MAT 121 E-Pack: Statistics I

Cheng Peng

West Chester University

Contents

Chapter 1

Introduction

This *E-Pack* is a self-contained homegrown Ebook that contains all topics covered in current MAT 121 at WCU.

All technical terms used in this Ebook are consistent with those used in the *required* textbook. There several benefits of using this *E-Pack*:

- All technical terms used in this eBook are consistent with those used in the *required* textbook.
- Three formats (PDF, HTML, and EPub) of the E-Pack are accessible from different devices.
- In the HTML version, there are many animated graphics as well as decorative visual aids such as font size, colors, etc. to make some concepts more intuitive.
- The supplement course website [https://pengdsci.github.io/MAT121/] has topic-wise online practice problems for most of the topics.
- 17 homegrown interactive statistics learning apps (ISLA) [https://pengdsci.github.io/ISLA/] covers all topics in this course. These apps can be used in two ways:
  - play around with the apps associated with each topic to enhance your understanding of the topics;
  - use these apps as **solution generator** to check your work. Yes. They can generator answers to your questions! I used these apps generated answers to many examples

Chapter 2

Frequency Tables and Charts

2.1 Basic Statistical Terminologies

In this note, we introduce basic terminology of statistics and methods for summarizing a given data set.

2.1.1 What is statistics?

Statistics is the science of collecting, organizing, visualizing, analyzing, and interpreting data in order to make decision.

2.1.2 Population vs Sample

- **Population**: The collection of **all** outcomes, responses, measurements, or counts that are of interest (the right group in the following figure).
- **Sample**: A subset of the population (the left group in the following figure).

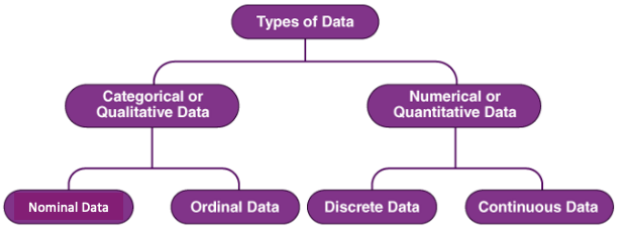
- **Parameter**: the numeric characteristic of the population. For example, the average height of **all** students at WCU. Here `all WCU students` is a

population.

- **Statistic**: the numeric characteristic of the sample (i.e., a subset of the population). For example, the average height of **subset of** students at WCU. Here `the subset of WCU students` is a sample taken from the population of all WCU students.

2.1.3 Types Statistics and Data

- **Descriptive Statistics** involves organizing, summarizing, and displaying data. For example, we can use tables, charts, averages, etc. All topics in this note and next note will focus on descriptive statistics.
- **Inferential Statistics** uses the sample data to make inferences about the underlying population. For example, all topics from week #3 are inferential statistics.
- **Data Types**: There different ways for classifying data in statistics. The following diagram given one of the simple bu widely used methods.



- **Data Types** examples
  - Nominal Data (also called unordered data): the place of birth, major, eye color, etc.
  - Ordinal Data: Military Rank (private, corporal, etc.), Course Grade (A, B, C, D, F), etc.
  - Discrete (a subset of which is "counting"): Number of children in a family, Shoe Size, etc.
  - Continuous: Weight, Height, temperature, income, GPA, etc.

2.2 Summarizing Qualitative Data

For a given categorical data, we can use frequency tables and charts to summarize the distribution of the data. Note that the given data set could either be a population or a sample.

2.2.1 Frequency Tables

Since each distinct data value represents a category, the number of values in each category is the frequency of the category. `An ordinary frequency table` is a

two-column table in which the left column lists the category labels and the right column lists the corresponding frequencies.

There are four types of frequency tables. The other three frequency tables are relative frequency table, cumulative frequency table, cumulative relative frequency table.

`a relative frequency = ordinary frequency / total`

`a cumulative frequeny table` is constructed based on the cumulatively combined categories. See the following example for more detail.

Example 1: In a class of 20 students, 3 students received a grade of "A", 6 students received a "B", 7 students received a "C", 3 students received a "D", and 1 student received an "F". These results are summarized, in a variety of ways, in the following table: (Note that for the ordered categorical variable "Grade" we also create the discrete quantitative variable "Grade-point".)

Solution: The raw data might be in the following form:

`A, C, B, F, D, B, C, B, C, C, A, D, C, B, C, B, A, C, B, D`

The resulting frequency tables is given by

Grade	Freq	Rel. Freq	Cum. Freq	Cum. Rel. Freq
F	1	.05	1	.05
D	3	.15	4	.20
C	7	.35	11	.55
B	6	.30	17	.85
A	3	.15	20	1.00
Total	**20**	**1.00**		

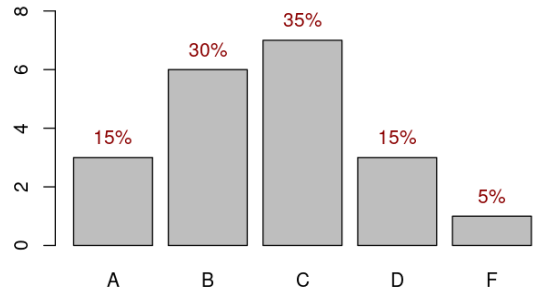
Remarks: (1). The cumulative categories are defined to be F, D or below, C or below, etc. (2). For a nominal data, the cumulative frequency table may not be practically meaningful because the combined categories may not make practical sense.

2.2.2 Charting Categorical Data

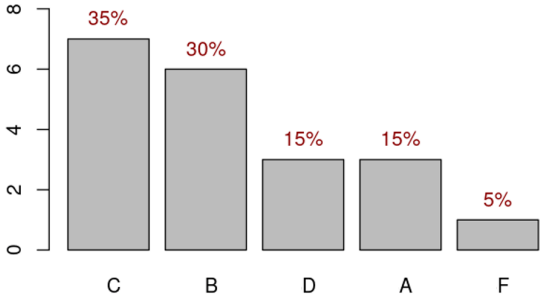
Two major charts are used to characterize the distribution of a given categorical data set: bar chart and pie chart. Both chart are geometric representations of the frequency table discussed earlier.

2.2.2.1 Bar Chart

Example 2: We convert the frequency table of the course grade data in the following



Remark: We can rearrange the vertical bars in ascending or descending order to get pare-to chart.



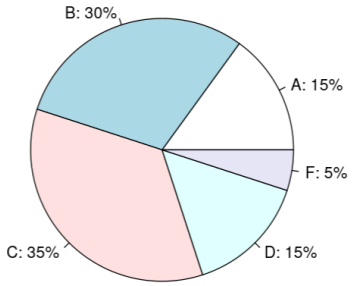
2.2.2.2 Pie Chart

To construct a pie chart manually, we need to calculate the degrees of the central angle of the circle and then slice it based on the degrees of the central angle.

Example 3: we still use the course grade frequency table (relative frequency) to to calculate the degrees of the corresponding central angle in the following table.

Grade	Relative Freq	Pie Chart Angle
F	.05	$0.05 \times 360^{o} = 18^{o}$
D	.15	$0.15 \times 360^{o} = 54^{o}$
C	.35	126^{o}
B	.30	108^{o}
A	.15	54^{o}
Total	1.00	360^{o}

The corresponding pie chart is given by



2.3 Summary of Numerical Data

There are primarily two methods that are commonly used to summarize a given numerical data: Frequency tables and histograms.

2.3.1 Frequency Table

A histogram displays numerical data by grouping data into "bins" of equal width. Each bin is plotted as a bar whose height corresponds to how many data points are in that bin. Bins are also sometimes called "intervals", "classes", or "buckets".

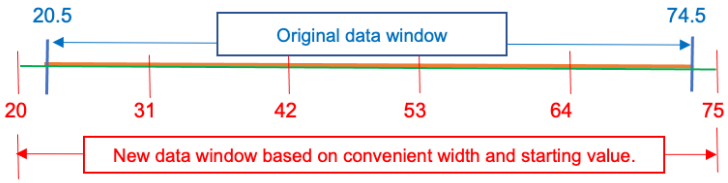
There are several steps to follow when creating bins (with equal width):

- Determine the number of bins
- Extend data window (from the smallest to the largest data values) if necessary to get "convenient" end values of the extended window. Caution: never shrink the data window because we must include all data values in one and only one of the bins!
- Find the boundary values (cut-offs) so that all bins have equal width which is equal to **data-window-width/number-of-bins**.
- The number of data values in each bin is the frequency of the bin.

Example 4 - Length of CD: Listed below are the lengths (in minutes) of randomly selected CDs of country, rock, and movie soundtracks.

`20.5, 29, 32, 32, 32, 33, 36, 37, 38, 39, 39, 43, 47, 48, 49, 49, 49, 50, 50, 51, 51, 52, 52, 52, 53, 54, 54, 54, 56, 56, 57, 58, 60, 61, 62, 62, 69, 73, 74, 74.5`

Solution: we follow the above suggested steps to define bins illustrated in the following figure.



- The number of bins chosen for this frequency table is 5.
- The original data window is '[20.5, 74.5]'. The two end values are decimals. We extended the data window on both sides and get a an extended window [20, 75].
- The bin width = (75-20)/5 = 11.
- The boundaries of the 5 bins are: 20, 31, 42, 53, 64, 75.
- the five bins are: [20, 31], (31, 42], (42, 53], (53, 64], (64, 75]. Note that the boundary values must be included in one and only one bins. We use `"["` and `")"` to denote inclusion and exclusion respectively. For example, in the second bin [31, 42), 31 is included in [31, 42) but 42 is NOT included in [31, 42). It is included in [42, 53).
- with the above defined bins, the resulting frequency tables are given by

Classes	Freq	Cum. Freq	Rel. Freq	Cum. Rel. Freq.
[20, 31]	2	2	0.05	0.05
(31, 42]	9	11	0.225	0.275
(42, 53]	14	25	0.35	0.625
(53, 64]	11	36	0.275	0.9
(64, 75]	4	40	0.1	1.0

2.3.2 Histogram

Similar to the bar chart and pie chat, a histogram is also a geometric representation of the frequency table constructed above. Since the bins are defined based on the numerical boundaries, they must be placed on the correct scales when constructing a histogram. This also means that the histogram is different from the bar chart in different perspectives:

- There is no gaps between the adjacent vertical bars because the horizontal axis is a numerical axis.
- We cannot rearrange the vertical bars as we did in a bar chart to make a pareto chart since we cannot shuffle the boundaries on the numerical axis.

Example 5 - Length of CD (cont'd): The histogram based on the frequency table is given in the following.

2.4 Exercises

Summarize the following data sets by using frequency tables (relative frequency, cumulative frequency, etc.) and histogram. You can use `IntroStatsApps` to check your work.

Exercise 1. Weights of 18- to 24- Year- Old Males. The U. S. National Center for Health Statistics publishes data on weights and heights by age and sex in the document Vital and Health Statistics. The weights shown in the following, given to the nearest tenth of a pound, were obtained from a sample of 18- to 24- year- old males. Use the cut-point grouping to organize these data into frequency and relative- frequency distributions. Use a class width of 20 and a first cut-point of 120.

```
129.2, 132.1, 136.7, 142.8, 145.6, 146.4, 149.9, 150.7, 151.3, 155.2,
158.5, 158.6, 161.0, 161.7, 165.0, 165.8, 167.3, 170.0, 170.1, 172.5,
173.6, 173.7, 175.4, 175.6, 178.2, 178.7, 182.0, 182.5, 185.3, 187.0,
187.5, 188.7, 191.1, 209.1, 214.6, 218.1, 278.8
```

Exercise 2. The following are the miles per gallon.

```
22.8, 22.9, 23.3, 23.4, 23.6, 23.7, 23.8, 23.9, 23.9, 24.1, 24.1, 24.2,
24.3, 24.4, 24.5, 24.5, 24.6, 24.6, 24.7, 24.7, 24.7, 24.8, 24.8, 24.9,
24.9, 25.0, 25.0, 25.1, 25.2, 25.3
```

Exercise 3. Following are 80 measurements of the iron-solution index of tin-plate specimens, designed to measure the corrosion resistance of tin-plated steel. The original data set has been sorted in an ascending order as:

```
14,  26,  28,  28,  28,  28,  30,  32,  34,  35,  36,  36,  37,  37,  40,
40,  40,  41,  41,  41,  42,  42,  42,  43,  43,  43,  44,  44,  44,  44,
45, 45,  45,  45,  45,  45,  46,  46,  46,  46,  47,  47,  47,  48,  49,
49,  49, 50,  50,  50,  51,  52,  52,  52,  52,  52,  53,  53,  54,  54,
54,  54,  55, 55,  55,  55,  55,  55,  55,  56,  56,  56,  56,  56,  56,
57,  57,  57,  57, 57,  58,  58,  58,  58,  58,  59,  59,  60,  60,  60,
60,  61,  61,  61,  61, 61,  62,  62,  62,  62,  62,  62,  63,  63,  63,
63,  63,  63,  64,  65,  66, 66,  67,  68,  68,  69,  69,  70,  70,  70,
70,  70,  70,  71,  71,  72,  72, 72,  73,  74,  74,  74,  76,  76,  77,
77,  79,  80,  81,  81,  83,  83,  84, 86,  86,  86,  87,  89,  92,  95
```

Exercises 4. From the 140 children whose urinary concentration of lead were investigated 40 were chosen who were aged at least 1 year but under 5 years. The following concentrations of copper were found.

```
0.70, 0.45, 0.72, 0.30, 1.16, 0.69, 0.83, 0.74, 1.24, 0.77, 0.65, 0.76,
0.42, 0.94, 0.36, 0.98, 0.64, 0.90, 0.63, 0.55, 0.78, 0.10, 0.52, 0.42,
```

`0.58, 0.62, 1.12, 0.86, 0.74, 1.04, 0.65, 0.66, 0.81, 0.48, 0.85, 0.75, 0.73, 0.50, 0.34, 0.88`

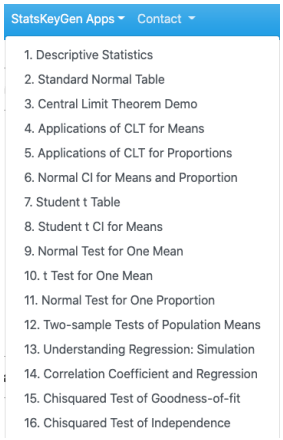
Exercise 5 The following data set represents the shoe sizes of 100 random selected students from a large university.

`8.0, 13.0, 8.5, 9.0, 11.0, 9.5, 10.0, 8.0, 11.0, 8.0, 10.0, 11.0, 10.0, 11.0, 6.0, 9.0, 8.0, 8.0, 12.0, 10.5, 9.5, 11.0, 6.0, 8.0, 10.0, 11.5, 11.0, 7.0, 10.5, 15.0, 12.0, 8.5, 8.0, 10.0, 8.0, 7.0, 10.5, 10.0, 5.0, 7.0, 10.0, 14.0, 14.0, 8.5, 8.0, 13.0, 11.0, 6.0, 8.0, 11.5, 8.5, 7.0, 12.5, 8.5, 15.0, 10.0, 6.0, 11.0, 11.0, 10.0, 10.5, 11.0, 7.5, 7.0, 7.5, 10.5, 10.0, 11.0, 9.5, 11.0, 9.5, 10.5, 7.5, 11.0, 13.0, 10.0, 9.0, 12.0, 8.0, 8.0, 9.0, 12.0, 8.5, 8.0, 11.0, 9.0, 9.0, 7.0, 9.0, 12.0, 5.5, 9.5, 8.0, 9.0, 12.0, 9.5, 9.0, 11.0, 13.0, 7.5`

`Caution`: The data values are given in the numeric form, but they are labels shoe sizes. Therefore, this is a categorical data set.

2.5 Use of Technology

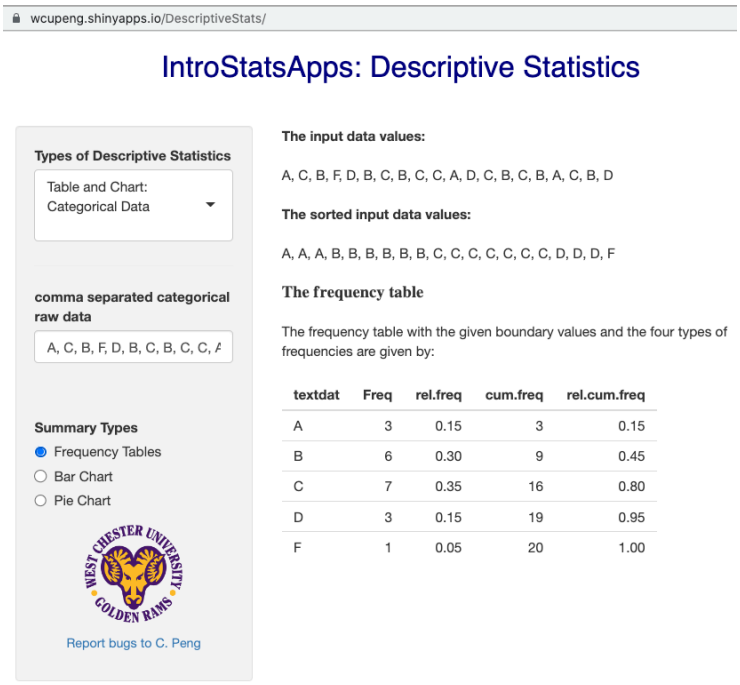
In this class, I created several StatsApps to facilitate your learning of statistical concepts. The following screenshot gives the list of the available apps that cover about 90% topics in this class.



Next we use the first App to produce the answers to the two examples that we used in this note. The App can found at https://wcupeng.shinyapps.io/DescriptiveStats/. You need to type in data values and relevant information to generate frequency tables and charts.

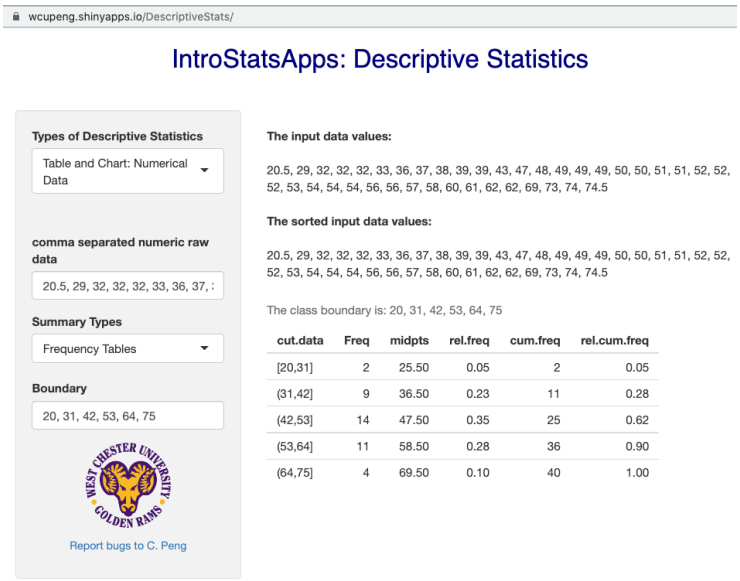
2.5.1 Course Grade Data

The following show the frequency table generated from the App using the course letter grades. You can try to generate bar charts and pie charts using this app.



2.5.2 Length of CD data

The following show the frequency table generated from the App using the length of CD data in the example. You can try to generate histograms using this app. Please keep in mind the boundary values MUST be provided to obtain the appropriate frequency tables and histograms.



cut.data	Freq	midpts	rel.freq	cum.freq	rel.cum.freq
[20,31]	2	25.50	0.05	2	0.05
(31,42]	9	36.50	0.23	11	0.28
(42,53]	14	47.50	0.35	25	0.62
(53,64]	11	58.50	0.28	36	0.90
(64,75]	4	69.50	0.10	40	1.00

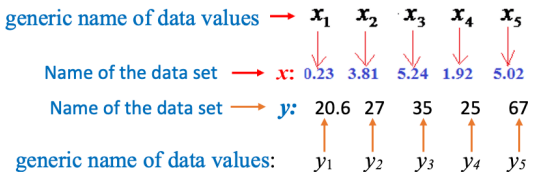
Chapter 3

Numerical Measures

This note focuses on using numerical measures to characterize numerical data sets. The numerical measures are used to describe the features such as mean, variance, and percentiles of a given numerical data set. These numeric measures are classified into three categories: central tendency, variation, and locations.

3.1 Notations Using Greek Letters for Parameters

Every data set has a name. For example, the set of heights of a group of WCU students is a data set that can be named `h` or `height`. We can give each data value has a "generic name" such as $h_1, h_2, \cdots, h_5$, etc. The following figure gives other examples of generic names of values in different data sets.



3.1.1 Big Sigma (Σ) Notation

These "generic names" were used to make compact formulas in some numeric measures. The sum of all data values in the data set with the name x (in the above figure) is given by the following **big sigma** notation.

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

$$= 0.23 + 3.81 + 5.24 + 1.92 + 5.02 = 16.22$$

Example 1: Consider the following two data sets with names **x** and **y**, we want to take the product of the corresponding values and sum up the product of the corresponding values. The following is the **big sigma** notation of the **sum of the cross-product**.

x	1.1	1.2	1.3	1.4	1.5	1.6
y	5	4	3	2	1	0

$$\sum_{i=1}^{4} x_i\, y_i = x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4$$
$$= 1.1 \times 5 + 1.2 \times 4 + 1.3 \times 3 + 1.4 \times 2 = 17$$

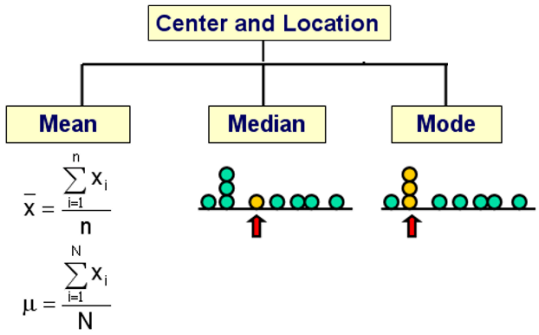
3.1.2 Notations for Parameters and Statistics

We use Greek letters to denote the population parameters of populations and English letters to denote statistics from the samples.

	Parameter (Always about population)	**Statistics** (Always about sample)
Definition	A number that describes the population characteristic.	A number that describes the sample characteristic.
Example	The average of PA population is 40 as of 2020. **Parameter value**: $\mu = 40$	A random sample of 5000 Pennsylvanians was collected and found the average age to be 39.5 **Sample value**: $\bar{x} = 39.5$

3.2 Measures of Center

Three measures are used as the center of a given numeric data set.



3.2.1 Mean

The mean of a given data set is defined as the average of all data values. The big sigma notations of sample and population means are given by

The Mean is the average of data values

Sample mean — n = Sample Size

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Population mean — N = Population Size

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Remark: the mean can be affected significantly by outliers (extreme values). For example,

3.2.2 Median

The middle value of a sorted data set is called the median of the data set.

- If a data set has an **odd number** of data values, there is a unique "middle" value in the **sorted data** set.
- If a data set has an **even number** of data values, there will be two **middle** values in the **sorted data** set, in this case, the **average** of the two "middle** values is defined to be the median.

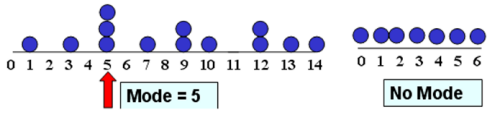
For example,

- {2, 6, 7} → median = 6
- {1, 2, 6, 7} → median = (2 + 6) / 2 = 4.

3.2.3 Mode

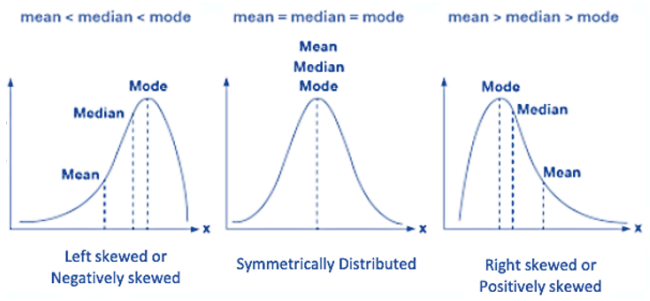
The mode(s) is (are) the data value(s) with highest frequency.

- If there is only one mode, the data set is unimodal.
- If there are two modes, the data set is bimodal.
- If there are more than two modes, the data is multi-modal.



3.2.4 Relationship between Mean, Median, and Mode

The relationship between mean, median, and mode is dependent on the shape of the distribution. The following figure illustrates this relationship.



3.3 Measures of Variation

Measures of variation are used to characterize the shape of the distribution. There are some different measures used in different situations. We only introduce the variance and the standard deviation in this course. We will also briefly introduce IQR in the applications of numerical measures.

3.3.1 Variance

Since the definitions of sample and population variances are different, we need to choose an appropriate formula based on whether the data set is a population or a sample. This information is provided to you before you select a formula to calculate the variances. The exact definitions using big sigma notation are given below

- Population Variance

$$\sigma^2 = \frac{\Sigma_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}$$

- Sample Variance

$$s^2 = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

We can see the only difference is in the denominator of the two definitions.

3.3.2 Standard Deviation

Once the variance is calculated, we simply take the square root to obtain the standard deviation

- Population standard deviation

$$\sigma = \sqrt{\frac{\Sigma_{i=1}^{N}(x_i - \mu)^2}{N}} = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}}$$

- Sample Standard Deviation

$$s = \sqrt{\frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}$$

3.3.3 Steps for Calculating Variance

The following are steps for calculating the variance of a data set.

Step 1: Compute the sample mean $\bar{x}$.

Step 2: For each sample value x, compute the difference $x - \bar{x}$. This quantity is called a deviation.

Step 3: Square the deviations, to obtain quantities $(x - \bar{x})^2$.

Step 4: Sum the squared deviations, obtaining $\Sigma (x - \bar{x})^2$.

Step 5: Divide the sum obtained in Step 4 by $n - 1$ to obtain the sample variance s^2.

Example 2 The following table illustrates how to use the above steps to calculate the variance of a small **sample** toy data set: A = {1,4,7}.

X (dataset)	Deviation: $X - \bar{X}$	Squared Deviation: $(X - \bar{X})^2$
1	$1 - 4 = -3$	$(-3)^2 = 9$
4	$4 - 4 = 0$	$0^2 = 0$
7	$7 - 4 = 3$	$3^2 = 9$
$\bar{X} = 12/3$		$s^2 = \frac{18}{3-1} = 9$

Based on the above table, we can see that the standard deviation is $\sqrt{9} = 3$.

3.4 Measures of Location

Two important types of measures of location will be introduced in this course: z-score and percentiles.

3.5 z-score

A Z-score of a value of a **sample** data set is a standardized score that is defined by

$$z = \frac{x - \bar{x}}{s}.$$

We can easily adjust the above formula for a population as

$$z = \frac{x - \mu}{\sigma}.$$

Example 3: We still use the same sample toy data, A = {1,4,7}, used in **Example 2** to illustrate how to find z-scores of corresponding data values.

Solution: We know from **Example 2** that $\bar{x} = 4$ and $s = 3$. Therefore, the z-scores of the corresponding data values are calculated in the following.

$x_1 = 1 \rightarrow z_1 = \frac{1-4}{3} = -1$

$x_2 = 4 \rightarrow z_2 = \frac{4-4}{3} = 0$

$x_3 = 7 \rightarrow z_3 = \frac{7-4}{3} = 1$

That is, the standardized set of z-scores is $\{-1, 0, 1\}$. Note this is a set of sample z-scores. We can easily verify that the mean and standard deviation of the above three z-scores are 0 and 1 respectively.

3.5.1 Percentile

A **percentile** indicates the percentage of scores that fall below a particular value.

Example 4 Consider the following PSAT percentile table.

PSAT PERCENTILES

PSAT SECTION SCORE	EVIDENCE-BASED READING & WRITING PERCENTILE	MATH PERCENTILE
760	99+	99+
750	99+	99
740	99	98
730	99	97
720	98	97
710	97	96
700	96	95
690	95	94
680	94	93
670	93	92
660	91	91
650	90	90
640	88	89
630	86	87
620	84	85

If you took PSAT and scored 640 in the MATH section, according to the above table, your MATH percentile is 89% meaning that 89% of all examinees scored below 640 in the PSAT. This also means that you did better than 89% of your peers on the PSAT.

Steps for Calculating Percentiles

Assume that we have a data set $\{x_1, x_2, \cdots, x_n\}$. we want to find $k-th$ percentile, denoted by P_k.

Step 1: Sort the data in ascending order: $\{x_{(1)}, x_{(2)}, \cdots, x_{(n)}\}$.

Step 2: Calculate the `rough location` of the $k-th$ percentile

$$L = \frac{k}{100} \times n.$$

Step 3: The k^{th} percentile is obtained depending on the form of L.

- if L is a whole number, then the k^{th} percentile is the average of the number in position L and the number in position $L+1$ in the sorted data set.
- if L is NOT a whole number, round it **up** to the next higher whole number. The k^{th} percentile is the number in the position corresponding to the **rounded-up** value.

Example 5: Consider the following data set

```
9, 13, 7, 7, 12, 15, 10, 10, 6, 19, 17, 10, 15, 9, 14, 12, 9, 13, 7,
7, 4, 8, 19, 5, 18, 20, 14, 1, 23, 10, 10, 7, 22, 9, 1
```

Find P_{40} and P_{55} percentiles respectively.

Solution we first sort the data in ascending order.

```
1, 1, 4, 5, 6, 7, 7, 7, 7, 7, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10, 12, 12,
13, 13, 14, 14, 15, 15, 17, 18, 19, 19, 20, 22, 23
```

To find 40th percentile,

$$L = \frac{40}{100} \times 35 = 14.$$

Since $L = 14$ is an integer, the 20th percentile is the average of 14th and 15th data values in the sorted data set. That is, $(9 + 9)/2 = 9$.

