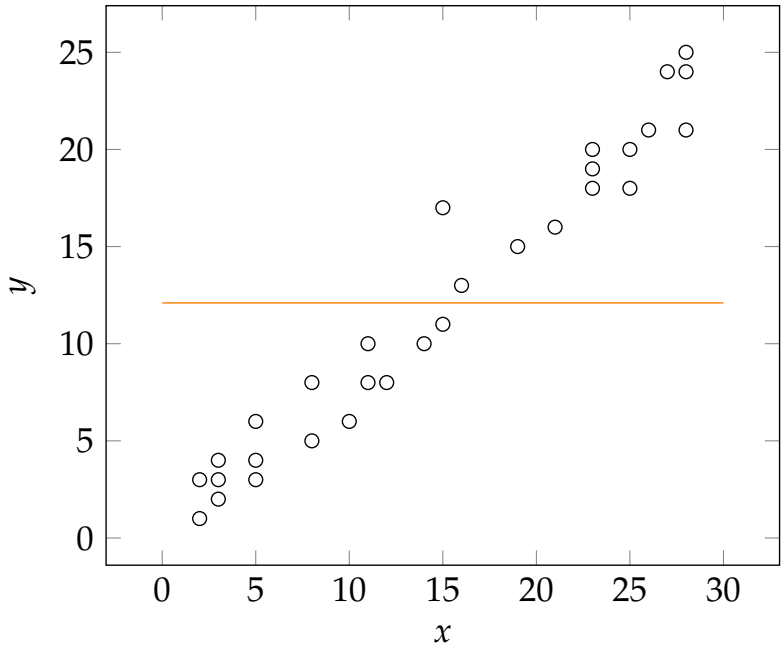
$$y_i = \beta_0 + e_i, i = 1, 2, \ldots, n$$

It is clear that our model does not include any independent (explanatory) variables on the right hand side, ie., values of y are scattered around β_0 (if they are not all accidentally equal to β_0).

Supposing there are K potential independent variables, $x_1, x_2, \ldots, x_k$, that might explain y, the unconditional model of mean can be viewed as:

$$y_i = \beta_0 + 0 \cdot x_{1i} + 0 \cdot x_{2i} + \cdots + 0 \cdot x_{ki} + e_i$$

where the researcher places zero weight on $x_1, x_2, \ldots, x_k$. In that, this model of mean turns out to be the simplest possible model or more like a non-model. When we plot y_i against one of the x's (say x_{ki}), the model of mean is to appear as a horizontal line (as the model disregards x 's). This is simply the orange line displayed below (observe that across the orange line $dy/dx = 0$:



To estimate β_0 in $y = \beta_0 + e$ we need two main ingredients:

- Data on y, a set of n observations $y_1, y_2, \ldots, y_n$ collected from the population $y_1, y_2, \ldots, y_N$. Our n observations as a whole is called a sample; recall from our discussions in earlier chapters that the sample should be randomly picked and large enough
- A formula to compute the desired numerical result; recall that this formula is called an estimator and the numerical result it yields is called an estimate, here $\hat{\beta}_0$. Note that we need a method (rule, criterion) to derive our estimator (formula). Here, we will use 'Least squares' as our method.

Now, suppose out estimator is $\hat{\beta}_0$. Then, the estimated values of y_i (denoted as $\hat{y}_i$) are written as:

$$\hat{y}_i = \hat{\beta}_0$$

Actual values of y_i, on the other hand, are:

$$y_i = \hat{\beta}_0 + \hat{e}_i$$

equivalently

$$y_i = \hat{y}_i + \hat{e}_i$$

The difference between y_i and $\hat{y}_i$ are the error terms:

$$\hat{e}_i = y_i - \hat{y}_i$$
$$\hat{e}_i = y_i - \hat{\beta}_0$$

Consider the function S :

$$S = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0\right)^2$$

The Least Squares method instructs us to minimize S by optimally choosing $\hat{\beta}_0$:

$$\min_{\{\hat{\beta}_0\}} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0\right)^2$$

The F.O.C. for this problem is:

$$\frac{dS}{d\hat{\beta}_0} = \sum_{i=1}^{n} 2\left(y_i - \hat{\beta}_0\right)(-1) = 0$$

which is followed by:

$$\begin{aligned}
&\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0\right) = 0 \\
&\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 = 0 \\
&\sum_{i=1}^{n} y_i - n\hat{\beta}_0 = 0
\end{aligned}$$

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$$

So, not surprisingly and as maybe called from our discussion of point estimators, the sample mean is the estimator of population mean. Namely, $\hat{\beta}_0 = \bar{y}$ estimates y.

A note on the function S may be useful here: As $\hat{\beta}_0$ is the estimated mean of y_i, the function S shows the variance of error terms multiplied by n. This is good to keep in mind: the least squares estimator is a 'minimum variance estimator' as we will formally discuss later. Statistical properties of the error terms e_i will also be covered in detail.

Returning to qualities of the sample mean $\hat{\beta}_0 = \bar{y}$ as an estimator of population mean β_0, one can be intellectually stunned by the beauty generated by simplicity. There are couple things to mention:

- Representing a variable y with its unconditional mean (β_0) is a meaningful alternative only when there is no good explanatory variables $(x_1, x_2, \ldots, x_k)$ to model y

- In that the unconditional model of mean simply provides us with a descriptive statistic

- Still, the unconditional model of mean is very valuable to us as a 'non-model'. This is the model when no explanatory variables work and we use this model as a benchmark in assessing the statistical significance of other (nonempty) models in the subsequent sections.

9.5 *Building and estimating a Simple Linear Regression model*

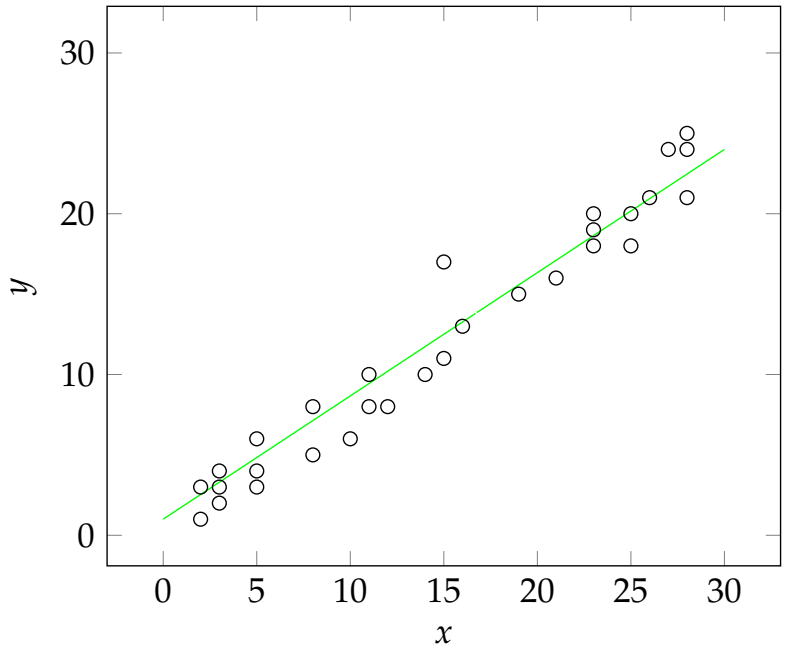
Consider a variable y which we believe is explained by another variable x via a linear relationship like:

$$y = \beta_0 + \beta_1 x + e$$

in this expression,

- β_0 stands for the autonomous / unconditional component of y
- $\beta_1 x$ stands for the part of y attributable to x; depending on the sign of β_1, an increase in x may induce an increase or a decrease in y
- A case of $\beta_1 = 0$ corresponds to our unconditional model of mean

Below, the green line is a good candidate to be a Simple Linear regression line:



