BUSAD40 - Busniess Statistics

Lecture Note 3

Norman Lo

Fall 2020

> In relation to any experiment we may speak of this hypothesis as the "null hypothesis," and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.
>
> The Design of Experiments, Edinburgh: Oliver and Boyd, 1935, p.18

Introduction to Statistical Inference

In this section, we are shifting our focus on statisical inference. One of the popular methods in statistical inference is **hypothesis testing**. Genearlly speaking, hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true.

Statistical Hypothesis:

The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.

There are two types of statistical hypotheses.

Null hypothesis: The null hypothesis, denoted by H_o, is usually the hypothesis that sample observations result purely from chance.

Alternative hypothesis: The alternative hypothesis, denoted by H_o or H_a, is the hypothesis that sample observations are influenced by some non-random cause.

For example, suppose we wanted to determine whether a coin was fair and balanced. A null hypothesis might be that half the flips would result in Heads and half, in Tails. The alternative hypothesis might be that the number of Heads and Tails would be very different. Symbolically, these hypotheses would be expressed as

$$H_o : P = 0.5$$
$$H_a : P \neq 0.5$$

Suppose we flipped the coin 50 times, resulting in 40 Heads and 10 Tails. Given this result, we would be inclined to reject the null hypothesis. We would conclude, based on the evidence, that the coin was probably not fair and balanced.

Hypothesis Tests:

Statisticians follow a formal process to determine whether to reject a null hypothesis, based on sample data. This process, called hypothesis testing, consists of four steps.

1. State the hypotheses. This involves stating the null and alternative hypotheses. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false.
2. Formulate an analysis plan. The analysis plan describes how to use sample data to evaluate the null hypothesis. The evaluation often focuses around a single test statistic.
3. Analyze sample data. Find the value of the test statistic (mean score, proportion, t statistic, z-score, etc.) described in the analysis plan.
4. Interpret results. Apply the decision rule described in the analysis plan. If the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis.

Decision Errors:

Two types of errors can result from a hypothesis test.

Type I error: A Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the significance level. This probability is also called alpha, and is often denoted by α.

Type II error: A Type II error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called Beta, and is often denoted by β. The probability of not committing a Type II error is called the Power of the test.

Sampling Distribution

Sampling theory plays a huge role in specifying the assumptions upon which your statistical inferences rely. And in order to talk about "making inferences" the way statisticians think about it, we need to be a bit more explicit about what it is that we're drawing inferences from (the sample) and what it is that we're drawing inferences about (the population).

In almost every situation of interest, what we have available to us as researchers is a sample of data. We might have run experiment with some number of participants; a polling company might have phoned some number of people to ask questions about voting intentions; etc. Regardless: the data set available to us is finite, and incomplete. We can't possibly get every person in the world to do our experiment; a polling company doesn't have the time or the money to ring up every voter in the country etc. Our only goal was to find ways of describing, summarising and graphing that sample.

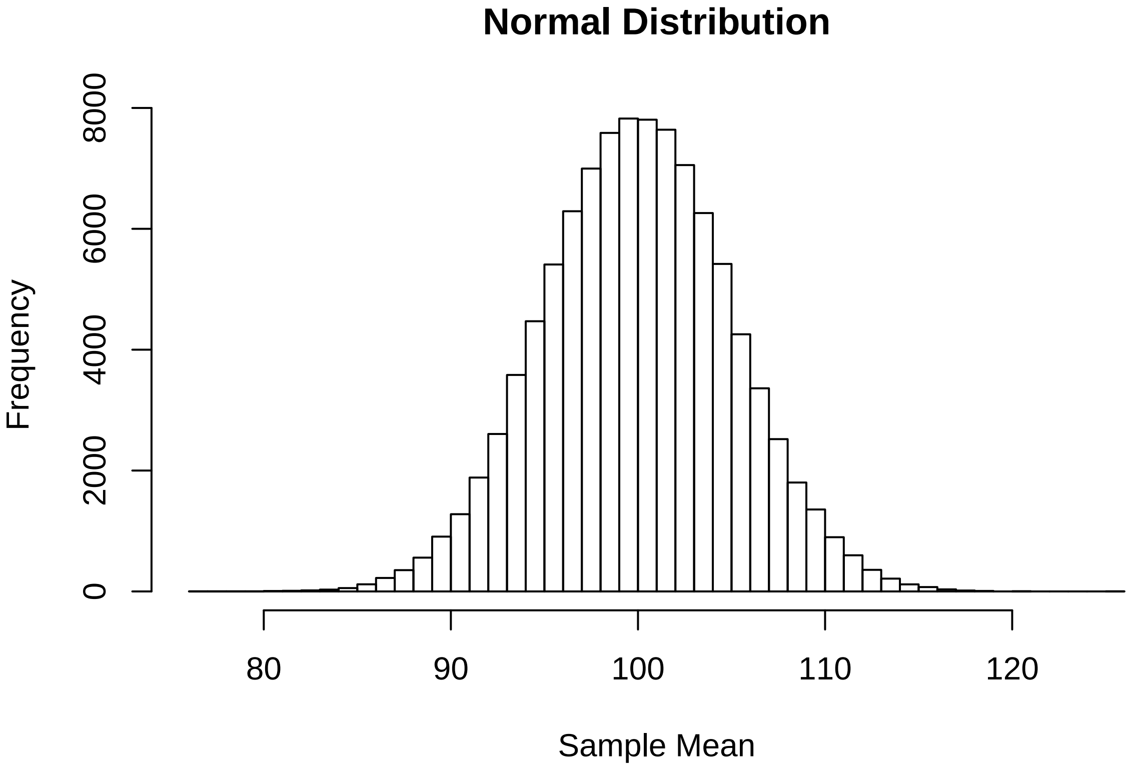
Unbias Estimation is a form of statistical inference, which we use the data from the sample to compute a value of a sample statistics that serves as an estimate of a population parameter. We refer $\bar{x}$ as the unbias estimator of the population mean μ. s is the unbias estimator of the population standard deviation σ. $\bar{p}$ is the unbias estimator of the population proportion p.

Central Limite Theorem:

When the population from which we are selecting a random sample does not have a normal distribution, the central limit theorem is helpful in identifying the shape of the sampling distribution of $\bar{x}$. In selecting random samples of size n from a population, the sampling distribution of the sample mean $\bar{x}$ can be approximated by a normal distribution as the sample size becomes large. The standard deviation of the sampling distribution is referred to the **standard error**. Here are the statistics for sampling distribution:

Finite Population	Infinite Population
$E(\bar{x}) = \mu$	$E(\bar{x}) = \mu$
$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}(\frac{\sigma}{\sqrt{n}})$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
$E(\bar{p}) = p$	$E(\bar{p}) = p$
$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{P(1-P)}{n}}$	$\sigma_{\bar{p}} = \sqrt{\frac{P(1-P)}{n}}$

