Contents

Chapter 1

A Review of Basic Concepts

Statistics is used in many aspects of life and has a wide spectrum of applications. It is a significant science and has a meticulous methodology that is used in almost all applied sciences such as business, economics, medicine, psychology, and actuarial science.

What is statistics?

Statistics is the science of data. It is the science of collecting, summarizing, analyzing, presenting, and interpreting data.

1.1 Statistics and Data

Eexperimental units or Observational units: Units are the entities or objects or people on which data are collected. They can also be called **subjects**, **cases**, **individuals**, or **items** such as people, cars, nations, and companies.

Variable: A variable is a characteristic or property of interest for the units which may vary among the units such as gender, height, household income, education level, and customer satisfaction.

Figure 1.1: Variable

Measurement: Measurement is the process we use to assign numbers, letters, or words to variables of units.

Observation: The set of measurements obtained for a particular unit is called an observation.

Data: Data are the facts, figures, or values obtained after the process of measurement. We collect, analyze, and summarize data for presentation and interpretation.

Data Set: All the data collected in a particular study are referred to as the data set for the study.

Note:

- A data set with n units contains n observations.

- The total number of data values in a complete data set is the number of units multiplied by the number of variables.

Types of Data

Data can be further classified as being qualitative (categorical) or quantitative (numerical). The statistical analysis that is appropriate depends on whether the data for the variable are qualitative or quantitative.

Categorical or Qualitative Data

Data that can be grouped by specific categories are referred to as categorical data. These data are measurements that cannot be measured on a natural numerical scale and we use either the nominal or ordinal scale of measurement. Labels or names are used to identify an attribute of each element. Each category can be assigned a numerical value so categorical data can be either numeric (with no numerical meaning) or non-numeric.

Quantitative or Numerical Data

Data that use numeric values to indicate how many or how much are referred to as quantitative data. Ordinary arithmetic operations are meaningful for quantitative data. Quantitative data are obtained using either the interval or ratio scale of measurement. Quantitative data are always numeric.

Categorical or Qualitative Variable

A variable with categorical or qualitative data. A categorical variable has two or more (limited and fixed) categories or classes.

Quantitative or Numerical Variable

A variable with quantitative data.

Note:

Typically, numerical variables have a unit of measurements, or simply units, like pounds (here, the unit) has to go with Weight (a numerical variable); or meters goes with Height, etc. A very good indicator is to check if the variable has units. If so, it is very likely to be a numerical variable.

Dependent or Response Variables

A researcher presents a question about a dependent variable and studies the variation of a dependent variable because of other variables.

Independent or Explanatory Variables

An independent variable explains changes in the dependent variable (or in the response).

Observational Study

In observational (nonexperimental) studies no attempt is made to control or influence the variables of interest (e.g., a survey which is the most common). For example, studies of smokers and nonsmokers are observational studies because researchers do not determine or control who will smoke and who will not smoke.

Experimental Study

In experimental studies the variable of interest is first identified. Then one or more other variables are identified and **controlled** so that data can be obtained about how they influence the variable of interest.

1.2 Populations, Samples, and Random Sampling

Population

The set of all units of interest in a particular study. Population Size, N is typically unknown.

Sample

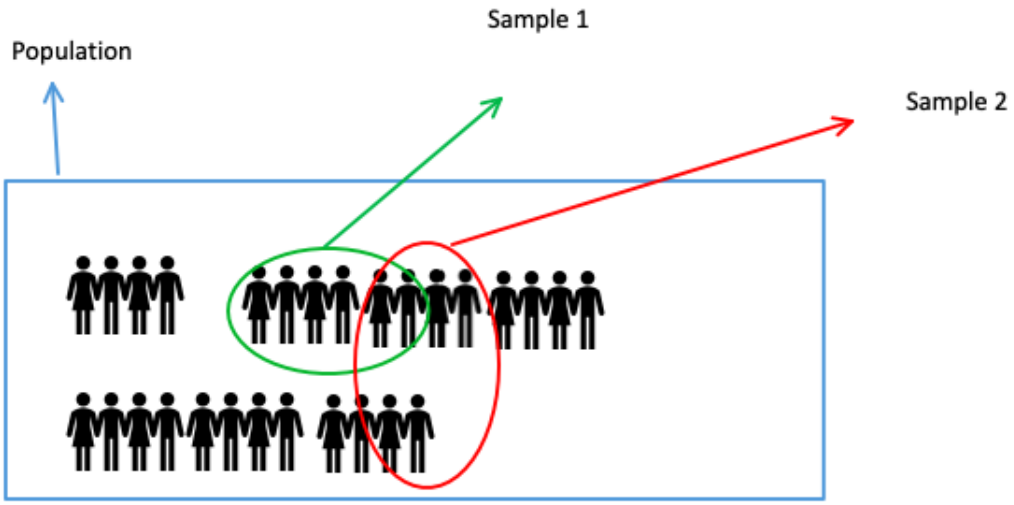
A subset of the population. Sample Size, n, is always known when getting the sample.

Census

Collecting data for the entire population.

Sample survey

Collecting data for a sample.



With proper sampling methods, the sample results can provide "good" estimates of the population characteristics.

Representative Sample

A representatives ample exhibits characteristics typical of those possessed by the population.

Random Smaple

A random sample of n observations is one selected from the population in such a way that every different sample of size n has an equal probability (chance) of selection.

Statistical Inference

The process of using data obtained from a sample to make estimates, predictions, or test hypotheses about the characteristics of a population.

Measure of Reliability

A measure of reliability is a statement (usually quantified with a probability value) about the degree of uncertainty associated with a statistical inference.

A **parameter** is a numerical characteristic of a population. (e.g population mean, μ, population standard deviation, σ, or population proportion, p.) The values of parameters are always constant or fixed. Why?

A **sample statistic** is a numerical value used as a summary measure for a sample. (e.g sample mean, $\bar{x}$, sample standard deviation, s, or sample proportion, $\hat{p}$.) The values of statistics are not constant or fixed and they vary from sample to sample.

Point Estimation

Point estimation is a form of statistical inference. In point estimation we use the data from the sample to compute a value of a sample statistic that serves as an estimate of a population parameter.

We refer to $\bar{x}$ as the point estimator of the population mean μ.

s is the point estimator of the population standard deviation, σ.

$\bar{p}$ or $\hat{p}$ is the point estimator of the population proportion P.

Sampling Error

The deviation of the sample from the population which happens when we study a sample of the population versus the entire population in order to estimate population parameters. Sampling error which always occurs in sampling usually decreases when the sample size is increased but does not disappear unless conducting a census instead. The sample design and the variability within the population can affect the sampling error.

Accuracy

Accuracy indicates how close a sample statistic (e.g. $\bar{x}$) is to a parameter of a population. (we will introduce statistic and parameter later)

1.3 Describing Qualitative Data

Summarizing Data for a Categorical Variable

Categorical data can be summarized by:

- Frequency Distribution
- Relative Frequency Distribution
- Percent Frequency Distribution
- Bar Chart

- Pie Chart

Example 1.3.1 *787 employees of a company were asked to complete a survey on their education level (some high school, high school graduate, some college, and college graduate). Here are the data on the percents and counts of employees who have different education levels.*

Bar Chart

A bar chart is a graphical display for depicting qualitative data. A frequency, relative frequency, or percent frequency scale can be used for the other axis (usually the vertical axis).