Notice that we need to estimate two parameters β_0 and β_1 this time. The Least squares method is again applicable. Let us go over its steps below:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i$$
$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$S = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

$$\min_{\left\{\hat{\beta}_0, \hat{\beta}_1\right\}} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

$$\frac{\partial S}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(-1) = 0$$
$$\frac{\partial S}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(-x_i) = 0$$

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$
$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0$$

$$\sum y_i - n\hat{\beta}_0 - \left(\sum x_i\right)\hat{\beta}_1 = 0$$
$$\sum x_i y_i - \left(\sum x_i\right)\hat{\beta}_0 - \left(\sum x_i^2\right)\hat{\beta}_1 = 0$$

$$n\hat{\beta}_0 + \left(\sum x_i\right)\hat{\beta}_1 = \sum y_i$$
$$\left(\sum x_i\right)\hat{\beta}_0 + \left(\sum x_i^2\right)\hat{\beta}_1 = \sum x_i y_i$$

$$\frac{n\sum x_i^2}{\sum x_i}\hat{\beta}_0 + \left(\sum x_i^2\right)\hat{\beta}_1 = \frac{\sum x_i^2 \sum y_i}{\sum x_i}$$
$$\left(\sum x_i\right)\hat{\beta}_0 + \left(\sum x_i^2\right)\hat{\beta}_1 = \sum x_i y_i$$

$$\left(\frac{n\sum x_i^2}{\sum x_i} - \sum x_i\right)\hat{\beta}_0 = \frac{\sum x_i^2 \sum y_i}{\sum x_i} - \sum x_i y_i$$
$$\left(\frac{n\sum x_i^2 - \left(\sum x_i\right)^2}{\sum x_i}\right)\hat{\beta}_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{\sum x_i}$$

$$\hat{\beta}_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$\sum y_i - n\hat{\beta}_0 - \left(\sum x_i\right)\hat{\beta}_1 = 0$$
$$n\hat{\beta}_0 = \sum y_i - \left(\sum x_i\right)\hat{\beta}_1 = 0$$
$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$$\sum x_i y_i - \left(\sum x_i\right)\left(\bar{y} - \hat{\beta}_1\bar{x}\right) - \left(\sum x_i^2\right)\hat{\beta}_1 = 0$$
$$\sum x_i y_i - \bar{y}\sum x_i - \hat{\beta}_1\bar{x}\sum x_i - \hat{\beta}_1\sum x_i^2 = 0$$
$$\frac{\sum x_i y_i}{n} - \bar{y}\frac{\sum x_i}{n} - \hat{\beta}_1\bar{x}\frac{\sum x_i}{n} - \hat{\beta}_1\frac{\sum x_i^2}{n} = 0$$
$$\frac{\sum x_i y_i}{n} - \bar{x}\bar{y} - \hat{\beta}_1\bar{x}^2 - \hat{\beta}_1\frac{\sum x_i^2}{n} = 0$$

$$\hat{\beta}_1 = \frac{\frac{\sum x_i y_i}{n} - \bar{x}\bar{y}}{\bar{x}^2 + \frac{\sum x_i^2}{n}}$$
$$= \frac{n\sum x_i y_i - n^2\bar{x}\bar{y}}{n^2\bar{x}^2 + n\sum x_i^2}$$

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 + \left(\sum x_i\right)^2}$$

Now, reconsider that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and that $\sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$. Substituting the first one into the second:

$$\begin{aligned}
\sum x_i y_i - \sum x_i \bar{y} + \hat{\beta}_1 \sum x_i \bar{x} - \hat{\beta}_1 \sum x_i^2 &= 0 \\
\hat{\beta}_1 \left(\sum x_i^2 - \sum x_i \bar{x}\right) &= \sum x_i \left(y_i - \bar{y}\right) \\
\hat{\beta}_1 \sum x_i \left(x_i - \bar{x}\right) &= \sum x_i \left(y_i - \bar{y}\right) \\
\hat{\beta}_1 &= \frac{\sum x_i \left(y_i - \bar{y}\right)}{\sum x_i \left(x_i - \bar{x}\right)}
\end{aligned}$$

Now notice the following:

$$\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \sum \left(x_i - \bar{x}\right) y_i = \sum \left(y_i - \bar{y}\right) x_i$$

as,

$$\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \sum \left(x_i - \bar{x}\right) y_i - \bar{y} \underbrace{\sum \left(x_i - \bar{x}\right)}_{0} = \sum \left(y_i - \bar{y}\right) x_i - \bar{x} \underbrace{\sum \left(y_i - \bar{y}\right)}_{0}$$

and, as the sum of the deviations from the mean is zero, i.e.,

$$\sum \left(x_i - \bar{x}\right) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x} = 0$$

and

$$\sum \left(y_i - \bar{y}\right) = \sum y_i - \sum \bar{y} = \sum y_i - n\bar{y} = 0$$

The same logic applies in:

$$\begin{aligned}
\sum \left(x_i - \bar{x}\right)^2 &= \sum \left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right) \\
&= \sum \left(x_i - \bar{x}\right) x_i - \bar{x} \underbrace{\sum \left(x_i - \bar{x}\right)}_{0}
\end{aligned}$$

At the end, the above-driven expression for $\hat{\beta}_1$, i.e.,

$$\hat{\beta}_1 = \frac{\sum x_i \left(y_i - \bar{y}\right)}{\sum x_i \left(x_i - \bar{x}\right)}$$

can be rewritten as:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

so, can be written as:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum (x_i - \bar{x})^2} = \frac{Cov(x,y)}{\text{Var}(x)}$$

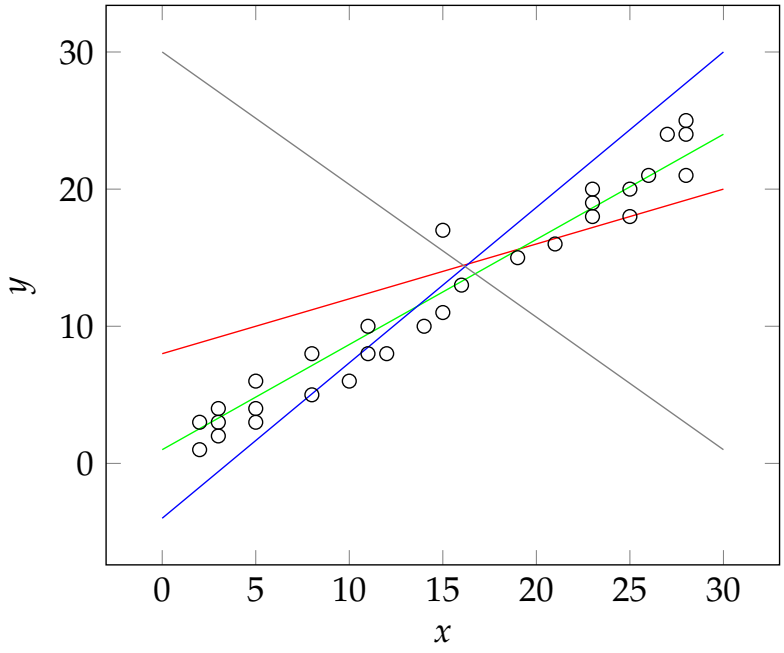
To sum up, our Least Squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model parameters β_0 and β_1 are found to be:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{Cov(x,y)}{\text{Var}(x)}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

In the graph given below, try to observe why the green line is superior to others in representing our data:



In a nutshell

Assumptions of the Simple Linear regression model
[SLR1.] The value of y, for each value of x, is

$$y = \beta_0 + \beta_1 x + e$$

[SLR2.] The average value of the random error e is $\mathrm{E}(e) = 0$ since we assume that

$$\mathrm{E}(y) = \beta_0 + \beta_1 x$$

[SLR3.] The variance of the random error e is

$$\mathrm{Var}\,(e) = \sigma^2 = \mathrm{Var}\,(y)$$

[SLR4.] The covariance between any pair of random errors, e_i and e_j is

$$Cov\,(e_i, e_j) = Cov\,(y_i, y_j) = 0, \quad i \neq j$$

[SLR5.] The variable x is not random and must take at least two different values.
[SLR6.] The values of e are normally distributed about their mean

$$e \sim \mathrm{Normal}(0, \sigma^2)$$

9.6 *Building and estimating a Multiple Linear Regression model: An increase in dimensionality*