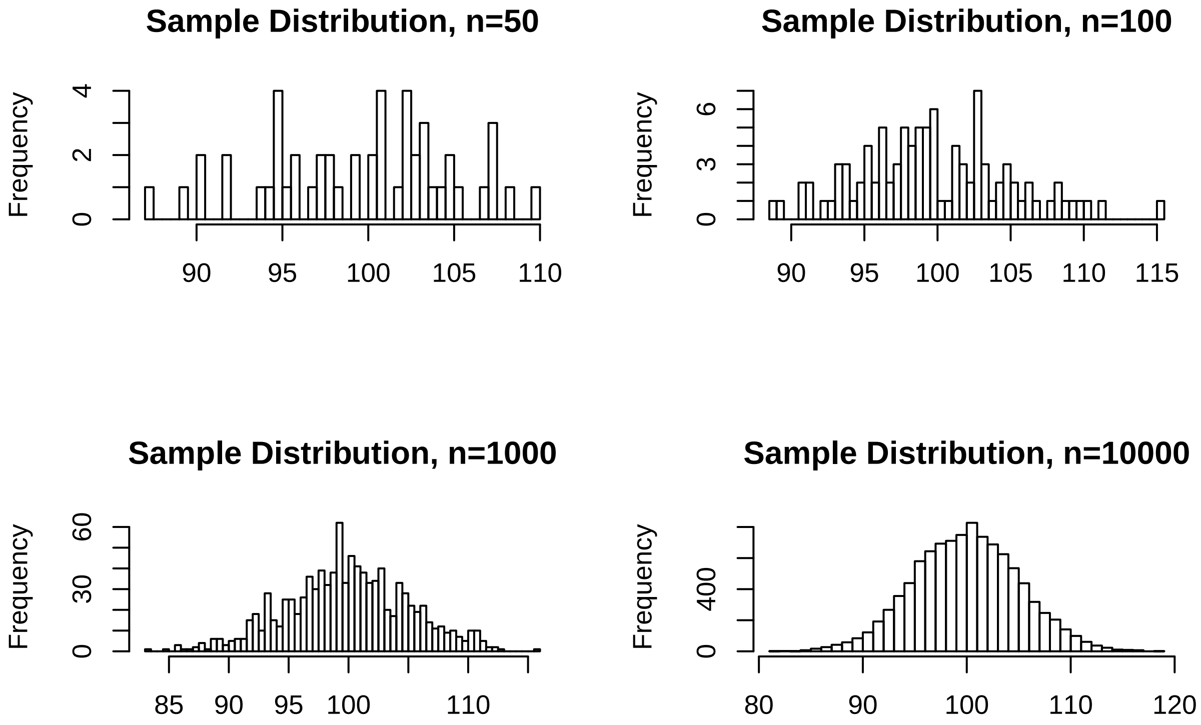
To demonstrates the **Central Limit Theorem**, we can simulate a normally distributed data with mean 100 and sd 5 as the population data for our study. Then, draw different size of sample from the population and see how the distribution shape getting closer to closer to normal distribution.



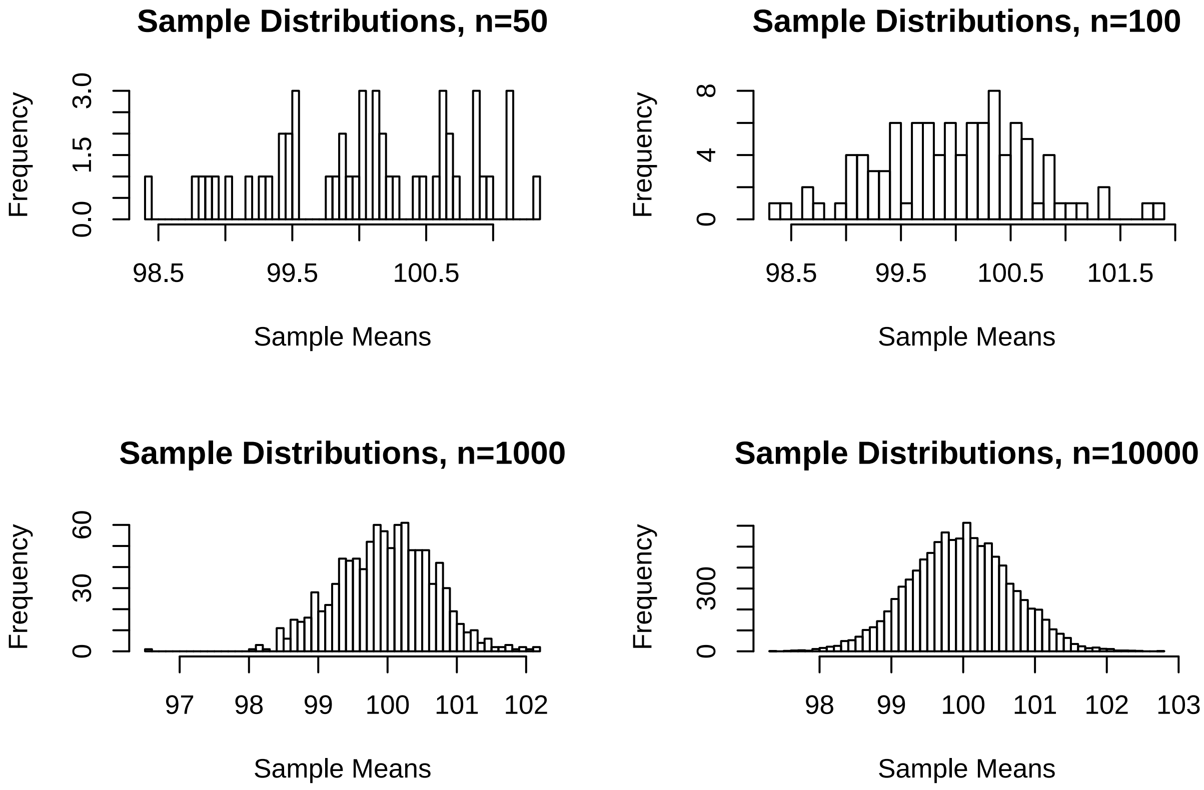
If we select 50 random sample from the population data, we can can calculate the mean of the samples.

```
## [1] "The sample mean from 50 randomly selected samples is  101.526"
```

If we increase the size of sample, the distributions of the sample data are going to get closer and closer to normal.



To be more accurate, we should randomly select sample sets with size of 50 observations multiple times from the population to check the distribution of the mean of the sample means to confirm the central limit theorem.



Overall, the mean of our samples, if we take enough, will equal the mean of the population. As we keep taking samples from the population data, the distribution of the means of those samples will stack up close to 100, forming a normal distribution.

Why do we want to know the sample distribution approximates the normal distribution?

One of the key objective for learning the central limit theorem is to know that the sampling distribution is normally distribute with mean $\bar{x}$ and standard deviation s, so that we can apply the probabiity distribution properties that we learned from the previous section with the replacement of σ by $\sigma_{\bar{x}}$.

Sample Means Normal Distribution Statistics:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

Example:

A production report indicates that the production lead time of the company is normally distributed with population mean $\mu = 13$minutes and standard deviation $\sigma = 4.6$minutes. What is the probability to randomly select 10 samples from the production that has the average production time greater than 15 minutes $\bar{x} > 15$?