To find 55th percentile,

$$L = \frac{55}{100} \times 35 = 19.25.$$

Since $L = 19.25$ is NOT an integer, we round up L to get **20**. The 55th percentile is the 20th data value in the sorted data set which is 10.

3.5.2 Applications of Numeric Measures

Three concepts based on the numeric measures will be introduced in the following.

3.5.2.1 Five Number Summary

The five number summary consists of the minimum, 25th, 50th, 75th percentiles, and the maximum. The 25th, 50th, and 75th percentiles are also called the first (Q_1), second (Q_2) and third quartiles (Q_3), respectively.

Example 6: We use the **length of CD** data to show the five-number-summary. The unit of data values is minute. The following is the sorted data.

```
20.5, 29, 32, 32, 32, 33, 36, 37, 38, 39, 39, 43, 47, 48, 49, 49, 49,
50, 50, 51, 51, 52, 52, 52, 53, 54, 54, 54, 56, 56, 57, 58, 60, 61, 62,
62, 69, 73, 74, 74.5
```

Solution, The minimum and maximum are 20.5 and 74.5 minutes. The quartiles are calculated in the following.

$Q_1 : L = (25/100) \times 40 = 10$, Q_1 is the average of the 10th and 11th data values $= (39 + 39)/2 = 39$.

$Q_2 : L = (50/100) \times 40 = 20$, Q_1 is the average of the 20th and 21st data values $= (51 + 51)/2 = 51$.

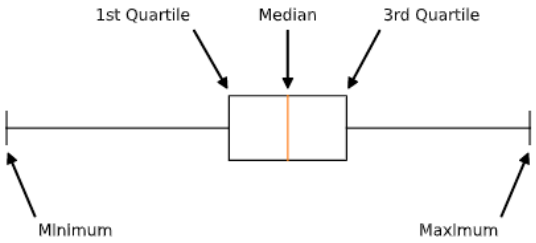
$Q_3 : L = (75/100) \times 40 = 30$, Q_1 is the average of the 30th and 31st data values $= (56 + 57)/2 = 56.5$.

Therefore, the five-number-summary is given by

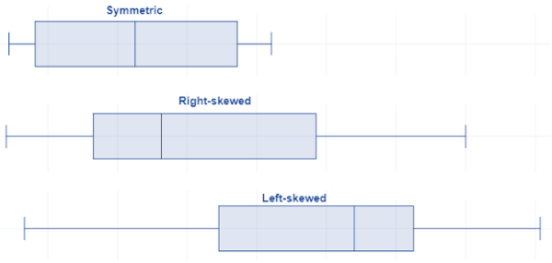
Min	Q1	Q2	Q3	Max
20.5	39	51	56.5	74.5

3.5.2.2 Box-plot

The box-plot is a geometric representation of the five-number-summary given in the following figure



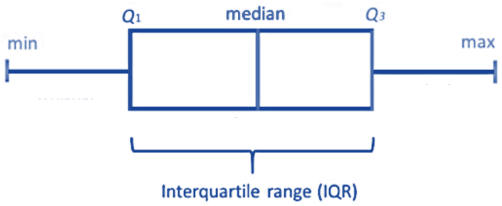
Box-plots are used to describe the distribution of data. The following three box-plots represent three different types of distributions:



Example 7: (Length of CS Continued) The box-plot is given by

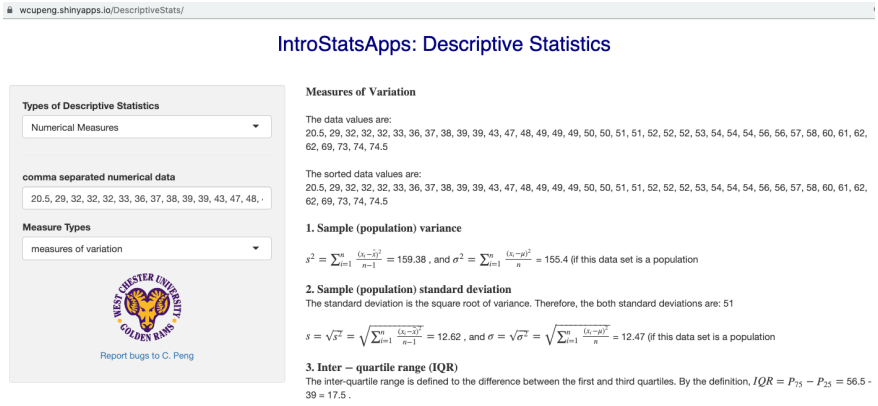
3.5.2.3 Inter-quartile Range (IQR)

The inter-quartile range of data is defined by $IQR = Q3 - Q1$. IQR is used to measure the variation of the data set. It is NOT sensitive to extremely large and small values since IQR is defined only based on the "middle 50% of data values".



3.6 Use of Technology

We still use the same app that we used in the previous note to find various numerical measures to summarize a given data set. The app can be found at https://wcupeng.shinyapps.io/DescriptiveStats/. The next screenshot illustrates the measures of variations in the length of CD data. You can choose to find other numerical measures of the data.



Chapter 4

Concepts of Probability

> **Q**: Why do we need to study probability? **A**: We can build a bridge between a sample and its population so that one can make inferences about the population from its samples.

In this note, we introduce the definitions of probability, random variables, and problems associated with a distribution based on a special distribution: uniform distribution.

4.1 Definitions of Probability

We need to use the following concepts to define or approximate the probability of an event.

4.1.1 Experiment and Event

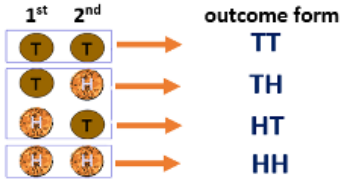
An experiment is any process that generates well-defined outcomes. On any single repetition of an experiment, one and only one of the possible experimental outcomes will occur.

Example 1. The following table gives several examples of experiments.

Experiments	*Experimental Outcomes*
Toss a coin	Heads, Tails
Select a part for inspection	Defective, non-defective
Conduct a sales call	Purchase, no purchase
Roll a dice (six-sided)	1, 2, 3, 4, 5, 6
Play a football game	win, lose, tie

The Sample Space of an experiment is the set of all possible outcomes. We usually use a capital letter in Greek or English to denote the sample space. For example, Ω, S, or U.

Example 2. Consider the experiment of tossing a balanced coin sequentially for 2 times. We use H to denote the heads and T to denote the tails.



Therefore, the sample space of this experiment is $\Omega = \{TT, TH, HT, HH\}$.

An Event is a subset of outcomes from the sample space. We usually use a capital letter to denote an event.

Example 3. We define Event E to be at least one H is observed in the coin tossing example.

Solution: The outcomes in the subset $\{HT, TH, HH\}$ satisfy the requirement in the definition. Therefore, event E is defined by $E = \{HT, TH, HH\}$.

4.2 Definition of Probabilities

Notations: We use P or Pr to denote the probability and $A, B, C, \cdot$, to denote specific events. $P(A)$ is the probability that event A occurs.

4.2.1 Definitions of Probability

- **Method 1:** Relative Frequency Approximation of Probability.

Conduct or observe an experiment a large number of times and count the number of times that event A occurs. Based on these actual results, P(A) is estimated as follows:

```
P(A)= (# of times of A being observed) / (# of repeated trials)
```

- **Method 2:** Classical Approach to Probability (Requires Equally Likely Outcomes of the experiment)

```
P(A) = (# of outcomes in A) / (# of outcome in S)
```

4.2.2 Some Examples

Example 4. **Guessing Answers on an ACT:** A typical multiple-choice question has 5 possible answers. If you make a random guess on one question, what is the probability that your response is wrong?

Solution: Method 2 classical approach applies to this question since each of the choices is random, that is, each of the 5 letters A, B, C, D, and E are equally likely to be selected as the correct answer.

```
Pr(wrong answer) = 4 / 5 =0.8
```

Example 5. **Gender of Children**: Find the probability that a randomly selected couple with 3 children will have exactly 2 boys. Assume that boys and girls are equally likely, and the gender of any child is not influenced by the gender of any other child.

Solution Based on the given condition, this is an equally likely experiment. We first use a table to list all possible outcomes in the sample space. Let B = boy and G = girl. Then the following table lists all possible outcomes of this experiment.

1st	2nd	3rd		Outcome form
B	B	B	→	BBB
B	B	G	→	BBG
B	G	B	→	BGB
G	B	B	→	GBB
G	G	B	→	GGB
G	B	G	→	GBG
B	G	G	→	BGG
G	G	G	→	GGG

We list the sample space as follows

```
S = { BBB   BBG   BGB   GBB   GGB   GBG   BGG   GGG }
E =  having exactly 2 boys = { BBG   BGB   GBB }.
```

The desired probability can be calculated as

```
Pr(E)= #E/#S  = 3 / 8 =0.375.
```

4.2.3 Discrete Probabilities on Contingency Tables

Contingency tables classify outcomes in rows and columns. Table cells at the intersections of rows and columns indicate frequencies of both events coinciding.

Example 6. The following table displays events for computer sales at a fictional store.

	PC	Mac	Row Totals
Male	66	40	106
Female	30	87	117
Column Totals	96	127	223

Specifically, it describes the frequencies of sales by the customer's gender and the type of computer purchased. The cells' counts represent the number of PCs and Macs purchased by both genders.

Finding the following probabilities:

(1). Randomly select a female customer, what is the probability that she bought a Mac computer?

(2). Randomly select a customer from those who bought a computer from the store, what is the probability that the customer is a male?

(3). Randomly select a customer who bought a computer from the store, what is the probability she/he bought a Windows computer?

(4). Randomly select a customer who bought a computer from the store, what is the probability that the customer is female and bought a Mac computer?

Solution (1). $P(\text{Mac among female}) = 87/117 = 0.744$

(2). $P(\text{male among customers who bought a computer}) = 106/223 = 0.4753$

(3). $P(\text{windows among all computers sold}) = 96/223 = 0.4305$

(4). $P(\text{female \& bought a mac computer}) = 87/223 = 0.3901$

4.3 Concepts of Random Variables

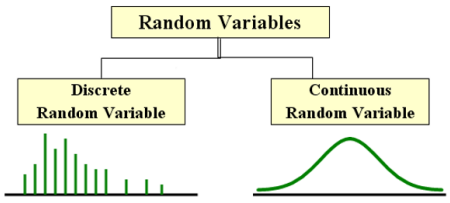
Random Variable (informal description): A random variable is a numerical description of the outcome of an experiment, its value depends on chance.

Example 7. Let Y be a variable denoting the height of WCU students. Before you measure the height of a selected student, Y is unknown. Furthermore, the unknown value is dependent on the chance since selecting a student involves uncertainty! This type of variable is a random variable.

Characterization of Random Variables: We study the behavior of a random variable through its probability distribution function.

Use of Random Variables: We use random variables to characterize the random behavior of real-world problems to inform statistical decisions using probability distributions.

Classification of Random Variables



4.3.1 Discrete Random Variable

Discrete Random Variable: A discrete random variable may assume either a finite number of values or an infinite sequence of values such as 1, 2, ..., etc.

Example 8. The following table gives several example of discrete random variables.

Experiment	Random Variable (x)	Possible values for the RV.
1. Inspect a shipment 50 radios	The number of defect radios	0, 1, ..., 50
2. Operate a restaurant for one day.	Number of customers	0, 1, ...
3. Sell an automobile	Gender of customer	0 if male, 1 if female.

We use the probability distribution to characterize a random variable. For convenience, we use a capital letter (such as X) to denote the name of the random variable and a lower-case letter (x) to denote the value of the random variable,

The probability distribution of a discrete random variable is a graph, table, or formula that gives the probability for each value of the random variable. For a probability distribution, the following requirements MUST be satisfied.

1. $\sum P(x) = 1$, where x assumes all possible values of random variable X.
2. $0 \leq P(x) \leq 1$, for every value of x.

Defining Events Based on Values of A Random Variable

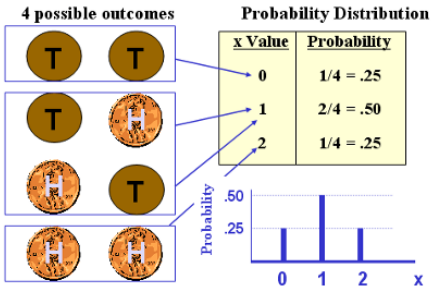
An event defined based on a discrete random variable is a subset of the collection of all possible values of the discrete random variable!

Example 9.. Let's consider the experiment of tossing a balanced coin two times. The sample space $\Omega = \{TT, TH, HT, HH\}$. We can define a random variable X = number of **heads** observed in an experiment.

As an illustration, we define several events based on the above random variable based on various values in the following.

```
E1 = {X = 1} = (TH, HT) ,  then   P(E1) = 2/4 = 0.5
E2= {X>0} ={TH, HT, HH},   then  P(E2) = 3/4
E3 = {X !=1} = {HH, TT},   then    P(E3) = 2/4 = 0.5
```

The complete probability distribution is outlined in the following figure.



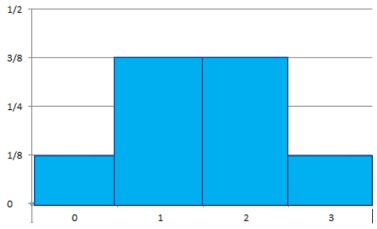
The top-right table is the probability distribution table, and the bottom right is the probability distribution table (geometric representation of the probability distribution table).

Example 10.. **Gender of Children** If a couple plans to have 3 children. Let X = the number of sons they may have. What is the probability distribution of X?

Solution: The possible values of the R.V. X are 0, 1, 2, and 3 which correspond to the four events. Using the definition of class probability, we have Pr(X =0) = 1/8, Pr(X = 1) = 3/8, Pr(X = 2) = 3/8 and Pr(X = 3) = 1/8. So, we have the following probability distribution table.

R.V. X	Pr(X = x)
0	1/8
1	3/8
2	3/8
3	1/8

The probability distribution histogram is given by



4.3.2 Continuous Random Variable and Density Functions

A continuous random variable takes on uncountably infinite values. For example, body temperatures, height, weight, etc.

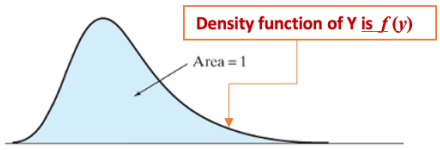
As opposed to the discrete variables that could be characterized by a probability

table or a probability function, for a continuous random variable, we use the probability density function to characterize the distribution.

The probability density function of Y, denoted by $f(y)$, MUST satisfy the following two conditions that are analogous to the two requirements in the probability distribution function:

(1) a non-negative real-valued function is defined on all the real numbers, and such that

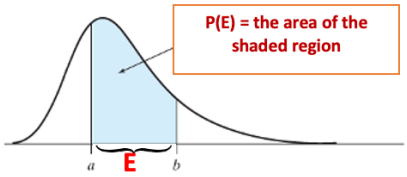
(2) the area between the curve and the horizontal axis is equal to 1.



Definition of Events Using Continuous Variables

An event defined based on a continuous random variable is one or more intervals of values of the continuous random variable. The probability of the event is equal to the area between the density curve and the interval(s).

The following figure illustrates the definition of an event and its corresponding probability.



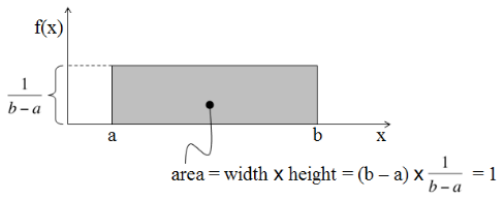
Since the `area` of a line segment is zero, the probability that a continuous random variable exactly equals any value is always zero.

4.4 Uniform Random Variable - A Special Continuous Distribution

A uniform distribution, also called a rectangular distribution, is a probability distribution that has constant probability. The general formula for the probability density function (pdf) for the uniform distribution is

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & elsewhere \end{cases}$$

The density curve is given by



Example 11.. Consider the following function defined on interval [1,9].

$$f(x) = \begin{cases} 1/8 & \text{if } 1 < x < 9 \\ 0 & \text{otherwise} \end{cases}$$

Is $f(x)$ a valid density function of the uniform distribution?

Solution: We only need to check whether the two requirements of continuous density functions are satisfied.

1. $f(x) \geq 0$ for all x;
2. The area between the density curve and the horizontal axis is equal to 1, This is obvious from the following density curve.

Therefore,

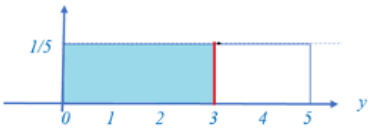
is a valid density function.

Example 11.. A shuttle train at a busy airport completes a circuit between two terminals every five minutes. Answer the following questions with the assumption that passengers arrive at a stop at any time in a 5-minute time window (we could assume a uniform arrival time).

1) What is the probability that a passenger will wait for less than three minutes for the train?
2) What is the probability that a passenger will wait for exactly three minutes for the train?
3) What is the probability that a passenger will wait for more than three minutes for a train?

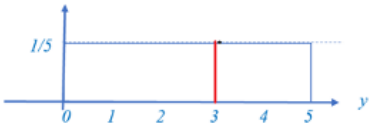
Solution: The brief answers to these questions are given below.

1).



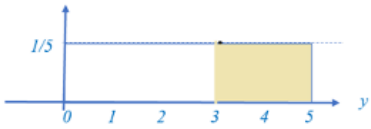
$$P(y < 3) = (3 - 0) \times \frac{1}{5} = \frac{3}{5} = 0.6.$$

2).



$$P(y = 3) = (3 - 3) \times \frac{1}{5} = \frac{0}{5} = 0.$$

3).



$$P(y > 3) = (5 - 3) \times \frac{1}{5} = \frac{2}{5} = 0.5.$$

Chapter 5

Normal Distributions

The standard normal distribution is the most frequently used distribution in this course. We will introduce the general normal distributions and focus on two basic questions: Finding probabilities and percentiles using the standard normal table.

5.1 Standard Normal Distribution

The standard normal distribution, Z, has a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$. Its probability density curve is

Two basic types of questions need to be answered for any distribution including the standard normal distribution:

1. Finding probability of events such as P(Z < a), P(Z > c), P(a < Z < b), etc.
2. Finding percentiles. For example, finding z_0 for given P(Z < z_0) = 0.90.

We have discussed how to find probabilities from the uniform distribution in topic #3 whose density curve is a rectangle. The probabilities of events defined based on uniform distributions are the areas of rectangles using the area formula of a rectangle. We still need to find the probability of events defined based on the standard normal distribution like P($-0.86 < Z < 0$) which is still the area of the following shaded region (as outlined in the previous note for any general distribution).

5.1.1 Finding Probabilities

Unlike uniform distributions whose density curves are rectangles, we can use the formula to calculate the areas of rectangles. In a standard normal distribution, there is no formula to calculate the area of the shaded irregular region in the above figure.

We will use a standard normal distribution table to find the area of the left-hand side tail regions shown below (part of the table).

Before doing examples, we point out the basic facts of the standard normal distribution.

1. The density curve is symmetric concerning the vertical axis.
2. The area between the curve and the horizontal axis is equal to 1. This means that the areas of the left and right regions are equal to 0.5.

Next, we use several examples to illustrate how to use the standard normal table to find the areas of different regions defined based on the standard normal distribution.

Example 1. Find the probabilities indicated, where as always Z denotes a standard normal random variable.

1). P(Z < - 1.48).

2). P(Z < 0.25).

Solution. First of all, we only keep two decimal places for the z value (also called z-score). When using the table, we first locate the integral part and the first decimal place of the z score in the **first column** and the second decimal place in the **top row**. The **left tail area** is on the interaction of the aforementioned row and column. This is explained in the following figure.

Example 2. Find the probabilities indicated, where as always Z denotes a standard normal random variable.

1). P(Z > - 1.96).

2). P(Z > 0.75).

Solution: The probabilities to be found represent the areas of right tail regions. We can use the table to find the area of the left tail region and then subtract it from 1 to get the desired probability. The following figure illustrates the idea to get the "right-tail" probabilities.

Example 3. Find the probabilities indicated, where as always Z denotes a standard normal random variable.

1). P(-1.96 < Z < 0.75).

Solution: The general idea is to find the two **left tail areas** and then take the difference to get the area of the region defined by the two z-scores as shown in

the following figure.

For 1), the probability is found in the following figure.

Summary of Finding Probabilities

5.1.2 Finding Percentiles

We have introduced how to find a percentile from a given data set. We basically do the same thing for the standard normal distribution.

Recall that the q-the percentile is the cut-off value such that 100q% data values are less than or equal to the cut-off value (q-th percentile). This means, we find a percentile, and the left tail area is always given. The general formulation of the problem is to find the cut-off k that satisfies

$$P(Z < k) = 0.90,$$

for a given q (such as 90%, etc). This is depicted in the following figure.

The process of finding a percentile is the opposite of the process of finding probability. If the given left tail probability itself is in the main body of the table, we then locate the row and the column to find the z-score (i.e., the percentile).

In general, the given left tail probability is not in the table but is closest to two values in the main boy of the table. Each of the two closed table values corresponds to a z-score. The average of the two z-scores is defined to be the desired percentile.

Example 4. Find 90th percentile of the standard normal distribution.

Solution: We go to the normal table and find two values in the main body of the table that is closest to 0.9 (see the figure below).

Example 5. **The Precision Scientific Instrument Company** manufactures thermometers that are supposed to give readings of 0^oC at the freezing point of water. Tests on a large sample of these instruments reveal that the freezing point of water is around zero (some thermometers give positive degrees, some thermometers give negative degrees), Assume that the mean reading is 0^oC and the standard deviation of the readings is 1.00^oC. Assume further that the readings are normally distributed.

1. Find the probability that, at the freezing point of water, the reading is between 0^oC and 1.58^oC.

2. Find the probability that the reading is between –2.43^oC and 0^oC.
3. Find the probability that the reading is between 0.5^oC and 2.5^oC.
4. Find the probability that the reading is between –1^oC and –2.5^oC.
5. Find the probability that the reading is between –1.5^oC and 1^oC.
6. Find the probability that the reading is exactly 0^oC.
7. Find the temperature z corresponding to P_{95}, the 95th percentile (95% of the readings less than z and 5% of the readings are greater than z).
8. Fin the 10th percentile.

Solution: Based on the given information, the thermometer readings follow the standard normal distribution. The standard normal distribution table will be used to answer the above questions. We only do questions 5 (finding probability) and 7 (finding percentile) to work and leave the rest of the questions to you to practice.

5). P(-1.5 < Z < 0) = P(Z < 0) - P(Z < -1.5) = 0.5 - P(Z < -1.5) = 0.5 - 0.0668 = 0.4332

7). We want to find P_{95}, or equivalently, to find k from $P(Z < k) = 0.95$. We can see from the normal table that 0.9495 and 0.9505 are the two values that are closest to 0.95. The two corresponding z-scores are 1.64 and 1.65. By the convention, the 95th percentile is the average of the two z-scores (see the figure below).

5.2 General Normal Distribution

In practice, we rarely have a standard normal distribution. Many real-world problems are associated with general normal distribution. We still need to answer the two basic types of questions: finding probabilities and percentile. The following figure illustrates the two types of questions based on the normal distribution with a mean of 500 ($\mu = 500$) and a standard deviation of 100 ($\sigma = 100$)

The question is whether we cannot use the standard normal distribution table to answer the above two types of questions associated with general normal distribution.

We can use z-score transformation to transform general normal distributions to the standard normal distribution to use the table and then transform back the original general normal distribution. The following figure outlines the above idea.

5.2.1 Finding Probabilities

We use the following example to show the steps for finding the left-tail probabilities.

Example 6. Consider the general normal distribution $N(500, 100)$. Find $P(X < 600) = ?$

Solution. The following figure shows the z-score transformation to obtain the answer.

5.2.2 Finding Percentiles

We continue to use the previous normal distribution as an example to show how to find a percentile of the general normal distribution.

Example 7. Consider the general normal distribution $N(500, 100)$. Find the 15th percentile.

Solution: We are given that the left tail area is 0.15. After z-score transformation, the left tail area of the standard normal density curve is also 0.15 (see the following figure). We can find Z_0 from $P(Z < Z_0) = 0.15$ using the standard normal table which is $Z_0 \approx -1.04$.

Using the relationship between Z_0 and K in the z-score transformation (see the above figure). We have

$$-1.04 = \frac{K - 500}{100}$$

Solve for K, we have $K = 500 - 1.04 \times 100 = 396$.

Example 8. Tomkins Associates reports that the mean clear height for a Class A warehouse in the United States is 22 feet. Suppose clear heights are normally distributed and that the standard deviation is 4 feet. A Class A warehouse in the United States is randomly selected

a). What is the probability that the clear height is greater than 17 feet?

b). What is the probability that the clear height is less than 13 feet?

c). What is the probability that the clear height is between 25 and 31 feet?

d). Find the clear height such that 10% of all clear heights are less than it.

Solution The following figures outline the process of finding the answers to each of the questions.

a). P(X > 17) = P(Z > -5/4) = 1 - P(Z < -5/4) = 0.8944.

b). P(X < 13) = P(Z < 9/4) = 0.012.

c). P(25 < X 31) = P(3/4 < Z < 9/4) = P(Z < 9/4) - P(Z < 3/4) = 0.9878 - 0.9734 = 0.2144.

d). Since P(Z < Zo) =0.10, we have Zo = - 1.28 (from the normal table). The desired clear height (10th percentile) is $X = 22 - 1.28 \times 4 = 16.88$ feet.

Example 9. In redesigning jet ejection seats to better accommodate women as pilots, it is found that women's weights are normally distributed with a mean of 143 lb and a standard deviation of 29 lb.

a). If a woman is randomly selected, what is the probability that she weighs between 163 lb and 201 lb?

b). If the current ejection seat for men weighs between 130 lb and 211 lb, what percentage of women have weights that are within those limits?

c). If a woman is randomly selected, what is the probability that she weighs less than 125 lb?

d). If a woman is randomly selected, what is the probability that she weighs exactly 143 lb?

e). If a woman is randomly selected, what is the probability that she weighs between 90 lb and 130 lb?

f). Find the 10th percentile P10, that is, the weight separating the bottom 10% from the top 90%.

Solution The following are brief solutions with graphical explanations.

a). P(163 < X < 201) = P(0.69 < Z < 2) = P(Z < 2) - P(Z < 0.69) = 0.9772 - 0.7549 = 0.2223.

b). P(130 < X < 211) = P(-0.45 < Z < 2.35) = P(Z < 2.35) - P(Z < -0.45) = 0.9906 - 0.3264 = 0.6642.

c). P(X < 125) = P(Z < -0.62) = 0.2676.

d). P(x=143) = 0.

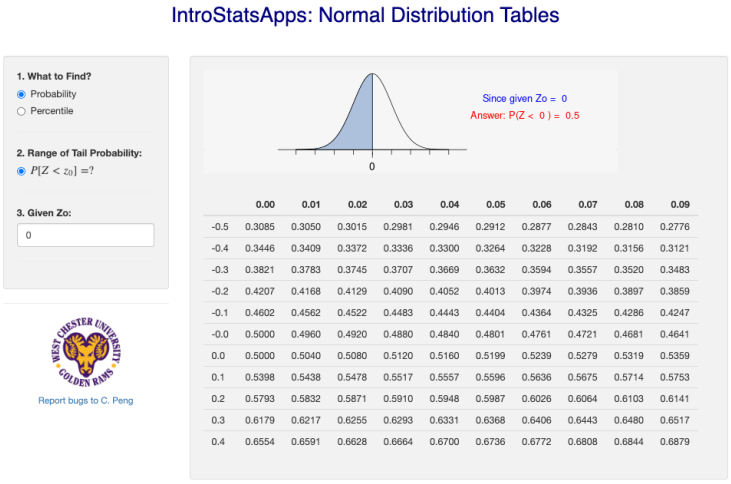
e). P(90 < X < 130) = P(-1.83 < Z < -0.45) = P(Z < -0.45) - P(Z < -1.83) = 0.3264 - 0.0336 = 0.2928.

f). P(z < z_0) = 0.1, so we get z_0 = -1.285. The P_{10} is calculated by $x = 143 - 1.285 \times 29 = 105.73$.

5.3 Use of Technology

The home-grown **IntroStatsApps** has a standard normal distribution we can use.

https://wcupeng.shinyapps.io/ZTable/



Chapter 6

CLT and Sampling Distributions

We have discussed both standard normal and general normal distributions as well as associated two types of questions. In this course, we primarily focus on the inference about the population mean means and proportions from random samples. We will use the sample mean and sample proportion (`both are statistics`) to approximate the population mean and proportion (`both are parameters`).

Since both sample mean and proportions are statistics, they are random. We need to discuss the distributions of sample means and sample proportions.

Some Terminologies: We have learned concepts of the population (parameters) and sample (statistics) as well as probability distributions of random variables.

- **Random Sample** is a subset of values that represents the population of interest.
- **Sampling Distribution** - the distributions of sample statistics are called sampling distributions.
- **Sampling Distribution of Sample Means** – A theoretical probability distribution of sample means that would be obtained by drawing from the population all possible samples of the same size.
- **The Standard Error Sample Mean** – The standard deviation of the sampling distribution of the mean. It describes how much dispersion there is in the sampling distribution of the mean

6.1 Central Limit Theorem (CLT)

> **Central Limit Theorem**: If all possible random samples of size n are drawn from a population with a mean μ and standard deviation σ, then as n increases, the sampling distribution of sample means approaches a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

Remarks

1. The population distribution is NOT specified in CLT. This implies that the CLT could be used for any population (continuous or discrete).
2. The sample size should be large to guarantee a good approximation of the sample mean to a normal distribution. **How large is large?** `By convention, in this course, if n > 30, the sample is called "large".`
3. The distribution of the sample is **approximately normally distributed**. The mean of the sample means ($\bar{X}$) is equal to the population mean (μ) and the standard deviation of the sample mean is equal to $\sigma/\sqrt{n}$.