Consider

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

where

- β_0 stands for the autonomous/ unconditional component of y
- Each $\beta_j x_j$ stands for the part of y attributable to x_j, sign of β_j determining the impact of x_j on y. $(j = 1, 2, \ldots, K)$
- Note again, a case of $\beta_1 = \beta_2 = \ldots = \beta_K = 0$ corresponds to our unconditional model of mean

$$\begin{aligned}
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_K x_{iK} \\
y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_K x_{iK} + \hat{e}_i \\
\hat{e}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} \cdots + -\hat{\beta}_K x_{iK}
\end{aligned}$$

$$S = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_K x_{iK}\right)^2$$
$$\left\{\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_K\right\}$$

As before, this minimization problem will give us $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K$, ie., the estimators of β_0, $\beta_1, \ldots, \beta_K$.

For future ease, let us restate our Multiple Linear model using matrix notation. To do this, let us first write our model equation for every single observation (for each $i = 1, 2, \ldots, n$):

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_K x_{1K} + e_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_K x_{2K} + e_2 \\
&\cdots \qquad\quad \cdots \qquad\quad \cdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_K x_{nK} + e_n
\end{aligned}$$

In matrix notation:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}}_{y_{(n\times 1)}} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1K} \\ 1 & x_{21} & x_{22} & & x_{2K} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nK} \end{bmatrix}}_{X_{(n\times(K+1))}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}}_{\beta_{((K+1)\times 1)}} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_n \end{bmatrix}}_{e_{(n\times 1)}}$$

can be written. Then,

$$y = X\beta + e$$

It is also possible to write each explanatory variable as a separate vector like:

$$x_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, x_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ \vdots \\ x_{n1} \end{bmatrix}, \ldots, x_K = \begin{bmatrix} x_{1K} \\ x_{2K} \\ \vdots \\ \vdots \\ x_{nK} \end{bmatrix}$$

so the model looks like:

$$y = x_0\beta_0 + x_1\beta_1 + \cdots + x_K\beta_K + e$$

When the matrix expression $y = X\beta + e$ is maintained, the function S becomes

$$S = e'e$$

where e' is the transpose of e.

Returning to our minimization problem written in classical notation, the following first order conditions are written:

$$\begin{aligned}
\frac{\partial S}{\partial \hat{\beta}_0} &= \sum_{i=1}^{n} -2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} \cdots - \hat{\beta}_K x_{iK}\right) = 0 \\
\frac{\partial S}{\partial \hat{\beta}_1} &= \sum_{i=1}^{n} -2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_K x_{iK}\right) x_{i1} = 0 \\
&\cdots \qquad \cdots \qquad \cdots \\
\frac{\partial S}{\partial \hat{\beta}_K} &= \sum_{i=1}^{n} -2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_K x_{iK}\right) x_{iK} = 0
\end{aligned}$$

Simplifying a little:

$$\begin{aligned}
\sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_K x_{iK}\right) &= 0 \\
\sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_K x_{iK}\right) x_{i1} &= 0 \\
\cdots \qquad \cdots \qquad \cdots \\
\sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_K x_{iK}\right) x_{iK} &= 0
\end{aligned}$$

Reorganizing the terms:

$$\begin{aligned}
n\hat{\beta}_0 + \sum x_{i1}\hat{\beta}_1 + \cdots + \sum x_{iK}\hat{\beta}_K &= \sum y_i \\
\sum x_{i1}\hat{\beta}_0 + \sum x_{i1}^2\hat{\beta}_1 + \cdots + \sum x_{i1}x_{iK}\hat{\beta}_K &= \sum x_{i1}y_i \\
\cdots \qquad \cdots \qquad \cdots \\
\sum x_{iK}\hat{\beta}_0 + \sum x_{i1}x_{iK}\hat{\beta}_1 + \cdots + \sum x_{iK}^2\hat{\beta}_k &= \sum x_{iK}y_i
\end{aligned}$$

Notice that this last set of equations can be written as:

$$\begin{bmatrix}
n & \sum x_{i1} & \cdots & \sum x_{iK} \\
\sum x_{i1} & \sum x_{i1}^2 & & \sum x_{i1}x_{iK} \\
\vdots & & & \\
\Sigma x_{iK} & \sum x_{i1}x_{iK} & \cdots & \sum x_{iK}^2
\end{bmatrix}
\begin{bmatrix}
\hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \vdots \\ \hat{\beta}_K
\end{bmatrix}
=
\begin{bmatrix}
\Sigma y_i \\ \Sigma x_{i1}y_i \\ \\ \\ x_{iK}y_i
\end{bmatrix}$$

In terms of our earlier definitions of x and y as well as β; what we have obtained is

$$X'X\hat{\beta} = X'y$$

So,

$$\hat{\beta} = \left(X'X\right)^{-1} X'y$$

Solves our minimization problem and $\hat{\beta} = \left[\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_K\right]'$ contains our parameter estimates.

9.2 EXERCISES

1. Reconsider the Simple Linear model $y = \beta_0 + \beta_1 x + e$ and show that the $\hat{\beta} = (X'X)^{-1} X'y$ works in estimating β_0 and β_1 (ie., while finding $\hat{\beta}_0$ and $\hat{\beta}_1$).
Solution:

$$X'X = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix}$$
$$X'y = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$
$$(X'X)^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix}$$

So,

$$\hat{\beta} = \frac{1}{n\sum x_i^2 - (\Sigma x_i)^2} \begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix} \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix}$$

So,

$$\hat{\beta}_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$
$$\hat{\beta}_1 = \frac{-\sum x_i \sum y_i + n\sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

Checking back our earlier solution, we verify that $\hat{\beta} = (X'X)^{-1} X'y$ works well.

2. Question: Write and solve the Least squares estimation problem for

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$$

that is, a model with a constant term (β_0) and two explanatory variables.
Solution: Solution: Left as self-study.

In a nutshell

As you have noticed, we used/devised the term $(X'X)^{-1}$ in solving our estimation problem: Think about what ensures the invertibility of $X'X$? In your future learning and practice this will be a central technical issue to address many times.

In a nutshell

Assumptions of the Multiple Linear regression model
[MLR1.]

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i$$

[MLR2.]

$$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \Longleftrightarrow \mathrm{E}(e_i) = 0$$

[MLR3.]

$$\mathrm{Var}\,(y_i) = \mathrm{Var}\,(e_i) = \sigma^2$$

[MLR4.]

$$Cov\,(y_i, y_j) = Cov\,(e_i, e_j) = 0, \quad i \neq j$$

[MLR5.] The values of x_{ik} are not random and are not exact linear functions of other explanatory variables.
[MLR6.]

$$y_i \sim \mathrm{Normal}\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}, \sigma^2\right) \Leftrightarrow e_i \sim \mathrm{Normal}(0, \sigma^2)$$

9.7 *Goodness of fit*

Suppose we have the following model:

$$y_i = \overbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}^{\hat{y}_i} + e_i = \hat{y}_i + e_i$$

Observe that

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + e_i$$

and consider the quantity $\sum (y_i - \bar{y})^2$: This quantity is called 'Total Sum of Squares'. In what follows, we decompose it into other useful quantities:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + 2 \sum (\hat{y}_i - \bar{y})\, e_i + \sum e_i^2$$