Solution:

$$P(\bar{x} > 15) = ?$$

Step 1: Convert $\bar{x}$ to the standard normal distribution

$$z = \frac{15 - 13}{\frac{4.6}{\sqrt{10}}} = \frac{15 - 13}{1.455} = 1.375$$

Step 2: Find the area under the standard normal curve to the right of z = 2.433 (using the pnorm() function in r)

```
# Calculate the probability with the z score
round(1 - pnorm(1.375), 3)
```

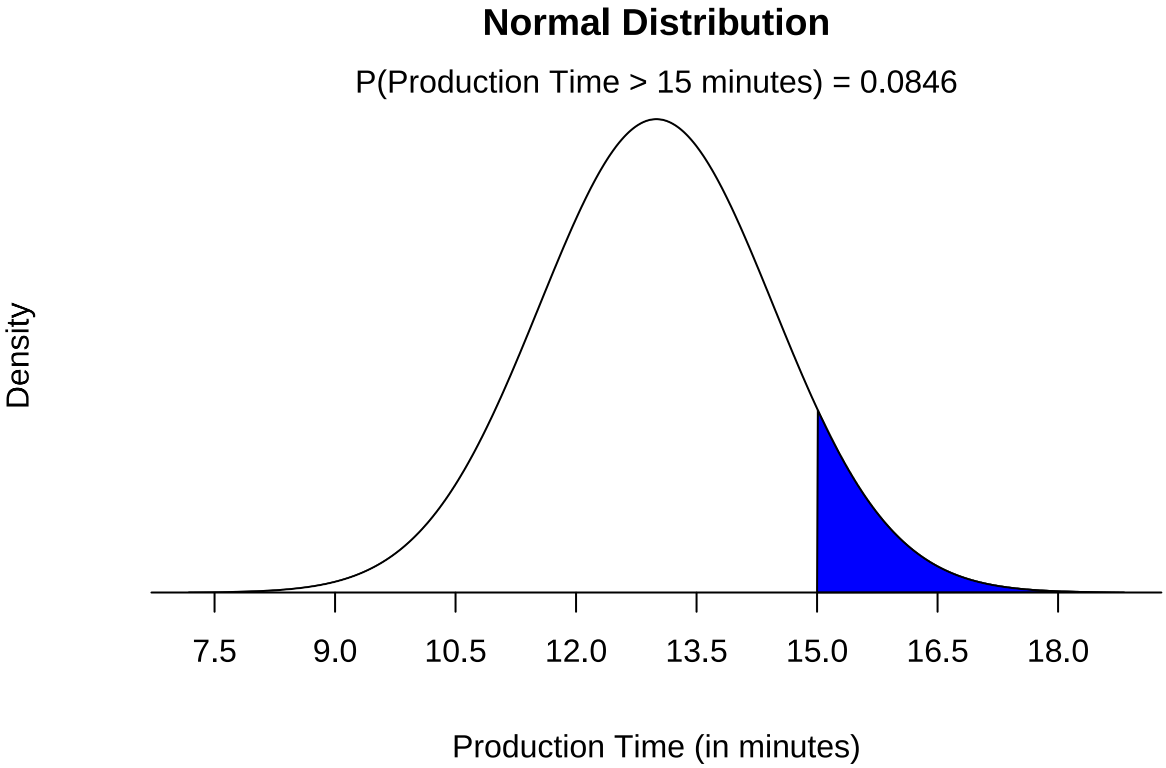
```
## [1] 0.085
```

```
# Alternative, calculate the probability with the average time
round(1 - pnorm(15, mean=13, sd=1.455), 3)
```

```
## [1] 0.085
```

$$P(z > 1.375) = 0.0846 \; or \; 8.46\%$$

Graphical Solution:



Sample Distribution without a Known Population Standard Deviation

In the last example, we assume that the population standard deviation σ is given. However, it is difficult to imagine we always know about the population parameters, especially σ. So, the problem is what if we only know the population mean μ, but not the standard deviation σ, can we still construct a distribution similar to the normal distribution for the probability problem?

Fortunately, the answer is "YES". A particular distribution we are going to apply is called the "**t-distribution**" or sometime referred to the "**Student's t-distribution**". The t-distribution is designed to approximate the normal distribution by estimating the unknown population standard deviation σ. The only tiny tweak to tranform the unbaised estimator is divide the sum of squares $\sum(x_i - \bar{x})^2$ by n minus 1, $(n-1)$, degree of freedom, to estimate the popluation variance.

Estimated Population Standard Deviation:

$$\hat{\sigma} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Estimated Standard Error:

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}$$

Let's use an example to demonstrate the different between the two situations.

Case 1: σ is known

Given the population data with $\mu = 10$ and $\sigma = 2.5$, calculate the standard error $\sigma_{\bar{x}}$ for the normal distribution with sample size n = 25.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{25}} = 0.5$$

Case 2: σ is unknown

Given the population data with $\mu = 10$, calculate the estimated standard error $\widehat{\sigma x}$ for the normal distribution with sample size n = 25.

Note: In this case, we need to calculate the estimated population standard deviation by the sample data.

Step 1: Calculate the estimated stardard deviation of the population.

$$\hat{\sigma} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Step 2: Calculate the estimated standard error of the t-distribution.

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}$$

```r
# Generate the population data
x <- rnorm(100, mean=10, sd=2.5)

# Extract 25 samples from the population data
s <- sample(x, 25)

# Calculate the mean of the sample
sBar <- mean(s)

# Calculate the estimated population standard deviation
sdHat <- round(sqrt((sum((s-sBar)^2)/(25-1))), 3)

print(paste("The estimated population standard deviation is ", sdHat))
```

```
## [1] "The estimated population standard deviation is  2.314"
```

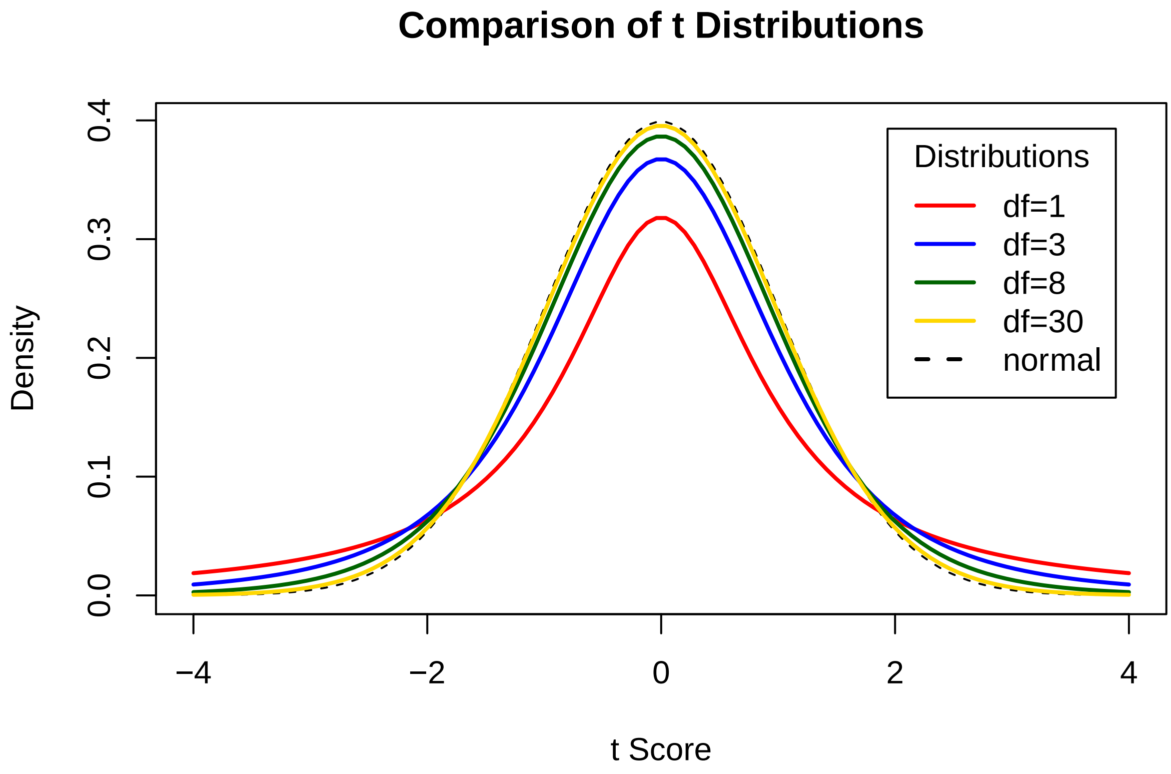
```
# Calculate the estimated standard error for the t-distribution
se <- round(sdHat/sqrt(25), 3)

print(paste("The estimated standard erorr of the t-distribution is ", se))
```

```
## [1] "The estimated standard erorr of the t-distribution is  0.463"
```