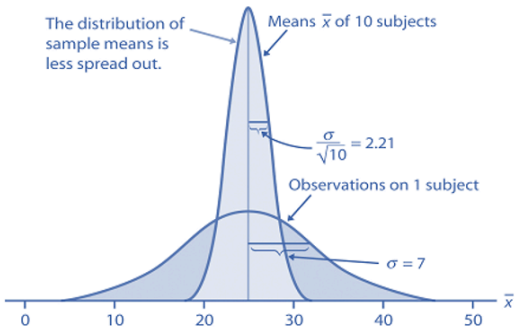
In summary, let $\{X_1, X_2, \cdots, X_n\}$ be a sample taken from a population (μ, σ), by convention, if $n > 30$,

$$\bar{X} \to N(\mu, \frac{\sigma}{\sqrt{n}})$$

or equivalently,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0, 1)$$

The following figure shows how the sample size impacts the variance of the sample mean.



An Important Fact

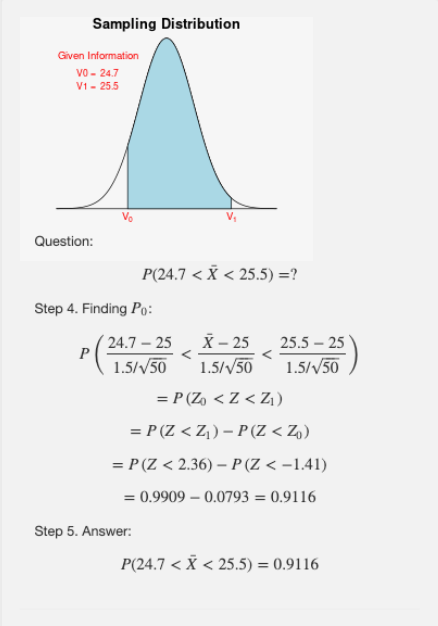
If the population is normal, then

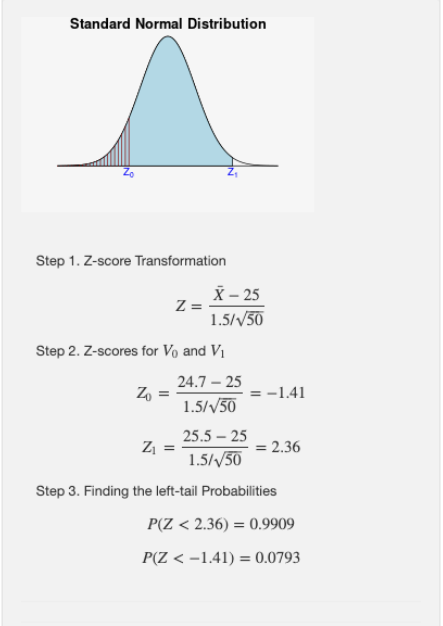
$$\bar{X} \to N(\mu, \frac{\sigma}{\sqrt{n}})$$

regardless of the sample size.

Example 1. The length of time people spend driving each day is different in different age groups. A previous study shows that drivers aged 15 to 19 drive on average $\mu = 25$ minutes a day and standard deviation $\sigma = 1.5$ minutes. A random sample of 50 drivers was selected What is the probability that the average time they spend driving each day is between 24.7 and 25.5 minutes?

Solution: Since $n = 50 > 30$, from the Central Limit Theorem, the sampling distribution of sample means is approximately normal with $(\mu, \sigma/\sqrt{n}) = (25, 0.21)$.

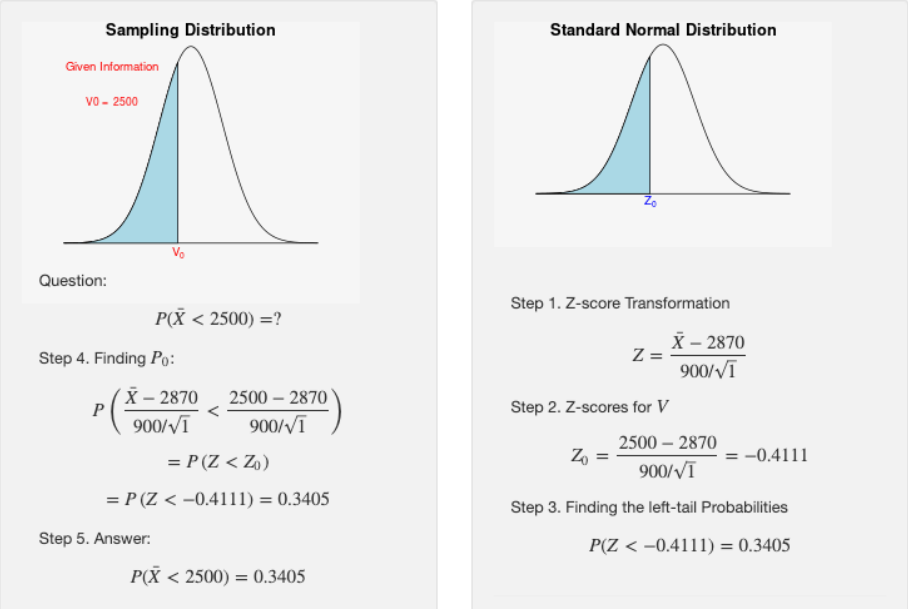




Example 2. A bank auditor claims that **credit card balances are normally distributed**, with a mean of \$2870 and a standard deviation of \$900.

1. What is the probability that a randomly selected credit card holder has a credit card balance of less than \$2500?

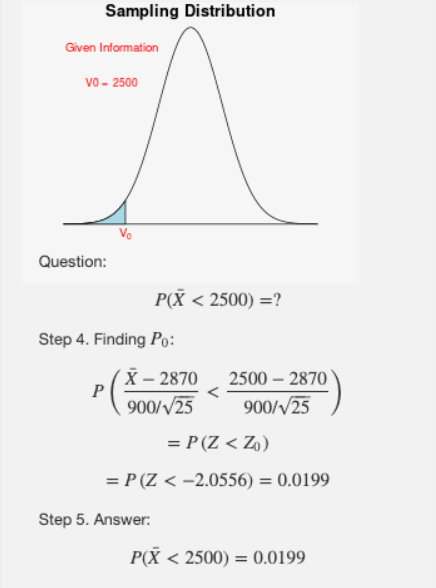
Solution: Since the card balances are normally distributed, we convert the general normal distribution to the standard normal distribution to find the probability (see the following figure)

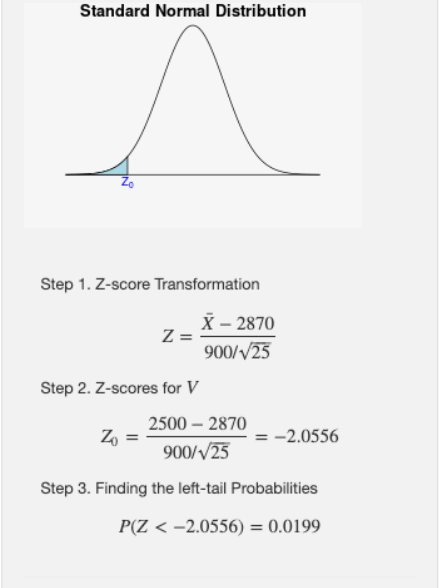


Therefore, there is a 34% chance that an individual will have a balance of less than $2500.

2. You randomly select 25 credit cardholders. What is the probability that their mean credit card balance is less than $2500?

Solution: Because the population is normal, we use the important fact that the sample mean $\bar{X}$ is normally distributed.





Therefore, there is only a 2% chance that the mean of a sample of 25 will have a balance of less than \$2500 (an unusual event).

3. Is it possible that the auditor's claim that the mean is exactly \$2870 is incorrect?

Solution

It is impossible since the probability of observing a single value from the population is always zero.

6.2 Sampling Distribution of Sample Proportion $\hat{p}$

As an application of the central limit theorem, we now discuss the sampling distribution of sample proportion ($\hat{p}$).

- In real-world problems, the responses of interest produce counts rather than measurements – gender (male, female), political preference (republican, democrat), and approval of the new proposal (yes, no).
- Our data will consist of counts or proportions based on the counts.
- We want to learn about population proportions based on the information provided from sample proportions.

A binary population: contains only two possible distinct values, say "success" and "failure".

Count: X = the number of successes in a sample of size n

Proportion: $\hat{p}$ = the proportion of successes in a sample of size n

Sampling Distribution Of A Proportion: if $np > 5$ and $n(1-p) > 5$, then

$$\hat{p} \to N\Big(p, \sqrt{\frac{p(1-p)}{n}}\Big)$$

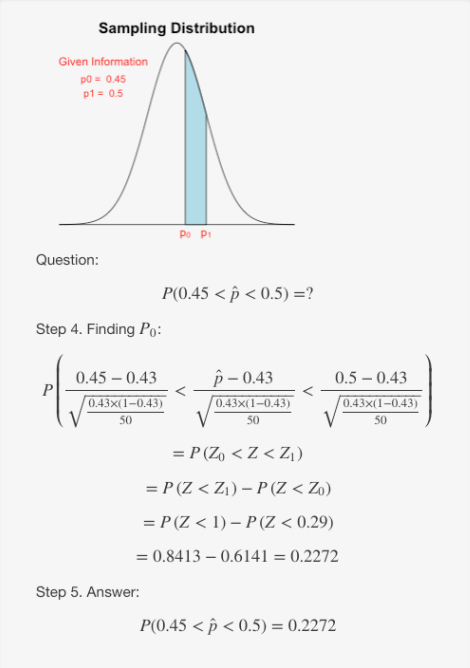
or equivalently

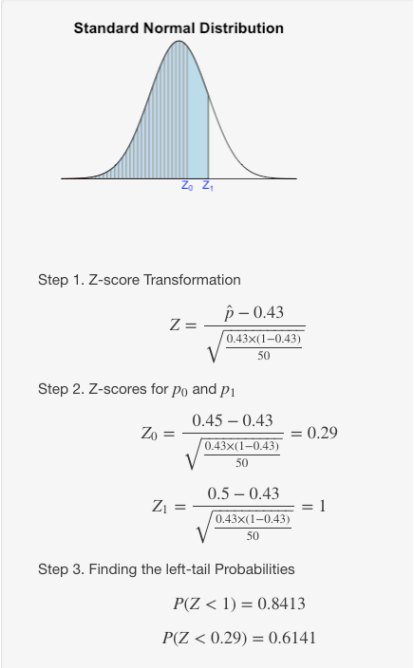
$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \to N(0,1)$$

A cautionary note about proportion: *Whenever working with proportion, we MUST use proportion in the form of decimal in all calculations.*

Example 3. Suppose it is known that 43% of Americans own an iPhone. If a random sample of 50 Americans were surveyed, what is the probability that the proportion of the sample who owned an iPhone is between 45% and 50%?

Solution: We are given that $n = 50$ and $p = 0.43$. Because $np = 50 \times 0.43 = 21.5 > 5$ and $n(1-p) = 50 \times (1 - 0.43) = 28.5 > 5$, we use the above sampling distribution. The following figure gives the steps for finding the probability.



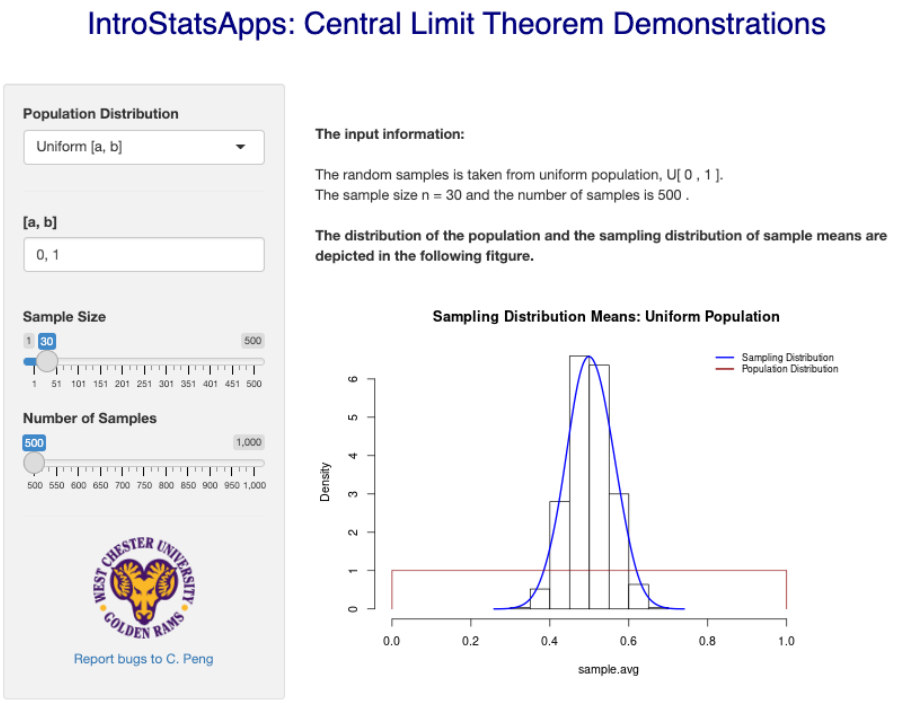


6.3 Use of Technology

Three **IntroStatsApps** were created to illustrate the central limit theorem (CLT) and its applications. The solutions of the above examples were produced using ** IntroStatsApps**.

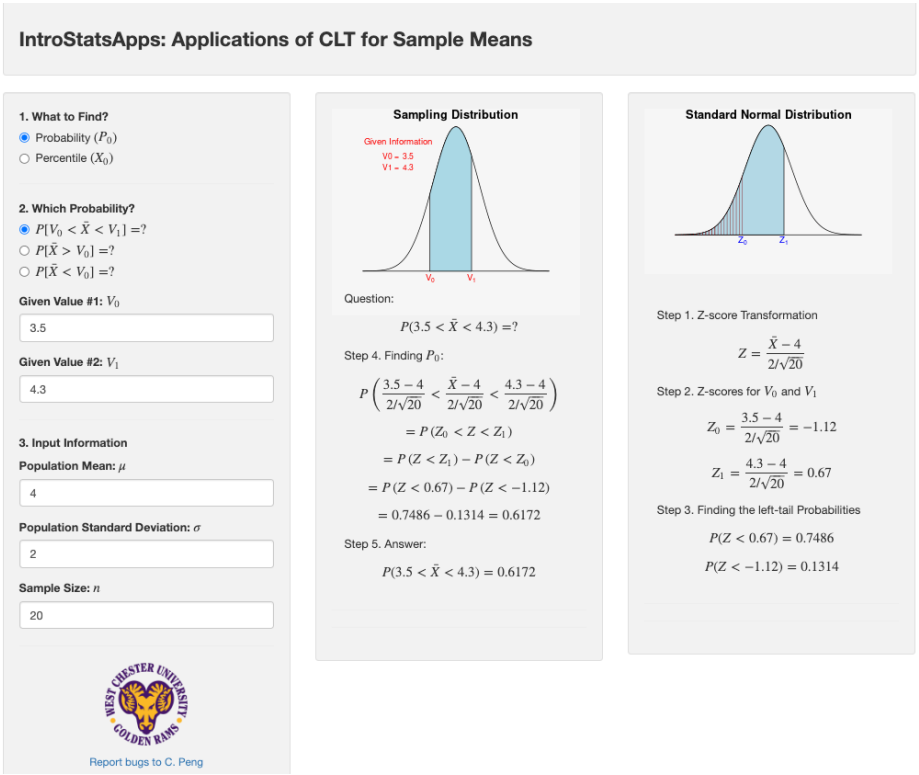
6.3.1 CLT Demo

IntroStatsApps-Central Limit Theorem Demo illustrates the CLT with various populations including the normal population. You can click the link (https://wcupeng.shinyapps.io/CLTdemo/) to explore the CLT under different populations. See the following screenshot of the demo.



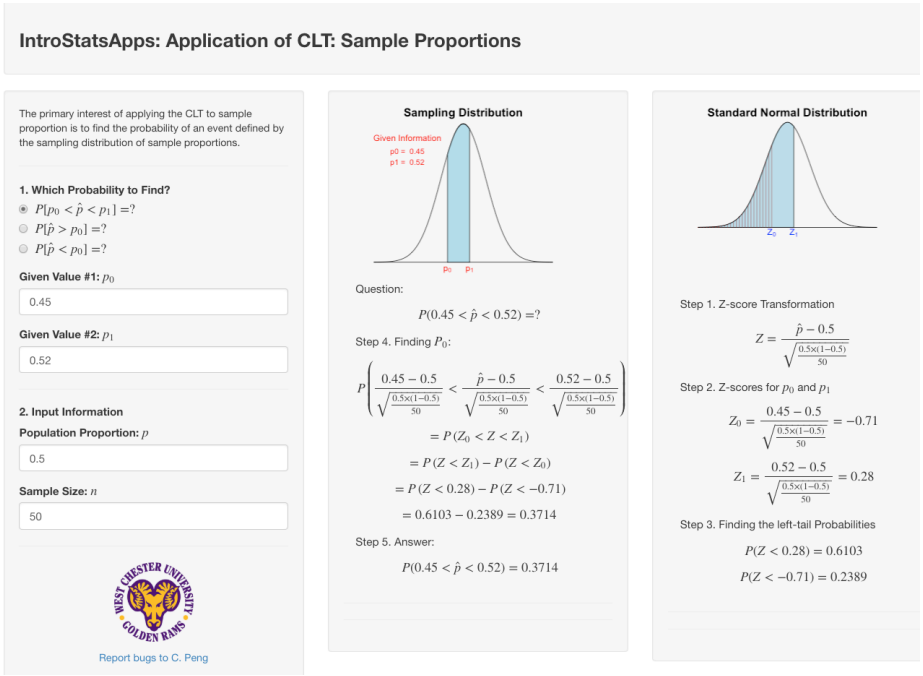
6.3.2 CLT for Mean

When the sampling distribution is normal (the cases of CLT or normal population), we can use this application to answer questions of probability and percentile. The link to this application is at: https://wcupeng.shinyapps.io/AppsCLT4Means/.



6.3.3 CLT for Proportion

For the sampling distribution of sample proportion, we use the following application to answer the questions about probability and percentile. The link to the applications is at: https://wcupeng.shinyapps.io/AppsCLT4Prop/.



6.4 Practice Exercises

1. The U.S. National Center for Health Statistics publishes information on the length of stay by patients in short-stay hospitals in Vital and Health Statistics. According to that publication, the mean stay of female patients in short-stay hospitals is 5.8 days and the standard deviation is 4.3 days. Let $\bar{x}$ denote the mean length of stay for a sample of discharged female patients.

   A). For a sample size of 81, find the mean and standard deviation of the sample mean. Interpret your results in words.

   B). Repeat part A) with n = 144. Find the percentage that the mean stay of those 144 female patients in short-stay hospitals is less than 5 days.

2. According to the U.S. Census Bureau publication Current Construction Reports, the mean price of new mobile homes is $43,800. The standard deviation of the prices is $7200. Let $\bar{x}$ denote the mean price of a sample of new mobile homes.

A). For a sample of 49 randomly selected mobile homes, find the mean and standard deviation of the sample mean.

B). Repeat part A) with n =100. Compare the results you obtained in A).

C). For a sample of 64 randomly selected mobile homes, find the probability that the mean is greater than $45,000.

D). For a sample of 64 randomly selected mobile homes, find the probability that the mean is exactly $45,000.

3. Suppose that the ages X of a certain population are normally distributed, with mean $\mu = 27.0$ years, and standard deviation $\sigma = 12.0$ years, i.e., $X \to N(27, 12)$.

   Find the probability that the mean age of a single sample of $n = 16$ randomly selected individuals is less than 30 years.

4. Fifty-one percent of adults in the U.S. whose New Year resolved to exercise more achieved their resolution. You randomly select 65 adults in the U.S. whose resolution was to exercise more and ask each if he or she achieved that resolution. What is the probability that the sample proportion is greater than 50%?

Chapter 7

Confidence Intervals for Population Means

We will discuss the general framework of confidence intervals of population means and proportions. In the next few sections, we learn how to **estimate the population parameters** such as mean, standard deviation, and proportion from a random sample and build a confidence interval (also called interval estimate) to show how good the estimate is, and finally, we should be able to interpret the estimate.

7.1 A General Framework

This section dedicates the basic framework to estimating population means using confidence intervals.

7.1.1 Some Technical Terms

The following is a list of several basic terms to use when discussing the estimate of population parameters.

- An **estimate** is *a specific value* or *range of values* obtained from a random sample that is used to approximate a population parameter.
- A **point estimate** is a single value (obtained from the random sample) that is used to approximate a population parameter.

Example 1. The sample mean, denoted by $\bar{x}$, is the best point estimate of the population mean μ.

Example 2. The sample variance, denoted by s^2 is the best point estimate of the population variance σ^2.

- The **bias of the point estimate** is equal to the difference between the estimate and the true parameter. For example, the bias of the sample mean is $\bar{x} - \mu$. The bias of an estimate measures *the accuracy of the estimate*.
- Since the point estimate of a parameter is obtained from an underlying random sample, it is a random variable. The variance of the point estimate measures the precision of the point estimate of the corresponding population parameter.
- **The goodness of an estimate** – *no bias and small variance*.

7.1.2 Issues of A Point Estimate and the Logic for Developing A Solution

For a point estimate, we can only say that it is close to the true population parameter. Three questions remain to answer:

1). How close is "close"?

2). How confident is your statement?

3). How precise is the estimate?

We use the following artificial example to explain the above issues about a **point estimate** of the population mean.

Example 3. Assume that the true mean height of WCU student population is 69 inches. To estimate the true mean, we randomly select a sample of 50 students and measure their heights to get a random sample of 50 heights. Assume that the sample average of 68.5 inches.

Now the question: **Is 68.5 close to 69?**

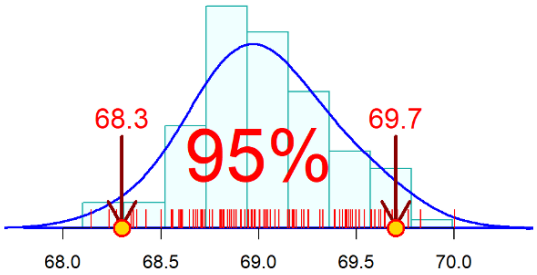
Apparently, we cannot provide an objective answer to the above question with a single sample mean without any additional information.

To address the issue, let's pretend sampling is free. We could take as many samples of the same size using the same protocol as we wish. Then we get many sample means. For example, pretending we obtained 100 sample means from 100 random samples taken from the WCU student population with the same size of 50.

```
68.2 68.3 68.3 68.4 68.4 68.5 68.6 68.6 68.6 68.6 68.6 68.7 68.7 68.7 68.7 68.7
68.7 68.7 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8
68.9 68.9 68.9 68.9 68.9 68.9 68.9 68.9 68.9 68.9 68.9 68.9 69.0 69.0 69.0 69.0
69.0 69.0 69.0 69.0 69.0 69.0 69.0 69.1 69.1 69.1 69.1 69.1 69.1 69.1 69.1 69.1
69.1 69.1 69.1 69.2 69.2 69.2 69.2 69.2 69.2 69.2 69.2 69.2 69.2 69.3 69.3 69.3
```

```
69.3 69.3 69.3 69.4 69.4 69.4 69.5 69.5 69.5 69.5 69.5 69.5 69.6 69.6 69.6 69.6
69.7 69.7 69.7 69.9
```

With these sample means, we make a histogram in the following



We use the definition of percentile to find 2.5% and 97.5% percentiles to be 68.3, and 69.7 respectively (shown in the above figure). This means that 95% of the sample means are within interval: [68.3, 69.7]. Clearly, 69 is inside the interval.

With the above interval, address the aforementioned three issues:

- **All values inside the interval are considered to be close to the true mean of 69.**
- The interval contains 95% of the sample means. If a sample mean is in the interval, we are 95% confident that the sample mean is close to the true mean of 69.
- The width of the interval ($69.7 - 68.3 = 1.4$) reflects the precision of the estimate.

Therefore, this interval contains all desired information, but what is more important is that **the interval was constructed from the distribution of sample means**.

Definition 1: The area of $0.95 = 95\%$ of the middle region in the above figure is called ** 95% Confidence Level**.

Definition 2: *The above interval is called the 95% confidence interval of the average height of WCU student population.* Furthermore, **68.3** is called lower confidence limit (LCL) and **69.7** is called upper confidence limit (UCL).

The above interval was found based on the histogram of 100 sample means. In real-world applications, taking multiple samples could be very costly. **The new and practical question is** *how to find an interval similar to the above one without taking multiple samples.*

7.1.3 General Framework of Confidence Intervals

The key information needed to find the confidence interval of a population mean (μ) is to know the distribution of the sample mean ($\bar{X}$). The distribution of $\bar{X}$ is also called the sampling distribution of the sample mean.

We have discussed the sampling distribution of sample $\bar{X}$ in the previous note.

- If the sample size $n > 30$, by the CLT, the distribution of the sample mean ($\bar{X}$) is *approximately* normal with the mean and standard deviation specified below

$$\bar{X} \to N\Big(\mu, \frac{\sigma}{\sqrt{n}}\Big)$$

- If the population is normally distributed as $N(\mu, \sigma)$, the

$$\bar{X} \to N\Big(\mu, \frac{\sigma}{\sqrt{n}}\Big)$$

regardless of the sample size.

- For a binary population with success probability p, the sampling distribution of sample proportion ($\hat{p}$) is

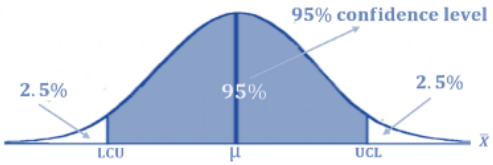
$$\hat{p} \to N\Big(p, \sqrt{\frac{p(1-p)}{n}}\Big)$$

Thought Process of Obtaining A Confidence Interval of Population Mean

With the above sampling distribution, we can find the confidence interval without taking multiple samples. Next, we use the sampling distribution based on the CLT to illustrate the logic for obtaining the confidence interval of the population mean (μ).

- when $n > 30$,

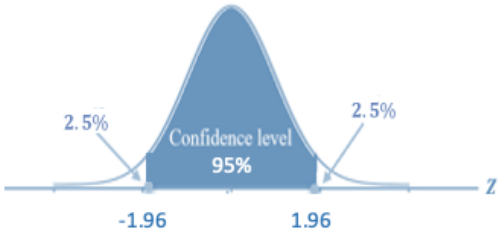
$$\bar{X} \to N\Big(\mu, \frac{\sigma}{\sqrt{n}}\Big)$$



For the given `95% confidence level` (the area of the middle region in the above figure), we cannot find the `confidence limits LCL and UCL` directly from the normal table since $\bar{X}$.

- However, we can use z-score transformation to transform the sampling distribution of $\bar{X}$ to the standard normal distribution in the following.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0,1)$$



When the area of the middle region in the above standard normal curve is given to be 0.95, we can find the 2.5% and 97.5% percentiles (-1.96 and 1.96 respectively) from the normal table directly.

Definition 3: The percentiles ± 1.96 are called **critical values (CV)** of 95% confidence level for the standard normal distribution. The two critical values are symmetric to the origin!

- Relationship between Critical Values and Confidence Limits

The critical values corresponding to the 95% confidence level are ± 1.96. Using the z-score transformation, we have the following relationship.

$$-1.96 = \frac{LCL - \mu}{\sigma/\sqrt{n}}$$

and

$$1.96 = \frac{UCL - \mu}{\sigma/\sqrt{n}}$$

We can solve the confidence limits

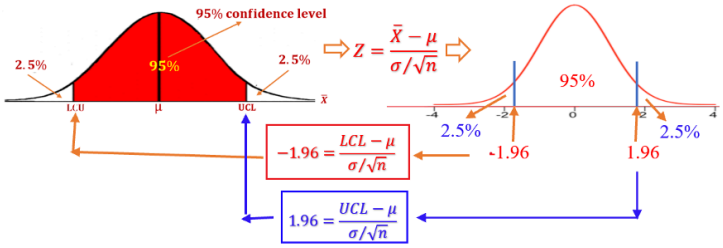
$$LCL = \mu - 1.96 \times \frac{\sigma}{\sqrt{n}}, \qquad UCL = \mu + 1.96 \times \frac{\sigma}{\sqrt{n}}$$

- If we replace μ and σ with sample mean $\bar{X}$ and sample standard deviation s, then LCL and UCL will be completely dependent on the sample data. Then 95% confidence interval can be written as

$$(LCL, UCL) = (\bar{X} - 1.96 \times \frac{s}{\sqrt{n}}, \qquad \bar{X} + 1.96 \times \frac{s}{\sqrt{n}})$$

- Since the confidence interval is constructed based on the random sample, it is random. We can interpret the confidence interval: **there is a 95% chance that the confidence interval [68.3, 69.7] contains the true population mean (μ).**

The above thought process is summarized in the following chart



Example 4. Suppose we want to estimate, with 95% confidence level, the mean (average) length of all walleye fingerlings in a fish hatchery pond. A random sample of 100 fingerlings was selected. The average length is 7.5 inches and the standard deviation is 2.3 inches. That is, $\bar{x} = 7.5$, $s = 2.3$, and $n = 100$.

Solution: Since $n = 100 > 30$, by the central limit theorem, the sample mean is approximately distributed. The 95% critical values are ± 1.96 (using normal table). Based on the above discussion, the lower and upper confidence limits are

$$(LCL, UCL) = (7.5 - 1.96 \times 2.3/\sqrt{100}, \quad 7.5 + 1.96 \times 2.3/\sqrt{100}) = (7.05, \ 7.95)$$

That is, interval $(7.05, \ 7.95)$ has a 95% chance of containing the true population mean length of all walleye fingerlings in a fish hatchery pond.