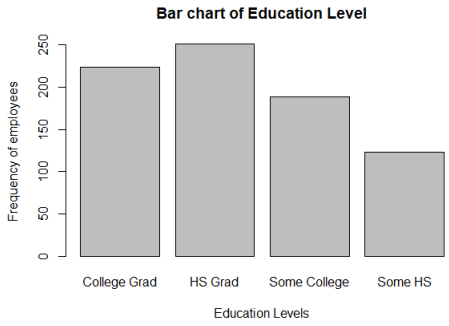


Figure 1.2: R bar graph for Education Level with y-axis in frequency

Pie Chart

The pie chart is a commonly used graphical display for presenting relative frequency and percent frequency distributions for categorical data.

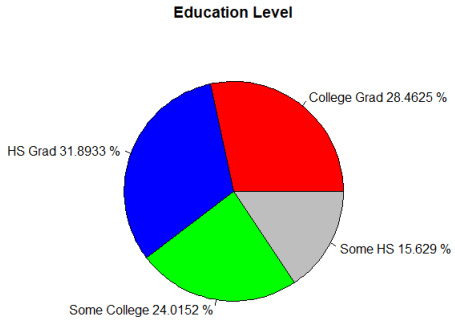


Figure 1.3: R pie chart for Education Level

Describing Quantitative Data Graphically

- Histogram
- Stem-and-Leaf Display

Example 1.3.2 *The following histogram summarizes the annual sales (in thousands of dollars) amount for some selected salespeople in a company for the last fiscal year.*

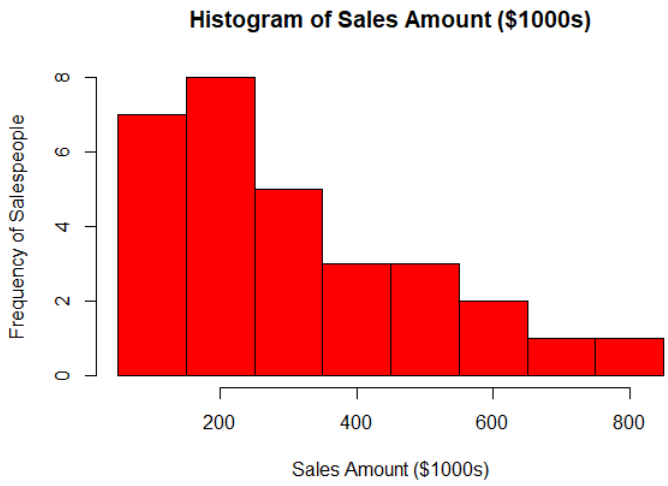


Figure 1.4: R histogram of Annual Sales

Histograms do not show the actual data (individual measurements).

Histograms Showing Skewness

One of the most important uses of a histogram is to provide information about the shape, or form, of a distribution.

- A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.
- A distribution is **skewed to the right** if the right side of the histogram (side with larger values) extends much farther out than the left side.
- A distribution is **skewed to the left** if the left side of the histogram extends much farther out than the right side.

Question: Describe the distribution (shape, center, and spread) of the parts Cost ($) for 50 tune-ups using the histogram.

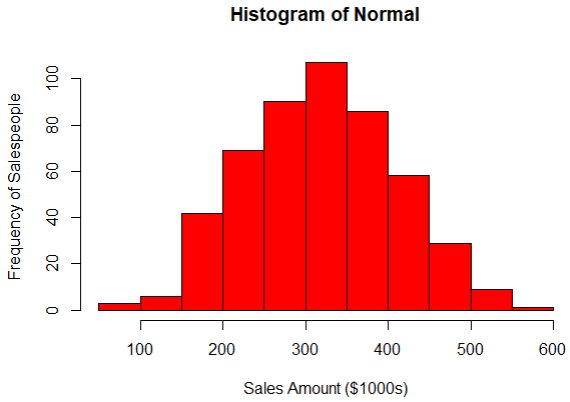


Figure 1.5: symmetric

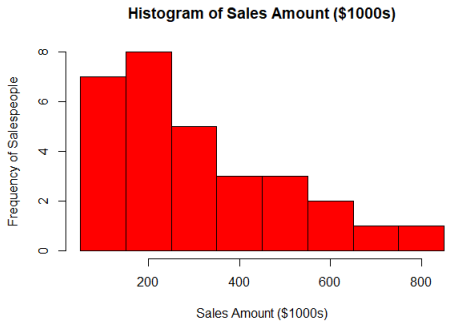


Figure 1.6: Right-skewed

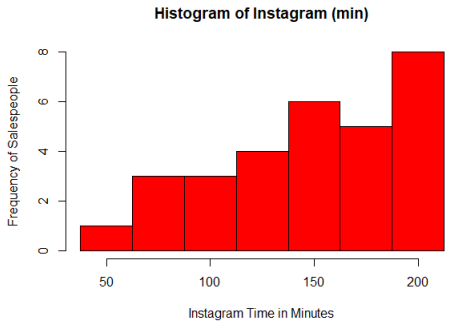


Figure 1.7: Left-skewed

1.4 Describing Quantitative Data Numerically

Measures of Centre

1- Mean

Perhaps the most important measure of centre is the mean (average). The mean of a data set is the average of all the data values. The sample mean $\bar{x}$ is the point estimator of the population mean, μ.

Formula for a Sample Mean

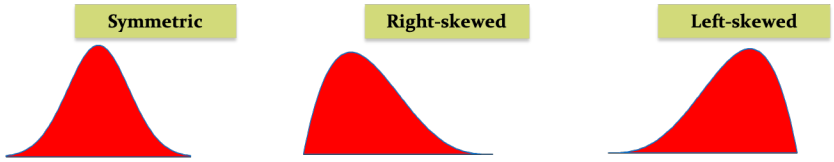


Figure 1.8: Symmetric and Skewed Distributions

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

(When: $\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_n$)

2- Median

The median of a data set is the value in the middle when the data items are arranged in ascending order. Whenever a data set has extreme values (skewed), the median is the preferred measure of central location. Most often, the median is reported for annual income and property value dat since a few extremely large incomes or property values can inflate the mean.

Calculation

1. Arrange the data in ascending order (smallest value to largest value) so that x_1 is the smallest observed value (data point) and x_n is the largest observed value (data point) in a sample of size n.

2. For an odd number of observations, the (sample) median is the middle value.

$$Median(M) = x_{\frac{n+1}{2}}$$

3. For an even number of observations, the median is the average of the two middle values.

$$Median(M) = \frac{1}{2} \times (x_{\frac{n}{2}} + x_{\frac{n+2}{2}})$$

Relative Location of Mean and Median

For a symmetric distribution, the mean and the median are equal. When the data are positively skewed, the mean will usually be greater than the median; when the data are negatively skewed, the mean will usually be less than the median.

Measures of Locations

Quartiles

Quartiles are specific percentiles. Quartiles divide all observed values into into 4 equal parts.

First Quartile = 25th Percentile denoted by Q_1

Second Quartile = 50th Percentile = Median = Q_2

Third Quartile = 75th Percentile denoted by Q_3

First part		Second part		Third part		Last part
25% of the data	Q_1	25% of the data	$M = Q_2$	25% of the data	Q_3	25% of the data

Measures of Variability (or Spread)

It is often desirable to consider measures of variability (dispersion), as well as measures of centre and location. Common measures of variability are:

1- Range Range is the simplest measure of variability which is also very sensitive to the smallest and largest data values. The range of a data set is the difference between the largest and smallest data value.

$$R = Max - Min$$

2- Interquartile Range

The interquartile range of a data set is the difference between the third quartile and the first quartile. It overcomes the sensitivity to extreme data values.

Note: The interquartile range is the range for the middle 50% of the data.

$$IQR = Q_3 - Q_1$$

Limits

Upper Limits, UL is located at $1.5(IQR)$ above Q_3.

$$\text{Upper Limit} = Q_3 + 1.5(IQR)$$

Lower Limits, LL is located at $1.5(IQR)$ below Q_1.

$$\text{Lower Limit} = Q_1 - 1.5(IQR)$$

Outlier

We consider an observation an outlier when it is unusually large or small relative to the other observations in the data set. Data outside the limits mentioned above are considered outliers.