Thanks to the t-distribution, now we can solve the probability problem with the new defined distribution. Since the shape of the t-distribution is mainly depends on the sample size (n), the probability distribution table and r function will be slightly different to the standardized normal distribution (z-distibution). The t-distribution has the standardized scores called "t statistic":

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}}$$

It is also worth to note that one of the properties of the t distribution is that the shape of the distribution is mainly determined by the sample size. Here we have a visualization to help to understand.



Example:

A production report indicates that the production lead time of the company is normally distributed with population mean $\mu = 13$minutes and sample sum of squares $\sum(x_i - \bar{x})^2 = 91.4$ minutes. What is the probability to randomly select 10 samples from the production that has the average production time greater than 15 minutes $\bar{x} > 15$?

Solution:

$$P(\bar{x} > 15) = ?$$

Step 1: Calculate the estimated population standard deviation $\hat{\sigma}$

$$\hat{\sigma} = \sqrt{\frac{91.4}{(10-1)}} = 3.187$$

Step 2: Calculate the estimated standard error of the t-distribution

$$\hat{\sigma}_{\bar{x}} = \frac{3.187}{\sqrt{10}} = 1.008$$

Step 3: Convert $\bar{x}$ to the t statistic score

$$t = \frac{15-13}{1.008} = 1.984$$

Step 4: Find the area under the t distribution to the right of t = 1.984 (using the pt() function in r)

```r
# Calculate the probability with the z score
round(1-pt(1.984, df=9), 3)
```

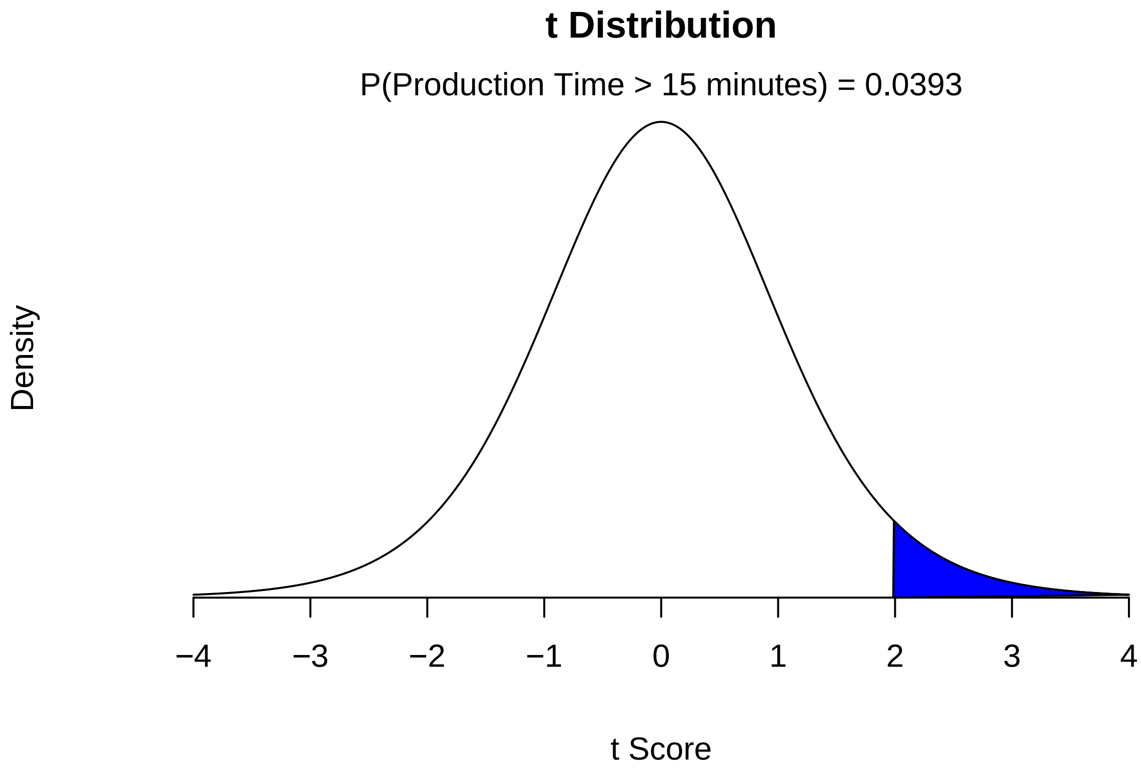
```
## [1] 0.039
```

```r
# Alternative, use the parameter lower.tail=FALSE
round(pt(1.984, df=9, lower.tail=FALSE), 3)
```

```
## [1] 0.039
```

$$P(t > 1.984) = 0.0393 \; or \; 3.93\%$$

Graphical Solution:



Confident Intervals

Confidence intervals are used to indicate how accurate a calculated statistic is likely to be. Confidence intervals can be calculated for a variety of statistics, such as the mean, median, or slope of a linear regression. In this section, we focus on confidences intervals for means.

Most of the statistics we use assume we are analyzing a sample which we are using to represent a larger population. If extension educators want to know about the caloric intake of 7^{th} graders, they would be hard-pressed to get the resources to have every 7^{th} grader in the U.S. keep a food diary. Instead they might collect data from one or two classrooms, and then treat the data sample as if it represents a larger population of students.

The mean caloric intake could be calculated for this sample, but this mean will not be exactly the same as the mean for the larger population. If we collect a large sample and the values aren't too variable, then the sample mean should be close to the population mean. But if we have few observations, or the values are highly variable, we are less confident our sample mean is close to the population mean.

We will use confidence intervals to give a sense of this confidence.

Our sample mean is a point estimate for the population parameter. A point estimate is a useful approximation for the parameter, but considering the confidence interval for the estimate gives us more information.

As a definition of confidence intervals, if we were to sample the same population many times and calculated a sample mean and a 95% confidence interval each time, then 95% of those intervals would contain the actual population mean.