Reordering the terms in the last expression:

$$\begin{aligned}\sum (y_i - \bar{y})^2 &= \sum (\hat{y}_i - \bar{y} + e_i)^2 \\ &= \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2 + 2\underbrace{\sum (\hat{y}_i - \bar{y})\, e_i}_{0} \\ \sum (y_i - \bar{y})^2 &= \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2\end{aligned}$$

is obtained. In this expression,

$$\underbrace{\sum (y_i - \bar{y})^2}_{TSS} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{ESS} + \underbrace{\sum e_i^2}_{RSS}$$

TSS, ESS and RSS stand for:

- TSS: Total Sum of Squares
- ESS: Explained Sum of Squares
- RSS: Residual Sum of Squares

Notice that the Total Sum of Squares $\sum (y_i - \bar{y})^2$ is nothing but the variance of y multiplied by n :

$$TSS = \sum (y_i - \bar{y})^2 = n\left(\frac{1}{n}\sum (y_i - \bar{y})^2\right)$$

Explained Sum of Squares $\sum (\hat{y}_i - \bar{y})^2$ measures the sum of squared deviations of our estimated values of y (namely $\hat{y}_i$) from $\bar{y}$ (namely the unconditional mean of our dependent variable y). As $\hat{y}_i$ values are implied by our model's explanatory variables $(x_1, x_2, \ldots, x_K)$, the ESS measures the portion of TSS that we explained. Residual Sum of Squares, then, measures the portion of TSS that could not be explained. The Coefficient of Determination R^2 is the fraction of variation in y explained by our knowledge of x:

$$\begin{aligned}R^2 = \frac{ESS}{TSS} &= \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \\ &= 1 - \frac{RSS}{TSS} \\ &= 1 - \frac{\sum \hat{e}_i^2}{\sum (y_i - \bar{y})^2}\end{aligned}$$

Note that, if the model does not have a constant term (that is β_0 is omitted), then the measure R^2 is not appropriate anymore. When the constant term is omitted,

$$\sum (y_i - \bar{y})^2 \neq \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2$$

A bad habit of R^2 is that it tends to somehow increase upon the inclusion of additional explanatory variables (in fact, when their t-statistics exceed 1 in absolute value: we will see in subsequent sections) in a model. Does this mean we should continue adding more and more explanatory variables to our model 'just to push up R^2'? The answer is quite the opposite: we must see the inclusion of more variables as a cost (after all we want to come up with a parsimonious model). Then, we need to balance the benefits of more explanatory variables (enhanced ESS) with the cost of including them.

The Adjusted Coefficient of Determination $\left(\bar{R}^2\right)$ serves that purpose:

$$\bar{R}^2 = 1 - \frac{RSS/(n-K-1)}{TSS/(n-1)}$$

Notice that:

$$\bar{R}^2 = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-K-1}\right)$$

Also keep in mind that neither R^2 nor $\bar{R}^2$ has a statistical distribution. So, they are not directly and formally testable. Though, a simple arithmetic reorganization of $\bar{R}^2$ resembles an F test score (test statistic) as we will consider very soon.

9.8 Handling statistical uncertainty: calculation of variances and covariances associated with a Multiple Linear Regression model

As stated before in "Our approach to teaching/learning', up to here we maintained a naive and mechanical view of the Linear Regression modeling. In that, we deliberately, avoided calculations and discussions of the measures of dispersion or co-dispersion associated with our models. Now, it is the time to turn to reality. After all, e_i sequence has a certain statistical distribution, so does y_i. As we will formally study under the heading of 'Ideal econometric conditions: Gauss-Markov assumptions', the e_i terms have:

$$e_i \sim \text{Normal}(0, \sigma^2)$$

that is, a Normal (Gaussian) distributton with a mean of zero (0) and constant (and preferably finite) variance.

As a consequence y_i values have:

$$y_i \sim \text{Normal}(\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K, \sigma^2)$$

Intuitively, the mean of y_i depends on (is conditional on) $x_1, x_2, \ldots, x_K$ (along with their parameters); while variance of y simply mimics that of e (by the very construction of our analytical framework).

The key thing to understand now is the variability of our parameter estimates: once they are obtained from a stochastic/ random data set, it is natural/ trivial to expect each of our estimators to have a nonzero variance and each pair of our estimators to have a covariance.

We devote this section to some rigorous treatment of what we call a 'variance-covariance' matrix.

Let us begin from $e \sim \text{Normal}(0, \sigma^2)$. Once we assume the error terms to have a Normal distribution with a mean of zero (0) and a variance of σ^2, we may proceed to the following **Q&A** style mathematical elaboration:

Q: Do we know the value of σ^2 ?

A: No, it belongs to the population of e_i's. But, we only have a sample of e_i 's, namely $\hat{e}_i$'s.

Q: Can we use those $\hat{e}_i$'s to estimate σ^2, that is to obtain $\hat{\sigma}^2$?

A: Yes, the formula for $\hat{\sigma}^2$ is:

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{n - (K+1)}$$

Q: can we express $\hat{\sigma}^2$ using matrix notation?

A: Yes, the expression is:

$$\hat{\sigma}^2 = \frac{\hat{e}'\hat{e}}{n - (K+1)} = \frac{(y - X\beta)'(y - X\beta)}{n - (K+1)}$$

Q: What about the $Cov\left(\hat{\beta}_i, \hat{\beta}_j\right)$ values, can we calculate them?

A: Sure, in matrix notation,

$$Cov(\hat{\beta}) = \text{E}\big((\hat{\beta} - \beta)(\hat{\beta} - \beta)'\big) = \sigma^2 \left(X'X\right)^{-1}$$

Q: What about the distribution of $\hat{\beta}$?

A:

$$\hat{\beta} \sim \text{Normal}(\beta, \sigma^2 \left(X'X\right)^{-1}\Big)$$

Q: What does this mean?

A: First, each parameter estimate is unbiased, $\text{E}\big(\hat{\beta}\big) = \beta$. Second, the variances are ruled by $\sigma^2 \left(X'X\right)^{-1}$.

Q: How is the structure of the variance-covariance matrix?

A:

$$\begin{aligned}\text{Cov}(\hat{\beta}) &= \text{E}\Big((\hat{\beta}-\beta)(\hat{\beta}-\beta)'\Big) \\ &= \begin{bmatrix} \text{Var}(\hat{\beta}_0) & Cov(\hat{\beta}_0,\hat{\beta}_1) & \cdots & Cov(\hat{\beta}_0,\hat{\beta}_K) \\ Cov(\hat{\beta}_0,\hat{\beta}_1) & \text{Var}(\hat{\beta}_1) & \cdots & Cov(\hat{\beta}_1,\hat{\beta}_K) \\ \cdots & & \cdots & \cdots \\ Cov(\hat{\beta}_0,\hat{\beta}_K) & Cov(\hat{\beta}_1,\hat{\beta}_K) & \cdots & \text{Var}(\hat{\beta}_K) \end{bmatrix}\end{aligned}$$

$$= \begin{bmatrix} E(\hat{\beta}_0-\beta_0)^2 & E(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1) & \cdots & E(\hat{\beta}_0-\beta_0)(\hat{\beta}_K-\beta_K) \\ E(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1) & E(\hat{\beta}_1-\beta_1)^2 & \cdots & E(\hat{\beta}_1-\beta_1)(\hat{\beta}_K-\beta_K) \\ E(\hat{\beta}_0-\beta_0)(\hat{\beta}_K-\beta_K) & E(\hat{\beta}_1-\beta_1)(\hat{\beta}_K-\beta_K) & \cdots & E(\hat{\beta}_K-\beta_K)^2 \end{bmatrix}$$

$$= \sigma^2 (X'X)^{-1} = \sigma^2 \begin{bmatrix} n & \sum x_{i1} & \cdots & \Sigma_{x_{iK}} \\ \Sigma x_{i1} & \Sigma x_{i1}^2 & \cdots & \Sigma_{x_{i1}x_{iK}} \\ \vdots & & & \\ \Sigma x_{iK} & \Sigma x_{i1}x_{iK} & \cdots & \Sigma x_{iK}^2 \end{bmatrix}^{-1}$$

Q: But, we do not know the value of σ^2?

