Interval Estimation

Lecture 6

1 Outline

- Estimation: Confidence Intervals
- The t-distribution
- Confidence Intervals for different parameters:
  - Population Mean
  - Difference Between Two Population Means
  - Population Proportion
  - Diffrence Between Two Population Proportions

2 Estimation

2.1 Statistical Inference

> **Note**
>
> **Statistical inference** is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample drawn from that population.

> **Important**
>
> The process of estimation entails calculating some statistic from the sample that is offered as an approximation of the corresponding parameter of the population from which the sample was drawn.

2.2 Point Estimation vs. Interval Estimation

Point Estimation

Point estimation is the process of finding a single number to use as our best guess for the value of an unknown population parameter. Point estimates do not provide a measure of uncertainty.

Interval Estimation

Interval estimation involves estimating two numerical values that define a range of values that, with a specified degree of confidence, most likely includes the parameter being estimated.

2.3 Point Estimation: Unbiased Estimators

Note

The single computed value has been referred to as an estimate. On the other hand, the rule that tells us how to compute this estimate is referred to as an estimator.

Unbiasedness

An estimator of a parameter Θ, T, is said to be an unbiased estimator of Θ if $E(T) = \Theta$. AN example of an unbiased estimator is the sample mean. The sample mean $\bar{X}$ is an unbiased estimator of the population mean.

2.4 Sampled vs. Target Populations

Target Population

The target population is the population about which one wishes to make an inference.

Sampled Population

The sampled population is the population from which one actually draws a sample.

⚠ Warning

If the sampled population and the target population are different, the researcher can reach conclusions about the target population only on the basis of nonstatistical considerations. Estimation from the sampled population could lead to biased estimates.

2.5 Random vs. Convenience Sampling

In most examples in this class, we assume that the data was collected from random samples. However, it is impossible or impractical to use truly random samples.

ℹ Convenience sampling

Convenience sampling involves sampling through volunteerism or readily available subjects to record data. While collecting data using convenience sampling is easy, the results might not be generalizable to the target population.

2.6 Interval Estimation: Confidence Intervals

There are many variants of interval estimation methods that can be used. One of the more popular interval estimates is the **confidence interval**.

❗ Confidence Interval

A confidence interval is a random interval that will contain the true value of a parameter with probability $1 - \alpha$. The value $1 - \alpha$ is called the **confidence level** of the interval.

❗ Important

The confidence interval provides all plausible values of the parameters. The parameter itself is treated as fixed, while the endpoints of the confidence interval are treated as random variables. The endpoints are random because it depends on the data collected.

⚠ Warning

The viewpoint for making a probability statement must be **before** the data are collected, not **after**.

2.7 Confidence Intervals

For the following tabs, supposed that the 95% confidence interval for the population mean, μ, was calculated to be the interval (80,100).

2.7.1 Misconception 1

> ⚠ Misconception 1
>
> Misconception 1: $P(80 < \mu < 100) = 0.95$

This is incorrect. The parameter μ is not random, hence there shouldn't be a probability assigned to it. 80 and 100 are also not random because we have already measured them. Either the interval contains the true value of μ or not. Hence, the probability interpretation does not make any sense.

2.7.2 Misconception 2

> ⚠ Misconception 2
>
> Misconception 2: 95% of the population lies in the interval (80,100).

The confidence interval is for the population mean, not the set of all data values from the population. A confidence interval for the mean will often include only a small fraction of the population.

2.8 Interpretation of Confidence Intervals

The confidence level is pertaining to the method that we calculated, not the particular interval. In the long run (or infinite repeated sampling), around 95% of the confidence intervals calculated will include μ.

> ❗ Important
>
> Since our method works 95% of the time, we believe that μ is in our 95% confidence interval. We don't have the time and means to calculate all confidence intervals.

3 Confidence Intervals

3.1 Confidence Interval Form

The confidence intervals in this chapter will follow the form:

$$estimate \pm reliability\ coefficient * standard\ error$$

3.2 Confidence Intervals for Population Means

We can estimate the population mean by using the mean of a representative sample from the population.

> **Warning**
>
> The method differs for cases with known and unknown population variance. Some approximations can be done when there is a large sample size.