3- Variance

The variance is a measure of variability that utilizes all the data. It is based on the difference between the value of each observation (x_i) and the mean ($\bar{x}$ for a sample, μ for a population). The variance is useful in comparing the variability of two or more variables. The variance is the average of the squared deviations between each data value and the mean.

Formula for the Sample Variance:

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

or

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

Formula for the Population Variance:

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

or

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

Example 1.4.1 *Calculate the variance of the following two data sets and compare the variability of the two data sets.*

46, 54, 42, 46, 32

46, 69, 32, 48, 71

4- Standard Deviation

The standard deviation of a data set is the positive square root of the variance. It is measured in the same units as the data, making it more easily interpreted than the variance.

$$s = \sqrt{s^2}$$

$$\sigma = \sqrt{\sigma^2}$$

Symbols for Variance and Standard Deviation

s^2= Sample variance σ^2 = Population variance

s= Sample standard deviation σ = Population standard deviation

Five-Number Summary and BoxPlot

Summary statistics and easy-to-draw graphs can be used to quickly summarize large quantities of data. Two tools that accomplish this are five-number summaries and box plots.

1- Minimum Value (Min)

2- First Quartile (Q_1)

3- Median (M)

4- Third Quartile (Q_3)

5- Maximum Value (Max)

Box Plot

A boxplot (box-and-whisker plot) is a graphical display of data that is based on a five-number summary (not all observations). A box is drawn with its ends (the hinges) located at the first and third quartiles. A vertical line is drawn in the box at the location of the median (second quartile). The horizontal lines extending from each end of the box are called whiskers. The whiskers are drawn from the ends of the box to the smallest and largest values inside the limits.

Box plot is a rectangle which describes the distribution, center (Median), spread (Range or IQR), shape of the data and identifies any **outliers**.

Example 1.4.2 *Consider the Sales Amount data set. Find the five-number summary and construct a boxplot.*

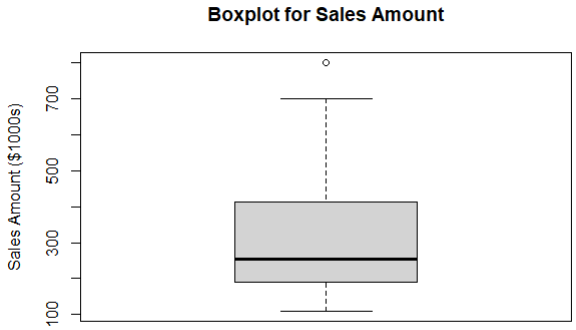


Figure 1.9: R Boxplot for Sales Amount

1.5 The Normal Probability Distribution

The normal distribution is one of the most important and common distribution for a theoretical population relative frequency distribution for a quantitative variable.

The **Normal curve (bell curve)** describes the **normal (gaussian) distribution** which is the most important distribution statistics.

- The normal curve is symmetric about its mean which is equal to the median.
- The normal curve is unimodal (single-picked) and bell-shaped.
- The Normal distribution depends only on its mean (μ) and its standard deviation (σ) which describes the spread of the distribution.
- Changing μ without changing σ moves the Normal curve along the horizontal axis without changing its variability.

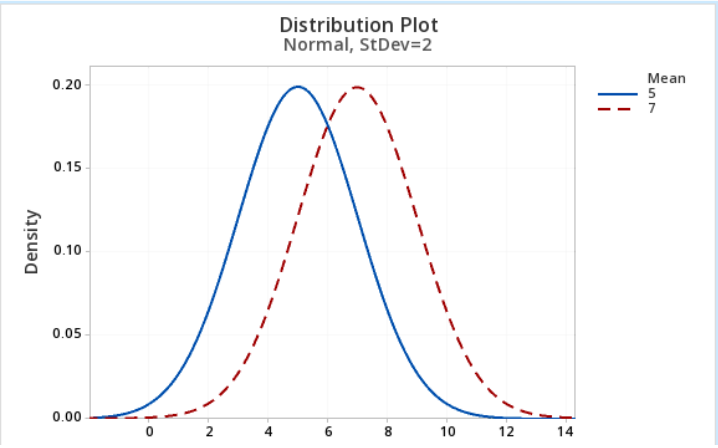
- The standard deviation σ controls the variability of a normal curve. When the standard deviation is larger, the area under the normal curve is less concentrated about the mean.

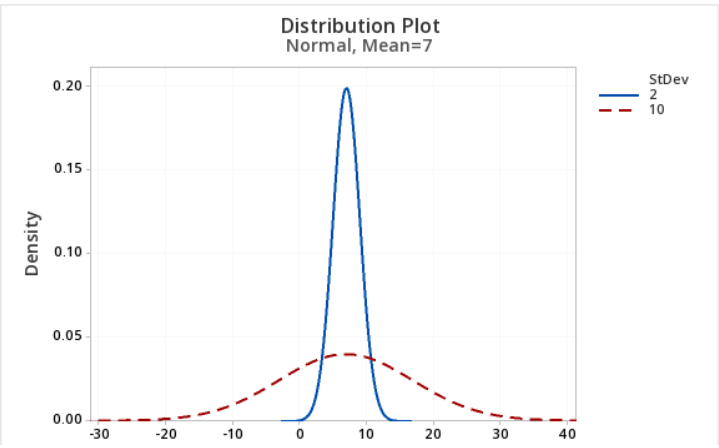
- The standard deviation is the distance from the center to the change-of-curvature points on either side.

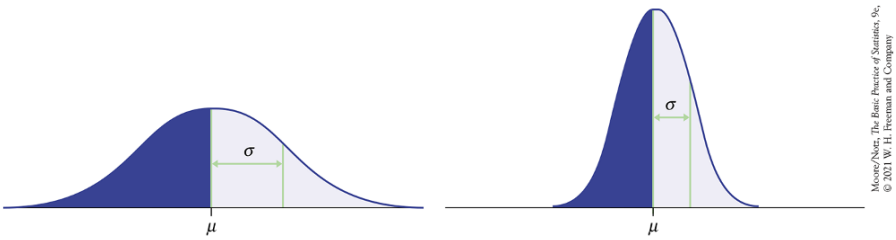
Probability density function: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\frac{1}{2}[\frac{(x-\mu)}{\sigma}]^2)}$

$\pi = 3.1416...$

$e = 2.71828...$







Standard Normal Distributions

All Normal distributions are the same if we measure in units of size σ from the mean μ as the center. Changing to these units is called standardizing.

Standardizing and $z-$scores

If x is an observation from a distribution with the mean μ and standard deviation σ, the standardized value of x called $z-$score is:

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

- The standard Normal distribution is the Normal distribution with mean 0 and standard deviation 1 ($N(0, 1)$).
- If a variable x has any Normal distribution with the mean μ and standard deviation σ, then the $z-$score has the standard Normal distribution.

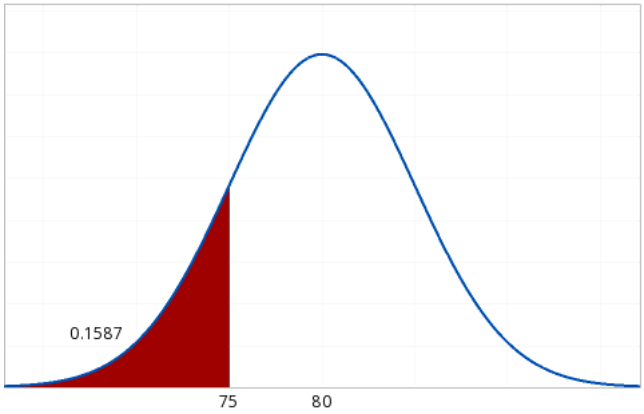
Finding Normal Proportion

In order to calculate the proportions of observations from a normal distribution that falls within a range, we calculate the area under the normal curve and above that range.