A: Then, substitute $\hat{\sigma}^2$ for it:

$$\hat{\text{Cov}}(\beta) = \hat{\sigma}^2 (X'X)^{-1}$$

Q: Does that mean we will be using the estimated values of variances and covariances?

A: Sure. This is what we have been doing since the beginning of our ECON 222 journey.

Q: Are we now ready to dive into the fascinating world of statistical inference over our estimated models?

A: Very much, indeed.

Q: Are you an AI?

A: No. Are you?

9.9 Statistical inference

We have studied/learned up to this point:

- Probability basics and a rich-enough collection of well-known statistical distributions in ECON 221 (Chapter 1, Chapter 2, Chapter 3, chapter 4)

- Point estimators of distributional parameters, and the fundamentals of statistical inference (confidence intervals and hypothesis testing) in ECON 222 (Chapter 5, Chapter 6, Chapter 7)

- Structure, formation and estimation of Simple Linear Regression and Multiple Linear Regression models in ECON 222 (earlier sections of Chapter 8)

Now, we are ready to place our estimated models under some serious scrutiny. Using the inferential tools that we learned, we will evaluate, test and scientifically question our regression models.

In a bold fashion, we can say that what we did up to here (i.e., estimating regression models) is no more than the half of the job. To have the job actually done, we need to delve into the following tasks:

1. Estimating confidence intervals for individual model parameters β_i
2. Estimating confidence intervals for linear combinations of (more than one) model parameters
3. Conducting hypothesis tests for individual model parameters β_i
4. Conducting hypothesis tests for linear combinations of (more than one) model parameters
5. Conducting hypothesis tests for all of our model parameters at once
6. Conducting hypothesis tests for specific subsets of our model parameters at once

Now, let us give examples to each category of tasks listed above. To do this, suppose we have the following economic model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$$

Recall that, this is our model written for the population and we turn it into a statistical model (written again for the population) by introducing the statistical error (disturbance, sometimes 'shock') terms:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i$$

where $e_i \sim \text{Normal}(0, \sigma^2)$. As you know well now, we do not know the true values (population values of β_j 's). So, we will estimate the model using a sample of n observations and the Least Squares technique.

$$\begin{gathered} y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i, \\ i = 1, 2, \ldots, n, e_i \sim \text{Normal}\left(0, \sigma^2\right) \end{gathered}$$

Provided that everything goes well on the paper and in the computer, we will end up with a rich set of estimates:

- Estimates of model parameters: $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$

- Estimated sequence of the dependent variable: $\hat{y}_i$
- Estimated sequence of error terms: $\hat{e}_i$
- Estimated model variance:

$$\hat{\sigma}^2 = \frac{\hat{e}'\hat{e}}{n-(K+1)} \quad \left(\text{here, } \frac{\hat{e}'\hat{e}}{n-5}\right)$$

- Estimated "variance-covariance matrix:

$$\hat{Cov}(\beta) = \hat{\sigma}^2 \left(X'X\right)^{-1}$$

Now, suppose the following claims and/or questions come from an academic/ technical colleague. (Needless to say, even when there is no criticizing colleague around, we need to put these claims on our own and heavily test our models):

1. Is 0.4 a viable value for β_1, with respect to a 95% confidence interval of β_1 ?
2. Is 0.7 a viable value for $\beta_1 + \beta_2$, with respect to a 95% confidence interval of $\beta_1 + \beta_2$?
3. Is β_3 equal zero or not; how do we know x_3 is an important/significant explanatory variable?
4. Is $\beta_3 + \beta_4$ equal one or not?
5. Is $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$; how do we know our explanatory variables x_1, x_2, x_3 and x_4 matter as a whole?
6. Is $\beta_1 = \beta_2 = 0$; how do we know the explanatory variables x_1 and x_2 matter together?

Our road map to assess these questions begins with formulating these questions/claims in some formal notation:

Following the same order as above:

1. We will calculate a 95% C.I. for β_1 and will check if 0.4 belongs to the calculated interval. This is simply done as:

$$\mathrm{P}\left(\hat{\beta}_1 - t_c \,\mathrm{se}\left(\hat{\beta}_1\right) \leq \beta_1 \leq \hat{\beta}_1 + t_c \,\mathrm{se}\left(\hat{\beta}_1\right)\right) = 1 - \alpha$$

2. We will calculate a 95% C.1. for $\beta_1 + \beta_2$ and will check if 0.7 belongs to the calculated interval.

$$\mathrm{P}\left(\hat{\beta}_1 + \hat{\beta}_2 - t_c \,\mathrm{se}\left(\hat{\beta}_1 + \hat{\beta}_2\right) \leq \beta_1 + \beta_2 \leq \hat{\beta}_1 + \hat{\beta}_2 + t_c \,\mathrm{se}\left(\hat{\beta}_1 + \hat{\beta}_2\right)\right) = 1 - \alpha$$

Here, we apparently need to calculate $\mathrm{Var}\left(\hat{\beta}_1 + \hat{\beta}_2\right)$. Using our knowledge from ECON 221:

$$\mathrm{Var}\left(\hat{\beta}_1 + \hat{\beta}_2\right) = \mathrm{Var}\left(\hat{\beta}_1\right) + 2Cov(\hat{\beta}_1, \hat{\beta}_2) + \mathrm{Var}\left(\hat{\beta}_2\right)$$

where Var $(\hat{\beta}_1)$, $Cov(\hat{\beta}_1, \hat{\beta}_2)$ and Var $(\hat{\beta}_2)$ are straightforwardly obtained during the estimation of the model. Once Var $(\hat{\beta}_1 + \hat{\beta}_2)$ is at hand, $se(\hat{\beta}_1 + \hat{\beta}_2) = \sqrt{\text{Var}\,(\hat{\beta}_1 + \hat{\beta}_2)}$ yields the required standard error.

3. We will conduct the test

$$H_0 : \beta_3 = 0$$
$$H_1 : \beta_3 \neq 0$$

Distribution of the test statistic:

$$\frac{\hat{\beta}_3 - \beta_3}{\sqrt{\text{Var}\,(\hat{\beta}_3)}} \sim t_{(n-K-1)}$$

Calculation of the test statistic:

$$\frac{\hat{\beta}_3 - \beta_3^0}{\text{se}\,(\hat{\beta}_3)} \sim t_{(n-K-1)}$$

$$\frac{\hat{\beta}_3 - 0}{\text{se}\,(\hat{\beta}_3)} \sim t_{(n-K-1)}$$

4. We will conduct the test

$$H_0 : \beta_3 + \beta_4 = 1$$
$$H_1 : \beta_3 + \beta_4 \neq 1$$

$$\frac{\hat{\beta}_3 + \hat{\beta}_4 - (\beta_3 + \beta_4)}{\sqrt{\text{Var}\,(\hat{\beta}_3 + \hat{\beta}_4)}} \sim t_{(n-K-1)}$$

$$\frac{\hat{\beta}_3 + \hat{\beta}_4 - (\beta_3 + \beta_4)^0}{\text{se}\,(\hat{\beta}_3 + \hat{\beta}_4)} \sim t_{(n-K-1)}$$

$$\frac{\hat{\beta}_3 + \hat{\beta}_4 - 1}{\text{se}\,(\hat{\beta}_3 + \hat{\beta}_4)} \sim t_{(n-K-1)}$$

Var $(\hat{\beta}_3 + \hat{\beta}_4)$ will be treated as outlined above for the case of Var $(\hat{\beta}_1 + \hat{\beta}_2)$. Note that this test can also be conducted as an F test, as we will cover in our class discussions.