3.3 CI for Mean: Known Population Variance

Cases in which population variances are known are rare. The best example of measurements with known population variances are standardized scales.

> **Note**
>
> The Wechsler Adult Intelligence Scale (WAIS) maintains a standard deviation of 15 points for its full scale IQ scores.

For a known population variance σ^2, sample mean $\bar{x}$, and sample size n, the 95% confidence limits for the $(1-\alpha)$ can be calculated using the following formula:

$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

3.4 CI for Mean: Notes

Confidence intervals are often written in parentheses notation as such:

$$\left(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$z_{1-\alpha/2}$ is referred to as the reliability coefficient and defined as the following:

$$P(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2$$

> **Tip**
>
> Recall that Z follows the standard normal distribution. $z_{1-\alpha/2}$ can be calculated using `qnorm(1-alpha/2,mean=0,sd=1)`. For a 95% confidence interval, $\alpha = 0.05$, then it follows that $1 - \alpha/2 = 1 - 0.05/2 = 0.975$. Therefore, the reliability coefficient for a 95% confidence interval can be calculated as:

```
alpha=0.05
qnorm(1-alpha/2,mean=0,sd=1)
```

```
[1] 1.959964
```

3.5 Example

The file `lec6example.csv` contains simulated values from a normal distribution with a variance of 4.

3.5.1 Question

Create a 95% confidence interval for the population mean. Is 1 a plausible value for the population mean?

3.5.2 Answer

```
# setwd("where your file is")
example <- read.csv("datasets/lec6example.csv")
sdpop <- sqrt(4)
```

We need to calculate the sample mean of the `Sims` column, the sample size, and the reliability coefficient.

```
xbar <- mean(example$Sims)
sampsize <- nrow(example)                                    ①
alpha <- 1-0.95
relcoeff <- qnorm(1-alpha/2,mean = 0, sd=1)                  ②
```

① `nrow()` calculates the number of rows of the data frame. If dealing with a singular vector, you can use `length()`. `nrow(example)` will yield the same answer as `length(example$Sims)`.

We can now calculate the 95% confidence interval limits.

```
lower <- xbar - relcoeff*sdpop/sqrt(sampsize)
upper <- xbar + relcoeff*sdpop/sqrt(sampsize)
lower
```

```
[1] 0.6637448
```

```
upper
```

```
[1] 1.303866
```

The resulting 95% confidence interval is (0.66, 1.3). Since 1 is in the confidence interval, then we can claim that 1 is a plausible value for the population mean.

3.6 Exercise

Consider the sleep health data uploaded on Canvas as `SleepHealthData.csv`. Suppose that in the population that this sample represents, the variance of the sleep duration is 0.5.

3.6.1 Question

Calculate the 95% confidence interval for the average sleep duration (`sleep_duration`) in the population this sample represents. Is 7 hours a plausible value for the average sleep distribution of the population?

3.6.2 Answer

```
sleep <- read.csv("SleepHealthData.csv")
sdpop <- sqrt(0.5)
xbar <- mean(sleep$sleep_duration)
sampsize <- nrow(sleep)
alpha <- 1-0.95
relcoeff <- qnorm(1-alpha/2)
# not specifying the mean and sd in qnorm assumes standard normal
```

```r
lower <- xbar - relcoeff*sdpop/sqrt(sampsize)
lower
```

```
[1] 7.060422
```

```r
upper <- xbar + relcoeff*sdpop/sqrt(sampsize)
upper
```

```
[1] 7.203749
```

The 95% confidence interval is (7.06,7.2). 7 is not a plausible value for the average sleep distribution of the population.

3.7 Precision

The precision of an interval estimate is related to its width.

> **Note**
>
> The precision, also known as the **margin of error**, can be expressed as the product of the reliability coefficient and the standard error.

For the case of the population mean with known population variance,

$$precision = z_{1-\alpha/2} * \frac{\sigma}{\sqrt{n}} = \frac{upper - lower}{2}$$

> **Note**
>
> Because of the symmetry of the confidence interval, the precision is half the width of the confidence interval.