Cumulative Proportions

The cumulative proportion for a value x in a distribution is the proportion of observations in the distribution that are less than or equal to x.

Example 1.5.1 *The following shaded area shows the cumulative proportion of students who are scored equal to or less than 75% in a class assuming the scores are normally distributed with the mean of 80% and standard deviation of 5%.*



a) What proportion of students are scored greater than 75%?

```
> pnorm(75, 80, 5, lower.tail = FALSE)
[1] 0.8413447
```

b) What proportion of students are scored exactly 75%?

1.6 Sampling Distributions and the Central Limit Theorem

As we saide, we use sample statistics such as $\bar{x}$ to make inferences about population parameters such as μ. However, our inferences depend on the sample. The value of the sample statistic can change from one sample to another which is called sample variability.

But if we repeat our sampling over and over agian (large number of times) from the population, the sample variability follow a predictable pattern. Hence, we will be able to know how close to μ our sample mean, $\bar{x}$ is likely to fall.

Sampling Distribution

Since we take a random sample from a population, the values of a statistic vary randomly from sample to sample. In other words, statistics are random variables and have probability distributions.

The probability distribution of a statistic is called **sampling distribution** when all possible samples of the same sample size, n, are taken from the same population.

The Sampling Distribution of a Sample Mean

Theorem 1.1

If $x_1, x_2, x_3, ..., x_n$ are a random sample of n measurements from a large (or infinite) population with mean μ and and standard deviation σ, then regardless of the population distribution, the mean and standard deviation of the sampling distribution (standard error) of estimate are:

$$E(\bar{x}) = \mu_{\bar{x}} = \mu$$

$$\sqrt{Var(\bar{x})} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Theorem 1.2 Central Limit Theorem

If the sample size, n is large (≥ 30), the sample mean, $\bar{x}$, is approximately normally distributed regardless of the probability distribution of the sampled population.

$$\bar{x} \quad \text{Approx.} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

If the probability distribution of the sampled population is normal ($x \sim N(\mu, \sigma)$), the sample mean, $\bar{x}$, is normally distributed regardless of the sample size.

$$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

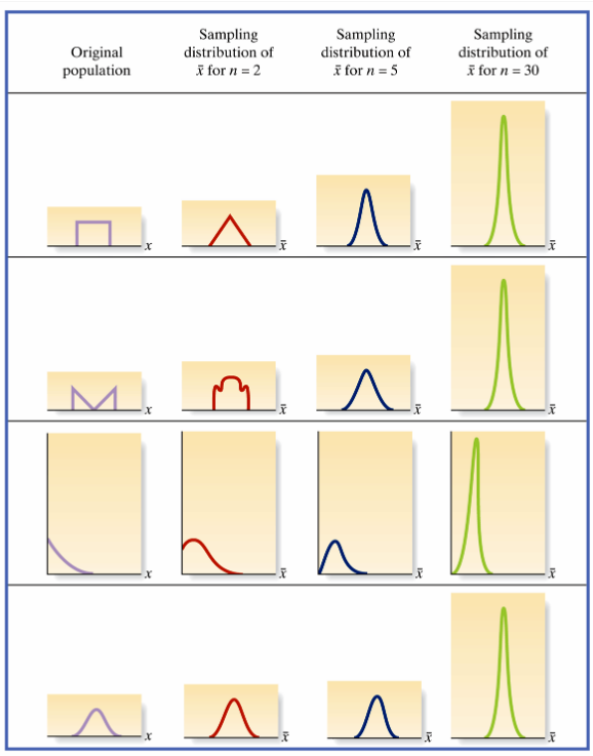


Figure 1.10: Sampling distributions of $\bar{x}$ for different populations and different sample sizes

1.7 Estimating a Population Mean

Making an inference about a population mean can be in two ways:

1. Estimate its value, $\hat{\mu}$
2. Make a decision about its value, Test of Hypothesis

Confidence Interval for a Population Mean: Normal (z) Statistic

Conditions:

- Random Sample Selected
- Large Sample Size
- Population Satndard Deviation, σ, is Known

Large-Sample $100(1-\alpha)\%$ **Confidence Interval for** μ **,Based on a Normal (z)Statistic**

- σ known:

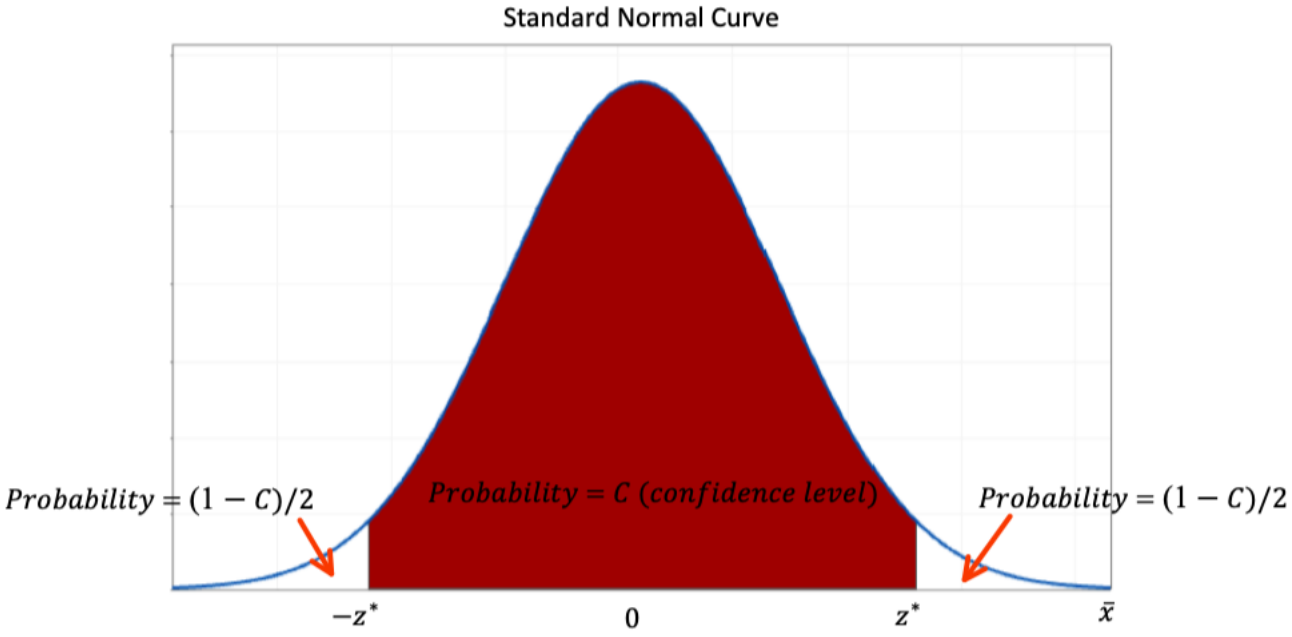
$$\bar{x} \pm \left(z_{\frac{\alpha}{2}}\right)\sigma_{\bar{x}} = \bar{x} \pm \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

- σ unknown:

$$\bar{x} \pm \left(z_{\frac{\alpha}{2}}\right)\sigma_{\bar{x}} \approx \bar{x} \pm \left(z_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$$