5. We will conduct the test

$$\begin{aligned} H_0 : &\quad \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \\ H_1 : &\quad \exists \beta_i \neq 0 \end{aligned}$$

Total sum of squares TSS being $\sum (y_i - \bar{y})^2$, explained sum of squares ESS being $\sum (\hat{y}_i - \bar{y})^2$ and residual sum of squares RSS being $\sum \hat{e}_i^2$:

$$\frac{(TSS - RSS)/K}{RSS/(n-K-1)} \sim F_{(K,n-K-1)}$$

6. We will conduct the test

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = 0 \\ H_1 &: \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \end{aligned}$$

J being the number of joint hypotheses, RSS_R being the RSS for the restricted model and RSS_U being the RSS for the unrestricted model:

$$\frac{(RSS_R - RSS_U)/J}{RSS_U/(n-K-1)} \sim F_{(J,n-K-1)}$$

Note / clarify again that RSS_U is the RSS value of the unrestricted, i.e., full model, which is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i$$

where, RSS_R is the RSS value of the restricted model, which is:

$$y_i = \beta_0 + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i$$

equivalently of:

$$y_i = \beta_0 + 0 \cdot x_{i1} + 0 \cdot x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i$$

Returning to our previous hypothesis test:

$$\begin{aligned} H_0 : &\quad \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \\ H_1 : &\quad \exists \beta_i \neq 0 \end{aligned}$$

you will notice that the restricted model is:

$$y_i = \beta_0 + e_i$$

or

$$y_i = \beta_0 + 0 \cdot x_{i1} + 0 \cdot x_{i2} + 0 \cdot x_{i3} + 0 \cdot x_{i4} + e_i$$

against the unrestricted (full) model of:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i$$

Herein, RSS_R becomes the TSS of the full model (verify yourself), RSS_U becomes the RSS of the full model (should be trivial) and J becomes K. Then, the equivalence between

$$\frac{(RSS_R - RSS_U)/J}{RSS_U/(n-K-1)} \sim F_{(J,n-K-1)}$$

and

$$\frac{(TSS - RSS)/K}{RSS/(n-K-1)} \sim F_{(K,n-K-1)}$$

becomes apparent.

We will now use a model estimated on a computer to exemplify each of the cases above:
[To be distributed as a handout]

9.10 Essence of the Gauss-Markov assumptions

Having studied the mechanical aspects of Linear Regression models, now it is the time to establish the conditions under which a linear regression model is viable with workable results. As we often call it 'ideal econometric conditions', the Gauss-Markov assumptions level the field for us. If a model abides by these assumptions, i.e., if a model has been formed so as to hold the Gauss-Markov assumptions, then it is a good econometric model.

In a nutshell

Gauss-Markov assumptions

A1. Linearity in parameters: The derivative of the y with respect to the parameters should not be a function of the parameters. Analytical solutions for the parameter estimates (coefficients) require this assumption and without it one needs numerical methods to solve for coefficients.

A2. Random sampling (non-stochastic x): The sample should be so randomly picked from the population that it is representative of the population. This has two advantages

- The results we get from the sample can be generalized to the whole population
- Our knowledge of x about the population can be applied in the sample so it is as if you know the sample x too.

A3. Variation in x: Econometrics analyzes how y changes with respect to x; for this x needs to change.

A4. Exogeneity: $\mathrm{E}(e \mid x) = \mathrm{E}(e) = 0$. Knowledge of x does not improve expectation of e as they will be independent of each other.

A5. The shocks to each observation are coming from the same distribution indepedently and identically: $\mathrm{Var}(e_i) = \sigma^2, \quad \forall i$ and $Cov(e_i, e_j) = 0, \quad \forall i \neq j$

A6. $\mathrm{Var}(e_i) \sim \mathrm{Normal}(0, \sigma^2) \Rightarrow y_i \sim \mathrm{Normal}(\beta_0 + \beta_1 x_i, \sigma^2)$

Now we can review how good is our LS estimator under these conditions. Consider the Simple Linear regression model $y_i = \beta_0 + \beta_1 x_i + e_i$ and consider the Gauss-Markov assumptions. Let us now try to see how good is our LS estimator under these assumptions. Recall that $\hat{\beta}_1$ is:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

which can also be written as:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum (x_i - \bar{x})\, y_i}{\sum (x_i - \bar{x})\, x_i} \\ &= \frac{\sum (x_i - \bar{x})\, (\beta_0 + \beta_1 x_i + e_i)}{\sum (x_i - \bar{x})\, x_i} \\ &= \frac{\beta_0 \sum (x_i - \bar{x}) + \beta_1 \sum (x_i - \bar{x})\, x_i + \sum (x_i - \bar{x})\, e_i}{\sum (x_i - \bar{x})\, x_i}\end{aligned}$$

As $\sum (x_i - \bar{x}) = 0$ (shown before), the expression becomes:

$$\begin{aligned}\hat{\beta}_1 &= \beta_1 + \frac{\sum (x_i - \bar{x})\, e_i}{\sum (x_i - \bar{x})\, x_i} \\ &= \beta_1 + \frac{\sum (x_i - \bar{x})\, e_i}{\sum (x_i - \bar{x})^2}\end{aligned}$$

Then,

$$\begin{aligned}E\left(\hat{\beta}_1 \mid x\right) &= E\left(\beta_1 + \frac{\sum (x_i - \bar{x})\, e_i}{\sum (x_i - \bar{x})^2} \,\middle|\, x\right) \\ &= \beta_1 + E\left(\frac{\sum (x_i - \bar{x})\, e_i}{\sum (x_i - \bar{x})^2} \,\middle|\, x\right) \\ &= \beta_1 + \frac{\sum (x_i - \bar{x})\, E\left(e_i \mid x\right)}{\sum (x_i - \bar{x})^2}\end{aligned}$$

can be written since our x (independent variable, explanatory variable) is non-stochastic.

We also know by the Gauss-Markov assumptions that $E\left(e_i \mid x\right) = 0$, i.e., our knowledge of x does not improve expectation of e. So,

$$E\left(\hat{\beta}_1 \mid x\right) = \beta_1$$

equivalently saying $\hat{\beta}_1$ is an unbiased estimator of β_1.

What about $E\left(\hat{\beta}_0 \mid x\right)$?

$$\begin{aligned}y_i = \beta_0 + \beta_1 x_i + e_i \rightarrow \bar{y} &= \beta_0 + \beta_1 \bar{x} + \bar{e} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= \beta_0 + \beta_1 \bar{x} + \bar{e} - \hat{\beta}_1 \bar{x} \\ &= \beta_0 - \left(\hat{\beta}_1 - \beta_1\right) \bar{x} + \bar{e}\end{aligned}$$

Then,

$$\begin{aligned}E\left(\hat{\beta}_0 \mid x\right) &= E\left(\beta_0 - \left(\hat{\beta}_1 - \beta_1\right) \bar{x} + \bar{e} \mid x\right) \\ &= \beta_0 - \bar{x} \underbrace{E\left(\left(\hat{\beta}_1 - \beta_1\right) \mid x\right)}_{0} + \underbrace{E\left(\frac{\sum e_i}{n} \,\middle|\, x\right)}_{0}\end{aligned}$$

So,

$$E\left(\hat{\beta}_0 \mid x\right) = \beta_0$$

equivalently saying $\hat{\beta}_0$ is an unbiased estimator of β_0.

$$\begin{aligned}\operatorname{Var}\left(\hat{\beta}_1\right) &= \mathrm{E}\left((\hat{\beta}_1 - \underbrace{E\left(\hat{\beta}_1\right)}_{\beta_1})^2\right) \\ &= E\left(\left(\frac{\sum\left(x_i - \bar{x}\right) e_i}{\sum\left(x_i - \bar{x}\right)^2}\right)^2\right)\end{aligned}$$

Expanding the expression and rearranging its terms:

$$\begin{aligned}\operatorname{Var}\left(\hat{\beta}_1\right) &= E\left(\left.\frac{\sum\left(x_i - \bar{x}\right)^2 e_i^2 + \sum\sum\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right) e_i e_j}{\left(\sum\left(x_i - \bar{x}\right)^2\right)^2}\right| x\right) \\ &= \frac{\sum\left(x_i - \bar{x}\right)^2 E\left(e_i^2 \mid x\right) + \sum\sum\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right) E\left(e_i e_j \mid x\right)}{\left(\sum\left(x_i - \bar{x}\right)^2\right)^2}\end{aligned}$$

As $E\left(e_i^2 \mid x\right) = \sigma^2$ and $E\left(e_i e_j \mid x\right) = 0$,

$$\operatorname{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2 \sum\left(x_i - \bar{x}\right)^2}{\left(\sum\left(x_i - \bar{x}\right)^2\right)^2} = \frac{\sigma^2}{\sum\left(x_i - \bar{x}\right)^2} = \frac{\sigma^2}{n \operatorname{Var}\left(x\right)}$$

This expression is a Noise/Signal (i.e., a noise-to-signal) ratio expression.