3.8 CI for Mean: Unknown Population Variance

It is more common to not know the population variance when estimating the population mean. Because of this, we are inclined to use the next best thing: an estimate of the population variance from the collected sample. This estimate is the **sample variance**

> **Warning**
>
> Using an approximate value for the population variance implies that our standardized statistic **might not** follow the standard normal distribution. We need a new distribution to calculate reliability coefficients for these confidence intervals.

3.9 The Student's t-distribution

The t distribution is symmetric and bell-shaped like the normal distribution, but has heavier tails.

> **Note**
>
> Heavier tails mean that the distribution is more likely to produce values that fall farther from the mean compared to the normal distribution. The heavier tails account for the extra uncertainty introduced by using the sample variance to estimate the population variance.

> **Important**
>
> The t-distribution is defined by the degrees of freedom denoted by ν or `df`, and specific t-distributions can be written as t_ν.

3.10 The t distribution: `R`

The `R` syntax for the t-distribution with degrees of freedom `df` consists of the following functions:

- PDF: `dt(x,df)`
- CDF: `pt(x,df)`
- Quantile: `qt(x,df)`
- Generate/Simulate `n` t-distribution points: `rt(n,df)`

3.11 CI for Mean: Unknown Population Variance

When the population variance is unknown OR the sample from a normally distributed population has a low sample size, the $(1-\alpha)*100$ confidence interval can be calculated using the following formula:

$$\bar{x} \pm t_{\nu,1-\alpha/2}\frac{s}{\sqrt{n}}$$

> **Important**
>
> s is the sample standard deviation $t_{\nu,1-\alpha/2}$ is defined as $P(t < t_{\nu,1-\alpha/2}) = 1 - \alpha/2$. The degrees of freedom ν can be calculated as $\nu = n - 1$.

The 95% confidence interval can be written as:

$$\left(\bar{x} - t_{\nu,1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{\nu,1-\alpha/2}\frac{s}{\sqrt{n}}\right)$$

3.12 Precision

The precision of an interval estimate is related to its width.

> **Note**
>
> The precision, also known as the **margin of error**, can be expressed as the product of the reliability coefficient and the standard error.

For the case of the population mean with unknown population variance,

$$precision = t_{\nu,1-\alpha/2}\frac{s}{\sqrt{n}} = \frac{upper - lower}{2}$$

> **Note**
>
> Because of the symmetry of the confidence interval, the precision is half the width of the confidence interval.

3.13 Example

The file `lec6example.csv` contains simulated values from a normal distribution.

3.13.1 Question

Supposed the population variance is unknown. Create a 95% confidence interval for the population mean. Is 1 a plausible value for the population mean?

3.13.2 Answer

```
# setwd("where your file is")
example <- read.csv("datasets/lec6example.csv")
```

We need to calculate the sample mean of the `Sims` column, the sample size, and the reliability coefficient.

```
xbar <- mean(example$Sims)
sampsize <- nrow(example)                                    ①
stdev <- sd(example$Sims)                                    ②
df <- sampsize-1                                             ③
alpha <- 1-0.95
relcoeff <- qt(p=1-alpha/2,df=df)
```

① `nrow()` calculates the number of rows of the data frame. If dealing with a singular vector, you can use `length()`. `nrow(example)` will yield the same answer as `length(example$Sims)`.
② We are calculating the sample variance using `sd()`.
③ Degrees of freedom `df = n-1`.

We can now calculate the 95% confidence interval limits.

```
lower <- xbar - relcoeff*stdev/sqrt(sampsize)
upper <- xbar + relcoeff*stdev/sqrt(sampsize)
lower
```

```
[1] 0.6642442
```

```
upper
```

```
[1] 1.303367
```

The resulting 95% confidence interval is (0.66, 1.3). Since 1 is in the confidence interval, then we can claim that 1 is a plausible value for the population mean.

3.14 Exercise

Consider the sleep health data uploaded on Canvas as `SleepHealthData.csv`.

3.14.1 Question

Calculate the **99%** confidence interval for the average heart rate (`heart_rate`) in bpm in the population this sample represents. Is 65 bpm a plausible value for the average sleep distribution of the population? Calculate the margin of error.