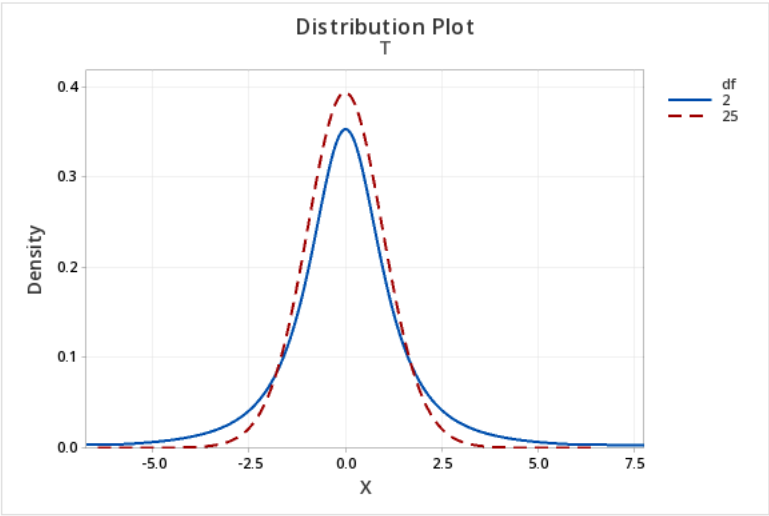
Where $z_{\frac{\alpha}{2}}$ is the z-value corresponding to an area $\frac{\alpha}{2}$ in the tail of a standard normal distribution, $\sigma_{\bar{x}}$ is the standard deviation of the sampling distribution of $\bar{x}, \sigma$ is the standard deviation of the population, and s is the standard deviation of the sample.

Where $z^* = z_{\frac{\alpha}{2}}$



Confidence Interval for a Population Mean: Student's t-Statistic

Since we rarely know a population satndard deviation and our sample size might not be large to apply CLT, we use the following statistic which is called $t-$statistic. The t-statistic has a sampling distribution very much like that of the z-statistic: mound shaped, symmetric, and with mean 0. It is however more variable than a z distribution. The variability depends on the number of degrees of freedom, df, which in turn depends on the number of measurements available for estimating σ^2.



Small-Sample $100(1-\alpha)\%$ Confidence Interval for μ , (t)-Statistic

σ unknown:

$$\bar{x} \pm \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$$

where $t_{\frac{\alpha}{2}}$ is the t-value corresponding to an area $\frac{\alpha}{2}$ in the upper tail of the Student's t-distribution based on $(n-1)$ degrees of freedom.

Conditions:

- Random Sample Selected
- Population is approximately Normally Distributed

Example 1.7.1 *Refer to Example 1.11 from the textbook. (Data set: ATTENTIMES)*

$n = 50 > 30, s = 13.41, C = .99,$ *We can use* $z_{\frac{\alpha}{2}}$.

```
> norm.interval = function(ï..AttentionTime, variance = var(ï..AttentionTime), conf.level = 0.99)
+   {z = qnorm((1 - conf.level)/2, lower.tail = FALSE)
+    xbar = mean(ï..AttentionTime)
+    sdx = sqrt(variance/length(ï..AttentionTime))
+    c(xbar - z * sdx, xbar + z * sdx)}
> norm.interval(ï..AttentionTime)
[1] 15.96165 25.73435
>
```

```
> sd(ï..Attention.Time)
[1] 13.41383
> z.test(ï..Attention.Time,stdev =13.41383, conf.level = 0.99 )

        One Sample z-test

data:  ï..Attention.Time
z = 10.99, n = 50.000, Std. Dev. = 13.414, Std. Dev. of the sample mean = 1.897, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 15.96165 25.73435
sample estimates:
mean of ï..Attention.Time
                   20.848
```

Example 1.7.2 *Refer to Example 1.12 from the textbook. (Data set: CILICA)*

$n = 5, s = 29.29505, C = .95$

```
> t.test(SiliconDioxide, conf.level = .95)

        One Sample t-test

data:  SiliconDioxide
t = 18.258, df = 4, p-value = 5.293e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 202.8254 275.5746
sample estimates:
mean of x
    239.2
```

1.8 Testing a Hypothesis About a Population Mean

Elements of a Statistical Test of Hypothesis

1. **Null Hypothesis, H_0:** The hypothesis that will be assumed to be true unless the data provide convincing evidence that it is false.

2. **Alternative Hypothesis, H_a:** The hypothesis that will be supported only if the data provide convincing evidence of its truth.

3. **Test Statistic**: A sample statistic, calculated from information provided in the sample, that the researcher uses to decide between the null and alternative hypotheses.

4. **Level of significance, α:** The probability of committing a Type I error is denoted by α. (Type I error: Rejecting H_0 given that H_0 is true.)

5. **Rejection Region:** The numerical values of the test statistic for which the null hypothesis will be rejected.

6. **p-Value:** The observed significance level, or p-value for a specific statistical test is the probability (assuming H_0 is true) of observing a value of the test statistic that is at least as contradictory to the null hypothesis, and supportive of the alternative hypothesis, as the actual one computed from the sample data.

7. **Decision:** Reject or Fail to Reject H_0.

8. **Conclusion:** Support or cannot support H_a.

Hypothesis Testing Steps

- State H_0 and H_a
- Compute the Test Statistics
- Compare the p-value with significancelevel, reject or do not reject H_0.
- Make Decision.
- Make Conclusion.

Large-Sample Test of Hypothesis about μ Based on a Normal (z) Statistic

Test statistic: σ known

$$z_c = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Test statistic: σ unknown

$$z_c = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Lower-Tailed Tests

$H_0 : \mu = \mu_0$

$H_a : \mu < \mu_0$

Rejection region: $z < -z_\alpha$

p-value: $P(z < z_c)$

Upper-Tailed Tests

$H_0 : \mu = \mu_0$

$H_a : \mu > \mu_0$

Rejection region: $z > z_\alpha$

p-value: $P(z > z_c)$

Two-Tailed Tests

$H_0 : \mu = \mu_0$

$H_a : \mu \neq \mu_0$

Rejection region: $|z| > z_{\frac{\alpha}{2}}$

p-value: if z_c is positive $2P(z > z_c)$ if z_c is negative $2P(z < z_c)$

Decision

Reject H_0 if $p - value < \alpha$ or if test statistic (z_c) falls in rejection region (the result is "statistically significant") where:

$P(z > z_\alpha) = \alpha, P(z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$

Conditions Required for a Valid Large-Sample Hypothesis Test for μ

1. A random sample is selected from the target population.

2. The sample size n is large (e.g.,$n \geq 30$).

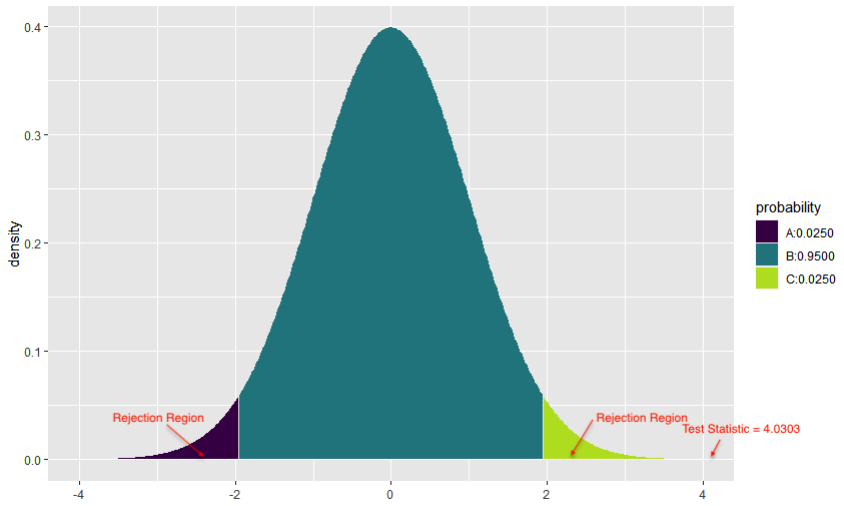
Example 1.8.1 *Refer to Example 1.14 from the textbook. (Data set: BONES)*