7.2 Formal Steps For Constructing C.I.

We now formulate the steps for constructing a CI based on the above framework and introduce a few new concepts.

Step 1: Identify the confidence level. If it is NOT given, our default confidence level $1 - \alpha = 0.95$ should be used in this class.

Step 2: Based on the confidence level to find the critical value from the normal table.

Step 3: Evaluate the margin of error, denoted by E and defined to be

$$E = CV \times \frac{s}{\sqrt{n}}$$

Step 4: Write out the confidence interval explicitly in the following form

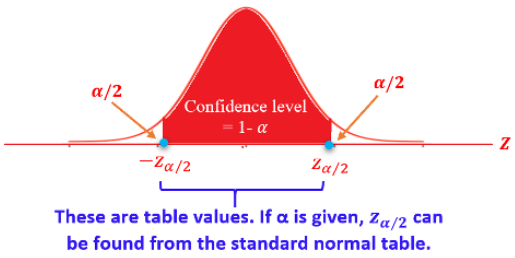
$$\bar{x} \pm E = (\bar{x} - E,\ \bar{x} + E)$$

Step 5: Interpretation of the confidence interval: Two versions of interpretations

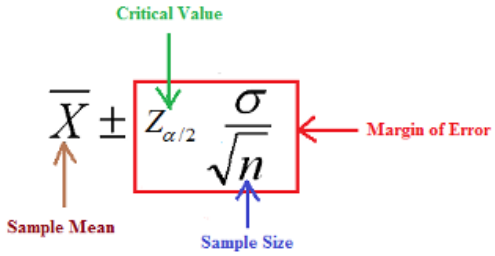
```
Version #1: We are 95% confident that the interval contains the population mean

Version #2: There is a 95% chance that the interval contains the population means.
```

Some Remarks: The following important general facts of confidence intervals can be visualized using **IntroStatsApps: Normal Confidence Interval** at: https://wcu-peng.shinyapps.io/NormalCI4MeanProp/

- By convention, we use $1 - \alpha$ to denote the confidence level (i.e., the area of the middle symmetric region on the density curve). This means that α is the sum of the two tail areas. In other words, both left and right tail areas are equal to $\alpha/2$.



- The margin of error, E, is equal to half of the width of the interval.
- Since sample size, n, is in the denominator of E, as sample size increases, E decreases. This implies that as the sample size increases, the interval gets narrower (See the following formula).



- As the level of confidence increases, the critical value (CV) increases. This implies that the width of the confidence interval increases as the confidence level increases (see the above formula for this relationship).

Example 5. Assume that we collect 81 measurements of the iron-solution index of tin-plate specimens, designed to measure the corrosion resistance of tin-plated steel. Assume that the sample mean is 57 and the sample standard deviation is 15. Construct a 95% confidence interval of the iron-solution index.

Solution We will follow the 5-step procedure to construct the confidence interval.

Step 1: The confidence level $1-\alpha = 0.95$, this means that $\alpha/2 = 0.025$.

Step 2: The critical value is $CV = Z_{0.025} = 1.96$.

Step 3: The margin of error

$$E = CV \times \frac{s}{\sqrt{n}} = 1.96 \times \frac{15}{\sqrt{81}} = 3.27.$$

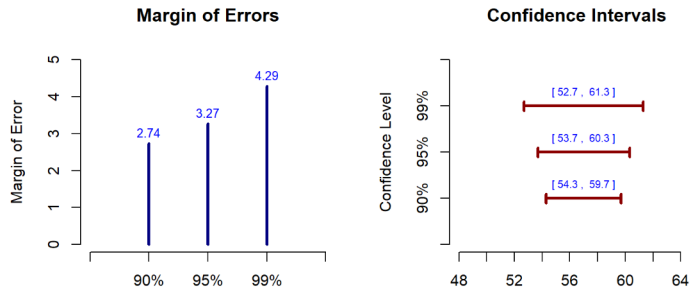
Step 4: The explicit expression of the 95% CI is given by

$$57 \pm E = (57 - 3.27,\ 57 + 3.27) = (53.73,\ 60.27).$$

Step 5: The interpretation of CI. We are 95% confident that the interval (53.73, 60.27) contains the true population mean of the iron-solution index.

Example 6 (continuation of Example 5). If we change the confidence level from 95% to 90% and 99% respectively, how will be the corresponding **margin of error** and **confidence interval** changed?

Solution. We follow the first four of the same 5-step procedure to calculate the margin of error and construct the confidence interval corresponding to the three significant levels and summarize the results in the following figure.



We can see the same patterns outlined earlier. as the confidence level increases, the margin of error gets bigger (the left panel in the above figure), hence, the corresponding confidence interval gets wider (the right panel of the figure).

7.3 Use of Technology

I also created an interactive application to generate confidence intervals for population mean and proportion based on the assumption of large samples. The app is at: https://wcu-peng.shinyapps.io/NormalCI4MeanProp/

The following is the screenshot of the solution of Example 5 generated by the app.

IntroStatsApps: Normal C.I.s for Population Mean (μ) and Proportion (p)

1. CI for μ or p?

- (•) Population Mean (μ)
- ( ) Population Proportion (p)

2. Sample Statistics

Sample Mean: $\bar{X}$

57

Standard Deviation: s **or** σ_0

15

3. Sample Size and Confidence Level

Sample Size: n

81

Confidence Level: α

80% 95% 99%

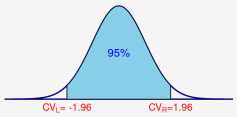
80 82 84 86 88 90 92 94 96 98 99

Report bugs to C. Peng

Steps for Constructing A C.I.

Step 1. The given confidence level.
conf.level = $1 - \alpha$ = 95%

Step 2. CV on the standard normal density curve.
$CV = Z_{\alpha/2}$ = 1.960



Step 3: Margin of Error

$$E = CV \times \frac{s}{\sqrt{n}} = 1.96 \times \frac{15}{\sqrt{81}} = 3.267$$

Step 4: Expression of Confidence Interval

$$(\bar{X} - E, \bar{X} + E) = (57 - 3.267, 57 + 3.267) = (53.733, 60.267)$$

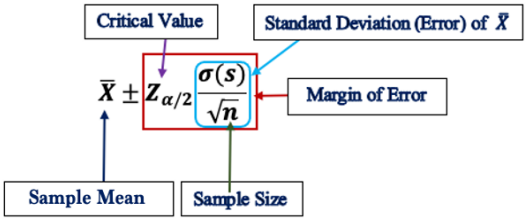
Step 5: Interpretation of Confidence Interval

There is a 95 % chance that the confidence interval (53.733 , 60.267) contains the true population mean.

Chapter 8

Confidence Intervals for Population Means and Proportions

We have developed the steps for constructing confidence intervals for population means based on large samples with which the central limit theorem can be used to find the sampling distribution of the sample mean.



As a review, we present the following example to highlight the importance of the steps for constructing a confidence interval.

Example 1. Market researchers use the number of sentences per advertisement as a measure of readability for magazine advertisements. The following represents a random sample of the number of sentences found in 50 advertisements. (Source: Journal of Advertising Research)

```
9   20   18   16   9   9   11   13   22   16    5   18   6    6   5   12   25
17  23   7   10   9   10   10   5   11   18   18   9   9   17   13   11   7
14  6   11   12   11   6   12   14   11   9   18   12   12   17   11   20
```

The summarized statistics are given in the following table.

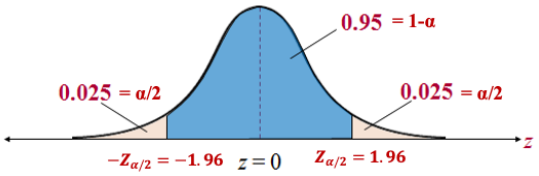
Population Parameter…	Sample Statistic
Mean: μ	$\overline{x} = 12.4$
Standard deviation: σ	s = 5

Based on the above data, construct a 95% confidence interval of the mean number of sentences (μ) in all magazine advertisements.

Solution We are going to use the 5-step procedure to construct the confidence interval in the following.

Step 1: We are given a confidence level of 95%.

Step 2: Since the sample size $n = 50 > 30$, using the CLT, we claim that the sample mean is approximately normally distributed. The critical value based on the standard normal distribution should be used. The critical value was found and labeled in the following figure.



Step 3: Find the margin of error E in the following

$$E = Z_{\alpha/2} \times \frac{s}{\sqrt{n}} \approx 1.96 \times \frac{5.0}{\sqrt{50}} \approx 1.4.$$

Step 4: The explicit form of the confidence interval is given by

$$(\bar{X} - E,\ \bar{X} - E) = (12.4 - 1.4,\ 12.4 + 1.4) = (11.0, 13.8).$$

Step 5: The confidence interval (11,13.8,) has a 95% chance to include the true average number of sentences in all magazine advertisements.

8.1 Confidence Interval of Proportion

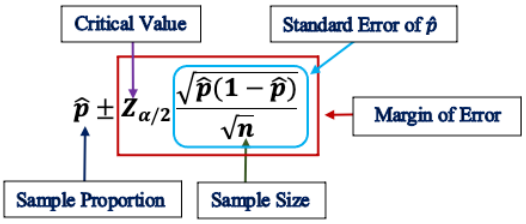
Recall that the sample proportion ($\hat{p}$) is approximately normally distributed if both $n\hat{p}$ and $n(1-\hat{p})$ are large. To be more specific, in this course, if $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$, we have

$$\hat{p} \to N\Big(p, \sqrt{\frac{p(1-p)}{n}}\Big),$$

which is equivalent to

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \to N(0,1).$$

Therefore, we can use the same 5-step procedure as used previously in constructing the confidence interval of the population mean for the large sample. The only difference is that the form of the margin of error involves the sampling error of the sample proportion. The following annotated formula explains the components of the confidence interval of the proportion.



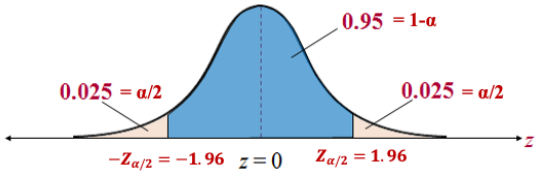
Next, we use an example to illustrate the 5-step procedure for constructing the confidence interval of proportion.

Example 2. In a survey of 1219 U.S. adults, 354 said that their favorite sport to watch is football. Construct a 95% confidence interval for the proportion of adults in the United States who say that their favorite sport to watch is football.

Solution: First of all, the sample proportion $\hat{p} = 354/1219 \approx 0.29$ and sample size $n = 1219$.

Step 1: The confidence level is $1-\alpha = 0.95$. $\alpha/2 = 0.025$.

Step 2: Since $n\hat{p} = 354 > 5$ and $n(1-\hat{p}) = 1219 \times 0.71 = 865 > 5$, the sampling distribution of $\hat{p}$ is normally distributed. The critical value corresponding to 95% confidence level is $Z_{0.025} = 1.96$ (from the normal table).



Step 3: The margin of error for the confidence interval of population proportion is given by

$$E = Z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \times \sqrt{\frac{0.29(1-0.29)}{1219}} \approx 0.025.$$

Step 4: The explicit form of the confidence interval is given by

$$(\hat{p} - E,\ \hat{p} + E) = (0.29 - 0.025,\ 0.29 + 0.025) = (0.365,\ 0.315).$$

Step 5: The confidence interval (0.265, 0.315) has a 95% chance to include the true proportion of adults who say football is their favorite sport.

8.2 t - Confidence Interval for Mean (μ)

We have constructed confidence intervals based on the CLT that require a large sample size. When we construct a confidence interval based on small samples, we need to make stronger assumptions about the population so that we have enough information for the interval.

8.2.1 t-distribution and t-Table

We have pointed out that

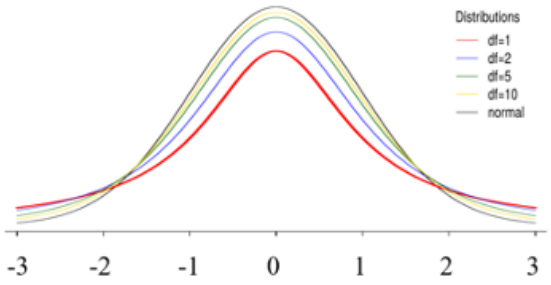
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0, 1)$$

if the sample is from a normal population regardless of the sample size. However, if the population is unknown, we need to use the sample standard deviation to estimate the population standard deviation. In this case,

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \nrightarrow N(0, 1)$$

if the sample size is small! The correct distribution is called t-distribution with $n - 1$ degrees of freedom.

The difference between standard normal and t-distributions are depicted in the following figure.

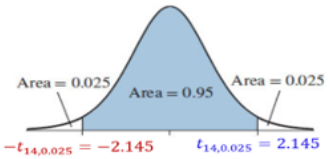


We can see from the above figure that t-distributions are also bell-shaped. As the degrees of freedom increase, the t-distributions approach the standard normal distribution! **However**, when the degree of freedom is small, the t-distribution and the standard normal distribution are VERY different.

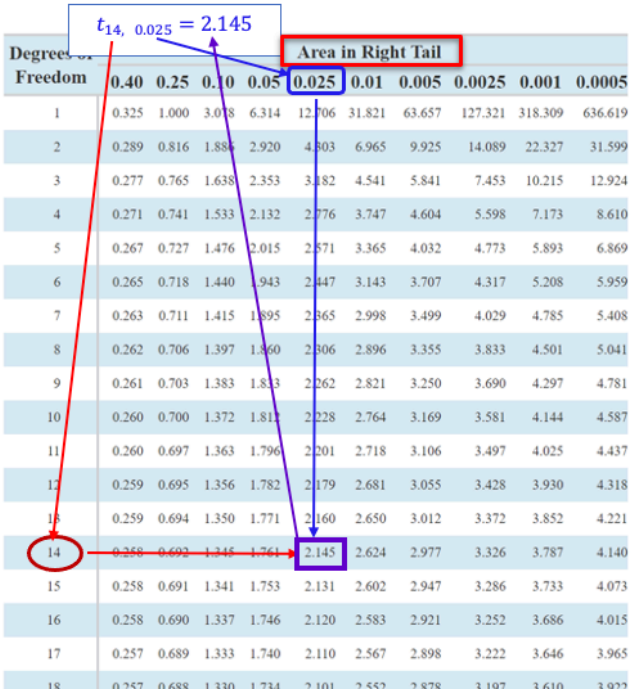
This implies that when a small sample from a normal population with unknown population variance, we should NOT use the normal critical value to construct the confidence interval for the unknown population mean. Instead, we MUST use the t-critical value that can be found in the t-table!

Example 3. Find the critical value for a 95% confidence when the sample size is 15.

Solution since sample size $n = 15$, the degrees of freedom $df = n - 1 = 14$. We use the notation $t_{df,\alpha/2}$ to denote the critical value based on the t-distribution. The following figure shows the location of the CV on the density curve.



The critical value $CV = t_{14,0.025} = 2.145$ which is found from the t-distribution table. The following table illustrates the structure of the t-table and how to use the t-table to find the critical value.



Degrees of Freedom	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922

8.2.2 t-confidence Interval

The steps for constructing t-confidence intervals are identical to those in the normal confidence intervals except for the critical value that is found in the t-table. We will use an example to show the steps for constructing a t-confidence interval.

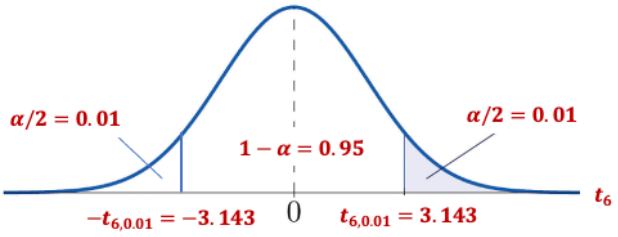
Example 4. . Estimating Car Pollution - In a sample of seven cars, each car was tested for nitrogen-oxide emissions (in grams per mile) and the following results were obtained: `0.06, 0.11, 0.16, 0.15, 0.14, 0.08, 0.15` (based on data from the Environmental Protection Agency). Assuming that this sample is representative of the cars in use. Further, the amounts of nitrogen-oxide emission for all cars are normally distributed. Construct a 98% confidence interval estimate of the mean amount of nitrogen-oxide emission for all cars

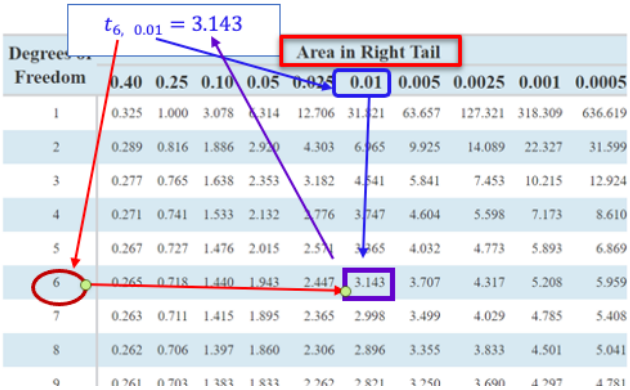
Solution: We first calculate the sample mean and sample standard deviation using the formulas introduced in the note of descriptive statistics.

$$\bar{x} = 0.1214, \quad s = 0.0389.$$

Step 1: The confidence level for the t-confidence interval is $1 - \alpha = 0.98$. The right-tailed area of the t-density curve is $\alpha/2 = 0.01$

Step 2: We need to use the t-critical value to calculate the confidence interval: CV $= t_{6,0.01} = 3.143$. The density curve and the table below.





Degrees of Freedom	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781

Step 3: The margin of error

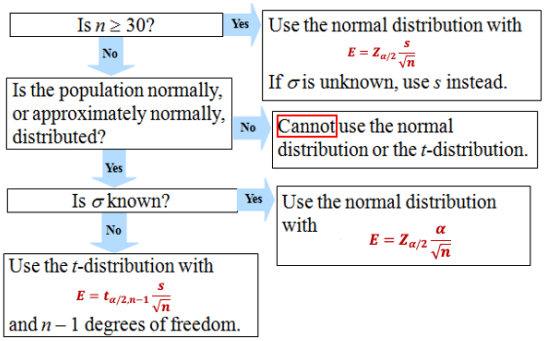
$$E = t_{n-1,\alpha/2} \times \frac{s}{\sqrt{n}} = 3.143 \times \frac{0.0398}{\sqrt{7}} \approx 0.0462.$$

Step 4: Therefore, the 98% confidence interval estimate of the population mean is
$(\bar{x} - E, \bar{x} + E) = (0.1214 - 0.0462, 0.1214 + 0.0462) = (0.075, 0.168)$.

Step 5: The interval (0.075, 0.168) has a 98% chance to include the true average amount of nitrogen-oxide emission of all cars in use.

8.2.3 Normal or t-Confidence Interval: A Summary

When a normal or t confidence interval should be constructed is dependent on the given amount of information. The following brief flow chart summarizes the selection of the two types of confidence intervals based on different conditions.

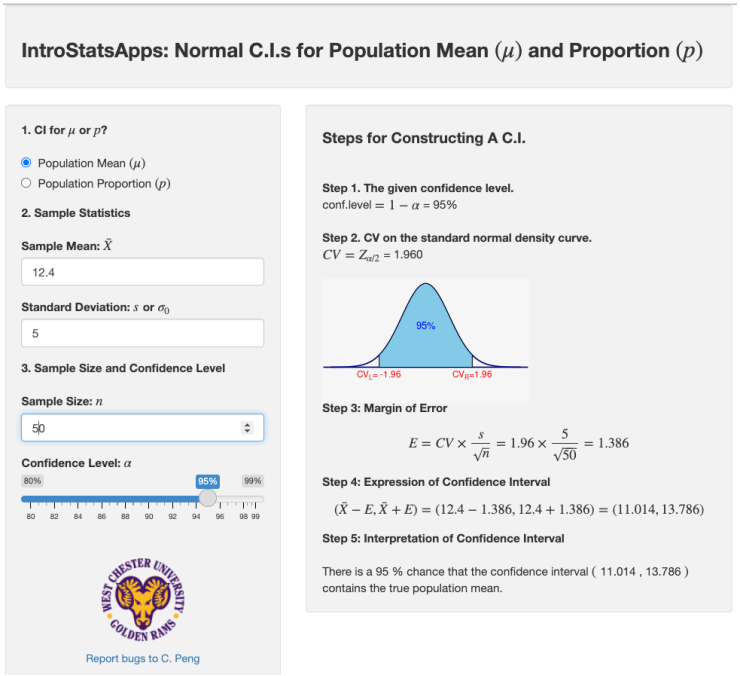


8.3 Use of Technology

Two **IntroStatsApps** can be used to generate solutions to most of the problems. You can these apps to check your work and make sure you correctly understand the concepts and steps for finding the confidence intervals of either population mean or proportion.

8.3.1 IntroStatsApps: Normal CI

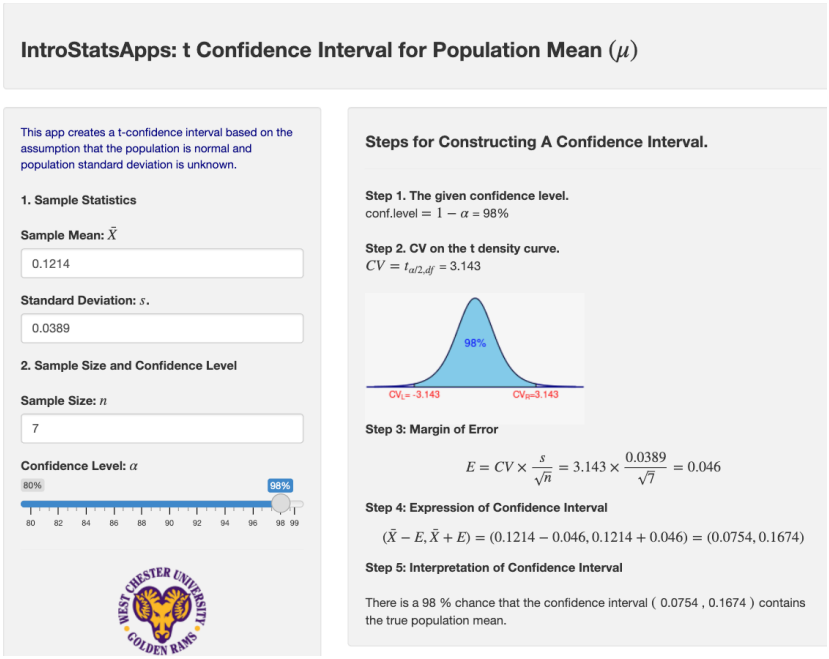
The direct link to the app is at: https://wcu-peng.shinyapps.io/NormalCI4MeanProp/. The following screenshot is the solution to **Example 1**.



8.3.2 IntroStatsApps: t-CI

The following apps generate solutions for t-confidence intervals. The direct link is at: https://wcu-peng.shinyapps.io/Student-t-CI/

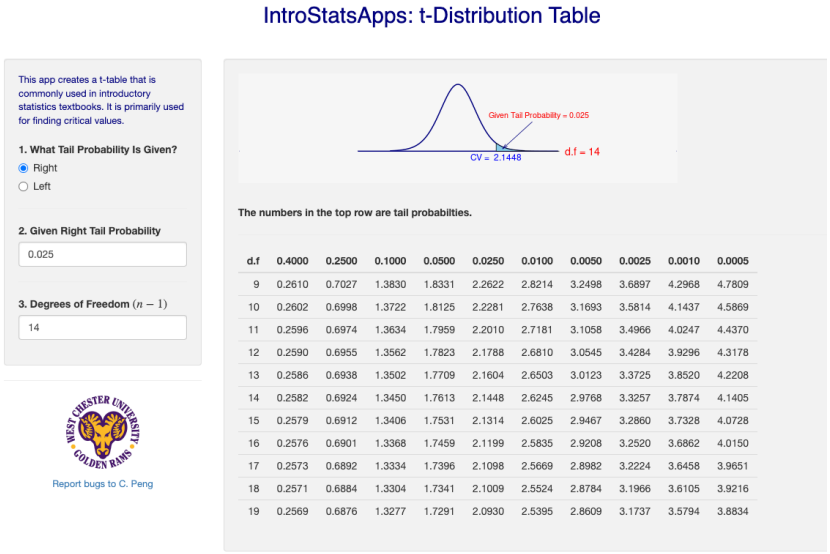
The following screenshot is the solution to **Example 4**.



8.3.3 IntroStatsApps: t-Table

For convenience, I also create an interactive t-table that provides the correct probability or percentile based on the provided information. The direct link is at: https://wcu-peng.shinyapps.io/t-Table/

The following screenshot is the solution to **Example 3**.



d.f	0.4000	0.2500	0.1000	0.0500	0.0250	0.0100	0.0050	0.0025	0.0010	0.0005
9	0.2610	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498	3.6897	4.2968	4.7809
10	0.2602	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437	4.5869
11	0.2596	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058	3.4966	4.0247	4.4370
12	0.2590	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545	3.4284	3.9296	4.3178
13	0.2586	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123	3.3725	3.8520	4.2208
14	0.2582	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874	4.1405
15	0.2579	0.6912	1.3406	1.7531	2.1314	2.6025	2.9467	3.2860	3.7328	4.0728
16	0.2576	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208	3.2520	3.6862	4.0150
17	0.2573	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982	3.2224	3.6458	3.9651
18	0.2571	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105	3.9216
19	0.2569	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609	3.1737	3.5794	3.8834

8.4 Practice Exercises

Please do the following problems manually and then use the **IntroStatsApps** to check your work.

1. A team of efficiency experts intends to use the mean of a random sample of size n =150 to estimate the average mechanical aptitude of assembly-line workers in a large industry (as measured by a certain standardized test) and found that the sample mean is 19.9 minutes, and the sample standard deviation is 5.73 minutes. Construct a 95% confidence interval for the average mechanical aptitude of assembly line workers in the given industry.

2. A college admissions director wishes to estimate the mean age of all students currently enrolled. In a random sample of 20 students, the mean age is found to be 22.9 years. From past studies, the standard deviation is known to be 1.5 years, and the population is normally distributed. Construct a 90% confidence interval of the population mean age.

3. Suppose you are interested in investigating the factors that affect the prevalence of tuberculosis among intravenous drug users. In a group of 97 individuals who admit to sharing needles, 24.7% had a positive tuberculin skin test result. Construct 95% confidence intervals for the population proportions.

4. Find the confidence interval for average SAT score for women given that n =15, the sample mean is 496, and the sample standard deviation is 108. Assume that the SAT scores are normally distributed.

Chapter 9

The Logic and Components of Hypothesis Testing

We have introduced how to construct confidence intervals to estimate the population means and proportion under various assumptions. We now introduce a new inference that justifies a statement about a population mean or proportion - **Testing Hypotheses**.

- In statistics, a hypothesis is a claim or statement about a property of a population. A hypothesis test is a process that uses sample statistics to test a claim about the value of a population parameter.

- The fundamental idea is to let data speak truthfully.

- **Rare Event Rule for Inferential Statistics**

  Under a given assumption (e.g., the null hypothesis is correct), if the probability of an observed event is exceptionally small, we conclude that the assumption (the null hypothesis) is probably not correct.

Example 1. If someone claims that the average height of WCU student population is at least 80 inches. To justify the claim, we take a random sample of 100 students (a set of representatives of the WCU student population). Assume that the sample mean is 68 inches and the standard deviation 15 inches.

9.1 Steps for Hypothesis Testing: Some Analogies

To better understand the logic and steps for conducting a statistical hypothesis and the types of inevitable errors in the decision process, we use the two analogies that are familiar to us.

Analogies of Statistical Hypothesis Testing		
Medical Diagnostics	**Test of Statistical Significance**	**Court Trial**
The patient experienced the symptoms of some disease and believe he is sick.	Someone claimed about the population parameter.	The prosecutor claimed the defendant committed a crime.
Every incoming patient is assumed to have a disease until proved disease-free	Null Hypothesis: claims the attainable values of the parameter.	Every defendant is assumed to be innocent until proved guilty
Order labs and gathering health info	data (information) collection	Investigation and evidence gathering
summarize clinical evidence	Information Aggregation: test Statistics	Summary of evidence exhibitions
Clinical standards	Statistical Decision Rules	Jury's instruction: cross-examination
Diagnostic decision	Statistical Decision on the null hypothesis	Verdict
Claim a diseased patient to be disease-free	**Type I Error**	Convict an innocent defendant
Claim a disease-free patient to be diseased	**Type II Error**	Acquit a criminal

No matter what decision will be made, there are two possible errors:

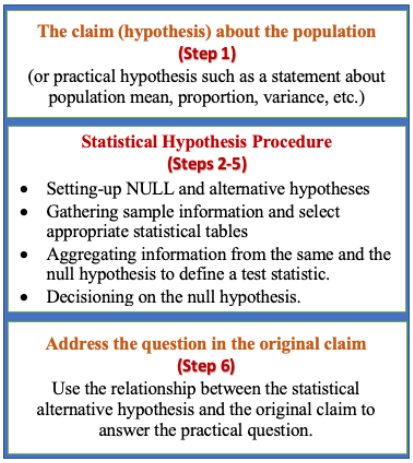
1. **Type I Error**: the patient has a disease, but the doctor claims NO disease.

2. **Type II Error**: the patient has NO disease, but the doctor claims a disease.

9.2 Components of Testing Hypothesis

The approach to formulating the process of hypothesis testing in this course is divided into three blocks: a practical claim (also called a practical hypothesis), steps for statistical hypothesis testing, and a conclusion addressing the claim.

The following chart depicts the process for testing a hypothesis.



Next, we introduce the components of hypothesis testing based on the above chart.