3.14.2 Answer

```
sleep <- read.csv("SleepHealthData.csv")
sdpop <- sd(sleep$heart_rate)
xbar <- mean(sleep$heart_rate)
sampsize <- nrow(sleep)
alpha <- 1-0.99
relcoeff <- qt(p=1-alpha/2,df=sampsize-1)
# not specifying the mean and sd in qnorm assumes standard normal

lower <- xbar - relcoeff*sdpop/sqrt(sampsize)
lower
```

```
[1] 69.6121
```

```
upper <- xbar + relcoeff*sdpop/sqrt(sampsize)
upper
```

```
[1] 70.71945
```

The 99% confidence interval is (69.61,70.72). 65 is not a plausible value for the average heart rate of the population.

The margin of error can be calculated as:

```
relcoeff*sdpop/sqrt(sampsize)
```

```
[1] 0.5536753
```

```
(upper-lower)/2
```

```
[1] 0.5536753
```

3.15 Confidence Interval for Population Proportions

Many questions of interest to the health worker relate to population proportions.

> **Note**
>
> Some questions could be:
>
> - What is the recovery rate of patients for a particular type of treatment?
> - What is the prevalence rate of a disease?
> - What proportion of medical providers agree with current health recommendations?

3.16 CI: Population Proportions

For a sample of size n and sample proportion $\hat{p}$, the $100(1-\alpha)\%$ confidence interval for the population proportion π is given by:

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Similar to the CI for means, $z_{1-\alpha/2}$ is defined as $P(Z \leq z_{1-\alpha/2}) = 1-\alpha/2$.

> **Important**
>
> The use of $z_{1-\alpha/2}$ can be justified using the central limit theorem (CLT). The most common guideline for the CLT to apply to proportions is that $n\pi$ and $n(1-\pi)$ are both greater than five.

The CI would then be reported as:

$$\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

3.17 Example

The data set `Liver_Steatosis.csv` comes from a retrospective cohort study to estimate the prevalence of liver steatosis (fatty liver disease) in those who had bariatric surgery (Wu et al, 2012). The data set also includes other covariates and comorbidities.

3.17.1 Question

One of the comorbidities studied was history of metabolic syndrome (1=Yes, 2=No, `NA` = missing). Calculate a 95% confidence interval for the prevalence of metabolic syndrome based on the data. Only consider **non-missing data** in your analysis.

> ⚠ Warning
>
> This data set was collected from patients who had bariatric surgery at the Cleveland Clinic between 2005 and 2009 and underwent liver biopsy. Is this result generalizable to all US adults who have had bariatric surgery?

3.17.2 Answer

We need to calculate the proportion of patients reported to have metabolic syndrome.

```
library(summarytools)
liver <- read.csv("datasets/Liver_Steatosis.csv")
summarytools::freq(liver$MET_Syndrome)
```

```
Frequencies
liver$MET_Syndrome
Type: Integer

             Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
---------- ------ --------- -------------- --------- --------------
         0    144     32.65          32.65     32.51          32.51
         1    297     67.35         100.00     67.04          99.55
     <NA>      2                               0.45         100.00
     Total    443    100.00         100.00    100.00         100.00
```

Based on `freq`, the proportion of those with metabolic syndrome is 0.6735 out of a sample size of 441 (removed two missing). We now have everything we need to calculate the confidence interval.

```
phat <- 297/443
sampsize <- 441
alpha <- 1-0.95
relcoeff <- qnorm(p=1-alpha/2,mean=0,sd=1)
se <- sqrt((phat*(1-phat)/sampsize))
```

```r
lower <- phat - relcoeff*se
lower
```

```
[1] 0.6265577
```

```r
upper <- phat + relcoeff*se
upper
```

```
[1] 0.7143001
```

The 95% confidence interval for the proportion of patients with metabolic syndrome is (0.6266,0.7143).

> **Warning**
>
> Caution should be exercised in generalizing this result to all patients from the country as this is a very specific sample.

3.18 Exercise

Consider the sleep health data uploaded on Canvas as `SleepHealthData.csv`.

3.18.1 Question

Use the `sleep_disorder` column to do the following:

- Estimate the prevalence rate of sleep apnea in the population this sample represents using the sample proportion.
- Calculate a 90% confidence interval for the prevalence rate of sleep apnea.