$H_0 : \mu = 8.5$

$H_a : \mu \neq 8.5$

```
> t.test(BONES,mu=8.5, alternative='two.sided')

        One Sample t-test

data:  BONES
t = 4.0303, df = 40, p-value = 0.0002427
alternative hypothesis: true mean is not equal to 8.5
95 percent confidence interval:
 8.877669 9.637453
sample estimates:
mean of x
 9.257561

> p_value=2*pnorm(4.0303, lower.tail = FALSE)
> p_value
[1] 5.570571e-05
> z = qnorm((0.05)/2, lower.tail = FALSE)
> z
[1] 1.959964
>
```

```
> z.test( ï..Ratio,stdev=1.203565,conf.level=0.95, mu=8.5, alternative=('two.sided'))

        One Sample z-test

data:  ï..Ratio
z = 4.0303, n = 41.00000, Std. Dev. = 1.20356, Std. Dev. of the sample mean =
0.18797, p-value = 5.57e-05
alternative hypothesis: true mean is not equal to 8.5
95 percent confidence interval:
 8.889156 9.625966
sample estimates:
mean of ï..Ratio
        9.257561
```



Decision: Since p-value$\approx 0 < \alpha = 0.05$, or test statistics falls in the rejection region, we reject H_0.

Conclusion: At 5% level of significance, we have sufficient evidence to conclude that the true mean length-to-width ratio of all humerus bones of this species differs from 8.5.

Small-Sample Test of Hypothesis about μ Based on a Student's t-Statistic

Test statistic:

$$t_c = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Lower-Tailed Tests

$H_0 : \mu = \mu_0$

$H_a : \mu < \mu_0$

Rejection region: $t < -t_\alpha$

p-value: $P(t < t_c)$

Upper-Tailed Tests

$H_0 : \mu = \mu_0$

$H_a : \mu > \mu_0$

Rejection region: $t > t_\alpha$

p-value: $P(t > t_c)$

Two-Tailed Tests

$H_0 : \mu = \mu_0$

$H_a : \mu \neq \mu_0$

Rejection region: $|t| > t_{\frac{\alpha}{2}}$

p-value: if t_c is positive $2P(t > t_c)$ if t_c is negative $2P(t < t_c)$

Decision

Reject H_0 if $p - value < \alpha$ or if test statistic (t_c) falls in rejection region (the result is "statistically significant") where:

$P(t > t_\alpha) = \alpha, P(t > t_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$

Conditions Required for a Valid Small-Sample Hypothesis Test for μ

1. A random sample is selected from the target population.

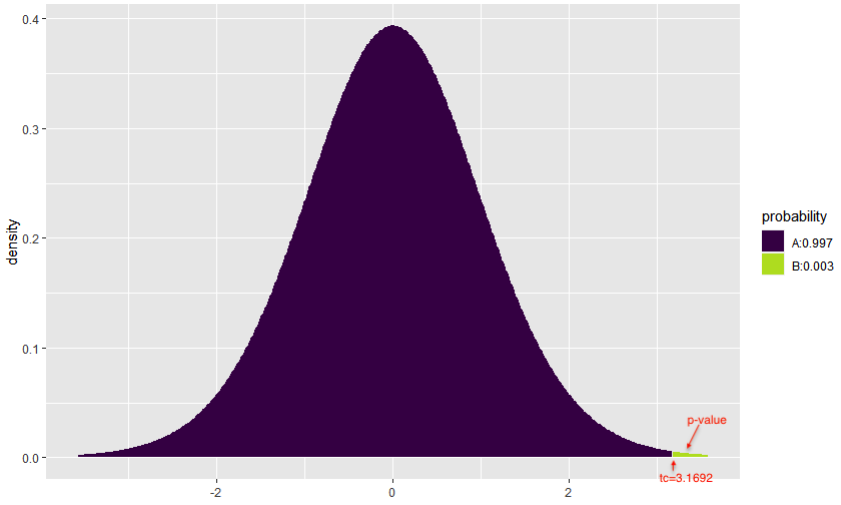
2. The population from which the sample is selected has a distribution that is approximately normal.

Example 1.8.2 *Refer to Example 1.15 from the textbook. (Data set: BENZENE)*

$H_0 : \mu = 1$

$H_a : \mu > 1$

```
> t.test(ï..Benzene,mu=1, alternative='greater')

        One Sample t-test

data:  ï..Benzene
t = 3.1692, df = 19, p-value = 0.002526
alternative hypothesis: true mean is greater than 1
95 percent confidence interval:
 1.49983     Inf
sample estimates:
mean of x
      2.1

> qt(0.05,19,lower.tail = FALSE)
[1] 1.729133
```



Decision: Since p-value$= 0.0025 < \alpha = 0.05$, or test statistics falls in the rejection region, we reject H_0.

Conclusion: At 5% level of significance, OSHA concludes with sufficient evidence that the plant is in violation of the revised government standards.

Note:

For a small sample, we need to have a population that is normally distributed which is not always the case. However, the good news is t procedure is **robust**, that yields valid results even when the data are nonnormal, as long as the population is not highly skewed.

1.9 Inferences About the Difference Between Two Population Means

The objective is to make an inference about the difference $(\mu_1 - \mu_2)$ between the two population means since we are interested in comparing two populations. The parameter of interest is mean difference; difference in averages, $(\mu_1 - \mu_2)$.

Large-Sample Confidence Interval for $(\mu_1 - \mu_2)$: Independent Samples

1. ${\sigma_1}^2$ and ${\sigma_2}^2$ known:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}}\sigma_{(\bar{x}_1 - \bar{x}_2)} = (\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

2. ${\sigma_1}^2$ and ${\sigma_2}^2$ unknown:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}}\sigma_{(\bar{x}_1 - \bar{x}_2)} \approx (\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

Assumptions: The two samples are **randomly** and **independently** selected from the two populations. The sample sizes, n_1 and n_2, are large enough so that $\bar{x}_1$ and $\bar{x}_2$ each have approximately normal sampling distributions and so that ${s_1}^2$ and ${s_2}^2$ provide good approximations to ${\sigma_1}^2$ and ${\sigma_2}^2$ This will be true if $n_1 \geq 30$ and $n_2 \geq 30$.

Large, Independent Samples Test of Hypothesis for $\mu_1 - \mu_2$ Normal (z) Statistic

Test statistic: ${\sigma_1}^2$ and ${\sigma_2}^2$ known

$$z_c = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

Test statistic: ${\sigma_1}^2$ and ${\sigma_2}^2$ unknown

$$z_c \approx \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

Lower-Tailed Tests

$H_0 : \mu_1 - \mu_2 = D_0$

$H_a : \mu_1 - \mu_2 < D_0$

Rejection region: $z < -z_\alpha$

p-value: $P(z < z_c)$

Upper-Tailed Tests

$H_0 : \mu_1 - \mu_2 = D_0$

$H_a : \mu_1 - \mu_2 > D_0$

Rejection region: $z > z_\alpha$

p-value: $P(z > z_c)$

Two-Tailed Tests

$H_0 : \mu_1 - \mu_2 = D_0$

$H_a : \mu_1 - \mu_2 \neq D_0$

Rejection region: $|z| > z_{\frac{\alpha}{2}}$

p-value: if z_c is positive $2P(z > z_c)$ if z_c is negative $2P(z < z_c)$

Decision

Reject H_0 if $p-value < \alpha$ or if test statistic (z_c) falls in rejection region (the result is "statistically significant") where:

$P(z > z_\alpha) = \alpha, P(z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$

Note

The symbol for the numerical value assigned to the difference $\mu_1 - \mu_2$ under the null hypothesis is D_0. For testing equal population means, $D_0 = 0$.