9.2.1 The claim (or practical hypothesis)

A claim (or a practical hypothesis) is a statement about the population parameter of interest such as population mean (μ) and proportion (p). It has six possible distinct forms. Using population mean (μ) as an example, we have the following six possible claims.

$$\mu \begin{cases} = \\ \geq \\ \leq \\ \neq \\ > \\ < \end{cases} \mu_0$$

where μ_0 is the claimed value. Three of them involve an equal sign ($=, \geq, \leq$) and the other three do not have an equal sign ($\neq, >, <$). For a given problem, we need to identify the claim and write it explicitly.

Some of the keywords are helpful for identifying the claim:

at most implies $\geq$;

at least implies $\leq$;

differs implies $\neq$;

majority implies that proportion is greater than 0.5, i.e., $p > 0.5$.

Example 1. A water faucet manufacturer announces that the mean flow rate of a certain type of faucet is at most 2.5 gallons per minute.

What is the claim about the population mean flow rate of the type of faucet?

Solution: we see the keyword **at most** in the claim. Therefore, the claim is $\mu \leq 2.5$.

Example 2. A cereal company advertises that the mean weight of the contents of its 20-ounce size cereal boxes is more than 20 ounces.

What is the claim about the population mean?

Solution: We see the statement has keyword **more than**, the claim is $\mu > 20$.

Example 3. A university publicizes that the proportion of its students who graduate in 4 years is 82%.

What is the claim about the population proportion?

Solution: Key word **is** implies that the claim is: $p = 0.82$. Caution: The proportion MUST be written as a decision in all formulas.

A Cautionary Note: If the claim in the problem is vague, we have to choose an explicit claim in order to set up the null and alternative hypotheses.

Example 4. A bakery machine fills boxes with crackers, averaging 454 grams (roughly one pound) of crackers per box. The quality control manager wants to know whether the production process still maintains the quality standard. What is the claim about the mean weight of crackers per box?

Solution: The **claim** is vague. The manager only wants to know whether the production line still maintains the same quality over time. Therefore, the claim could be either $\mu = 454$ or $\mu \neq 454$. No matter what claim is used, we eventually draw a conclusion to justify it.

9.2.2 Null and Alternative Hypotheses

Definition 1: *Null Hypothesis* - The null hypothesis (H0) is a statement about a single value of a population parameter (such as the mean). It MUST take one of the following forms:

$$H_0 : \mu = \mu_0, \qquad H_0 : \mu \geq \mu_0, \qquad H_0 : \mu \leq \mu_0$$

where μ_0 is the value in the claim to be justified.

Important Observation: "=" must be included in H_0.

Definition 2: *Alternative Hypothesis*: The alternative hypothesis (H_a or H_1) is the statement that must be true if the null hypothesis is false. That is, the corresponding alternative hypothesis of each of the above three null hypotheses must be:

$$H_a : \mu \neq \mu_0, \qquad H_a : \mu < \mu_0, \qquad H_a : \mu > \mu_0$$

Important Observation: H_a must NOT have "="!

Relationship between H_0 and H_a

The null and alternative hypotheses are mutually complementary. This means that only one of the hypotheses is correct. For example, if we reject H_0, then we must accept H_a, and vice versa.

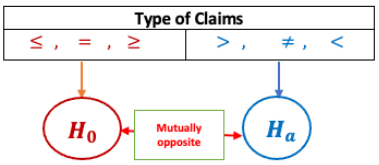
Because of the relationship between H_0 and H_a, there are ONLY three possible pairs of NULL and alternative hypotheses. Next, we use population mean as an example to explicitly write these three possible pairs of H_0 and H_a.

$$H_0 : \mu = \mu_0 \quad v.s. \quad H_a : \mu \neq \mu_0 \ ,$$
$$H_0 : \mu \geq \mu_0 \quad v.s. \quad H_a : \mu < \mu_0,$$

and

$$H_0 : \mu \leq \mu_0 \quad v.s. \quad H_a : \mu > \mu_0.$$

The above relationship is depicted in the following.



Relationship between the claim and H_a.

The following table illustrates the relationship.

The guideline for setting up Ha is given below:

a). If **the original claim** contains an "=" (i.e., ≤, =, ≥), we choose the opposite of the original claim (i.e., >, ≠, <) as **the alternative hypothesis** H_a.

b). If **the original claim** doesn't contain an "="(i.e., >, ≠, <), we simply choose the original claim (i.e., >, ≠, <) as **the alternative hypothesis** H_a.

Example 5. A bakery machine fills boxes with crackers, averaging 454 grams (roughly one pound) of crackers per box

(a). If the management of the bakery is concerned about the possibility that the actual average is different from 454 grams, what null hypothesis and what alternative hypothesis should it use to put this to test?

(b). If the management of the bakery is concerned about the possibility that the actual average is at most 454 grams, what null hypothesis and what alternative hypothesis should it use to put this to test?

Solution: We use the relationship between claim, null and alternative hypotheses.

(a). The original claim: "the actual average is different from 454 grams", i.e., $\mu \neq 454$. Since claim $\mu \neq 454$ doesn't have an "=" sign, it is used as H_a. Therefore, the null and alternative hypotheses are specified in the following.

$$H_0 : \mu = 454 \quad v.s. \quad H_a : \mu \neq 454.$$

(b). The original claim: " the actual average is at most 454 grams", i.e., $\mu \leq 454$. Since the "=" sign is contained in the claim, the claim $\mu \leq 454$ is used as H_0. Therefore, the null and alternative hypotheses are specified in the following.

$$H_0 : \mu \leq 454 \quad v.s \quad H_a : \mu > 454$$

Exercises Read examples 1 - 4 and set up the null and alternative hypotheses for these examples.

9.2.3 Test Statistic

Some characteristics of test statistics.

- The **test statistic** is a value computed from the sample data that is used to decide whether **null hypothesis** is rejected.
- The **test statistic** converts the sample statistic (such as sample mean) to a score (such as z-score, t-score, etc.) with the assumption that the null hypothesis is true.

- When we test the claims about a population mean (μ) in a large sample case, we can use the central limit theorem (CLT) to derive the distribution of the **test statistic**.

In other words, a test statistic is an **information aggregator** that consolidates information from multiple sources. For example, the following **test statistic** is used to test a population mean μ.

$$TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

The above **test statistic** contains information from the sample ($\bar{x}$, s, and sample size n) and the hypotheses (the claimed population mean μ_0).

Recall that the above **test statistic** is an approximate standard normal distribution if the sample size is large (n $>$ 30 in our course).

An Intuitive Interpretation of TS: TS is a standardized "distance" that measures the discrepancy between the observed value ($\bar{x}$) and the claimed value (μ_0).

For example, if the above test statistic is used to test

$$H_0 : \mu = \mu_0 \quad v.s. \quad H_a : \mu \neq \mu_0.$$

If TS is close to 0, the observed value and the claimed value are close to each other implying that the sample evidence supports the null hypothesis (H_0) and rejects the alternative hypothesis (H_a). Otherwise, if TS is NOT close to 0, there is a difference between the observed value and the claimed value implying that the sample evidence **does not** support the null hypothesis (H_0) but supports the alternative hypothesis (H_a).

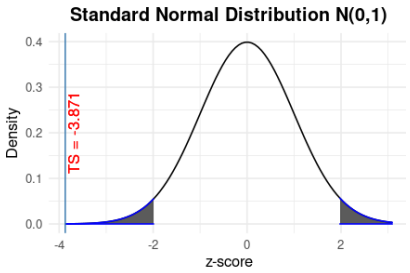
A natural question: *how close* is called close?

Example 6. We use the body temperature data collected by the researchers at the University of Maryland with characteristics: $n = 36$, $\bar{x} = 98.20^0F$, $s = 0.62^0F$. Find the value of the test statistic for the claim that the population mean is $\mu = 98.6^0F$.

Solution: Based on the given information and the formula, we calculate the test statistic in the following

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{98.2 - 98.6}{0.62/\sqrt{36}} = -3.87.$$

The "closeness" question: Is $TS = -3.87$ close to 0?



We can see from the above figure that TS is not close to 0 since it is located far away from the center. The next question is how to find a threshold that can tell **closeness**. The threshold is called **critical value**.

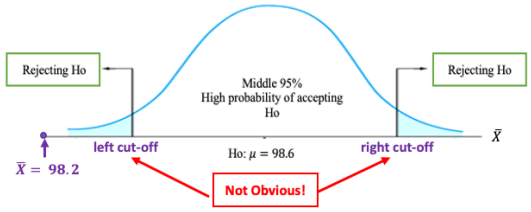
9.2.4 Significance Level and Critical Value(s)

- The **Rejection Region (RR)** is the set of all values of the score converted from the statistic that causes us to reject the **null hypothesis**. The cut-off value(s) is(are) called Critical Value(s).
- The **significance level** (denoted by α) is the probability that the test statistic will fall into the critical region while the null hypothesis is true. This α is the same α in confidence level $1-\alpha$ used in constructing confidence intervals.

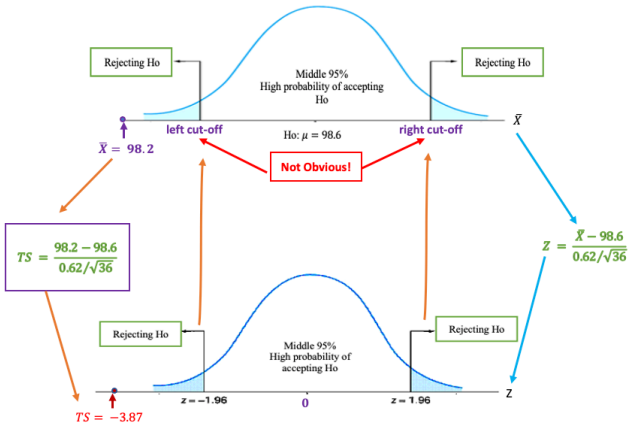
Example 7. (continuation of Example 6) Assume we null and alternative hypotheses are

$$H_0 : \mu = 98.6 \quad v.s. \quad H_a : \mu \neq 98.6.$$

we have found that $TS = -3.87$. Under Ho, we have the following density curve



Since left and right cut-offs are unknown, we still cannot tell whether $\bar{X} = 98.2^0F$ is close to $H_0 : \mu_0 = 98.6^0F$. However, we can easily see from the standardized test statistic on the density curve of the standard normal curve is NOT close to 0 (see the following figure).



As we have discussed in the confidence interval, all numbers in interval $(-1.96, 1.96)$ are considered to be close to 0. Therefore, $TS = -3.87$ is NOT close to 0. This implies that $\bar{x} = 98.2^0F$ is NOT close to $H_0 : \mu_0 = 98.6^0F$. Since the bottom panel of the above figure is independent of the specific normal distribution, we use the distribution of TS to define rejection region(s) and critical value(s) for a given significance level. In other words, for testing

$$H_0 : \mu = 98.6 \quad v.s. \quad H_a : \mu \neq 98.6.$$

at **significance level** $\alpha = 0.05$,

Rejection Region (RR): consists of two tail regions.

Critical Values (CV): are the values that are defined in such a way that both tail areas are equal to $\alpha/2 = 0.05/2 = 0.025$. That is, the critical value for this testing hypothesis is $CV = \pm 1.96$ which can be found in the normal table.

9.2.5 Statistical Decision on Ho

Once we found the critical value and the value of the test statistic, we can make a statistical decision using the following rule:

- If TS falls into the rejection region (RR), we **reject** null hypothesis **Ho**;
- If TS does NOT fall into the rejection region (RR), we **fail to reject** null hypothesis **Ho**.

9.2.6 Conclusion about the Claim

We use the relationship between **the claim**, **null hypothesis**, and **alternative hypothesis** to conclude the claim (supporting or rejecting the claim.)

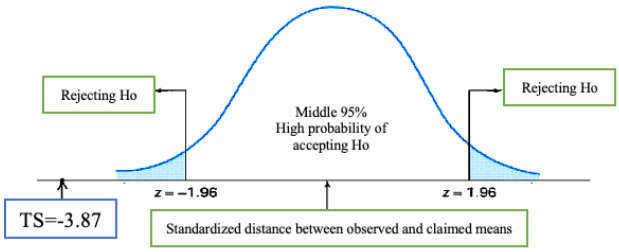
Example 8. (continuation of Example 6) The claim is *the population mean is* $\mu = 98.6^0 F$. Recall that null and alternative hypotheses are

$$H_0 : \mu = 98.6 \quad v.s. \quad H_a : \mu \neq 98.6.$$

and

$$TS = -3.87.$$

For this specific hypothesis test. The rejection region (RR) has two parts located on each tail (see the figure below).



Apparently, the test statistic $TS = -3.87$ falls into the rejection. Therefore, we reject the null hypothesis $H_0 : \mu = 98.6$. Since the null hypothesis and the claim are identical, the sample does not have evidence to support the claim.

9.3 Formal Steps for Hypothesis Testing

Based on the components discussed in the previous section, we formulate the following **six-step procedure** for testing the hypothesis about the mean or proportion of a population.

Step 1: Identify the claim about a population parameter such as the population mean or proportion.

Step 2: Set up null and alternative hypotheses.

Step 3: Evaluate the test statistic.

Step 4: Find the critical value(s) based on the given significance level (α) and specify the rejection region (RR).

Step 5: Make a statistical decision: reject or fail to reject H_0.

Step 6: Draw a conclusion about the population parameter in the claim.

These steps will be used in all examples of hypothesis testing problems throughout the course.

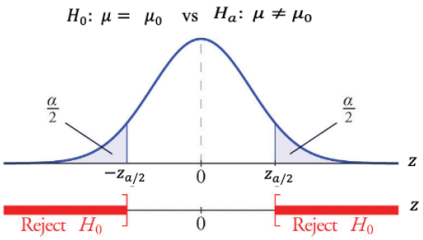
9.4 Types of Hypothesis Tests

In examples 6 - 8, we test the null hypothesis $H_0 : \mu_0 = 98.6$. This means that the null hypothesis should be rejected if the test statistic TS is far from zero (in either a positive or negative direction), the rejection region has two parts: one on the left tail and one on the right tail. This type of test is called the **two-tailed** test.

The form of the alternative hypothesis tells the type of test. For example, for a two-tailed test, the alternative hypothesis test takes the form of $H_0 : mu \neq \mu_0$. Note that $\neq$ means either $>$ or $<$. That is, **the direction of the inequality sign points to the location of the rejection region (RR)**.

9.4.1 Two-tailed Test

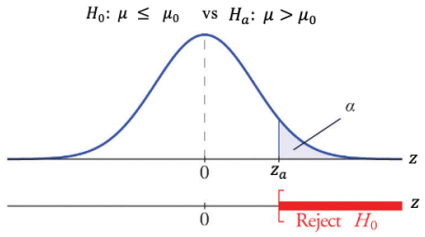
The following figure shows the locations of the rejection regions of a two-tailed test. The form of the alternative hypothesis $H_a : \mu \neq \mu_0$ implies that TS is not close to 0. Therefore, the rejection region of H_0 is **evenly** split into right and left sub-rejection-regions.



both tail areas in the density curve are equal to half of the significance level.

9.4.2 Right-tailed Test

When the alternative hypothesis takes the form of $H_a : \mu > \mu_0$, the corresponding test is called a **right-tailed test** with the inequality sign in H_a pointing to the right. Consequently, there is only **one** critical value on the right tail of the density curve.

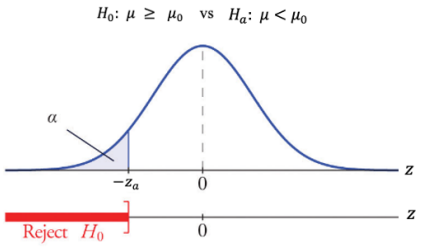


This is intuitive since the null hypothesis of a right-tailed test takes the form of $H_0 : \mu \leq \mu_0$. This implies that H_0 is supported if the observed **test statistic** is far left. Otherwise, we reject the null hypothesis H_0. Consequently, there are **two** critical values: one on the right tail and one on the left tail of the density curve.

The right tail area of the density curve is equal to the significance level.

9.4.3 Left-tailed Test

When the alternative hypothesis takes the form of $H_a : \mu < \mu_0$, the corresponding test is called a **left-tailed test** with the inequality sign in H_a pointing to the left. Consequently, there is only **one** critical value on the left tail of the density curve.



Similarly, the null hypothesis of a right-tailed test takes the form of $H_0 : \mu \geq \mu_0$. This implies that H_0 is supported if the observed **test statistic** is far right. Otherwise, we reject the null hypothesis H_0.

The left tail area of the density curve is equal to the significance level.

Example 9. A professor wants to investigate the effectiveness of a new pedagogical method. Assume the average score for the mean population is 80 before the experiment. With a new training method, the professor believes that the score might change. Professor tested randomly 36 students' scores. The average score of the sample is 88 and the standard deviation is 10. With a 5% significance level, is there enough evidence to suggest the average score changed?

Solution: We follow the 6-step procedure to perform the hypothesis to answer the above question.

Step 1: Since the statement **'professor believes the mean score *might change*'** does not specifically mention the direction of the change, we can take either one of the two forms: $\mu = 80$ or $\mu \neq 80$.

In this solution, we take the form: $\mu \neq 80$.

Step 2: Since the claim $\mu \neq 80$ has no '=' in it, it will be the alternative hypothesis and the opposite of it will be the null hypothesis. More specifically,

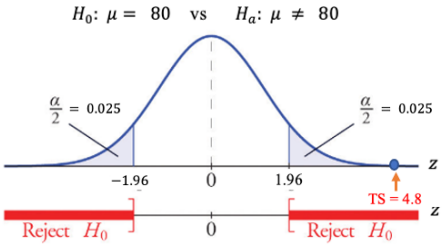
$$H_0 : \ \mu = 80 \ \ v.s. \ \ H_a : \mu \neq 80$$

This is a two-tailed test. Therefore, there will be two rejection regions.

Step 3: The test statistic is defined by

$$TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{88 - 80}{10/\sqrt{36}} = 4.8.$$

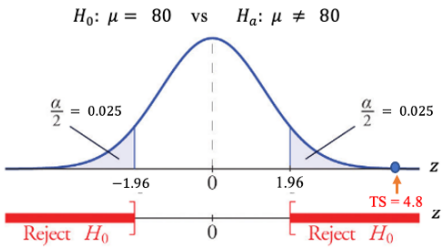
Step 4: Since $n = 36 > 30$, by the central limit theorem (CLT), the test statistic is approximately normally distributed. We use the given confidence level $\alpha = 0.05$ and the normal table to find the two critical values: $\pm z_{0.05/2} = \pm z_{0.025} \pm 1.96$. All above information is labeled in the following figure.



Step 5: Since $TS = 4.8$ is in the rejection region (right-hand side), the null hypothesis is rejected.

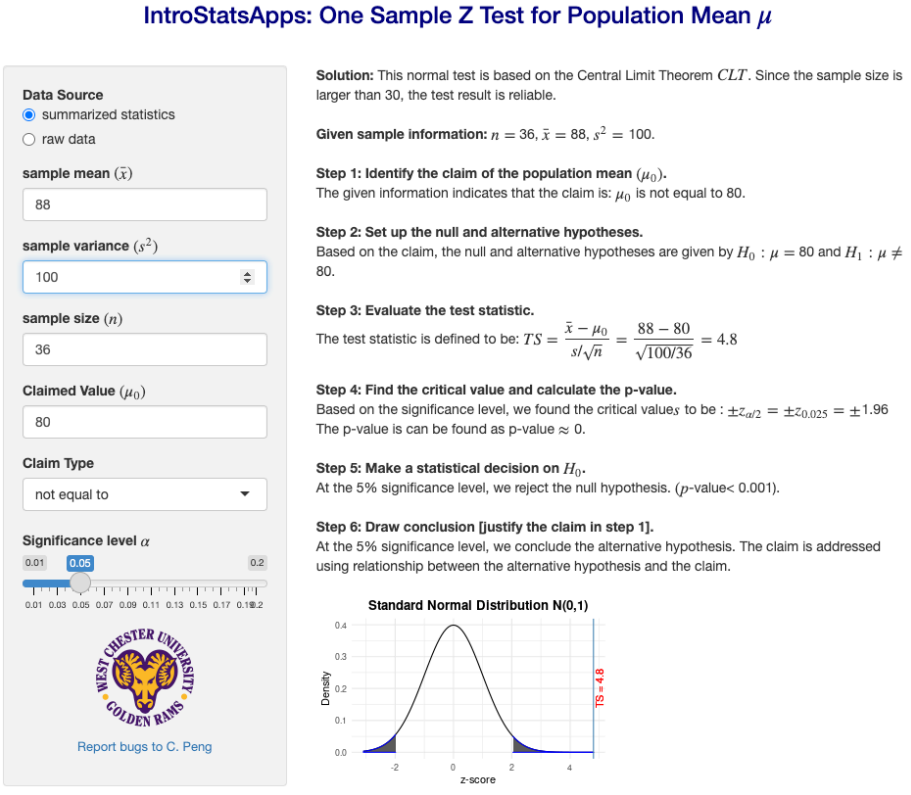
Step 6: We conclude that the population score has changed after the training program.

Critical values of Z-tests with commonly used significance level: Three commonly used significance levels are $\alpha = 0.10, 0.05$ and 0.01. Their corresponding critical values associated with different types of tests are summarized in the following table.



9.5 Use of Technology

The Stats Apps you can use for generating the solution to Example 09 is at https://wcu-peng.shinyapps.io/oneMean-z-Test/. The following screenshot shows the input information.



Chapter 10

Hypothesis Testing: Normal Tests

10.1 Introduction

This note focuses on the normal tests for the population means and proportions using the central limit theorem. We have formulated a general six-step testing procedure **for all hypothesis tests** in the previous topic. Also introduced was the critical value method for making a statistical decision on H_0. For normal tests using the normal table, we can also use another method of statistical decision: the p-value method.

We will use the same six-step procedure with both p-value and critical value methods to test the population means and proportions.

10.1.1 Recall the 6-step procedure of testing hypotheses

Step 1: Identify the claim about a population parameter such as the population mean or proportion.

Step 2: Set up null and alternative hypotheses.

Step 3: Evaluate the test statistic.

Step 4: Find the critical value(s) based on the given significance level (α) and specify the rejection region (RR).

Step 5: Make a statistical decision: reject or fail to reject H_0.

Step 6: Draw a conclusion about the population parameter in the claim.

10.1.2 A Review Example

Example 1. According to the salary survey by the National Association of Colleges and Employers, the average salary offered to computer science majors who graduated in May 2002 was \$50,352. Suppose this result is true for all computer science majors who graduated in May 2002. A random sample of 200 computer science majors who graduated this year showed that they were offered a mean salary of \$51,750 with a standard deviation of \$5240. Use significance level 5%, can you conclude that the mean salary of this year's computer science graduates is **higher than** \$50,352?

Solution Note that sample statistics given in the above story problem are: $n = 200$, $\bar{x} = 51750$, and $s = 5240$. We are also given the significance level $\alpha = 0.05$. Next, we explicitly write out the 6-step procedure for this hypothesis test.

Step 1: Since the last sentence contains the keyword *higher than*, the original claim takes the form $\mu > 50352$.

Step 2: Based on the above claim, the null and alternative hypotheses are given by

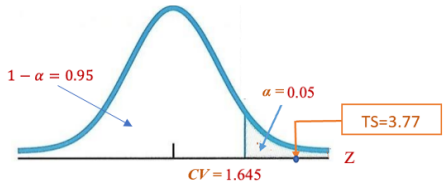
$$H_0 : \ \mu \leq 50352 \ \ vs. \ \ H_a : \ \ \mu > 50352.$$

The form of H_a indicates that this is a right-tailed test with a rejection region on the right-hand side tail. We

Step 3: Evaluate the test statistic for testing the population mean in the following

$$TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{51750 - 50352}{5240/\sqrt{200}} \approx 3.773$$

Step 4: Since $n = 200 > 30$, by the central limit theorem (CLT), TS approximately follows the standard normal distribution. For the given significance level, we go to the standard normal distribution table to find the critical value (i.e., 95th percentile) on the right tail: $CV = z_{0.05} = 1.645$.. The above information is labeled in the following figure.



Step 5: Statistical decision on **Ho** (always on Ho). Since the test statistics $TS = 3.773$ is in the rejections, we **REJECT** the null hypothesis **Ho:** $\mu \leq 50352$ and **CONCLUDE** the alternative hypothesis **Ha:** $\mu > 50352$.

Step 6: The sample evidence supports the claim that the mean salary is **higher than** \$50352.

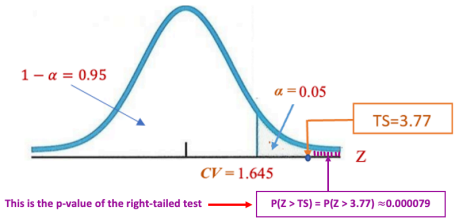
10.2 Testing Population Means Using P-value Method

We first introduce the method of p-value method for testing population means.

10.2.1 What is P-value?

In the critical value method, the decision rule is simple: if the test statistic TS is inside the rejection region (RR) defined by the critical value(s), we reject the null hypothesis (**Ho**); otherwise we fail to reject the null hypothesis (**Ho**).

Next we use the **right-tailed test** in the above example to demonstrate an equivalent decision rule. First of all, once we obtained the value of the test statistic, we can use the normal table to find the area to **the right-hand side of the tail region**. In the above example, the right-tailed area is 0.000079 (see the following figure).



The area to the right of the test statistic for the right-tailed test is called **p-value**.

Important Observations:

1. If the test statistic is in the reject region, the **p-value** is **less than** the significance level $\alpha = 0.05$. That is, if the **p-value** is **less than** the significance level, we REJECT **Ho**;

2. If the test statistic is outside the rejection region, the **p-value** (*area to the right of the test statistic*) is **greater than** the significance level $\alpha = 0.05$. That is, if the **p-value** is **greater than** the significance level, we FAIL TO REJECT **Ho**.

The above observations established a general decision method – the p-value method:

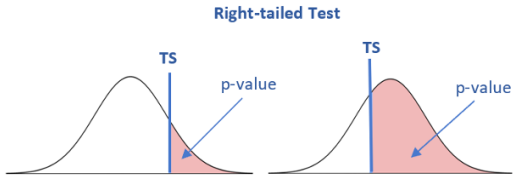
- If the p-value is less than α, H_0 is REJECTED.
- If the p-value is greater than α, H_0 is concluded.

The p-value defined above is based on the right-tailed test. The general definition of the p-value for all three types of tests.

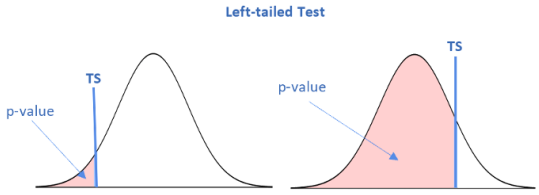
10.2.2 Definitions of P-values

The p-value is defined based on the type of tests. The following figures **show** the p-value of the three types of tests.

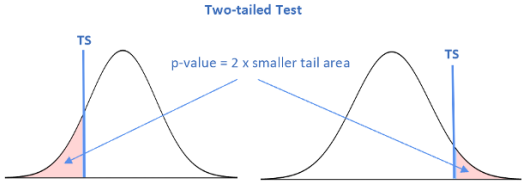
1. For a right-tailed test, the p-value is defined to be the area to the **right** of the test statistic **TS**.



2. For a left-tailed test, the p-value is defined to be the area to the **left** of the test statistic **TS**;



3. For a two-tailed test, the p-value is defined as the smaller tail area times two (i.e., double the smaller tail area).



10.2.3 Summary of p-Value

We can see from the above discussion that the definition of p-value requires

1. The sampling distribution of the test statistic;
2. The value of the test statistic.

In terms of changes in the testing procedure, (1) instead of finding the critical value in **step 4**, we will find the **p-value**; (2) the decision rule in **step 5** will be changed from the **critical value** method to **p-value** Method.

In this course, we only use the p-value method when the sampling distribution is normal.

Example 2. [p-value Method] A manufacturer of salad dressings uses machines to dispense liquid ingredients into bottles that move along a filling line. The machine that dispenses salad dressings is **working properly when 8 ounces are dispensed**. Suppose that the average amount dispensed in a particular sample of 36 bottles is 7.91 ounces with a variance of 0.04 ounces. Is there evidence that the machine should be stopped for repairs?

Solution: We use the p-value method to answer this question.

Step 1: The question "*Is there evidence that the machine should be stopped for repairs?*" implies that the claim is $\mu \neq 8$.

Step 2: The null and alternative hypotheses are

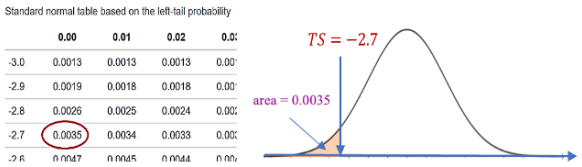
$$H_o\ \mu = 8 \quad v.s. \quad Ha : \mu \neq 8$$

The form of the alternative hypothesis indicates that this is a two-tailed test.

Step 3: The test statistic is evaluated below

$$TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{7.91 - 8}{0.2/\sqrt{36}} = 2.7$$

Step 4: Since the sample size $n = 36$, by the CLT, the sampling distribution of TS is a standard normal distribution. Since this is a two-tailed test, the p-value is equal to the smaller tail area times 2. That is, p-value $= 2 \times 0.0035 = 0.007$. The following figure shows the way of finding the p-value.



Step 5: Decision rule. Since the p-value $= 0.07 < 0.05$ (default significance), we reject $H_o : \ \mu = 8$.

Step 6: Based on the decision made in **step 5** and the relationship between the claim in **step 1** and the hypotheses in **step 2**, the sample evidence supports the claim. This implies that the machine should be stopped for repairs.