Examining

$$\operatorname{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{n \operatorname{Var}\left(x\right)}$$

We see that, to decrease $\operatorname{Var}\left(\hat{\beta}_1\right)$, a larger sample size n, a larger $\operatorname{Var}\left(x\right)$ and a smaller σ^2 would help. Among these, the researcher's choice of the sample data affects n and $\operatorname{Var}\left(x\right)$. σ^2, on the other hand, is out of the researcher's reach.

$$\begin{aligned}
\text{Var}\left(\hat{\beta}_0\right) &= E\left(\left(\hat{\beta}_0 - E\left(\hat{\beta}_0\right)^2\right)\right. \\
As\hat{\beta}_0 &= \beta_0 - \left(\hat{\beta}_1 - \beta_1\right)\bar{x} + \bar{e}: \\
\text{Var}\left(\hat{\beta}_0\right) &= E\left(\left(\beta_0 - \left(\hat{\beta}_1 - \beta_1\right)\bar{x} + \bar{e} - E\left(\hat{\beta}_0\right)^2\right)\right. \\
&= E\left(\left(\beta_0 - \left(\hat{\beta}_1 - \beta_1\right)\bar{x} + \bar{e} - \beta_0\right)^2\right) \\
&= E\left(\left(-\left(\hat{\beta}_1 - \beta_1\right)\bar{x} + \bar{e}\right)^2\right) \\
&= \bar{x}^2 E\left(\hat{\beta}_1 - \beta_1\right)^2 + E\left(\bar{e}^2\right) - 2\bar{x}E\left(\left(\hat{\beta}_1 - \beta_1\right)\bar{e}\right)
\end{aligned}$$

To simplify this expression observe/elaborate:
(1)

$$E\left(\hat{\beta}_1 - \beta_1\right)^2 = \text{Var}\left(\hat{\beta}_1\right)$$

(2)

$$\begin{aligned}
E\left(e^2\right) = E\left(\frac{\sum e_i}{n}\right)^2 &= \frac{1}{n^2}E\left(\sum e_i\right)^2 \\
&= \frac{1}{n^2}(E\left(\sum e_i^2\right) + \underbrace{E\left(\sum e_i e_j\right)}_{0}) \\
&= \frac{E\left(\sum e_i^2\right)}{n^2} \\
&= \frac{n\sigma^2}{n^2} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

(3) $E\left(\hat{\beta}_1 - \beta_1\right)\mathrm{E}(\bar{e}) = 0$
Then,

$$\text{Var}\left(\hat{\beta}_0\right) = \frac{\bar{x}^2\sigma^2}{\sum\left(x_i - \bar{x}\right)^2} + \frac{\sigma^2}{n}$$

is reached. Rearranging:

$$\begin{aligned}
\text{Var}\left(\hat{\beta}_0\right) &= \frac{\sigma^2\left(n\bar{x}^2 + \sum\left(x_i - \bar{x}\right)^2\right)}{n\sum\left(x_i - \bar{x}\right)^2} \\
&= \frac{\sigma^2}{n\sum\left(x_i - \bar{x}\right)^2}\left(n\bar{x}^2 + \sum x_i^2 - 2\bar{x}\sum x_i + \sum \bar{x}^2\right) \\
&= \frac{\sigma^2}{n\sum\left(x_i - \bar{x}\right)^2}\left(nx^2 + \sum x_i^2 - 2n\bar{x}^2 + nx^2\right) \\
\text{Var}\left(\hat{\beta}_0\right) &= \sigma^2\left(\frac{\sum x_i^2}{n\sum\left(x_i - \bar{x}\right)^2}\right)
\end{aligned}$$

is obtained.

$$\text{Var}\left(\hat{\beta}_0\right) = \sigma^2 \left(\frac{\sum x_i^2}{n\sum\left(x_i - \bar{x}\right)^2}\right)$$
$$\text{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum\left(x_i - \bar{x}\right)^2}$$
$$\text{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = \sigma^{-2}\left(\frac{-\bar{x}}{\sum\left(x_i - \bar{x}\right)^2}\right)$$

(1) The larger the value of σ^2 the larger will be the variances of the estimators.

(2) Var $\left(\hat{\beta}_1\right)$ will be smaller, the larger the value of $\sum\left(x_i - \bar{x}\right)^2$. This is also true for Var $\left(\hat{\beta}_0\right)$, but it is less evident as $\sum x_i^2$ appears in the numerator of Var $\left(\hat{\beta}_0\right)$ expression.

(3) Because the number of terms in $\sum\left(x_i - \bar{x}\right)^2$ increases in n (sample size), an increase in n generally leads to an increase in precision.

9.11 Model Specification

There are two main approaches to model specification:

- Starting out small, with one or few explanatory variables; retaining statistically significant ones and expanding the variables when needed
- Starting out large and throwing out insignificant variables to reach the true model.

Regarding either of the approaches, we need a good methodological basis. The material of the section entitled 'Statistical inference', luckily, provides us with the toolset to establish that. The task of model specification involves a systematic sequence of hypothesis tests and evaluation of models with respect to some *ad hoc* criteria. While the t tests and F tests equip us to assess our models, $R^2, \bar{R}^2$, AIC, BIC (or SIC) and HQ information criteria further strengthen our hand to come up with parsimonious model specifications.

Akaike Information Criterion:

$$AIC = \ln(\sigma^2) + \frac{2k}{n}$$

Bayesian Information Criterion or **Schwarz Information Criterion** or **Schwarz Criterion** or **Schwarz-Bayesian Criterion:**

$$BIC = SIC = SC = SBC = \ln(\sigma^2) + \frac{k\ln n}{n}$$

Hannan-Quin Criterion:

$$HQ = \ln(\sigma^2) + \frac{k\ln(\ln(n))}{n}$$

Among the rival models, the ones with lower information criterion values are preferable to others. Therein, it is a good practice to use the same sample size while comparing models via information criteria.

In a nutshell

As of this point, we have a sufficient knowledge base to proceed to our first econometric practice. In what follows through the lecture notes, we mobilize our statistical knowledge on field. Note that some new formulations and/or theoretical elements can be introduced if a need arises.

9.12 *Regression analysis at work*

In this section we will put our theoretical knowledge into practice. The modeling exercises that we will consider maintain a manageable pedagogical standard, they are somehow downsized and sometimes oversimplified. Yet, they are designed to deliver the intended message of the chapter with regard to applied statistical/ econometric research.