3.18.2 Answer

```r
freq(sleep$sleep_disorder)
```

```
Frequencies
sleep$_disorder
Type: Character

                   Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
----------------- ------ --------- -------------- --------- --------------
         Insomnia     77     20.59          20.59     20.59          20.59
             None    219     58.56          79.14     58.56          79.14
      Sleep Apnea     78     20.86         100.00     20.86         100.00
             <NA>      0                                0.00         100.00
            Total    374    100.00         100.00    100.00         100.00
```

The proportion of participants who reported sleep apnea is 0.2086. The sample size is 374. We can now calculate a 90% confidence interval for the prevalence rate of sleep apnea based on this sample.

```
phat <- 78/374
sampsize <- 374
alpha = 1-0.9
relcoeff = qnorm(p=1-alpha/2,mean=0,sd=1)
se <- sqrt((phat*(1-phat)/sampsize))

lower <- phat - relcoeff*se
lower
```

```
[1] 0.174001
```

```
upper <- phat + relcoeff*se
upper
```

```
[1] 0.2431113
```

The 90% confidence interval for the prevalence rate of sleep apnea in the population this sample represents is (0.174,0.2431).

4 Sample Size Calculation

4.1 How to determine sample sizes

Before data are collected, we often need to make a decision about how large of a sample will be required to estimate population parameters.

> **!** Important
>
> The sample size depends on the precision/margin of error we would like for our resulting confidence intervals.

4.2 Sample size for population proportion estimation

Suppose we want our margin of error (MOE) to be equal to E for our planned study. The corresponding estimate for the sample size needed to reach this level of precision is given by

$$n = z^2_{1-\alpha/2}\frac{\hat{\pi}(1-\hat{\pi})}{E^2}$$

> **!** Important
>
> Note that the population proportion AND sample proportion are unknown before data collection. However, we need to have an idea what the expected proportion is for our study, denoted by $\hat{\pi}$ in this case. This proportion can be estimated from prior or pilot studies. If there are no prior studies available (which is the case for some pilot studies), we can assume $\hat{\pi} = 0.5$ to create a conservative confidence interval.

4.3 Example

Suppose we want to design a pilot study to estimate the proportion of young adults in NV who use anabolic-androgenic steroids for muscle growth.

4.3.1 Question

Calculate the sample size required so the resulting 95% confidence interval estimates will have a margin of error of 5%.

4.3.2 Answer

```
E <- 0.05
pi_hat <-0.5
alpha <- 1-0.95
z <- qnorm(p=1-alpha/2)
```

```r
sampsize <- z^2*pi_hat*(1-pi_hat)/E^2
sampsize
```

```
[1] 384.1459
```

> **!** Important
>
> When doing sample size calculations, always **round up** your final answer. For this study, you would recommend recruiting 385 participants for the study so that the MOE for the interval estimate of the usage rate of anabolic-androgenic steroids for muscle growth will be approximately 5%.

4.4 Exercise

Suppose we performed the pilot study on steroid use and we plan to use the results to design a large scale study.

4.4.1 Question

The pilot study yielded an estimated usage rate of 32%. Calculate the sample size required so the resulting 95% confidence interval estimates will have a margin of error of 1%.

4.4.2 Answer

```r
E <- 0.01
pi_hat <-0.32
alpha <- 1-0.95
z <- qnorm(p=1-alpha/2)

sampsize <- z^2*pi_hat*(1-pi_hat)/E^2
sampsize
```

```
[1] 8359.014
```

> **Important**
>
> For this study, you would recommend recruiting 8360 participants for the study so that the MOE for the interval estimate of the usage rate of anabolic-androgenic steroids for muscle growth will be approximately 1%.

4.5 Sample size for population mean estimation

Suppose we want our margin of error (MOE) to be equal to E for our planned study. The corresponding estimate for the sample size needed to reach this level of precision is given by

$$n = z^2_{1-\alpha/2}\frac{\hat{\sigma}^2}{E^2}$$

where $\hat{\sigma}$ is an a priori estimate of the population standard deviation. As with the estimated population proportion, this value can be obtained from prior studies or estimated using pilot studies.