10.3 Testing Hypotheses About Population Proportions

We have discussed the sampling distribution of sample proportions.

$$\hat{p} \to N\Big(p, \sqrt{\frac{p(1-p)}{n}}\Big)$$

when $np > 5$ and $n(1-p) > 5$. We standardize the above sampling distribution to get

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \to N(0,1)$$

The above equation measures the "distance" between the observed proportion from the sample ($\hat{p}$) and the true proportion of the population. Therefore, for testing $H_0 : \ p = p_0$, we use the following test statistic

$$TS = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \to N(0,1)$$

when the same conditions, $np_0 > 5$ and $n(1-p_0) > 5$, are satisfied.

Note that, another more commonly used test statistic is given by

$$TS = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \to N(0,1)$$

From theory, both are valid test statistics, and both have an approximate standard normal distribution when the above two conditions are satisfied. The latter one was used in Stats Apps and examples.

With the above test statistic and its sampling distribution, we can perform hypothesis testing about population proportion using the same 6-step procedure as we used in testing population means. The only difference is the form of the test statistic and the conditions of the sampling distribution of the test statistic.

Next, we use an example to illustrate the steps for the hypothesis test.

Example 3. Suppose the previous example is stated a little bit differently. The CEO claims that at least 80 percent of the company's customers are very satisfied. To justify the CEO's claim, 166 customers are surveyed using simple random sampling and found that 73 percent are very satisfied. Based on these

results, should we accept or reject the CEO's claim? Assume a significance level of 0.05.

Solution: We use the p-value method to perform this test.

Step 1: The claim is clearly given in the statement "*The CEO claims that at least 80 percent of the company's customers are very satisfied.* ": $p \geq 0.8$ (the proportion MUST be written in the decimal form).

Step 2: The claim contains an "=" sign, it is used as the null hypothesis and its opposite is the alternative hypothesis.

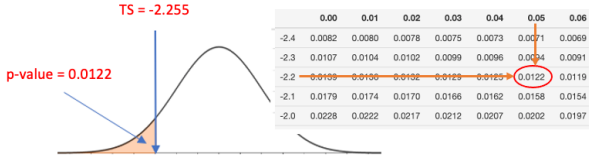
$$H_0 : \ p \geq 0.8 \ \ v.s. \ \ H_a : \ p < 0.8.$$

This means that this is a left-tailed test.

Step 3: The test statistic is evaluated below using the following form

$$TS = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.73 - 0.8}{\sqrt{\frac{0.8\times(1-0.8)}{166}}} \approx -2.255.$$

Step 4: Since this is a left-tailed test, the p-value is the left-tailed area labeled in the following density curve. Note also that $166 \times 0.73 = 122 > 5$ and $166 \times (1 - 0.73) = 45 > 5$, therefore, the test statistic is normally distributed.



	0.00	0.01	0.02	0.03	0.04	0.05	0.06
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197

The p-value is equal to the left-tail area = 0.0122.

Step 5: Since the p-value is less than the significance level of 0.05, we reject the null hypothesis.

Step 6: Based on the decision in **Step 5**, the same evidence supports the CEO's claim that at least 80 percent of the company's customers are very satisfied.

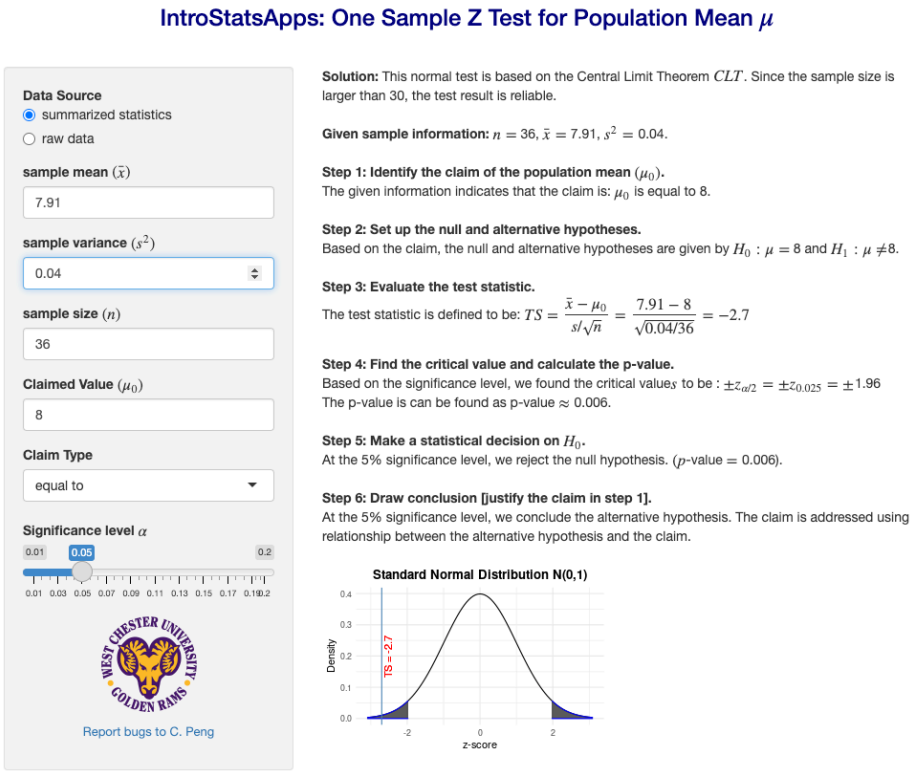
Remark We can change steps 4 and 5 to use the critical value method.

10.4 Use of Technology

We can use the Stats Apps to perform the **normal tests** for population means and proportion.

10.4.1 Normal Test for Means

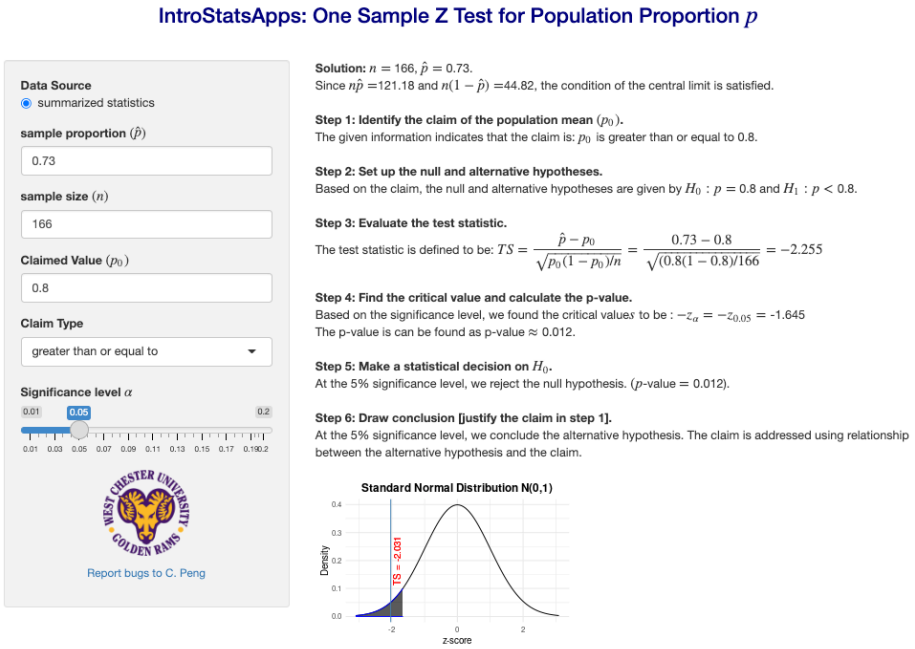
The App for a normal test of the population mean is at: (https://wcu-peng.shinyapps.io/oneMean-z-Test/). We use the app to generate the solution of Example 2 (see the following screenshot). We need to provide information on the left navigation panel.



Note that the App requires the **variance of the sample**. If you are given the sample standard deviation, you need to square it to get the variance.

10.4.2 Normal Test for Proportion

The App for a normal test of the population proportion is at: (https://chpeng.shinyapps.io/oneProp-z-testR/). We next generate the solution of example 3.



10.5 Practice Exercises

Please do the following exercises manually first and then check your answers using Stats Apps.

1. An industrial company claims that the mean pH level of the water in a nearby river is 6.8. You randomly select 31 water samples and measure the pH of each. The sample mean and standard deviation are 6.7 and 0.24, respectively. Is there enough evidence to reject the company's claim at $\alpha = 0.05$? Assume the population is normally distributed.

2. A used car dealer says that the mean price of a 2005 Honda Pilot LX is at least \$23,900. You suspect this claim is incorrect and find that a random sample of 36 similar vehicles has a mean price of \$23,500 and a standard deviation of \$1250. Is there enough evidence to reject the dealer's claim at $\alpha = 0.05$?

3. The mayor of a large city claims that the average net worth of families living in this city is at least $300,000. A random sample of 100 families selected from this city produced a mean net worth of $288,000 with a standard deviation of $80,000. Using the 2.5% significance level, can you conclude that the mayor's claim is false?

4. In an advertisement, a pizza shop claims that its mean delivery time is less than 30 minutes. A random selection of 36 delivery times has a sample mean of 28.5 minutes and a standard deviation of 3.5 minutes. Is there enough evidence to support the claim at $\alpha = 0.01$?

5. Survey of Voters In a survey of 1002 people, 701 said that they voted in the recent presidential election (based on data from ICR Research Group). Use a 0.05 significant level to test the claim that when surveyed, the proportion of people who say that they voted is equal to 0.61.

Chapter 11

Hypothesis Testing: t-tests

In this note, we discuss the hypothesis test of the means of **normal populations** with the assumption that the population variances are unknown. Unlike the normal test we discussed earlier, the sampling distribution of the sample means is NOT a moral distribution. We will introduce a new distribution to characterize the random behavior of the test statistic.

As an application, we also introduce a special procedure to test the difference between paired samples based on the t-test.

11.1 t-test for Normal Population Means

In the previous topic, we test means of unspecified populations based on large samples using the central limit theorem to derive the normal distribution of the test statistic. When the population is normal and population variance is unknown, then the test statistic

$$TS = \frac{\bar{x} - \mu}{s/\sqrt{n}} \to t_{n-1}$$

We introduced t-distribution when we constructed confidence intervals for normal population means in an earlier topic. The critical value(s) used to define rejection region(s) are found from the t-table.

Next, we use a numerical example to show the steps for a t-test.

Example 1. The yield of alfalfa from a random sample of six test plots is 1.4, 1.6, 0.9, 1.9, 2.2 and 1.2 tons per acre. Assume that the random sample comes from a normal population. Test at the 0.05 level of significance whether this supports the contention that the average yield for this kind of alfalfa is 1.5 tons per acre.

Solution: We are given a small raw data set of alfalfa yields $\{1.4, 1.6, 0.9, 1.9, 2.2, 1.2\}$. To perform the test, we calculate the sample statistics from the sample. $n = 6$, $\bar{x} = (1.4 + 1.6 + 0.9 + 1.9 + 2.2 + 1.2)/5 = 1.533$, and $s = 0.472$.

Step 1: The claim is clearly specified in the question **" the average yield for this kind of alfalfa is 1.5 tons per acre"**, that is, $\mu = 1.5$.

Step 2: The null and alternative hypotheses are given by

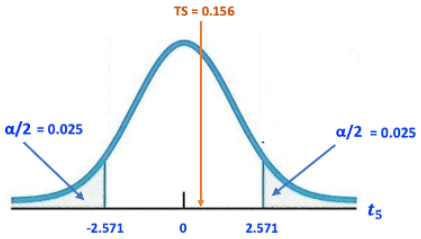
$$H_0 : \ \mu = 1.5 \ \ v.s. \ \ H_a : \ \mu \neq 1.5.$$

This is a two-tailed test. There are two rejection regions.

Step 3: The test statistic is given below

$$TS = \frac{\bar{x} - 1.5}{s/\sqrt{n}} = \frac{1.533 - 1.5}{0.472/\sqrt{6}} \approx 0.171.$$

Step 4 Since $n = 6$ and the population is normal with an unknown variance. The above test statistic is a t distribution with 5 degrees of freedom. We use the t-table to find the critical values: $CV = \pm t_{0.05/2,6-1} = \pm t_{0.025,\ 5} = \pm 2.571$



d.f	0.4000	0.2500	0.1000	0.0500	0.0250	0.0100	0.0050
1	0.3249	1.0000	3.0777	6.3138	12.7062	31.8205	63.6567
2	0.2887	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.2767	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.2707	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.2672	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.2648	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.2632	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995

Step 5: Since the test statistic $TS = 0.156$ is NOT in the rejection region, we fail to reject the null hypothesis $H_0 : \ \mu = 1.5$.

Step 6: We do not have enough evidence to reject the null hypothesis that the mean yield of the given kind of alfalfa is 1.5 tons. The data tend to support the contention.

Remarks: Here are some remarks about the t-test.

1. The steps of the t-test are identical to those in the normal test except for the distribution table used in the test.
2. If the sample size is large, the t-test and the normal test based on the central limit theorem will yield essentially the same result.
3. If the sample size is small and the normal population variance is unknown, we must use the t-test.

4. If the small size is small and the population is NOT normal, we cannot perform any test in this course.

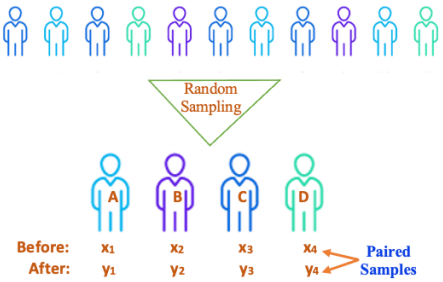
11.2 t-test for Paired Samples

This is a special hypothesis test that involves two samples. The general two-sample tests will be discussed later as a new topic. We can also use the one-sample t-test to generate the solution to the paired t-test.

11.2.1 What Are Paired Samples?

Paired samples are samples taken from the set of subjects under two different conditions such that each observation in one sample can be paired with an observation in the other sample.

To understand the idea of the paired sample, let's consider a hypothetical example. A pharmaceutical manufacturer is developing a new blood pressure drug that requires the FDA's evaluation of safety and effectiveness. To show FDA's review panel the effectiveness of the potential new drug, the clinical trial team recruits a group of subjects following the regulatory requirements. For example, the team selected 4 subjects to participate in the clinical trial. **BEFORE** they receive the new drug, the team took the blood pressure readings from the group of subjects $S_1 = \{x_1, x_2, x_3, x_4\}$ and took another blood pressure readings **AFTER** they received the new drug denoted by $S_2 = \{y_1, y_2, y_3, y_4\}$.



S_1 and S_2 are called paired samples because x_1 and y_1, x_2 and y_2, x_3 and y_3, x_4 and y_4 are paired. These paired readings were taken from the same subjects **before** and **after** receiving the treatment.

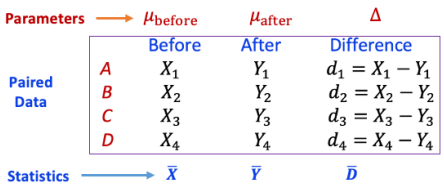
The paired sample method is widely used in many real-world applications.

11.2.2 The Logic of Paired t-test

The object of the paired t-test is to compare the mean measurement between two groups under different conditions in which each observation in one sample can be paired with an observation in the other sample.

If the "before" and "after" means are equal to each other, the drug has no treatment effect. This implies that we assess the treatment effect by comparing the two means of "before" and "after" sample means. For paired samples, we can convert this "two-sample" problem to a one-sample problem and use the regular t-test to compare the two means.

The following figure depicts the structure and notations related to the paired data and the testing procedure.



Recall that our primary interest is test a claim related to $(\mu_{before} - \mu_{after})$ to see the difference between the two means. With the above notation and fact that $\Delta = \mu_{before} - \mu_{after}$, we only need to test claims associated with Δ.

For example, testing the following hypotheses

$$H_0 \ (\mu_{before} - \mu_{after}) = 0 \ \ v.s. \ \ H_a: \ (\mu_{before} - \mu_{after}) \neq 0$$

is equivalent to testing

$$H_0: \ \Delta = 0 \ \ v.s. \ \ H_a: \ \Delta \neq 0.$$

However, the test based on Δ only needs to use the single sample data of differences between the paired measurements in the "before" and "after" samples. Therefore, we can use the regular t-test introduced earlier to perform the paired t-test.

11.2.3 Steps for Paired t-test

The steps for paired t-test are the same as those we used before except for the data preparation step added to the 6-step procedure. Due to the nature of the design of the paired sample problems, the size of the sample is usually small. We need to assume the differences of paired measurements, $\{d_1, d_2, d_3, \cdots, d_n\}$, are normally distributed.

In the initial step of data preparation, we need to calculate the mean and standard deviation of the data:

$$D = \frac{d_1 + d_2 + \cdots + d_n}{n}, \quad s_d = \sqrt{\frac{(d_1 - D)^2 + (d_2 - D)^2 + \cdots + (d_n - D)^2}{n-1}}$$

We will use the following example to illustrate the paired t-test.

Example 2. To determine whether applying a protective coating to the exterior of a printer increases its operating temperature, 6 printers were selected, and the operating temperature was recorded before and after treatment.

Printer	no coating	coating
1	186	189
2	185	186
3	179	183
4	184	188
5	183	181
6	186	188

Assuming that the temperatures are normally distributed. Does the data support the theory that the coating increases the mean operating temperature? $\alpha = 0.05$.

Solution: We first calculate the mean and standard deviation of the differences (the last column in the following table).

No-coating	coating	Difference
186	189	3
185	186	1
179	183	4
184	188	4
183	181	-2
186	188	2

After some algebra (using the given formulas), we have $D = 2$ and $s_d = 2.28$.

Step 1. The claim is that *coating increases the mean operating temperature.* That is D = coating. temp – no-coating.temp > 0.

Step 2. Setting up the null and alternative hypotheses

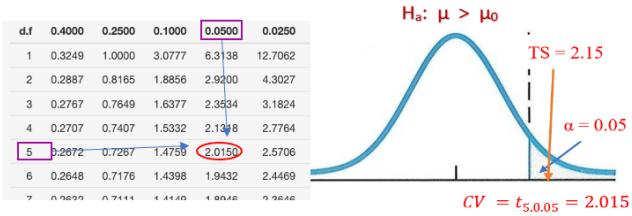
$$H_0 : \ \Delta \leq 0 \ \ versus \ \ H_a : \ \Delta > 0$$

This is a right-tailed test.

Step 3. The test statistic is defined by

$$TS = \frac{D-0}{s_d/\sqrt{n}} = \frac{2-0}{2.28/\sqrt{6}} \approx 2.15.$$

Step 4. The critical value of this right-tailed test is given by $CV = t_{5,0.05} = 2.015$. This means that the rejection region is $RR = (2.015, \infty)$. All information is summarized in the following figure.



Step 5. Since the test statistic is in the rejection region, we **reject** the null hypothesis $H_0: \Delta \leq 0$ and **conclude** the alternative hypothesis $H_a: \Delta > 0$.

Step 6. We conclude that the protective coating on the exterior of a printer increases its operating temperature.

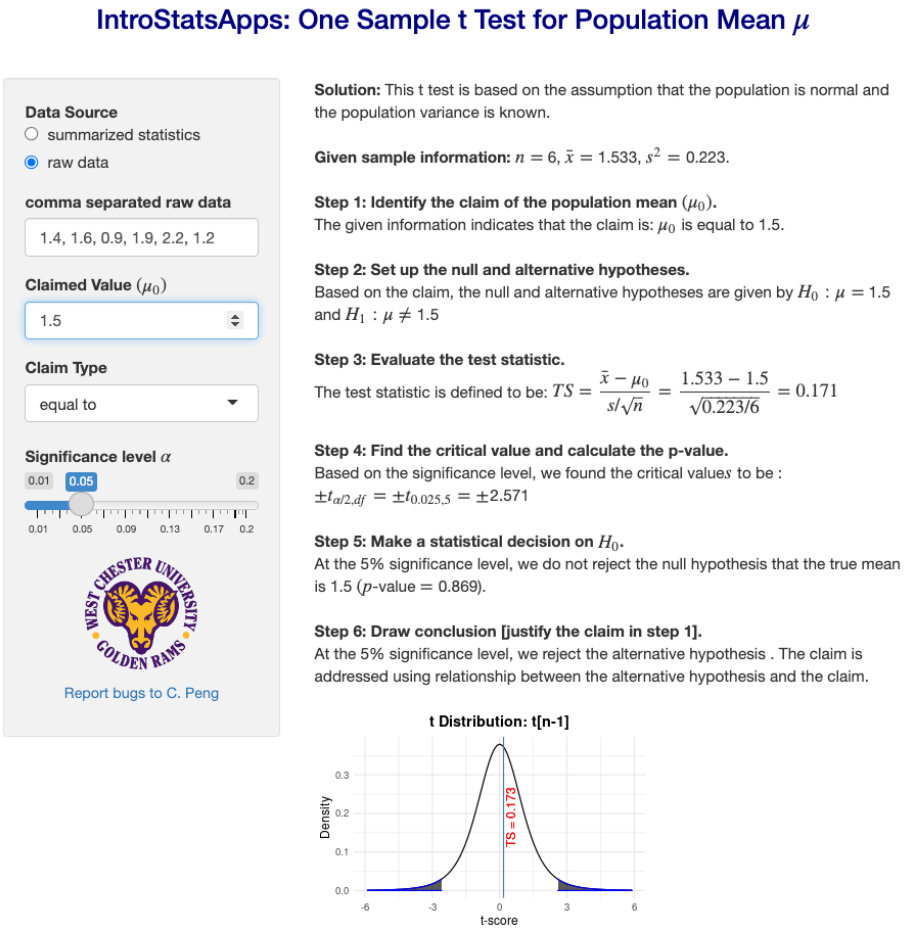
A Cautionary Remark. When calculating the differences in paired measurements, we take the form of either *before - after* or *after-before*. However, the form of the difference will impact the form of the claim. In the above example, the claim is **coating increases the mean operating temperature**, if d = *coating - no.coating*, the form of the claim is $\Delta > 0$; if d = *no.coating - coating*, the form of the claim is $\Delta < 0$.

11.3 Use of Technology

Stats App *one sample t-test* is available at: (https://wcu-peng.shinyapps.io/oneMean-ttest/). You can use this app to check your work.

11.3.1 One-sample t-test

We use the Apps to generate the solution of the above **Example 1**. Since we have raw data (i.e., individual data values), we need to type in all values with adjacent values separated by a comma.

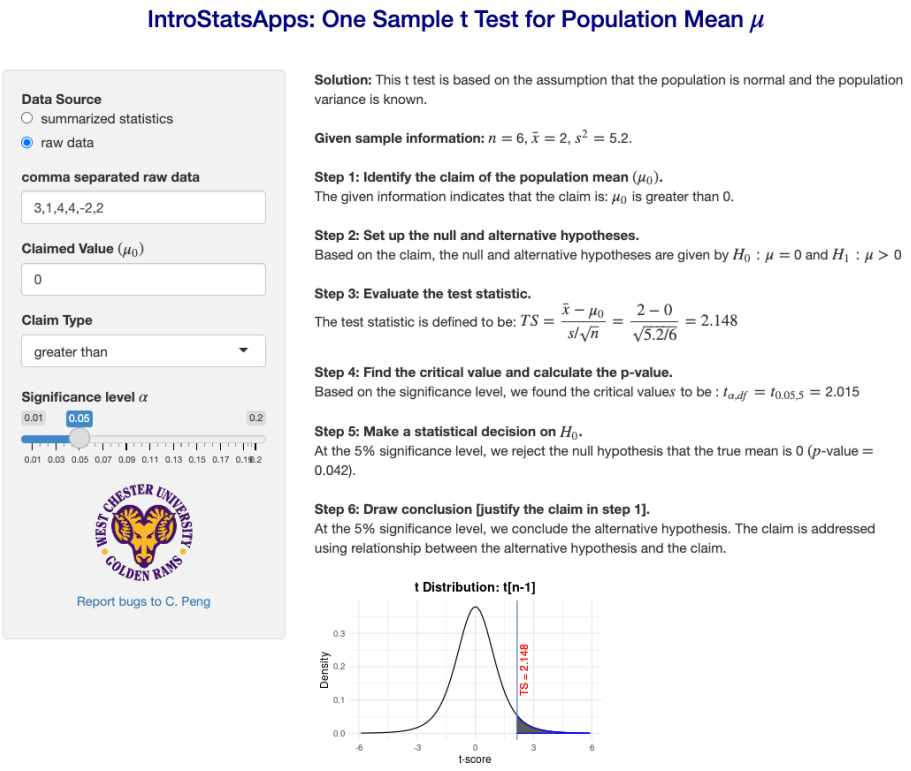


The generated solution is essentially the same as the manual solution except for small rounding-up errors.

11.3.2 Paired t-test

Before using this for paired t-test, we first calculate the difference. Then based on the form of the difference specified in the claim and provide this information to the app. To illustrate this application, we generate the solution of the above example 2.

Recall the claim and the form of difference used in example 2: D = coating. temp – no-coating.temp > 0. The set of differences is $\{3, 1, 4, 4, -2, 2\}$ and will be typed in the raw data input box.



11.4 Practice Exercises

Practice the following exercises and use the app to check your work.

1. We have the potato yield from 12 different farms. We know that the standard potato yield for the given variety is $\mu = 20$. Test if the potato yield from these farms is significantly better than the standard yield using the following random sample.

   21.5, 24.5, 18.5, 17.2, 14.5, 23.2, 22.1, 20.5, 19.4, 18.1, 24.1, 18.5

2. Two different tips are available for a hardness-testing machine. The machine operates by pressing the tip into a metal specimen and then measuring the depth of the resulting depression. Eight metal specimens are chosen, and each specimen is tested with both tips. Assuming that depths are

normally distributed, and the resulting depths are shown below (coded). At level $\alpha = 0.05$, is there any difference between the two tips?

	S1	S2	S3	S4	S5	S6	S7	S8
Tip 1	16	3	10	12	4	7	3	9
Tip 2	13	3	9	9	4	4	4	5

3. A golf club manufacturer claims that golfers can lower their scores by using the manufacturer's newly designed golf clubs. Eight golfers are randomly selected, and each is asked to give his or her most recent score. After using the new clubs for one month, the golfers are again asked to give their most recent score. The scores for each golfer are shown in the table. Assuming the golf scores are normally distributed, is there enough evidence to support the manufacturer's claim at $\alpha = 0.10$?

Golfer	1	2	3	4	5	6	7	8
Score (old)	89	84	96	82	74	92	85	91
Score (new)	83	83	92	84	76	91	80	91

4. To assess whether or not a certain training program can increase the max vertical jump (in inches) of college basketball players, we recruit a simple random sample of 20 college basketball players and measure each of their max vertical jumps and then have each player use the training program for one month. At the end of the month, we measure their max vertical jump again. The following table records the measurements.

Player	Max Vertical Jump Before Training Program	Max Vertical Jump After Training Program
Player 1	22	24
Player 2	20	22
Player 3	19	19
Player 4	24	22
Player 5	25	28
Player 6	25	26
Player 7	28	28
Player 8	22	24
Player 9	30	30
Player 10	27	29
Player 11	24	25
Player 12	18	20
Player 13	16	17
Player 14	19	18
Player 15	19	18
Player 16	28	28
Player 17	24	26
Player 18	25	27
Player 19	25	27
Player 20	23	24

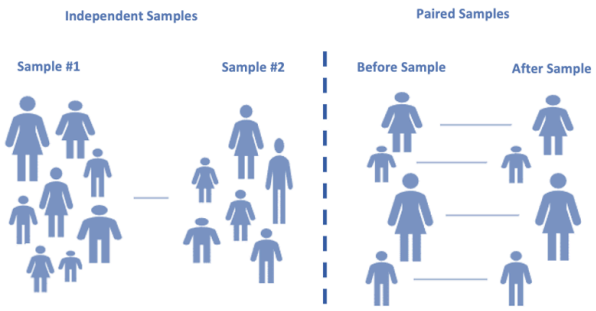
5. A professor wants to know whether the average scores of quiz 1 and quiz 2 are different. The scores of the two quizzes are given below.

Student ID	Quiz 1	Quiz 2
001	98	94
002	100	98
003	95	98
004	90	88
005	90	89
006	92	91
007	80	84
008	78	80
009	88	88

Chapter 12

Two-sample Tests

In practice, we may want to compare some characteristics of two populations. Two-sample tests are used for this purpose. In the previous topic, we introduced a special two-sample test - paired t-test in which the two populations are defined based on the same group of subjects, but measurements are taken under different conditions. In this note, we discuss two independent samples that are taken from two independent populations. For example, we want to know whether the percentages of STEM majors at WCU and Bloomsburg University are different. We take a random sample for WCU and one from Bloomsburg. There is no way to pair observations from WCU with those from Bloomsburg.



The objective of this topic is to test various hypotheses about the difference between the two population means based on both large sample and small sample scenarios.

12.1 Testing Two Populations Means: Large Samples

The general 6-step procedure will be used in the two-sample test. Before we use examples to illustrate the steps, we need to know how to define the test statistic and what is the sampling distribution of the test statistic.

12.1.1 Test Statistic and Its Sampling Distribution

The claimed the difference $\mu_1 - \mu_2$ can be estimated by $\bar{x}_1 - \bar{x}_2$. In order to define the test statistic, we need the variance of $\bar{x}_1 - \bar{x}_2$ which has form $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. In practice, population variances σ_1^2 and σ_2^2 are estimated by their corresponding sample variances s_1^2 and s_2^2, respectively.

The test statistic for testing $\mu_1 - \mu_2 = 0$ can be defined to be of the following form

$$TS = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \to N(0, 1)$$

12.1.2 Steps for Testing Two Means

The 6-step procedure for testing the difference between two population means. The next example shows the detailed steps for the two-sample test based on large samples.