Assumptions: Same as for the previous large-sample confidence interval.

Example 1.9.1 *Refer to Example 1.16 from the textbook. (Data set: DIETS). Develope*

a) Form a 95% confidence interval for the difference between the population mean weight losses for the two diets. Interpret the result.

b) Determine if the low-fat diet is more effective than the regular diet.

a) 95% CI:

```
> t.test(WTLOSS~DIET,conf.level=0.95)

        Welch Two Sample t-test

data:  WTLOSS by DIET
t = 3.0954, df = 193.94, p-value = 0.002256
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6930167 3.1269833
sample estimates:
 mean in group LOWFAT mean in group REGULAR
                9.31                 7.40

> |
```

We are we are 95% confident that the mean weight loss for the low-fat diet is between .69 and 3.13 pounds more than the mean weight loss for the other diet.

b) Test of Hypothesis

$H_0 : \mu_1 - \mu_2 = 0$ or $\mu_1 = \mu_2$

$H_a : \mu_1 - \mu_2 > 0$ or $\mu_1 > \mu_2$

Decision: Since p-value$= 0.001 < \alpha = 0.05$, or test statistics falls in the rejection region, we reject H_0.

```
> t.test(WTLOSS~DIET,conf.level=0.95,alternative = c("greater"))

        Welch Two Sample t-test

data:  WTLOSS by DIET
t = 3.0954, df = 193.94, p-value = 0.001128
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.8901773       Inf
sample estimates:
 mean in group LOWFAT mean in group REGULAR
                9.31                 7.40
```

Conclusion: At 5% level of significance, the dietitian has sufficient evidence to conclude that the mean weight loss for the low-fat diet, μ_1, will exceed the mean weight loss for the regular diet, μ_2,.

Approximate Small-Sample Procedures when $\sigma_1^2 \neq \sigma_2^2$

Confidence interval:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}} \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

Test statistic for $H_0 : \mu_1 - \mu_2 = 0$

$$t_c = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

where t is based on degrees of freedom equal to

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

Note: The value of ν will generally not be an integer. Round ν down to the nearest integer to use the t-table.

Conditions Required for Valid Small-Sample Inferences about $\mu_1 - \mu_2$

1. The two samples are randomly selected in an independent manner from the two target populations.

2. Both sampled populations have distributions that are approximately normal.

Small-Sample Confidence Interval for $(\mu_1 - \mu_2)$: Independent Samples

Pooled t-test

If $\sigma_1 = \sigma_2 = \sigma$, the two sample standard deviations are combined to construct the following **pooled sample estimator** σ^2.

Pooled Estimator of σ^2, s_p^2

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

Interval Estimate

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}}\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}$$

$t_{\frac{\alpha}{2}}$ has a $df = n_1 + n_2 - 2$

Test Statistic t

$$t_c = \frac{(\bar{x}_1-\bar{x}_2)-D_0}{s_p\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}}$$

With the $df = n_1 + n_2 - 2$

Conditions Required for Valid Inferences about $\mu_1 - \mu_2$ using t distribution

- The samples are randomly and independently selected from the populations.
- Both populations have a **Normal** distribution or $n_1 \geq 30$ and $n_2 \geq 30$
- σ_1 and σ_2 Unknown
- $\sigma_1 = \sigma_2$

Example 1.9.2 *Refer to Example 1.17 from the textbook. (Data set: READING)*

a) Use the data in the table to test whether the true mean test scores differ for the new method and the standard method. Use $\alpha = .05$.

b) What assumptions must be made in order that the estimate be valid?

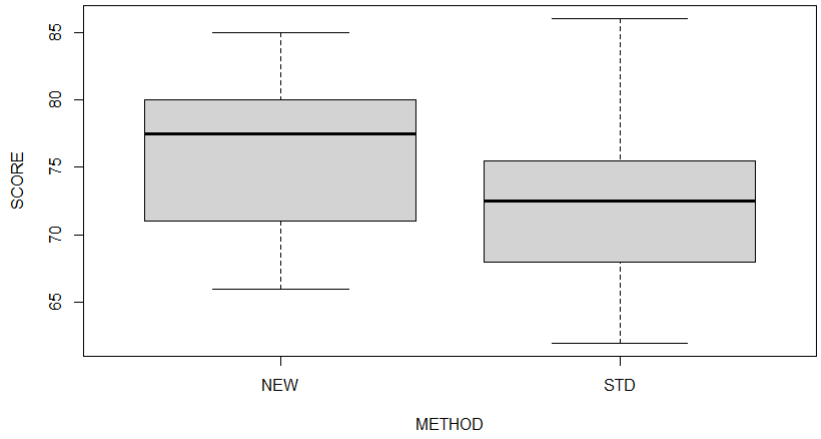
μ_1: Mean reading test scores of slow learners taught with the new method

μ_2: Mean reading test scores of slow learners taught with the standard method

$H_0 : \mu_1 - \mu_2 = 0$ or $\mu_1 = \mu_2$

$H_a : \mu_1 - \mu_2 \neq 0$ or $\mu_1 \neq \mu_2$

```
> t.test(SCORE~METHOD, conf.level=0.95, alternative = c("two.sided"), var.equal=TRUE)

        Two Sample t-test

data:  SCORE by METHOD
t = 1.5519, df = 20, p-value = 0.1364
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.399371  9.532705
sample estimates:
mean in group NEW mean in group STD
         76.40000          72.33333
```



Decision: Since p-value$= 0.1364 > \alpha = 0.05$, or test statistics does not fall in the rejection region, we fail to reject H_0.

Conclusion: At 5% level of significance, there is insufficient evidence of a difference between the true mean test scores for the two reading methods.

Note: The two-sample t procedure is more robust against nonnormal data than the one-sample method. And, when the sample sizes are equal, the assumption of equal population variances can be relaxed. That is, when $n_1 = n_2$, ${\sigma_1}^2$ and ${\sigma_2}^2$ can be quite different and the test statistic will still have (approximately) a Student's t distribution.

Paired Difference Experiment

An experiment, in which observations are paired (ie.dependent) and the differences are analyzed, is called a paired difference

experiment. In many cases, a paired difference experiment can provide more information about the difference between population means than an independent samples experiment can. The idea is to compare population means by comparing the differences between pairs of experimental units (objects, people, etc.) that were similar prior to the experiment.