If this definition of confidence intervals doesn't make much intuitive sense to you at this point, don't worry about it. Working through some of the examples in this section will help you understand their usefulness.

One use of confidence intervals is to give a sense of how accurate our calculated statistic is relative to the population parameter.

Interval Estimate of a Population Mean: (σ is known)

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

where: $\bar{x}$ is the sample mean

1 - α is the confidence coefficient

$z_{^{\alpha}/_{2}}$ is the z value providing an area of $^{\alpha}/_{2}$ in the upper tail of the standard normal probability distribution

σ is the population standard deviation

n is the sample size

Most Commonly Used Confidence Levels:

Confidence Level	α	α	Table Look-up Area	$z_{^{\alpha}/_{2}}$
90%	0.10	0.05	0.9500	1.645
95%	0.05	0.025	0.9750	1.960
99%	0.01	0.005	0.9950	2.576

Because 90% of all the intervals constructed using $\bar{x} \pm 1.645_{\sigma_{\bar{x}}}$ will contain the population mean, we say we are 90% confident that the interval $\bar{x} \pm 1.645_{\sigma_{\bar{x}}}$ includes the population mean μ.

We say that this interval has been established at the 90% confidence level. The value .90 is referred to as the confidence coefficient.

Example:

Discount Sounds has 260 retail outlets throughout the United States. The firm is evaluating a potential location for a new outlet, based in part, on the mean annual income of the individuals in the marketing area of the new location.

A sample of size n = 36 was taken; the sample mean income is $41,100. The population is not believed to be highly skewed. The population standard deviation is estimated to be $4,500, and the confidence coefficient to be used in the interval estimate is 0.95.

Solution:

$$z_{^{\alpha}/_{2}} \frac{\sigma}{\sqrt{n}} = 1.96(\frac{4500}{\sqrt{36}}) = 1470$$

Thus at 95% confidence, the margin of error (marginal difference) is $1,470.

Interval Estimate of μ is:

$$41,100 \pm 1,470$$

or

$$39,630 \; \textit{to} \; 42,570$$

We are 95% confident that the interval between $39,630 to $42,570 contains the population mean.

Interval Estimate of a Population Mean: (σ is unknown)

If an estimate of the population standard deviation σ cannot be developed prior to sampling, we use the sample standard deviation s to estimate σ. In this case, the interval estimate for μ is based on the t distribution. A specific t distribution depends on a parameter known as the degrees of freedom. Degrees of freedom refer to the number of independent pieces of information that go into the computation of s.

A t distribution with more degrees of freedom has less dispersion. As the degrees of freedom increases, the difference between the t distribution and the standard normal probability distribution becomes smaller and smaller. For more than 100 degrees of freedom, the standard normal z value provides a good approximation to the t value. Usually, a sample size of n $\geq$ 30 is adequate when using the expression $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ to develop an interval estimate of a population mean. If the population distribution is highly skewed or contains outliers, a sample size of 50 or more is recommended. If the population is not normally distributed but is roughly symmetric, a sample size as small as 15 will suffice. If the population is believed to be at least approximately normal, a sample size of less than 15 can be used.

Example:

A reporter for a student newspaper is writing an article on the cost of off-campus housing. A sample of 16 one-bedroom apartments within a half-mile of campus resulted in a sample mean of $750 per month and a sample standard deviation of $55.

Let us provide a 95% confidence interval estimate of the mean rent per month for the population of one-bedroom apartments within a half-mile of campus. We will assume this population to be normally distributed.

Solution:

At 95% confidence, $\alpha = 0.05$, and $\alpha/2 = 0.025$

$t_{0.025}$ is based on n - 1 => 16 - 1 = 15 degrees of freedom

Interval Estimate:

$$\bar{x} \pm t_{0.025} \frac{s}{\sqrt{n}}$$

$$750 \pm 2.131 \frac{55}{\sqrt{16}} = 750 \pm 29.30$$

We are 95% confident that the mean rent per month for the population of one-bedroom apartments within a half-mile of campus is between $720.70 and $779.30.

Example:

National Discount, Inc. has general merchandise retail outlets in 260 locations throughout the U.S. In considering possible new retail outlet locations, National evaluates each potential new location on several factors, one of which is the mean annual-income of the individuals in the marketing area serviced by the new outlet.

To estimate the population mean annual income m, National decided to use a simple random sample of size n = 64. In this sample, $\bar{x}$ = $21,000 and s = $5,600.

Based on similar annual income surveys, the standard deviation of annual incomes in the entire population is considered known with σ = $5,000.

Develop a 95% confidence interval estimate.

Solution:

Equation: $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

```
z <- qnorm(0.025, lower.tail = FALSE)
margin_of_error <- z * 5000 / sqrt(64)
Upper <- 21000 + margin_of_error
Lower <- 21000 - margin_of_error
print(paste("The 95% confidence interval of the average income is $", round(Lower, digits = 0),"and $",
```

```
## [1] "The 95% confidence interval of the average income is $ 19775 and $ 22225 ."
```

Example:

A reporter for a student newspaper is writing an article on the cost of attending college. A portion of the article deals with the cost of off-campus housing. A sample of 30 one-bedroom units within one-half mile of campus shows a sample mean = $700 per month. The sample standard deviation is s = $60.

Provide a 95% confidence interval estimate of the population mean cost per month for one-bedroom units within one-half mile of campus.

Use formula $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$. Because we do not know the population standard deviation.

```
t <- qt(0.025, df = 29, lower.tail = FALSE)
margin_of_error <- t * 60 / sqrt(30)
Upper <- 700 + margin_of_error
Lower <- 700 - margin_of_error
print(paste("95% C.I. of the average monthly cost for one-bedroom units is $", round(Lower, digits = 2)
```

```
## [1] "95% C.I. of the average monthly cost for one-bedroom units is $ 677.6 and $ 722.4 ."
```

In some cases, we may want to find out if a sample data significantly different (higher or lower) to the other sample given a different conditions. The confidence intervals is useful to define how significantly different between the two samples.

Example:

Imagine that we consider a town with a mean household income of greater than $100,000 to be high-income.

For Town A we sample some households, and calculate the mean household income and the 95% confidence interval for this statistic. The mean is $125,000, but the data are quiet variable, and the 95% confidence interval is from $75,000 to $175,000. In this case, we don't have much confidence that Town A is actually a high-income town. The point estimate for the population mean is greater than $100,000, but the confidence interval extends considerably lower than this threshold.

For Town B, we also get a mean of $125,000, so the point estimate is the same as for Town A. But the 95% confidence interval is from $105,000 to $145,000. Here, we have some confidence that Town B is actually a high-income town, because the whole 95% confidence interval lies higher than the $100,000 threshold.

Hypothesis Testing

Statistical hypothesis testing is used to assess the strength of the evidence in a random sample against a stated null hypothesis concerning a population parameter. A null hypothesis is a conjecture about a population parameter that is stated as a mathematical equation. The usual process of hypothesis testing consists of four steps.

1. Formulate the **null hypothesis** H_0 (commonly, that the observations are the result of pure chance) and the **alternative hypothesis** H_a (commonly, that the observations show a real effect combined with a component of chance variation).