Example 1. The American Automobile Association claims that the average daily cost for meals and lodging for vacationing in Texas is less than the same average cost for vacationing in Virginia. The table shows the results of a random survey of vacationers in each state. The two samples are independent. At $\alpha = 0.01$, is there enough evidence to support the claim?

Texas (1)	Virginia (2)
$\bar{x}_1 = \$248$	$\bar{x}_2 = \$252$
$s_1 = \$15$	$s_2 = \$22$
$n_1 = 50$	$n_2 = 35$

Solution: We follow the 6-step to perform the hypothesis testing.

Step 1: The statement *" the average daily cost for meals and lodging for vacationing in Texas (μ_1) is less than the same average cost for vacationing in Virginia (μ_2)."* Therefore, the claim is $\mu_1 - \mu_2 < 0$.

Step 2: The null and alternative hypotheses are given by

$$H_0: \ \mu_1 - \mu_2 \geq 0 \ \ v.s. \ \ H_a: \ \mu_1 - \mu_2 < 0.$$

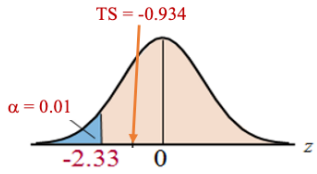
The alternative hypothesis indicates that this is left tailed test.

Step 3: The test statistic is defined to be

$$TS = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(248 - 252) - 0}{\sqrt{\frac{15^2}{50} + \frac{22^2}{35}}} = -0.934$$

Step 4: Since the test statistic is normally distributed. The critical value of this left-tailed test is $CV = -z_{0.01} = -2.33$

	0.00	0.01	0.02	0.03	0.04	0.05
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158



We can also find the p-value $= P(Z < -0.934) \approx 0.1752$.

Step 5: Both critical value and p-value methods indicate that the null hypothesis is NOT rejected. This implies that the alternative hypothesis is supported.

Step 6: The sample evidence supports the claim that the average daily cost for meals and lodging for vacationing in Texas is less than the same average cost for vacationing in Virginia.

Remark: For all two-sample tests of the difference between two population means, we need to keep the form of difference of the two means consistent in (1) claim; (2) null and alternative hypotheses; and (3) the test statistic.

12.2 Two-sample t-tests

In the previous section, we test the difference between two population means based on a large sample assumption so that the test statistic is approximately normally distributed.

Now, we want to test the difference of means of **two normal populations with unknown but equal variances**. Since the two population variances

are assumed to be equal, we need to combine the two samples to estimate the common variance.

$$s^2_{pool} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The test statistic is defined by

$$TS = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s^2_{pool}/n_1 + s^2_{pool}/n_2}} \to t_{n_1+n_2-2}$$

The sampling distribution of TS is t-distribution with $n_1 + n2 - 2$ degrees of freedom. We next perform the two-sample t-test using the above test statistic and its sampling distribution with a numerical example.

Example 2. The braking distances of 8 Volkswagen GTIs and 10 Ford Focuses were tested when traveling at 60 miles per hour on dry pavement. The results are shown below. Can you conclude that there is a difference in the mean braking distances of the two types of cars? Use $\alpha = 0.01$. Assume the populations are normally distributed and the population variances are equal.

GTI (1)	Focus (2)
$\bar{x}_1 = 134$ ft	$\bar{x}_2 = 143$ ft
$s_1 = 6.9$ ft	$s_2 = 6.6$ ft
$n_1 = 8$	$n_2 = 10$

Solution: The 6-step procedure is given below.

Step 1: The claim *there is a difference in the mean braking distances of the two types of cars* implies that $\mu_{GTI} - \mu_{Ford} \neq 0$.

Step 2: The null and alternative hypotheses are given below.

$$H_0 : \ \mu_{GTI} - \mu_{Ford} = 0 \ \ v.s. \ \ H_a : \ \mu_{GTI} - \mu_{Ford} \neq 0$$

This is a two-tailed test.

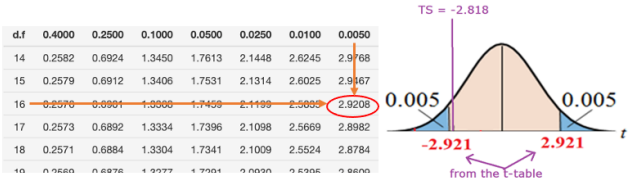
Step 3: The pooled sample variance is calculated as follows

$$s^2_{pool} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(8-1)6.9^2 + (10-1)6.6^2}{8 + 10 - 2} = 45.33.$$

The test statistic is given by

$$TS = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s^2_{pool}/n_1 + s^2_{pool}/n_2}} = \frac{(134 - 143) - 0}{\sqrt{45.33/8 + 45.33/10}} \approx -2.818.$$

Step 4: The t-critical value of this two-tailed test with 16 degrees of freedom is $CV = \pm 2.921$.



d.f	0.4000	0.2500	0.1000	0.0500	0.0250	0.0100	0.0050
14	0.2582	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.2579	0.6912	1.3406	1.7531	2.1314	2.6025	2.9467
16	0.2576	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.2573	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.2571	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784

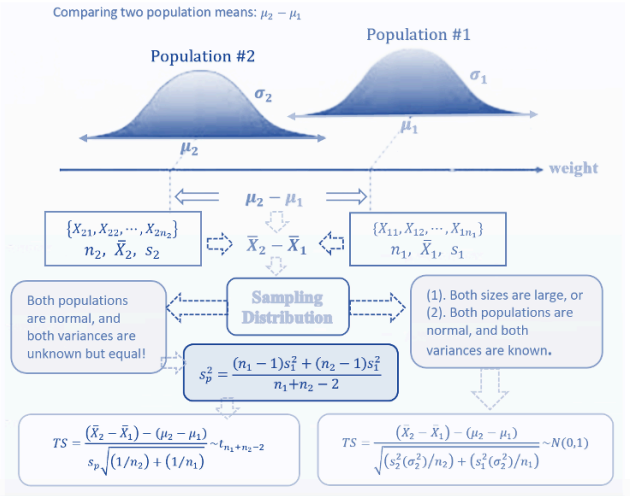
Step 5: Since the test statistic is NOT inside the rejection, we fail to reject the null hypothesis.

Step 6: We have enough sample evidence to support the claim that there is a difference in the mean braking distances of the two types of cars.

Remarks: (1). If one of the sample sizes is small, we have to assume both populations to be normal and variance are unknown but equal; (2). If any of the assumptions are not satisfied, we cannot perform any two-sample test in this class.

12.3 Two-sample Test Workflow: Summary

The following flow chart shows the workflow of two-sample tests.



12.4 Practice Exercises

1. Suppose we have a dataset containing 130 observations of body temperature, along with the gender of each individual and his or her heart rate. Is there a significant difference between the mean body temperatures for men and women? Summarized sample statistics are:

	n	Mean	Stdev
Women	65	98.105	0.699
Men	65	98.395	0.743

2. A consumer education organization claims that there is a difference in the mean credit card debt of males and females in the United States. The results of a random survey of 200 individuals from each group are shown below. The two samples are independent. Do the results support the organization's claim? Use $\alpha = 0.05$.

Females (1)	Males (2)
$\bar{x}_1 = \$2290$ $s_1 = \$750$ $n_1 = 200$	$\bar{x}_2 = \$2370$ $s_2 = \$800$ $n_2 = 200$

3. A manufacturer claims that the calling range (in feet) of its 2.4-GHz cordless telephone is greater than that of its leading competitor. You perform a study using 14 randomly selected phones from the manufacturer and 16 selected similar phones from its competitor. The results are shown below. At $\alpha = 0.05$, can you support the manufacturer's claim? Assume the populations are normally distributed and the population variances are equal.

Manufacturer (1)	Competition (2)
$\bar{x}_1 = 1275$ ft $s_1 = 45$ ft $n_1 = 14$	$\bar{x}_2 = 1250$ ft $s_2 = 30$ ft $n_2 = 16$

12.5 Use of Technology

The Stats Apps for the two-sample test is at: (https://chpeng.shinyapps.io/twoSampleTests/).

12.5.1 Two-sample Test: Large Samples

IntroStatsApps: Two-sample Test About $\mu_1 - \mu_2$

Sample #1

sample mean $(\bar{x}_1)$
248

sample variance (s_1^2)
225

sample size (n)
50

Sample #2

sample mean $(\bar{x}_2)$
252

sample variance (s_2^2)
484

sample size (n_2)
35

Claimed Value $(\mu_1 - \mu_2)$
0

Claim Type
less than

Significance level α
0.01 0.2
0.01 0.03 0.05 0.07 0.09 0.11 0.13 0.15 0.17 0.19 0.2

Solution: Since both sample sizes are greater than 30, this normal test is based on the Central Limit Theorem (CLT).

Given sample information: $n_1 = 50$, $\bar{x}_1 = 248$, $s_1^2 = 225$; $n_2 = 35$, $\bar{x}_2 = 252$, $s_2^2 = 484$.

Step 1: Identify the claim of the population mean $(\mu_1 - \mu_2)$**.**
The given information indicates that the claim is: $\mu_1 - \mu_2$ is less than 0.

Step 2: Set up the null and alternative hypotheses.
Based on the claim, the null and alternative hypotheses are given by $H_0 : \mu_1 - \mu_2 = 0$ and $H_1 : \mu_1 - \mu_2 < 0$.

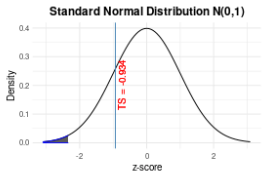
Step 3: Evaluate the test statistic.
The test statistic is defined to be:

$$TS = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{(248 - 252) - 0}{\sqrt{225/50 + 484/35}} = -0.934$$

Step 4: Find the critical value and calculate the p-value.
Based on the significance level, we found the critical value$s$ to be : $-z_\alpha = -z_{0.01}$ = -2.326
The p-value is can be found as p-value ≈ 0.175.

Step 5: Make a statistical decision on H_0**.**
At the 1% significance level, we do not reject the null hypothesis. (p-value = 0.175).

Step 6: Draw conclusion [justify the claim in step 1].
At the 1% significance level, we reject the alternative hypothesis . The claim is addressed using relationship between the alternative hypothesis and the claim.



12.5.2 Two-sample t-test

IntroStatsApps: Two-sample Test About $\mu_1 - \mu_2$

Sample #1

sample mean $(\bar{x}_1)$
134

sample variance (s_1^2)
47.61

sample size (n)
8

Sample #2

sample mean $(\bar{x}_2)$
143

sample variance (s_2^2)
43.56

sample size (n_2)
10

Claimed Value $(\mu_1 - \mu_2)$
0

Claim Type
equal to

Significance level α
0.01 0.2
0.01 0.03 0.05 0.07 0.09 0.11 0.13 0.15 0.17 0.19 0.2

Solution: Since one of the sample sizes is less than 31. The following normal test assumes both populations are normal and the two unknown population variances are equal.

Given sample information: $n_1 = 8, \bar{x}_1 = 134, s_1^2 = 47.61$.
$n_2 = 10, \bar{x}_2 = 143, s_2^2 = 43.56$.

Step 1: Identify the claim of the population mean $(\mu_1 - \mu_2)$**.**
The given information indicates that the claim is: $\mu_1 - \mu_2$ is equal to 0.

Step 2: Set up the null and alternative hypotheses.
Based on the claim, the null and alternative hypotheses are given by $H_0 : \mu_1 - \mu_2 = 0$ and $H_1 : \mu_1 - \mu_2 \neq 0$

Step 3: Evaluate the test statistic.
We first find the pooled sample variance in the following

$$s_{pool}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(8-1)47.61 + (10-1)43.56}{8 + 10 - 2} = 45.332$$

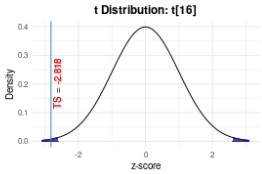
The test statistic is defined to be:

$$TS = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{(134 - 143) - 0}{\sqrt{45.332/8 + 45.332/10}} = -2.818$$

Step 4: Find the critical value and calculate the p-value.
Based on the significance level, we found the critical values to be : $\pm t_{\alpha/2,df} = \pm t_{0.005,16} = \pm 2.921$

Step 5: Make a statistical decision on H_0**.**
At the 1% significance level, we do not reject the null hypothesis that the true mean is 0 (p-value = 0.012).

Step 6: Draw conclusion [justify the claim in step 1].
At the 1% significance level, we reject the alternative hypothesis . The claim is addressed using relationship between the alternative hypothesis and the claim.



Chapter 13

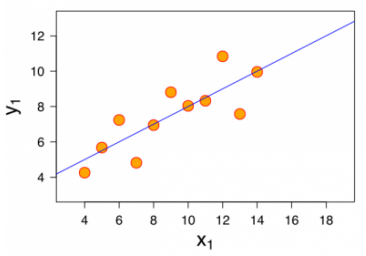
Correlation and Least Square Regression

In this note, we focus on the relationship between continuous numeric variables. There are different types of relationships between two numeric variables. The relationship we are interested in is the linear relationship. Two specific topics to be covered in this note are

- **Correlation coefficient** - determining if there is a relationship between these two variables.
- **Linear regression** - Describing how the values of one variable change when the corresponding changes in the other variable.

13.1 Correlation Coefficient

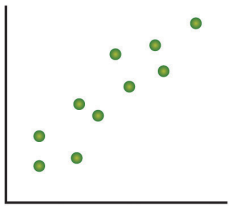
A correlation exists between two numeric variables when one of them is related to the other in some ways. To visualize the relational pattern, we use a graphic tool - scatter plot (or scatter diagram), which is a graph of the paired (x, y) data with a horizontal x-axis and a vertical y-axis.



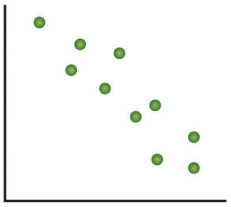
13.1.1 Linear Correlations

We now look at a few scatter plots that demonstrate different general relationships.

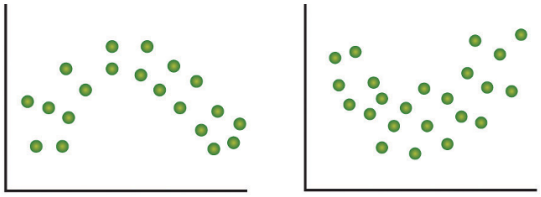
- A relationship is **linear** when the points on a scatter plot follow a somewhat straight-line pattern.
  - **Positive linear association**: The scatter plot has points that incline upwards to the right: As x values increase, y values increase. As x values decrease, y values decrease.



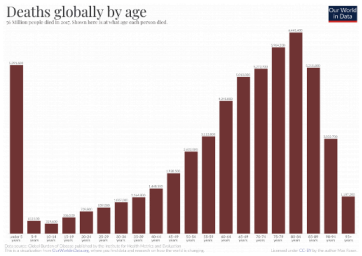
- **Example 1**. There's a positive **correlation** between height and weight. In general, as the weight increases, the height increases.
  - **Negative linear association**: The scatter plot has points that decline down to the right. As x values increase, y values decrease. As x values decrease, y values increase.
  - **Example 2**. There exist a negative correlation between absence and the scores obtained by the student in the exams, i.e., the more lectures a student missed, the lower the scores he/she will obtain in the exam.

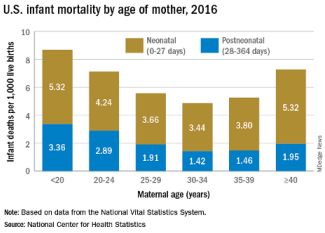


- **Non-linear Relationships** have an apparent pattern, just not linear. The following two figures represent a quadratic relationship between two numeric variables.

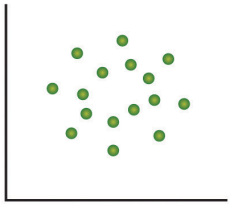


+ **Example 3**. Non-linear relationship is everywhere in the real world. For example, the following figure based on real-world data shows two special non-linear relationships. *Left Panel*: the relationship between age and death rate worldwide in 2007. *Right Panel*: the US infant mortality rate and maternal age.





* When two variables have **no relationship**, there is no straight-line relationship or non-linear relationship. When one variable changes, it does not influence the other variable.



- **Example 4**. The amount of coffee that individuals consume and their shoe sizes have no relationship with each other.

13.1.2 Linear Correlation Coefficient

The above visual representations and examples demonstrate the various relationship between two numeric variables. To quantify the **strength** and **direction** of the relationship between two variables, we use the **linear** correlation coefficient that can be estimated from sample data using the following formula

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

The structure of data used in estimating the correlation coefficient is something like the following data which will be used in the following Example 5.

Heightcm	Weightkg
150.00	49.44
159.00	62.60
172.00	75.75
153.00	48.99
166.00	53.09
161.00	52.62
156.00	47.97
150.00	45.59
167.00	57.85

We can think about **Height(cm)** and **Weight(kg)** to be X and Y. The components in the formula of the correlation coefficient are important 'sum of squares which are used in the parameters in the simple linear regression (with only one independent variable).

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad and \quad SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

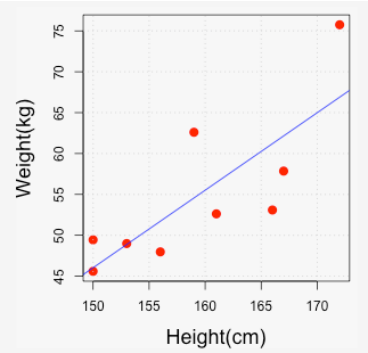
With the above notation, we can re-express the correlation coefficient as

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}}\sqrt{SS_{yy}}}$$

The calculation of the sum of squares is not difficult but could be time-consuming.

Example 5. Let's consider the relationship between height and weight. A sample data set is given in the above data table. Make a scatter plot and calculate the correlation coefficient.

Solution: We first make the scatter plot in the following.



Using the above formula, we calculate the coefficient of correlation between weight and height and obtain $r = 0.789$. You can use IntroStatsApps (https://chengpeng.shinyapps.io/correlation-reg/) to calculate the correlation coefficient on any other data. This data was used in the App as a default example.

The interpretation of the correlation coefficient is summarized in the following table.

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

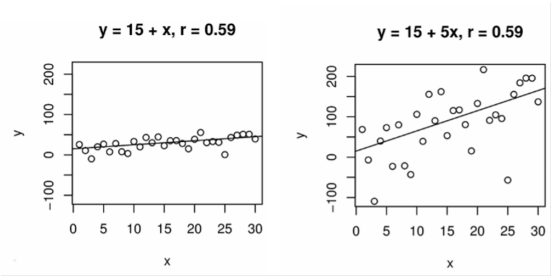
Important Remarks

1. Correlation coefficient is defined to measure the strength of the **linear** correlation between two numeric variables. Therefore, the correlation coefficient should never be used to measure the non-linear relationship between two numeric variables.
2. In general, a linear correlation does not necessarily imply causation.
3. If we use notation **corr(X, Y)** to denote the correlation coefficient between X and Y, then $cor(X, Y) = cor(Y, X)$.

13.2 Least Square Regression Lines

The linear correlation coefficient provides us with the strength and the direction of the association between two numeric variables. However, it does not tell how the change of one variable is impacted by the change of the other variable. For

example, in the following figure, if we increase x by one unit, the change of y in the left plot is less than the change in y in the right plot. However, the correlation coefficients of the two variables are the same.



13.2.1 Linear Regression and Interpretations

The equation of the linear regression line is, in general, given by

$$y = b + mx$$

It gives the explicit relationship between two variables. b is the intercept and m is the slope. Variable x is called a predictor or explanatory variable that explains the other variable y called the response or dependent variable.

- If $m > 0$, x and y are positively (linearly) correlated.
- If $m < 0$, x and y are negatively (linearly) correlated.
- If $m = 0$, x and y are NOT **linearly** correlated.

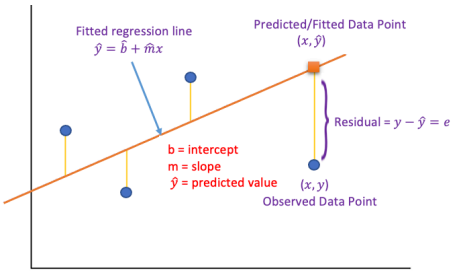
Note that both b and m are estimated from the. Once their estimated values are obtained, the estimated regression model is written in the following form

$$\hat{y} = \hat{b} + \hat{m}x$$

where

- $\hat{y}$ = predicted (or fitted) value.
- $\hat{b}$ and $\hat{m}$ are estimated intercept and slope.

The following figure shows the concepts given above.



Example 6. A hydrologist creates a model to predict the volume flow for a stream at a bridge crossing with a predictor variable of daily rainfall in inches.

$$\hat{y} = 1.6 + 29x.$$

The **y-intercept** $b = 1.6$ can be interpreted this way: On a day with no rainfall, there will be 1.6 gal. of water/min. flowing in the stream at that bridge crossing.

The **slope** $m = 29$ tells us that if it rained one inch that day the flow in the stream would increase by an additional 29 gal./min. If it rained 2 inches that day, the flow would increase by an additional 58 gal./min.

Prediction: What would be the average stream flow if it rained 0.45 inches that day?

$\hat{y} = 1.6 + 29x = 1.6 + 29(0.45) = 14.65$ gal./min.

13.2.2 Estimating Regression Coefficients

The structure of the data set for the regression is the same as the one used in calculating the correlation coefficient (see the Weight and Height data set). In fact, we can use the **sum of squares** introduced above to estimate the regression coefficients in the following.

$$m = \frac{SS_{xy}}{SS_{xx}} \quad and \quad b = \bar{y} - m\bar{x}$$

With the above explicit expression of the regression coefficient, we can estimate the intercept and slope from given data sets.

Example 7. **Determining If There Is a Relationship**: Is there a relationship between the alcohol content and the number of calories in 12-ounce beer? To determine if there is one a random sample was taken of beer's alcohol content and calories and the data is in the following table.

Brand	Brewery	Alcohol Content	Calories in 12 oz
Big Sky Scape Goat Pale Ale	Big Sky Brewing	4.70%	163
Sierra Nevada Harvest Ale	Sierra Nevada	6.70%	215
Steel Reserve	MillerCoors	8.10%	222
Coors Light	MillerCoors	4.15%	104
Genesee Cream Ale	High Falls Brewing	5.10%	162
Sierra Nevada Summerfest Beer	Sierra Nevada	5.00%	158
Michelob Beer	Anheuser Busch	5.00%	155
Flying Dog Doggie Style	Flying Dog Brewery	4.70%	158
Big Sky I.P.A.	Big Sky Brewing	6.20%	195

Solution: The objective of least square regression is to find the intercept b and the slope m to uniquely determine the regression line based on the data set and then use the fitted regression equation to answer the questions.

We use the following table to calculate the sum of squares that are used to estimate the regression coefficients.

Alcohol Content	Calories	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
4.70	163	-0.8167	-7.2222	0.6669	52.1605	5.8981
6.70	215	1.1833	44.7778	1.4003	2005.0494	52.9870
8.10	222	2.5833	51.7778	6.6736	2680.9383	133.7593
4.15	104	-1.3667	-66.2222	1.8678	4385.3827	90.5037
5.10	162	-0.4167	-8.2222	0.1736	67.6049	3.4259
5.00	158	-0.5167	-12.2222	0.2669	149.3827	6.3148
5.00	155	-0.5167	-15.2222	0.2669	231.7160	7.8648
4.70	158	-0.8167	-12.2222	0.6669	149.3827	9.9815
6.20	195	0.6833	24.7778	0.4669	613.9383	16.9315
5.516667 $=\bar{x}$	170.2222 $=\bar{y}$			12.45 $=SS_{xx}$	10335.5556 $=SS_{yy}$	327.6667 $=SS_{xy}$

Based on the sum of squares in the above table and the formulas for the regression coefficients, we have

$$\hat{m} = \frac{SS_{xy}}{SS_{xx}} = \frac{327.667}{12.45} \approx 26.3.$$

$$\hat{b} = \bar{y} - \hat{m}\bar{x} = 170.222 - 26.3 \times 5.51667 \approx 25.0$$

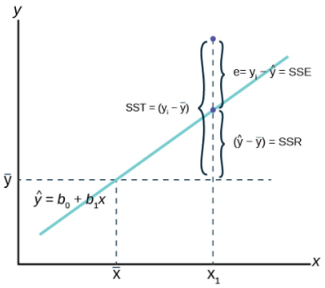
Therefore, the estimated (also called fitted) regression line is given by

$$\hat{y} = 25 + 26.3x.$$

The above regression indicates that if we increase the alcohol content by 1 unit, the corresponding number of calories increases by 26.3 units. The above regression equation can also be used as a prediction model when a new

13.2.3 Coefficient of Determination

The coefficient of determination assesses the goodness of the regression line by measuring the amount of variation in the response captured by the regression model. To develop a formula to calculate the coefficient of determination, we need the following sum of squares of errors depicted in the following figure.



where

- **The sum of squares of total variability about the mean (SST)**: $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ measures the difference between the observed and the mean.
- **the sum of squares due to regression (SSR)**: $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ represents the variability `explained by the regression line`.
- **the sum of squares due to error (SSE)**: $SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$ represents the prediction errors.

Note that, **Total Variation (SST) = Explained Variation (SSR) + Unexplained Variation (SSE)**

Therefore, we have the following definition of the coefficient of determination

$$R^2 = \frac{Variation\ \ Explained}{Total\ \ Variation} = \frac{SSR}{SST}$$

Interpretation of R^2: The percentage of total variation (in the response) by the regression.

Relationship between the correlation coefficient (r) and the coefficient of determination (R^2): $R^2 = r^2$

13.2.4 Inference of Regression Coefficients

Two major applications of regression models are association analysis and predictive analysis. The inference will focus on these two applications. Although the following discussions are valid for more general regression models, we restrict our discussion to simple linear regression: $y = b + mx$.

- **Association Analysis**: The goal of association analysis is to assess the relationship between the two numeric variables through slope coefficient(s). As usual, confidence intervals and testing hypotheses are inferential tools for analyzing regression coefficients.
  - **Confidence Interval Method**: The confidence intervals we discussed in this class are called two-sided confidence intervals. We can

also construct two-sided confidence intervals for slope m in the above regression model.

  * if 0 is inside the confidence interval, x and y do NOT have a significant linear correlation.
  * if 0 is NOT in the confidence interval, x and y have a significant linear correlation. To explore the direction of the linear correlation, one-sided confidence intervals are needed. This is not covered in this course.

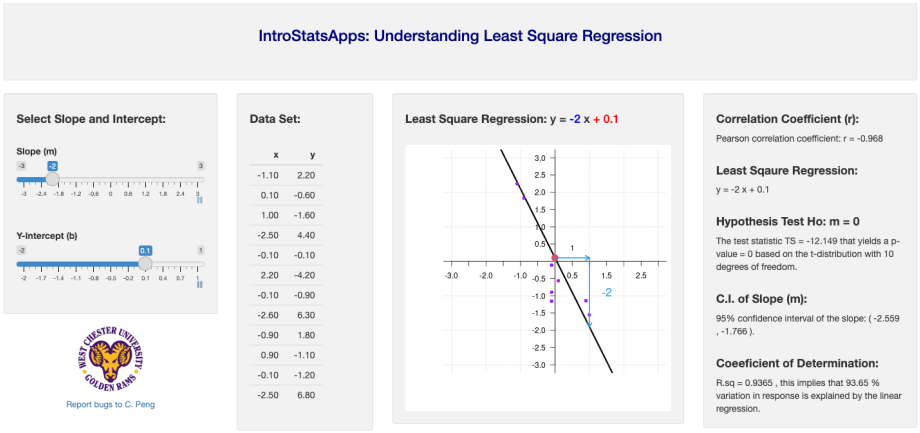
– **Testing Hypothesis**: By default, computer programs test $H_0 : m = 0$ v.s. $H_a : m \neq 0$ and report the p-value for a statistical decision on whether the two variables are correlated. For testing the positive/negative correlation between x and y, we need to redefine the p-value based on the output from computer programs.

13.3 Use of Technology

Two StatsApps were created for studying the linear relationship between two numeric variables.

13.3.1 Understanding Correlation and Linear Regression - Simulation

This simulation demonstrates the correlation between x and y through simulated data sets. The app is at (https://chpeng.shinyapps.io/LSE-Reg/). The following is the screenshot of the simulator. You can click the **arrows** under the slider bar to automatically select different intercepts and slopes as well as the random *y-values*.

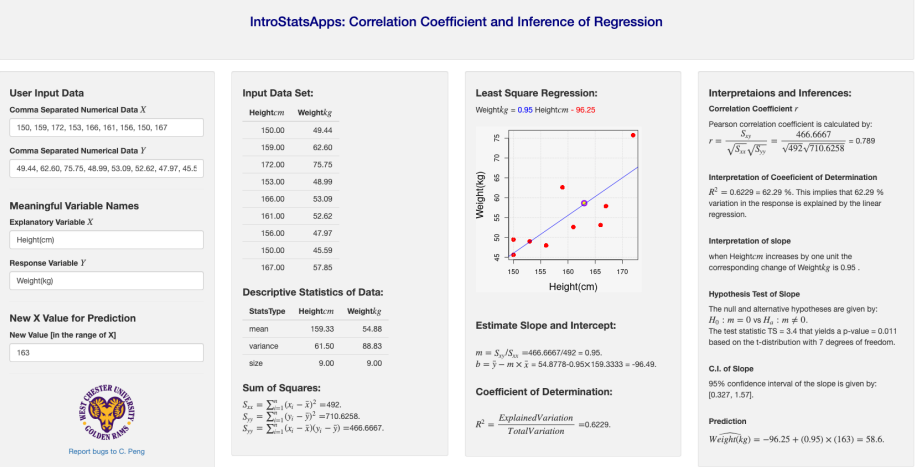


You can watch the animation in the video.