The cases we will consider are as follows:

- **Case 01** State public expenditures in the US: A public finance model (Economics)

  Data reference:

  - U.S. Department of Commerce, Bureau of the Census, Government Finances in 1960, Census of Population, 1960, Census of Manufactures, 1958, Statistical Abstract of the United States, 1961.
  - U.S. Department of Agriculture, Agricultural Statistics, 1961.
  - U.S. Department of the Interior, Minerals Yearbook, 1960.
  - Authorization: for educational use

  Variables:

  1. EX: Per capita state and local public expenditures (USD)
  2. ECAB: Economic ability index, in which income, retail sales, and the value of output (manufactures, mineral, and agricultural) per capita are equally weighted.

3. MET: Percentage of population living in standard metropolitan areas
4. GROW: Percent change in population, 1950-1960
5. YOUNG: Percent of population aged 5-19 years
6. OLD: Percent of population over 65 years of age
7. WEST: Western state (1) or not (0)

- **Case 02** Home prices in Albuquerque: what determines home prices? (Economics, Real estate, Business)

  Data reference:

  – Albuquerque Board of Realtors
  – Authorization: for educational use

  Variables:

  1. PRICE: Selling price (USD, hundreds)
  2. SQFT: Square feet of living space
  3. AGE: Age of home (years)
  4. FEATS: Number out of 11 features (dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access
  5. NE: Located in northeast sector of city (1) or not (0)
  6. COR: Corner location (1) or not (0)
  7. TAX: Annual taxes (USD)

- **Case 03** Taste of cheese: An assessment of subjective scores (Product development, Business)

  Data reference:

  – Moore, David S., and George P. McCabe (1989). Introduction to the Practice of Statistics.
  – Authorization: for educational use

  Variables:

  1. TASTE: Subjective taste test score, obtained by combining the scores of several tasters
  2. ACETIC: Natural log of concentration of acetic acid
  3. H2S: Natural log of concentration of hydrogen sulfide
  4. LACTIC: Concentration of lactic acid

- **Case 04** Consumption of soft drinks: Practicing categorical determinants (Consumer research)

  Data reference:

  - Artificial data - Eray Yucel
  - Authorization: for educational use

  Variables:

  1. GENDER (0: male, 1: female)
  2. URBAN: 1 for urban, 0 for rural
  3. RURAL: 1for rural, 0 for urban
  4. AGE
  5. INCOME (TL)
  6. CONS2: consumption of soft drinks per month

- **Case 05** A promotion for soda consumers: The Linear Probability Model - simple and still useful (Business)

  Data reference:

  - Artificial data - Eray Yucel
  - Authorization: for educational use

  Variables:

  1. INCOME2
  2. U: 1 for Urban, 0 for Rural
  3. F: 1 for Female, 0 for Male
  4. W: 1 for Working, 0 for Non-working
  5. SODA: Monthly soda consumption (bottles)
  6. SODA20: 1 if SODA is at least 20

- **Case 06** A demonstration of the effect of omitted variables (Simpson's paradox)

  Data reference:

  - Artificial data - Eray Yucel
  - Authorization: for educational use

  Variables:

  1. Y
  2. X
  3. D1, D2, D3

While these cases are being examined, we will concurrently be learning the use of '**Dummy variables**' in an embedded fashion: The theoretical knowledge needed will be provided when/as necessary.

Cross-section versus Time series data

Our choice of theoretical exposition in ECON 222 maintained/kept cross-section data at a central position. In that, we often referred to our observations $y_i, x_{i1}, x_{i2}, \ldots, x_{iK}$ using the observation index ' i '. When this is the case, note that there is no natural ordering of observations. For example, writing the USA's inflation rate in Row 2 of a data file, while we write the UK's inflation rate in Row 7 for the same year and 'switching their rows' do not yield different results.

Time series data, on the other hand, do have a natural ordering of observations, merely by the definition of time: before comes before now, now comes before tomorrow, so tomorrow comes after both. This underlines the importance of time as the primary key of our dataset when analyzing time series data and especially when we do it via dynamic models. Indifference/silence of this book, ECON 221 and ECON 222 about time series notation and data was of course intentional from a pedagogical viewpoint. Once you proceed to ECON 301 and ECON 302 (Econometrics sequence) be prepared to replace ' i ' with ' t ' as your new (and naturally ordered, $t = 1, 2, \cdots, T$) observation index. Note that, all our formulations are rock solid / robust up to this change.

In the set of cases/exercises of this section, we make use of cross-section data sets.

NOTICE: Until a proper typeset is prepared, the cases/exercises of this section will be handled using Handouts. These Handouts will follow and summarize what is to be done in class lectures and they are available through the "Handouts" link under https://sites.google.com/view/erayyucel/probability-and-statistics. To have the latest available material and stay informed, keep a keen eye on this page.

9.13 Frisch-Waugh-Lovell theorem (FWL theorem)

FWL theorem shows how to decompose a regression of y on a set of variables x into two pieces. If we divide x into two sets of variables x_1 and x_2 and regress y on x_1 and x_2, the coefficient estimates on x_2 can also be obtained through the following steps:

1. Regress all variables in x_2 on x_1 and take the residuals.
2. Regress y on x_1 and take the residuals.

3. Regress the residuals from step 2 on the residuals from step 1.

To demonstrate what the FWL theorem says, consider our Case02, Home prices, again:

Dependent Variable: LP

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.652384	0.350431	1.861662	0.0655
LS	0.521313	0.085217	6.117444	0.0000
LT	0.368324	0.065693	5.606762	0.0000

In `case02.wf1`, page `case02s2`, we have the regression equation

$$LP = \beta_0 + \beta_1 LS + \beta_2 LT + e$$

estimated as above. So,

$$\hat{\beta}_0 = 0.6523$$
$$\hat{\beta}_1 = 0.5213$$
$$\hat{\beta}_2 = 0.3683$$

Focus on $\hat{\beta}_2 = 0.3683$, *i.e.*, the coefficient estimate of taxes, LT.

As to our application of the FWL theorem,

$$x = \{LS, LT\}$$
$$x_1 = \{LS\}$$
$$x_2 = \{LT\}$$

and $y = LP$.

First, regress x_2 on x_1, that is, regress LT on LS, extract the residuals, and name them `E_LT_ON_LS`. You can view this series in `case02.wf1`.

Dependent Variable: LT

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−1.473070	0.500340	−2.944136	0.0040
LS	1.095477	0.067801	16.15724	0.0000

Second, regress y on x_1, that is, regress LP on LS, extract the residuals, and name them `E_LP_ON_LS`. You can view this series in `case02.wf1`.

Finally, regress `E_LP_ON_LS` on `E_LT_ON_LS` and obtain the coefficient estimate for `E_LT_ON_LS` as 0.3683.

Dependent Variable: `E_LP_ON_LS`

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.007499	0.013445	0.557758	0.5782
`E_LT_ON_LS`	0.368324	0.065399	5.631914	0.0000

Notice that the coefficient estimate of LT in the very first regression is identical with the coefficient estimate of `E_LT_ON_LS` in the final regression.

This is how the FWL theorem functions.

References

Chris Brooks. *Introductory Econometrics for Finance*. Cambridge University Press, Cambridge, 1st edition, 2002. ISBN 9780521793674.

Damodar N. Gujarati. *Basic Econometrics*. McGraw-Hill, Boston, MA, 4th edition, 2003. ISBN 0072335424.

R. Carter Hill, William E. Griffiths, and Guay C. Lim. *Principles of Econometrics*. John Wiley & Sons, Hoboken, NJ, 4th edition, 2011. ISBN 9780470626733.

James T. McClave, P. George Benson, and Terry Sincich. *Statistics for Business and Economics*. Pearson, Boston, MA, 13th edition, 2018. ISBN 9780134506593.

Paul Newbold, William L. Carlson, and Betty Thorne. *Statistics for Business and Economics*. Pearson, Boston, MA, 8th edition, 2013. ISBN 9780132745659.

Murray R. Spiegel and Larry J. Stephens. *Schaum's Outline of Statistics*. Schaum's Outline Series. McGraw-Hill, New York, NY, 4th edition, 2008. ISBN 9780071485845.

Index