2. Identify a **test statistic** that can be used to assess the truth of the null hypothesis.
3. Compute the **p-value**, which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the null hypothesis were true. The smaller the p-value, the stronger the evidence against the null hypothesis.
4. Compare the p-value to an acceptable significance value alpha (sometimes called an alpha value). If $p \leq \alpha$, that the observed effect is statistically significant, the null hypothesis is ruled out, and the alternative hypothesis is valid.

In fixed level testing, a significance level is chosen prior to collecting the sample, and the following decision rule is used:

If the p-value of the test statistic is less than or equal to the significance level (α), reject the null hypothesis
If the p-value of the test statistic is greater than the significance level (α), fail to reject the null hypothesis

Note: In this section, we mainly focus on using the t distribution for the hypothesis testing because population standard deviation σ is hardly be given in most of the real world problems.

Example:

The average US household spends $90 per day. Assume a sample of 30 households in Corning, NY, showed a sample mean daily expenditure of $84.50 with a sample standard deviation of $14.50.

Test the hypothesis $Ho : \mu = 90$ and $Ha : \mu \neq 90$ to see whether the population mean in Corning, NY, differs from the U.S. mean. Use α=0.05 significance level. What is your conclusion?

What is the p-value?

Solution:

$H_0 : \mu = 90$
$H_a : \mu \neq 90$

Method 1: Critical Value Approach

```r
# Calculate the t statistics with the given information
t = (84.50 - 90) / (14.50 / sqrt(30))

# Identify the critical t statistics with df = 29
tcrit = qt(0.025, 29)

# Print the results
print(paste("t value is", round(t, digits = 3)))
```

```
## [1] "t value is -2.078"
```

```r
print(paste("t critical value is", round(tcrit, digits = 3)))
```

```
## [1] "t critical value is -2.045"
```

```r
print("Is t value less than t critical value?")
```

```
## [1] "Is t value less than t critical value?"
```

```
t < tcrit
```

```
## [1] TRUE
```

Conclusion:

Reject H_0 at a 0.05 level of significance. The population mean in Corning, NY, differs from the U.S. mean at a 0.05 level of significance.

Method 2: p-value Approach

```
# Calculate the t statistics with the given information
t = (84.50 - 90) / (14.50 / sqrt(30))

# Calculate the probability on two tails given the t statistics with df = 29
p = 2 * pt(t, 29)

# Print the results
print(paste("p value= ", round(p, digits = 4)))
```

```
## [1] "p value=  0.0467"
```

```
print("Is p value less than the significance level 0.05?")
```

```
## [1] "Is p value less than the significance level 0.05?"
```

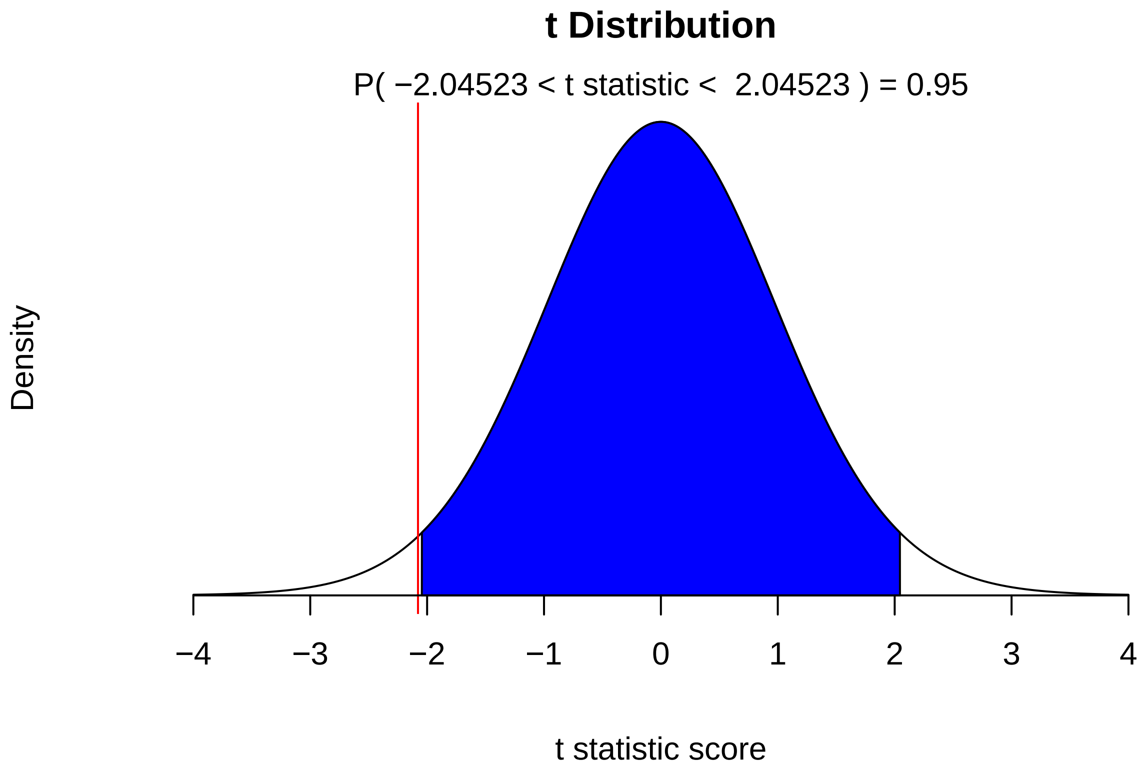
```
p < 0.05
```

```
## [1] TRUE
```

Conclusion:

Reject H_0 at a 0.05 level of significance. The population mean in Corning, NY, differs from the U.S. mean at a 0.05 level of significance.

Graphical Solution:



Example:

The Coca-Cola Company reported that the mean per capita annual sales of its beverages in the United States was 423 eight-ounce servings (Coca-Coca Company Website, February 3, 2009).

Suppose you are curious whether the consumption of Coca-Cola beverages is higher in Atlanta, Georgia, the location of Coca-Cola's corporate headquarters. A sample of 36 individuals from the Atlanta area showed a sample mean annual consumption of 460.4 eight-ounce servings with a standard deviation of s=101.9 ounces. Using α =0.05, do the sample results support the conclusion that mean annual consumption of Coca-Cola beverage products is higher in Atlanta?

Solution:

Let μ denote the mean annual consumption of Coca-Cola beverage products in Atlanta.

$H_0 : \mu \leq 423$

$H_a : \mu > 423$

Method 1: Critical Value Approach

```r
# Calculate the t statistics from the given information
t = (460.4 - 423) / (101.9 / sqrt(36))

# Calculate the critical region locations
tcrit = qt(0.05, 35, lower.tail=F)

# Print the results
print(paste("t value is", round(t, digits = 3)))
```

```
## [1] "t value is 2.202"
```

```
print(paste("t critical value is", round(tcrit, digits = 3)))
```

```
## [1] "t critical value is 1.69"
```

```
print("Is t value greater than t critical value?")
```

```
## [1] "Is t value greater than t critical value?"
```

```
t > tcrit
```

```
## [1] TRUE
```

Conclusion:

Reject H_0 at a 0.05 level of significance. The sample results support the conclusion that mean annual consumption of Coca-Cola beverage products is higher in Atlanta at a 0.05 level of significance.