(https://github.com/pengdsci/MAT121/raw/main/notes/video/MAT121-corRegDemo.mp4)

13.3.2 Interactive Apps

This app analyzes user input data. You can click this link https://chengpeng.shinyapps.io/correlation-reg/ to use it. The following screenshot of the app.



13.4 Practice Exercises

1. When an anthropologist finds skeletal remains, they need to figure out the height of the person. The height of a person (in cm) and the length of their metacarpal bone (in cm) were collected and are in the following table.

Length of Metacarpal (cm)	Height of Person (cm)
45	171
51	178
39	157
41	163
48	172

2. The World Bank collected data on the percentage of GDP that a country spends on health expenditures ("Health expenditure," 2013) and also the percentage of women receiving prenatal care ("Pregnant woman receiving," 2013). The part of the data for the countries where this information is available for the year 2011 is in the following table.

Health Expenditure (% of GDP)	Prenatal Care (%)
9.6	47.9
3.7	54.6
5.2	93.7
5.2	84.7
10.0	100.0
4.7	42.5
4.8	96.4
6.0	77.1
5.4	58.3

(1). Create a scatter plot of the data and find a regression equation between the percentage spent on health expenditure and the percentage of women receiving prenatal care.

(2). Use the regression equation to find the percent of women receiving prenatal care for a country that spends 5.0% of GDP on health expenditure and for a country that spends 12.0% of GDP.

(3). Which prenatal care percentage that you calculated do you think is closer to the true percentage? Why?

3. A random sample of beef hotdogs was taken, and the amount of sodium (in mg) and calories were measured ("Data hotdogs," 2013). The data are in the following table.

Calories	Sodium
186	495
181	477
176	425
149	322
184	482
190	587
158	370
139	322
175	479
148	375
152	330
111	300

(1). Create a scatter plot and find a regression equation between the number of calories and the amount of sodium.

(2). Use the regression equation to find the amount of sodium a beef hotdog has if it is 170 calories and if it is 120 calories. Which sodium level that you calculated do you think is closer to the true sodium level? Why?

Chapter 14

Chi-square Tests

We have discussed the relationship between two numerical variables using linear correlation coefficient and linear regression. We now explore the relationship between two categorical variables. The idea is to make an assumption (hypothesis) about the variable (s) and use the assumption to construct a frequency table (called the expected table). At the same time, we tabulated the data to obtain an observed table. The discrepancy between the expected table and the observed table can be used to make the inference about the relationship between categorical variables.

14.1 Chi-square Test of Goodness-of-fit

A goodness-of-fit test of a distribution is a testing procedure that justifies whether the null hypothesis that specified distribution is correct based on sample information.

For a single categorical variable, the null hypothesis should specify the cell probabilities. In other words, if the category has k categories, then $(p_1, p_2 \cdots, p_k)$ must be specified in the null hypothesis.

14.1.1 A Motivational Example

As a special case, we look at the following example of testing the proportion problem.

Example 1. We want to justify a claim that about 30% of WCU students are STEM majors. That is, we test the following hypotheses.

$$H_0: \quad p = 0.3 \quad v.s. \quad H_a: \quad p \neq 0.3.$$

We take a random sample of 100 students and record the majors and found that 33 of them claimed a major in STEM. This means 67 of them are non-STEM

majors. We have introduced a procedure to test the above hypotheses with the test statistic

$$TS = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

that compares the claimed proportion with the sample proportion.

Note that the proportion of STEM majors contains the number of majors (frequencies) in both STEM and non-STEM disciplines. we can think about using (observed)sample frequencies and null (expected) frequencies to define the test statistic.

- **Under** H_0, we would **expect** to have 30 STEM majors and 70 non-STEM majors.
- We **observed** 33 STEM majors and 67 STEM majors in the random sample.

The above observed and expected number of STEM and non-STEM majors are summarized in the following table.

	Observed Values (from the sample)	Expected Value (under Ho)
STEM	$O_1 = 33$	$E_1 = 30$
non-STEM	$O_2 = 67$	$E_2 = 70$

In fact, a test statistic that measures the "distance" between the observed and expected frequency tables and has a χ^2 (chi-square) distribution is defined below

$$TS = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \to \chi^2_1.$$

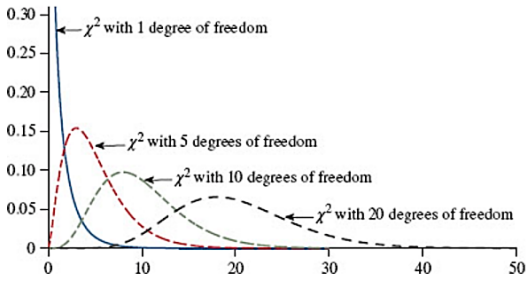
The value of the test statistic in this example

$$TS = \frac{(33-30)^2}{30} + \frac{(67-70)^2}{70} = 9/30 + 49/70 = 3/10 + 7/10 = 1$$

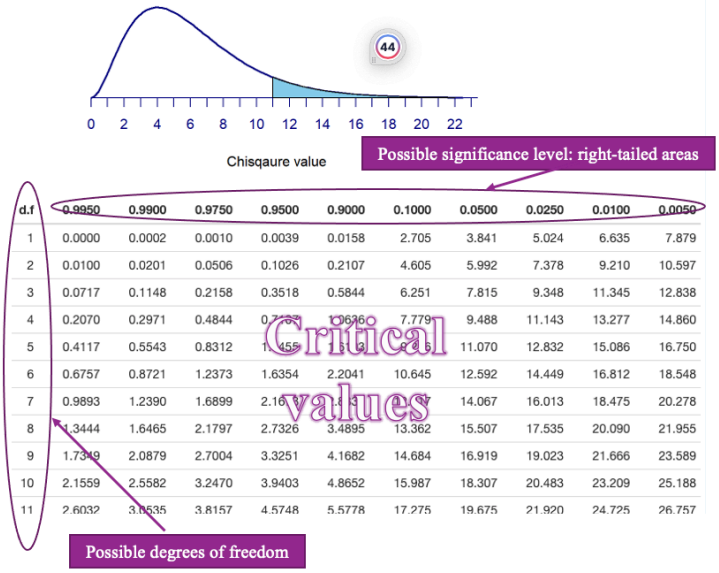
With the above value of test statistic, we can make a statistical decision on H_0 based on a given significance level.

14.1.2 Chi-square Distribution

The chi-square distribution is used to characterize the positive random variable. Unlike normal and t distributions that have symmetric density curves, the chi-square distributions (dependent on the degrees of freedom) have skewed density curves.



We can find the critical value of chi-square distribution from the chi-square table that is available on the course web page. The structure of the chi-square table is similar to the t-table.

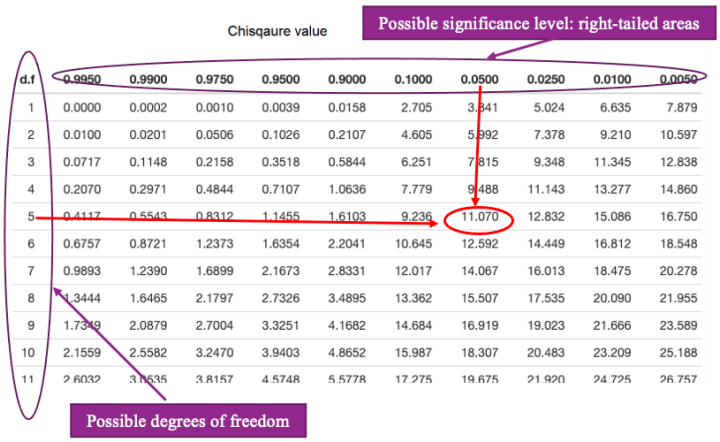


d.f	0.9950	0.9900	0.9750	0.9500	0.9000	0.1000	0.0500	0.0250	0.0100	0.0050
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.705	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.605	5.992	7.378	9.210	10.597
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.345	12.838
4	0.2070	0.2971	0.4844	0.7[illegible]	[illegible]0636	7.779	9.488	11.143	13.277	14.860
5	0.4117	0.5543	0.8312	1.[illegible]455	[illegible]	[illegible]	11.070	12.832	15.086	16.750
6	0.6757	0.8721	1.2373	1.6354	2.2041	10.645	12.592	14.449	16.812	18.548
7	0.9893	1.2390	1.6899	2.16[illegible]	[illegible]	[illegible]	14.067	16.013	18.475	20.278
8	1.3444	1.6465	2.1797	2.7326	3.4895	13.362	15.507	17.535	20.090	21.955
9	1.7349	2.0879	2.7004	3.3251	4.1682	14.684	16.919	19.023	21.666	23.589
10	2.1559	2.5582	3.2470	3.9403	4.8652	15.987	18.307	20.483	23.209	25.188
11	2.6032	3.0535	3.8157	4.5748	5.5778	17.275	19.675	21.920	24.725	26.757

The possible degrees of freedom are listed in the first column, the possible right-tail areas are listed in the top row, and the critical values are listed in the main body of the table.

The steps for finding the critical values are the same as those we followed for finding t- critical values.

Example 2. Find the critical value of the chi-square distribution with 5 degrees of freedom with significance level 0.05.



d.f	0.9950	0.9900	0.9750	0.9500	0.9000	0.1000	0.0500	0.0250	0.0100	0.0050
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.705	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.605	5.992	7.378	9.210	10.597
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.345	12.838
4	0.2070	0.2971	0.4844	0.7107	1.0636	7.779	9.488	11.143	13.277	14.860
5	0.4117	0.5543	0.8312	1.1455	1.6103	9.236	11.070	12.832	15.086	16.750
6	0.6757	0.8721	1.2373	1.6354	2.2041	10.645	12.592	14.449	16.812	18.548
7	0.9893	1.2390	1.6899	2.1673	2.8331	12.017	14.067	16.013	18.475	20.278
8	1.3444	1.6465	2.1797	2.7326	3.4895	13.362	15.507	17.535	20.090	21.955
9	1.7349	2.0879	2.7004	3.3251	4.1682	14.684	16.919	19.023	21.666	23.589
10	2.1559	2.5582	3.2470	3.9403	4.8652	15.987	18.307	20.483	23.209	25.188
11	2.6032	3.0535	3.8157	4.5748	5.5778	17.275	19.675	21.920	24.725	26.757

The above figure shows how to find the critical value, denoted by $CV = \chi^2_{5,0.05} = 11.071$. The first subscript denotes 5 degrees of freedom and the second subscript is the significance level of 0.05.

14.1.3 Formulation of Chi-square Test of Goodness-of-fit

Let k be the number of categories of categorical variable Y. The category labels are $C_1, C_2, \cdots, C_k$. Let $P_1 = Pr(C_1), P_2 = Pr(C_2), \cdots, P_k = Pr(C_k)$. The null hypothesis claims that the categorical follows a specific distribution, and the alternative hypothesis claims that the categorical distribution does NOT follow the distribution specified in the null hypothesis. That is,

$$H_0: \ P_1 = p_1, P_2 = p_2, \cdots, P_k = p_k \quad v.s. \quad H_a: \ \textit{the distribution in } H_0 \textit{ is not correct.}$$

The N is the sample size. We can then calculate the **expected cell frequency** of each category using formulas: $E_1 = N \times p_1, E_2 = N \times p_2, \cdots, E_k = N \times p_k$. The **observed cell frequency**, denoted by O_i (for $i = 1, 2, \cdots, k$), of each category is obtained from the data set. The **expected** and **observed** frequencies are summarized in the following table.

	Category 1	Category 1	⋯	Category (k-1)	Category k
Observed	O_1	O_2	...	O_{k-1}	O_k
Expected	E_1	E_2	...	E_{k-1}	E_k

The chi-square statistic is

$$G^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k} \to \chi^2_{k-1}$$

A small G^2 indicates a lack of evidence for rejecting the null hypothesis. This implies that **the Pearson chi-square test of goodness is always a right-tailed test**. The degrees of freedom are always $(k-1)$ if the categorical factor variable has k levels.

Example 3. A gambler wants to test a die to determine whether it is fair. The gambler rolls a die that has six possible outcomes: 1, 2, 3, 4, 5, and 6; and the die is fair if each of these outcomes is equally likely. The gambler rolls the die 60 times and counts the number of times each number comes up. These counts, which are called the observed frequencies, are

Outcome	1	2	3	4	5	6
Observed	12	7	14	15	4	8

Solution: The null hypothesis is that the six-sided die is fair. That is equivalent to

$$H_0: \ p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6 \quad v.s. \quad H_a: \ The\ die\ is\ NOT\ fair.$$

Based on the observed frequency table, the size of the sample is 60. Using the *cell probabilities* in H_0, we have **expected frequencies** of the 6 categories to be equal to 10. We summarize the **expected** and **observed** frequencies in the following table.

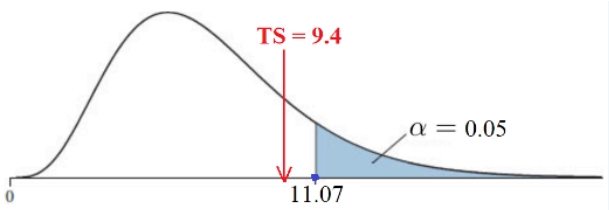
Outcome	1	2	3	4	5	6	Total
Observed	12	7	14	15	4	8	N=60
Expected	10	10	10	10	10	10	

$Np_i = E_i$ ⇨ $60\times\frac{1}{6}$ | $60\times\frac{1}{6}$ | $60\times\frac{1}{6}$ | $60\times\frac{1}{6}$ | $60\times\frac{1}{6}$ | $60\times\frac{1}{6}$

The test statistic for testing H_0 is defined by

$$G^2 = \frac{(12-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(14-10)^2}{10} + \frac{(15-10)^2}{10} + \frac{(4-10)^2}{10} + \frac{(8-10)^2}{10} = (4+9+16+25+36+4)/10 = 9.4$$

Since the categorical variable (number of dots on the faces of the 6-sided die) has 6 categories. The test statistic has 5 degrees of freedom. The critical value based on the significance level of 0.05 is given in the following figure



Degrees of Freedom	Area in Right Tail									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278

The test statistic is NOT in the rejection region. We fail to reject the null hypothesis. The die is a fair die.

Example 4. Grade distribution: A statistics teacher claims that, on average, 20% of her students get a grade of A, 35% get a B, 25% get a C, 10% get a D, and 10% get an F. The grades of a random sample of 100 students were recorded. The following table presents the sample frequencies of each grade.

Grade	A	B	C	D	F
Observed	29	42	20	5	4

Solution: Based on the given information. $N = 100$. The null hypothesis and the alternative hypothesis are

$$H_0: \ p_A = 0.2, p_B = 0.35, p_C = 0.25, p_D = 0.1, p_F = 0.1 \quad v.s. H_a: \ \textit{The claimed distribution is wrong.}$$

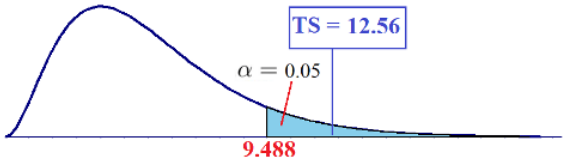
Under the null hypothesis, the expected table is given by

Grade	A	B	C	D	F
Expected	20	35	25	10	10

The test statistic has 4 degrees of freedom and the value is calculated as follows

$$G^2 = \frac{(29-20)^2}{20} + \frac{(42-35)^2}{35} + \frac{(20-25)^2}{25} + \frac{(5-10)^2}{10} + \frac{(4-10)^2}{10} = 12.56$$

The critical value with a significance level of 0.05 is given by



d.f	0.9950	0.9900	0.9750	0.9500	0.9000	0.1000	0.0500	0.0250
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.705	3.841	5.024
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.605	5.992	7.378
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348
4	0.2070	0.2971	0.4844	0.7107	1.0636	7.779	9.488	11.143
5	0.4117	0.5543	0.8312	1.1455	1.6103	9.236	11.070	12.832

Since the test statistic is inside the rejection region, we reject the null hypothesis. That is, the sample does not support the claimed grade distribution in H_0.

Remark: In the chi-square goodness-of-fit test, the crucial step is to find the *expected** frequency table under the null hypothesis.

14.2 Chi-square Test of Independence

Let X and Y be two categorical variables with k and m categories respectively. Their relationship between X and Y is characterized by their joint distribution (table). For simplicity, we use the following two special categorical to explain the ideas of statistical testing of independence.

14.2.1 Independence of Two Categorical Variables

We use the following example to illustrate **independence** and **dependence** between two categorical variables.

Example 5. *Joint probabilities and contingency tables.* Let X = political preference (Democrat vs Republican) and Y =gender (Male and Female). Let's **assume** their joint distribution to be of the following *contingency table.*

	Democrat	Republican	Row total
Male	$p_{11} = 0.3$	$p_{12} = 0.2$	$p_{1*} = 0.50$
Female	$p_{21} = 0.3$	$p_{22} = 0.3$	$p_{2*} = 0.50$
Column Total	$p_{*1} = 0.50$	$p_{*2} = 0.50$	$p_{**} = 1.00$

Column marginal probabilities

Row marginal probabilities

The cell numbers are joint probabilities. For example, $p_{12} = 0.2 = 20\%$ says 20% of the *study population* are male republicans. The row and column totals represent the percentage of male/female and democrats/republicans in the

study population. Any observed data table is **governed** by the above joint distribution table.

Definition Two categorical variables are **independent** if and only if their joint probabilities are equal to the product of their corresponding marginal probabilities.

With this definition, we can see that X and Y with joint distribution specified in the above table (in Example 5) ar NOT independent since $p_{11} = 0.3 \neq 0.5 \times 0.5 = 0.25$.

Example 6. We consider two variables X = preference of hair color (Blonde and Brunette) and Y = gender (Male and Female). Assume the joint distribution of the two variables is given by

	Male	Female	**total**
Blonde	0.18	0.27	**0.45**
Brunette	0.22	0.33	**0.55**
total	**0.40**	**0.60**	**1.00**

Based on the definition of independence. The preference for hair color is independent of gender. Since **all** joint probabilities are equal to the product of their corresponding marginal probabilities.

$0.45 \times 0.40 = 0.18, 0.45 \times 0.60 = 0.27, 0.55 \times 0.40 = 0.22$, and $0.55 \times 0.60 = 0.33$.

14.2.2 Expected Table Under Independence Assumption (H_0)

We construct the **expected table** under the **null hypothesis of independence** and the **observed contingency table**. For ease of interpretation, we use an example to illustrate the steps for obtaining the expected table.

Example 7. Consider the potential dependence between the attendance (good vs poor) and course grade (pass vs fail). We take 50 students from a population and obtain the following observed table.

	Pass	Fail	total
Good	25	2	27
Poor	8	15	23
total	33	17	50

Question: Whether the attendance is independent of class performance?

$$Ho: \ attendance \ is \ \ independent \ \ of \ the \ performance$$

$$versus$$

$$Ha: \ attendance \ is \ \ dependent \ of \ the \ performance$$

To obtain the expected table, we follow the next few steps.

1. Estimate the marginal probabilities

	Pass	Fail	Marginal Probability
Good			0.54
Poor			0.46
Marginal Probability	0.66	0.34	1.00

where marginal probabilities are calculated by Pr(Good) = 27/50 = 0.54, Pr(Poor) = 23/50 = 0.46, Pr(Pass) = 33/50 = 0.66, Pr(Fail) = 17/50 = 0.34.

2. Estimate the joint probability under the null hypothesis of independence

	Pass	Fail	Marginal Probability
Good	0.3564	0.1836	0.54
Poor	0.3036	0.1564	0.46
Marginal Probability	0.66	0.34	1.00

where joint probabilities under the independence assumption (H_0) are calculated by taking the product of the corresponding marginal probabilities. For example, $0.54 \times 0.66 = 0.3564$.

3. Calculate Expected Table

The expected frequencies are calculated in the following table (with detailed steps).

	Pass	Fail	Row Total
Good	0.3564 x 50 =**17.82**	0.1836 x 50 = **9.18**	27
Poor	0.3036 x 50 =**15.18**	0.1564 x 50 =**7.82**	23
Column Total	33	17	Sample size = 50

Remark. For categorical variables with **more than two categories**, the expected table can be found using ** the same 3 steps** as those used in the above example.

14.2.3 Formulation of Chi-squares Test of Independence

The test statistic used to test the independence of two categorical variables is the same as that used in the goodness-of-fit test. That is the standardized "distance" between the observed and the expected table (under H_0).

Assume that the two categorical variables have k and m categories respectively, then the resulting test statistic has a chi-square distribution with $(k-1) \times (m-1)$ degrees of freedom.

Example 8. [Continuation of **Example 7**]. Test whether **attendance** and **class performance**.

Solution: We have found the expected table under H_0 in **Example 7**, we put the observed and expected tables in the following.

Observed Table

	Pass	Fail	total
Good	**25**	**2**	27
Poor	**8**	**15**	23
total	33	17	50

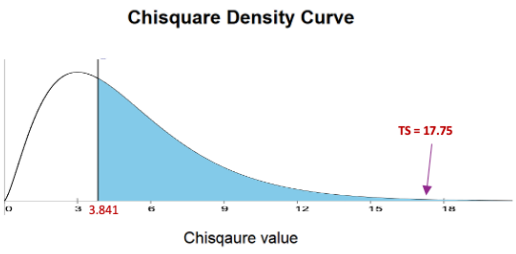
Expected Table

	Pass	Fail	total
Good	**17.82**	**9.18**	27
Poor	**15.18**	**7.82**	23
total	33	17	50

The test statistic is given by

$$TS = \frac{(25-18.82)^2}{17.82} + \frac{(2-9.18)^2}{9.18} + \frac{(8-15.18)^2}{15.18} + \frac{(15-7.82)^2}{7.82} = 17.75$$

The test statistic has a chi-square distribution with $(2-1)\times(2-1) = 1$ degrees of freedom. The critical value at the significance level of 0.05 is found in the following figure.



d.f	0.9950	0.9900	0.9750	0.9500	0.9000	0.1000	0.0500	0.0250	0.01
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.705	3.841	5.024	6.6
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.605	5.992	7.378	9.2
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.3
4	0.2070	0.2971	0.4844	0.7107	1.0636	7.779	9.488	11.143	13.2
5	0.4117	0.5543	0.8312	1.1455	1.6103	9.236	11.070	12.832	15.0

Since the test statistic is inside the rejection region, we reject the null hypothesis that attendance and class performance are independent.

Example 9. Do some college majors require more studying than others? The National Survey of Student Engagement asked a number of college freshmen what their major was and how many hours per week they spent studying, on average. A sample of 1000 of these students was chosen, and the numbers of students in each category are tabulated in the following two-way contingency table.

Hours Studying Per Week	Major				Total
	Humanities	Social Science	Business	Engineering	
0–10	68	106	131	40	345
11–20	119	103	127	81	430
More Than 20	70	52	51	52	225
Total	257	261	309	173	1000

Solution: The null and alternative hypotheses are given by

```
Ho: studying time is INDEPENDENT on majors
                versus
Ha: studying time is DEPENDENT on majors.
```

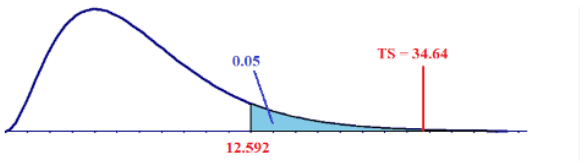
Under the null hypothesis, we obtained the expected table using the same steps in Example 7 in the following.

	Humanities	Soc Sci	Business	Engineering	Total
0-10	88.7	90.0	106.6	59.7	345
11-20	110.5	112.2	132.87	74.4	430
>20	57.8	58.7	69.5	38.9	225
total	457	261	309	173	1000

The test statistic is given by

$$TS = \frac{(68-88.7)^2}{88.7} + \frac{(106-90.0)^2}{90.0} + \cdots + \frac{(52-38.9)^2}{38.9} \approx 34.64$$

The critical value and rejection region based on significance level 0.05 is given by



d.f	0.9950	0.9900	0.9750	0.9500	0.9000	0.1000	0.0500	0.0250	0.010
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.705	3.841	5.024	6.63
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.605	5.992	7.378	9.21
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.34
4	0.2070	0.2971	0.4844	0.7107	1.0636	7.779	9.488	11.143	13.27
5	0.4117	0.5543	0.8312	1.1455	1.6103	9.236	11.070	12.832	15.08
6	0.6757	0.8721	1.2373	1.6354	2.2041	10.645	12.592	14.449	16.81
7	0.9893	1.2390	1.6899	2.1673	2.8331	12.017	14.067	16.013	18.47

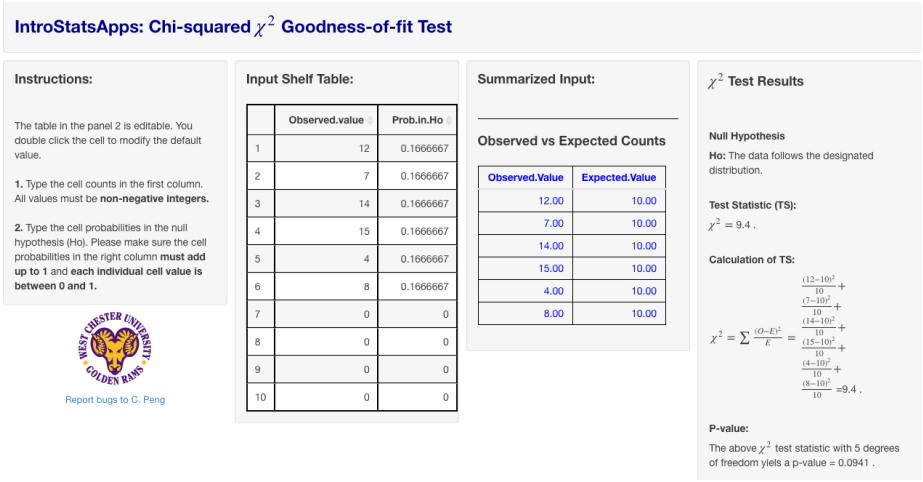
Conclusion: Since the test statistic is inside the rejection region, we reject the null hypothesis and conclude that the studying time is dependent on the majors.

14.3 Use of Technology

Two apps were created for the two chi-square tests of goodness-of-fit and independence respectively.

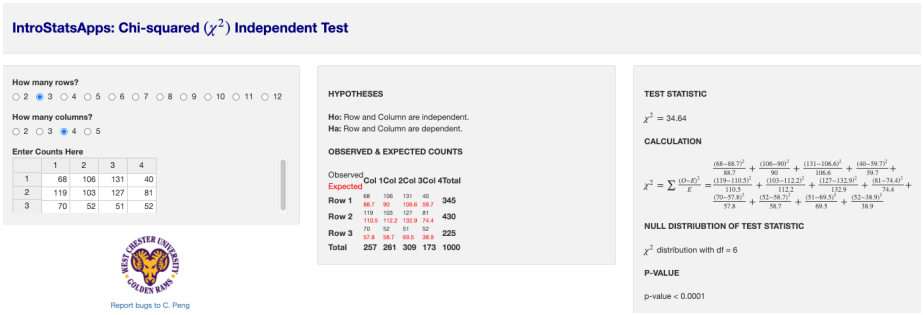
14.3.1 Goodness-of-fit Chi-square

The app is at: https://chpeng.shinyapps.io/chisq-gof/. The following screenshot illustrates the use of this app using Example 03 in this note.



14.4 Chi-square Test of Independence

The app is at https://chpeng.shinyapps.io/chisq-independence/. We use this app with Example 09.



14.5 Practice Exercises

You can use the apps to do the following exercises.

1. **College Sports**

A University conducted a survey of its recent graduates to collect demographic and health information for future planning purposes as well as to assess students' satisfaction with their undergraduate experiences. The survey revealed that a substantial proportion of students were not engaging in regular exercise, many felt their nutrition was poor and a substantial number were smoking. In response to a question on regular exercise, 60% of all graduates reported getting no regular exercise, 25% reported exercising sporadically and 15% reported exercising regularly as undergraduates. The next year the University launched a health promotion campaign on campus in an attempt to increase health behaviors among undergraduates. The program included modules on exercise, nutrition, and smoking cessation. To evaluate the impact of the program, the University again surveyed graduates and asked the same questions. The survey was completed by 470 graduates and the following data were collected on the exercise question:

	No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
Number of Students	255	125	90	470

We specifically want to compare the distribution of responses in the sample to the distribution reported the previous year (i.e., 60%, 25%, 15% reporting no, sporadic and regular exercise, respectively). Whether the data supports the above distribution at a significance level of 0.05.

2. **Political Affiliation and Opinion**

The following table based on the sample will be used to explore the relationship between Party Affiliation and Opinion on Tax Reform.

	favor	indifferent	opposed	total
democrat	138	83	64	285
republican	64	67	84	215
total	202	150	148	500

Find the expected counts for all of the cells.

3. **Tire Quality**

The operations manager of a company that manufactures tires wants to determine whether there are any differences in the quality of work among the three daily shifts. She randomly selects 496 tires and carefully inspects them. Each tire is either classified as perfect, satisfactory, or defective, and the shift that produced it is also recorded. The two categorical variables of interest are the shift and condition of the tire produced. The data can be summarized by the

accompanying two-way table. Does the data provide sufficient evidence at the 5% significance level to infer that there are differences in quality among the three shifts?

	Perfect	Satisfactory	Defective	Total
Shift 1	106	124	1	231
Shift 2	67	85	1	153
Shift 3	37	72	3	112
Total	210	281	5	496

4. **Condiment preference and gender**

A food services manager for a baseball park wants to know if there is a relationship between gender (male or female) and the preferred condiment on a hot dog. The following table summarizes the results. Test the hypothesis with a significance level of 10%.

		Condiment			
		Ketchup	Mustard	Relish	Total
Gender	Male	15	23	10	48
	Female	25	19	8	52
	Total	40	42	18	100