Paired Difference Confidence Interval for $\mu_d = \mu_1 - \mu_2$

Large Sample, Normal (z) Statistic

$\bar{x}_d \pm z_{\frac{\alpha}{2}} \frac{\sigma_d}{\sqrt{n_d}} \approx \bar{x}_d \pm z_{\frac{\alpha}{2}} \frac{s_d}{\sqrt{n_d}}$

Small Sample, Student's t-Statistic

$\bar{x}_d \pm t_{\frac{\alpha}{2}} \frac{s_d}{\sqrt{n_d}}$

where $t_{\frac{\alpha}{2}}$ is based on $(n_d - 1)$ degrees of freedom

Paired Difference Test of Hypothesis for $\mu_d = \mu_1 - \mu_2$

Large Sample, Normal (z) Test Statistic:

$$z_c = \frac{\bar{x}_d - D_0}{\frac{\sigma_d}{\sqrt{n_d}}} \approx \frac{\bar{x}_d - D_0}{\frac{s_d}{\sqrt{n_d}}}$$

Lower-Tailed Tests

$H_0 : \mu_d = D_0$

$H_a : \mu_d < D0$

Rejection region: $z < -z_\alpha$

p-value: $P(z < z_c)$

Upper-Tailed Tests

$H_0 : \mu_d = D_0$

$H_a : \mu_d > D0$

Rejection region: $z > z_\alpha$

p-value: $P(z > z_c)$

Two-Tailed Tests

$H_0 : \mu_d = D_0$

$H_a : \mu_d \neq D0$

Rejection region: $|z| > z_{\frac{\alpha}{2}}$

p-value: if z_c is positive $2P(z > z_c)$ if z_c is negative $2P(z < z_c)$

Small Sample, Student's t-Test Statistic:

$$t_c = \frac{\bar{x}_d - D_0}{\frac{s_d}{\sqrt{n_d}}}$$

Lower-Tailed Tests

$H_0 : \mu_d = D_0$

$H_a : \mu_d < D0$

Rejection region: $t < -t_\alpha$

p-value: $P(t < t_c)$

Upper-Tailed Tests

$H_0 : \mu_d = D_0$

$H_a : \mu_d > D0$

Rejection region: $t > t_\alpha$

p-value: $P(t > t_c)$

Two-Tailed Tests

$H_0 : \mu_d = D_0$

$H_a : \mu_d \neq D0$

Rejection region: $|t| > t_{\frac{\alpha}{2}}$

p-value: if t_c is positive $2P(t > t_c)$ if t_c is negative $2P(t < t_c)$

Decision

Reject H_0 if $p - value < \alpha$ or if test statistic falls in rejection region (the result is "statistically significant")

Conditions Required for Valid Large-Sample Inferences about μ_d

1. A random sample of differences is selected from the target population of differences.

2. The sample size n_d is large (i.e., $n_d \geq 30$).

Conditions Required for Valid Small-Sample Inferences about μ_d

1. A random sample of differences is selected from the target population of differences.

2. The population of differences has a distribution that is approximately normal.

Example 1.9.3 *Data set: PAIRED shows the data on reading test scores for eight pairs of slow learners with similar reading IQs of which one member is randomly assigned to the standard teaching method while the other is assigned to the new method. Do the data support the hypothesis that the population mean reading test score for slow learners taught by the new method is greater than the mean reading test score for those taught by the standard method?*

μ_1: Mean reading test scores of slow learners taught with the new method

μ_2: Mean reading test scores of slow learners taught with the standard method

$\mu_d = \mu_1 - \mu_2$

$H_0 : \mu_d = 0$

$H_a : \mu_d > 0$

Decision: Since p-value$= 0 < \alpha = 0.05$, or test statistics falls in the rejection region, we fail to reject H_0.

```
> t.test(NEW, STANDARD, conf.level=0.95, alternative = c("greater"), paired=TRUE)

        Paired t-test

data:  NEW and STANDARD
t = 7.3438, df = 7, p-value = 7.838e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 3.246316      Inf
sample estimates:
mean of the differences
                 4.375
```

Conclusion: At 5% level of significance, there is sufficient evidence to conclude that population mean test score for slow learners taught by the new method exceeds the population mean score for those taught by the standard method.

1.10 Comparing Two Population Variances

We use data collected from two independent random samples, one from population 1 and another from population 2. The two sample variances will be the basis for making inferences about the two population variances.

We also need to conduct a test, F-test, to determine whether the two populations have unequal standard deviations before applying a t-test.

Hypothesis Tests about σ_1^2 and σ_2^2

Test Statistic (F Distribution):

$$F_c = \frac{s_1^2}{s_2^2}$$

Where: $df = (n_1 - 1, n_2 - 1)$ and

where s_1^2 is the larger of the two sample variances, $n_1 - 1$ is the degrees of freedom for the numerator, and $n_2 - 1$ is the degrees of freedom for the denominator.

Upper-Tailed Tests

$H_0 : \sigma_1^2 \leq \sigma_2^2$

$H_a : \sigma_1^2 > \sigma_2^2$

Rejection region: $F \geq F_\alpha$

p-value: $P(F > F_c)$

Two-Tailed Tests

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_a : \sigma_1^2 \neq \sigma_2^2$

Rejection region: $F \geq F_{\frac{\alpha}{2}}$

p-value: $2 * P(F > F_c)$

For each type of test,

- σ_1^2 and σ_2^2 are the variances of populations 1 and 2 respectively.
- The test statistic is $F_c = \frac{s_1^2}{s_2^2}$ where s_1^2 is larger sample variance.
- The value of F_α is based on an F distribution with $n_1 \breve{}1$ degrees of freedom for the numerator, and $n_2 \breve{}1$ degrees of freedom for the denominator.

F Distribution

Whenever independent simple random samples of sizes n_1 and n_2 are selected from two Normal populations with equal variances, the sampling distribution of $\frac{s_1^2}{s_2^2}$ is and F distribution with $n_1 - 1$ degrees of freedom for the numerator, and $n_2 \breve{}1$ is the degrees of freedom for the denominator.

The F distribution is right-skewed with positive values.

The shape of the F distribution depends on its numerator and denominator degrees of freedom.

Conditions Required for Valid Inferences about σ_1^2 and σ_2^2

- Samples are selected separately, independently, and randomly
- Both populations have a **Normal** distribution.

Example 1.10.1 *Refer to Example 1.18 from the textbook. (Data set: READING)*

The use of the t statistic was based on the assumption that the population variances of the test scores were equal for the two methods. Conduct a test of hypothesis to check this assumption at $\alpha = .10$.

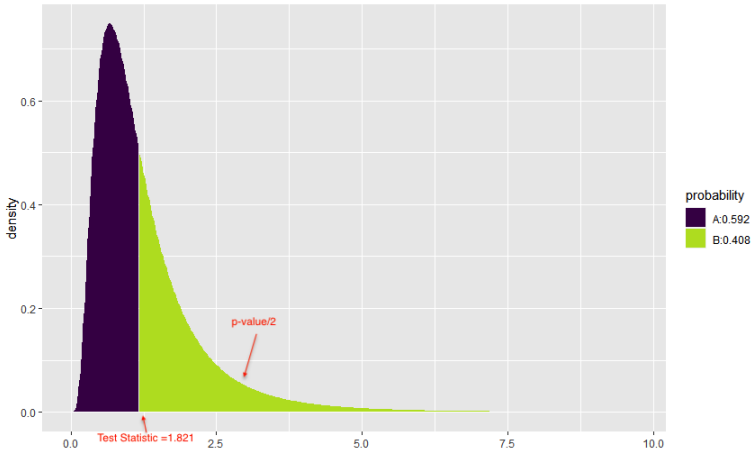
$H_0 : \sigma_1^2 = \sigma_2^2$

$H_a : \sigma_1^2 \neq \sigma_2^2$

```
> var.test(SCORE[METHOD=="STD"],SCORE[METHOD=="NEW"], alternative = "two.sided")

        F test to compare two variances

data:  SCORE[METHOD == "STD"] and SCORE[METHOD == "NEW"]
F = 1.1821, num df = 11, denom df = 9, p-value = 0.8148
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3021557 4.2410955
sample estimates:
ratio of variances
          1.182056
```

Decision: Since p-value$= 0.8148 > \alpha = 0.1$, or test statistics does not fall in the rejection region, we fail to reject H_0.

Conclusion: At 10% level of significance, there is insufficient evidence to conclude that the population variances of the reading test scores are not equal.



Acknowledgement

The core content of the slides are from the textbook of this course;

A Second Course in Statistics: Regression Analysis (7th Edition)

by

Mendenhall, William and Sincich, Terry; Pearson Education.