Method 2: p-value Approach

```
# Calculate the t statistic
t = (460.4 - 423) / (101.9 / sqrt(36))

# Calculate the probability for rejecting H0 based on the given t statistics
p = pt(t, 35, lower.tail=F)

# Print the results
print(paste("p value= ", round(p,digits=4)))
```

```
## [1] "p value=  0.0172"
```

```
print("Is p value less than the significance level 0.05?")
```

```
## [1] "Is p value less than the significance level 0.05?"
```

```
p < 0.05
```

```
## [1] TRUE
```

Conclusion:

Reject H_0 at a 0.05 level of significance. The sample results support the conclusion that mean annual consumption of Coca-Cola beverage products is higher in Atlanta at a 0.05 level of significance.

Example:

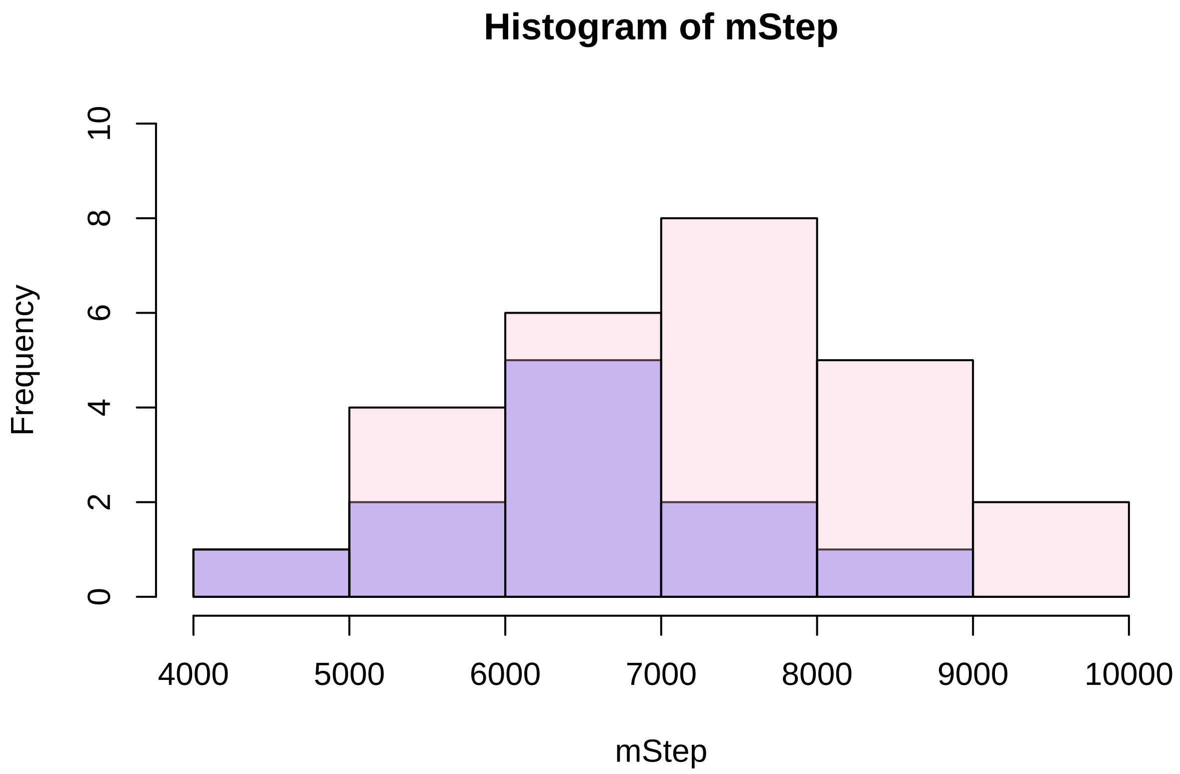
In this example, extension educators had students wear pedometers to count their number of steps over the course of a day. The following data are the result. Rating is the rating each student gave about the usefulness of the program, on a 1-to-10 scale.

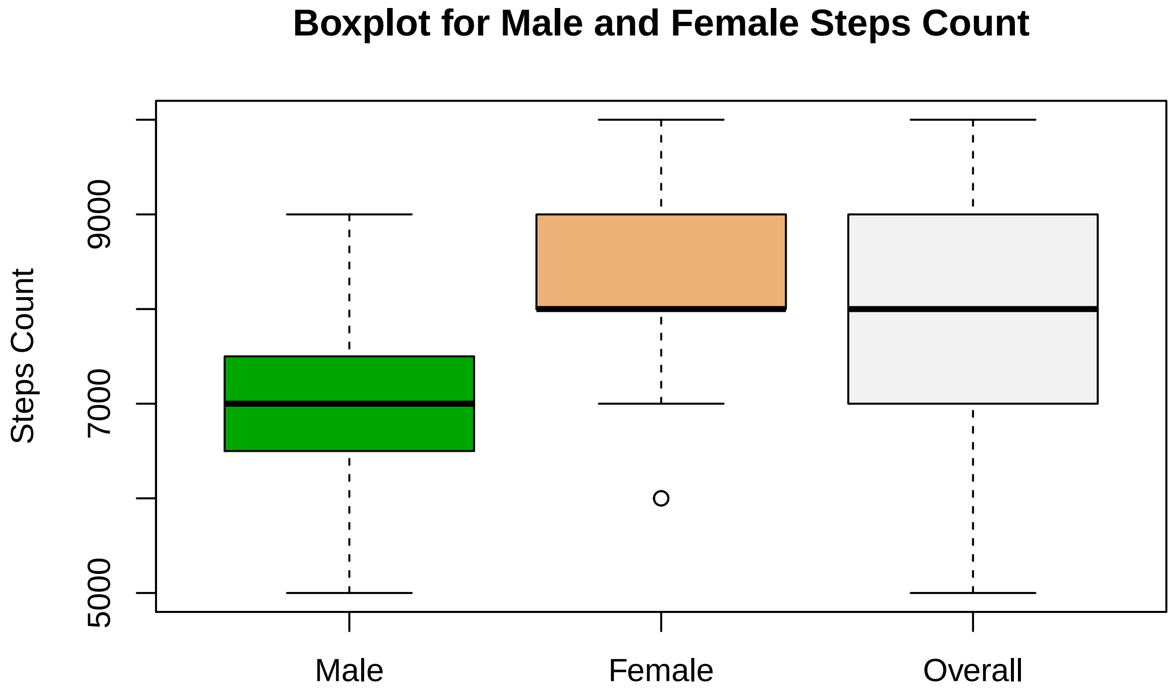
Suppose we are interested to test the hypothesis that male students, on average, recorded **less** steps in a day then an average students from the sample. Using $\alpha = 0.05$ or at 95% confident level, we are going to construct the test by calculating the sample means and standard deviations from both the female and male students. Then, use the t.test() function in R to see the result.

```
##    Student    Sex Teacher Steps Rating
## 1        a female  Catbus  8000      7
## 2        b female  Catbus  9000     10
## 3        c female  Catbus 10000      9
## 4        d female  Catbus  7000      5
## 5        e female  Catbus  6000      4
## 6        f female  Catbus  8000      8
## 7        g   male  Catbus  7000      6
## 8        h   male  Catbus  5000      5
## 9        i   male  Catbus  9000     10
## 10       j   male  Catbus  7000      8
## 11       k female Satsuki  8000      7
## 12       l female Satsuki  9000      8
## 13       m female Satsuki  9000      8
## 14       n female Satsuki  8000      9
## 15       o   male Satsuki  6000      5
## 16       p   male Satsuki  8000      9
## 17       q   male Satsuki  7000      6
## 18       r female  Totoro 10000     10
## 19       s female  Totoro  9000     10
## 20       t female  Totoro  8000      8
## 21       u female  Totoro  8000      7
## 22       v female  Totoro  6000      7
## 23       w   male  Totoro  6000      8
## 24       x   male  Totoro  8000     10
## 25       y   male  Totoro  7000      7
## 26       z   male  Totoro  7000      7
```

```
##     Student        Sex        Teacher       Steps           Rating
##  a      : 1   female:15   Catbus :10   Min.   : 5000   Min.   : 4.000
##  b      : 1   male  :11   Satsuki: 7   1st Qu.: 7000   1st Qu.: 7.000
##  c      : 1               Totoro : 9   Median : 8000   Median : 8.000
##  d      : 1                            Mean   : 7692   Mean   : 7.615
##  e      : 1                            3rd Qu.: 8750   3rd Qu.: 9.000
##  f      : 1                            Max.   :10000   Max.   :10.000
##  (Other):20
```