> **Tip**
>
> Recall that for a Gaussian distribution, approximately 99% of the population can be found within the range $(\mu - 3\sigma, \mu + 3\sigma)$. The width of this region is 6σ, which could be approximated by the range of the data. Hence, an approximate of the population standard deviation could be expressed as: $\sigma = R/6$, where R is the range.

4.6 Example

Suppose we want to estimate the average number of views that health provider social media pages in Nevada receive weekly.

4.6.1 Question

If the range of mean views was estimated to be 300,000, how many social media pages should be sampled so that our 99% confidence interval for the average number of views of Nevada health provider social media pages has a margin of error of 10,000?

4.6.2 Answer

```r
alpha <- 1-0.99
z <- qnorm(p=1-alpha/2)
sigma <- 300000/6
E <- 10000
sampsize <-z^2*sigma^2/E^2
sampsize
```

```
[1] 165.8724
```

We need to sample 166 to achieve a margin of error of 10,000 for our 99% confidence interval.

4.7 Example

You designed a study to estimate the average number of views that health provider social media pages in Nevada receive weekly.

4.7.1 Question

Suppose your boss was not happy with the margin of error for the 99% confidence interval of 10,000 and wanted to decrease the margin of error to 500. How many samples would be recommended for the study if the range of mean views was estimated to be 300,000.

4.7.2 Answer

```r
alpha <- 1-0.99
z <- qnorm(p=1-alpha/2)
sigma <- 300000/6
E <- 500
sampsize <-z^2*sigma^2/E^2
sampsize
```

```
[1] 66348.97
```

We need to sample 66349 to achieve a margin of error of 500 for our 99% confidence interval.

5 Confidence Intervals for Difference in Statistics

5.1 CI for Difference in Statistics

Often, we are more interested in comparing between two groups.

> **Note**
>
> Recall that in Chapter 5, we showed how differences between two parameters had a different sampling distribution that individual parameters.

5.2 CI for Difference in Population Means

For two groups with known population variances σ_1 and σ_2, the $100(1-\alpha)\%$ can be calculated using the following formula:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

5.3 CI for Difference in Population Means

For two groups with unknown population variances, we will still use the sample variances to estimate the population variances of each group. However, we have to consider two cases:

5.3.1 Equal Pop. Variances

When the population variances are assumed to be equal, we need to introduce a pooled variance to estimate the variance for each group.

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

This results to the following $100(1-\alpha)\%$ confidence interval:

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

where $t_{1-\alpha/2,df}$ has degrees of freedom equal to $df = n_1 + n_2 - 2$.

5.3.2 Unequal Pop. Variances

When the population variances are not equal, then we don't need to pool the variances. However, the t-distribution has to account for the fact that we are using estimated variances for both groups.

We introduce the **Welch-Satterthwaite** procedure to construct this confidence interval such that the $100(1-\alpha)\%$ confidence interval is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{1-\alpha/2,df}\sqrt{s_1^2/n_1 + s_2^2/n_2}$$

where

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\left(\frac{s_1^2}{n_1}\right)^2/(n_1-1) + \left(\frac{s_2^2}{n_2}\right)^2/(n_2-1)}$$

5.4 `R` implementation

We can use `R` to manually calculate the confidence intervals using the formulas provided. However, there is an `R` function called `t.test(formula,conf.level=conf.level, var.equal=TRUE)` that we can use to calculate these confidence intervals.

> **Note**
>
> This approach will not work with aggregated data (Mean and SD provided). When data is aggregated, you would need to use the formulas.

> **Note**
>
> The `formula` should be of the form `response_variable~group_variable`. It can also be of the form `x = df$col1, y=df$col2` if the data has two separate columns for the two groups.

5.5 Example

Consider the sleep health data uploaded on Canvas as `SleepHealthData.csv`.

5.5.1 Question

Calculate a 95% confidence interval for the difference in average age of males and females who participated in the study assuming equal population variances.