Visualization of the Data:





Solution:

Step 1: State the hypothesis for testing.

$H_0 : \mu_{male} \geq \mu$

$H_a : \mu_{male} > \mu$

Step 2: Calculate the overall sample mean and standard deviation in R.

```r
# Calculate the mean of the overall sample
M <- mean(Data$Steps)

# Calcualte the mean of male students
m <- mean(Data$Steps[Data$Sex == "male"])

# Calculate the standard deviation of the male students.
s <- sd(Data$Steps[Data$Sex == "male"])
```

Step 3: Calculate the t statistics for testing.

```r
# Calculate the t statistics by sample size n = 11 (male students)
t <- (m - M) / (s / sqrt(11))
```

Step 4: Calculate the p value and compare to our acceptance level (0.05)

```r
# Calculate the p value with degree of freedom, df = 11 - 1 = 10
p <- pt(t, 10, lower.tail = TRUE)

# Print the results
print(paste("p value= ", round(p, digits = 4)))
```

```
## [1] "p value=  0.0312"
```

```r
print("Is p value less than the significance level 0.05?")
```

```
## [1] "Is p value less than the significance level 0.05?"
```

```r
p < 0.05
```

```
## [1] TRUE
```

Using the t.test() function, we can get to the solution much faster and the function return with all the statistics we just calculated.

```r
# Use t.test() function for the sample problem
t.test(Data$Steps[Data$Sex == "male"], mu = mean(Data$Steps),
       alternative = "less", conf.level = 0.95)
```

```
##
##  One Sample t-test
##
```

```
## data:  Data$Steps[Data$Sex == "male"]
## t = -2.0961, df = 10, p-value = 0.03124
## alternative hypothesis: true mean is less than 7692.308
## 95 percent confidence interval:
##      -Inf 7598.636
## sample estimates:
## mean of x
##      7000
```

Testing Two Independent Samples

In many cases, we may not be interested in just testing the sample against the population mean. Sometime we are interested to test the different between two subsets of sample given an unique characteristic. For example, we may be interested to test the hypothesis that the female workers in some area of the U.S. get paid less than male workers; or the sales of the company from the west coast market is significantly higher than the sales from the east coast market; or family household is more likely to pay for the high speed internet service than an individual household. We can set up the hypothesis testing with student's t distribution with a little tweak.

For instance, if we want to setup the hypothesis for a two independent sample case,

Step 1: State the hypothesis for testing $H_o : \mu_1 = \mu_2$
$H_a : \mu_1 \neq \mu_2$

Step 2: Set the rejection criteria, α $\alpha = 0.05$ (Can be any accpetance level)

Step 3: Calculate the statistics for testing

Pooled Variance:

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

where,

n_1 and n_2 are the sample size for sample 1 and 2.
s_1^2 and s_2^2 are the sample variance for sample 1 and 2.

Student's t Statistic: (Two Independent Samples)

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

.

where,

$(\bar{x}_1 - \bar{x}_2)$ is the differece of the two sample means.
d_0 is the null hypothesis setting assume the difference between the two mean is zero.
$s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ is the standard error of the t distribution (weighted standard error).

Degree of Freedom:

$$df = n_1 + n_2 - 2$$

Step 4: Make conclusion about the test

The decision rule is if the p value greater than α, we fail to reject the null hypothesis. If the p value is smaller than α, we reject the null hypothesis at 5% significant level.

Example:

In a wage discrimination case involving male and female employees, independent samples of male and female employees with 5 years of experience or more provided the following hourly wage results. The null hypothesis is that male employees have a mean hourly wage less than or equal to that of the female employees. Rejection of H_0 leads to the conclusion that male employees have a mean hourly wage exceeding that of the female employees. Test the hypothesis with $\alpha = 0.01$. Does the discrimination appear to be present in this case?

Male: $n_1 = 44$, $\bar{x}_1 = \$9.25$, $s_1 = \$1.00$.

Female:$n_2 = 32$, $\bar{x}_2 = \$8.70$, $s_2 = \$0.80$.

Solution:

Define μ_1 as the mean hourly wage of male employees. Define μ_2 as the mean hourly wage of female employees. $H_0 : \mu_1 - \mu_2 \leq 0$
$H_a : \mu_1 - \mu_2 > 0$.

Use equations:

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

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

Also degrees of freedom $= n_1 + n_2 - 2$

```r
# Calculate the pooled variance
s <- ((44 - 1) * 1^2 + (32 - 1) * 0.8^2) / ((44 - 1) + (32 - 1))

# Calculate the t statistic
t <- ((9.25 - 8.70) - 0) / (s * sqrt(1/44 + 0.80/32))
print(paste("t value is", round(t,digits=2),"."))
```

```
## [1] "t value is 2.96 ."
```

```r
# Calculate the degree of freedom for the two samples
df <- (44 - 1) + (32 - 1)
print(paste("The degrees of freedom is", df,"."))
```

```
## [1] "The degrees of freedom is 74 ."
```

Critical Value Approach:

```r
# Find the critical value at 1% significant level
t_critical = qt(0.01, df, lower.tail = F)
print(paste("t critical value is", round(t_critical, digits = 3), "."))
```

```
## [1] "t critical value is 2.378 ."
```

Since t value > t critical value, reject H_0 at a 0.01 level of significance. Male employees have a mean hourly wage exceeding that of the female employees. The discrimination appears to be present in this case at a 0.01 level of significance.

p Value Appraoch:

```r
# Calculate the p value based on the t statistic
p = pt(t, df , lower.tail = F)
print(paste("p value is", round(p, digits = 5), "."))
```

```
## [1] "p value is 0.00204 ."
```

Since p value < 0.01, reject H_0 at a 0.01 level of significance. Male employees have a mean hourly wage exceeding that of the female employees. The discrimination appears to be present in this case at a 0.01 level of significance.