5.5.2 Answer

The groups are not presented as two separate columns, but rather defined by another column.

```
confidence_int <- t.test(age~gender,
       data=sleep,
       var.equal=T,
       conf.level=0.95)

confidence_int
```

```
    Two Sample t-test

data:  age by gender
t = 14.329, df = 372, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Female and group Male is not e
95 percent confidence interval:
  8.913564 11.749098
sample estimates:
mean in group Female   mean in group Male
            47.40541             37.07407
```

The 95% confidence interval was calculated to be: (8.9136,11.7491)

5.6 Example

The file `40yarddash.csv` includes the 40-yard dash times from randomly selected football players from a D1 and D3 university.

5.6.1 Question

Calculate a 90% confidence interval for the difference in average 40-yard dash times between the two groups assuming unequal variances.

5.6.2 Answer

The groups are presented as two separate columns, hence we cannot use the formula notation. We have to define `x` and `y` in `t.test` to estimate the difference of averages in `x` and `y`

```
yard<- read.csv("40yarddash.csv")
confidence_int <- t.test(x=yard$D1,y=yard$D3,
       var.equal=F,
       conf.level=0.9)

confidence_int
```

```
    Welch Two Sample t-test

data:  yard$D1 and yard$D3
t = -4.5203, df = 41.952, p-value = 4.979e-05
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -0.7112031 -0.3254636
sample estimates:
mean of x mean of y
 4.717667  5.236000
```

The 90% confidence interval for the difference in 40-yard dash times was calculated to be: (-0.7112,-0.3255)

5.7 Exercise

Consider the sleep health data uploaded on Canvas as `SleepHealthData.csv`.

5.7.1 Question

Calculate a 95% confidence interval for the difference in average sleep duration of males and females who participated in the study assuming equal population variances.

5.7.2 Answer

```r
confidence_int <- t.test(sleep_duration~gender,
       data=sleep,
       var.equal=T,
       conf.level=0.95)

confidence_int
```

```
    Two Sample t-test

data:  sleep_duration by gender
t = 2.3624, df = 372, p-value = 0.01867
alternative hypothesis: true difference in means between group Female and group Male is not e
95 percent confidence interval:
 0.03239537 0.35404821
sample estimates:
mean in group Female   mean in group Male
            7.229730             7.036508
```

The 95% confidence interval was calculated to be: (0.0324,0.354)

5.8 CI for Difference in Proportions

For populations with n_1 and n_2 are large and the population proportions π_1 and π_2 are not too close to 0 or 1, we can implement the central limit theorem and assume the Gaussian distribution for the sampling distribution of the difference in proportions.

The resulting $100(1-\alpha)\%$ confidence interval for $\pi_1 - \pi_2$ can be expressed as:

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

> **Note**
>
> This does not include a continuity correction.

5.9 R implementation

The function `prop.test()` can calculate the confidence intervals for the difference in proportions, even adding the continuity correction. Suppose the successes for each group are denoted by `x1` and `x2`, while the totals for each group are `n1` and `n2`.

```
prop.test(x=c(x1,x2), n=c(n1,n2),conf.level=0.95, correct=FALSE)
```

For the resulting CI with the continuity correction, set `correct=TRUE`.

5.10 Example

Consider the liver steatosis data set.

```
liver <- read.csv("Liver_Steatosis.csv")
```

5.10.1 Question

Calculate the 95% confidence interval of the difference between the proportion of positive liver steatosis detections using ultrasound (`LS_US=1`) and through biopsy (`LS_Biopsy=1`). Is it plausible that the long-run detection rates for the two methods are equal? Omit missing data and apply the continuity correction.

5.10.2 Answer

```
library(summarytools)
freq(liver$LS_US)
```

```
Frequencies
liver$LS_US
Type: Numeric

              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
----------- ------ --------- -------------- --------- --------------
          0    122     28.98          28.98     27.54          27.54
        0.5      4      0.95          29.93      0.90          28.44
          1    295     70.07         100.00     66.59          95.03
       <NA>     22                               4.97         100.00
      Total    443    100.00         100.00    100.00         100.00
```

```
freq(liver$LS_Biopsy)
```

```
Frequencies
liver$LS_Biopsy
Type: Numeric

             Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
----------- ------ --------- -------------- --------- --------------
          0    124     27.99          27.99     27.99          27.99
        0.5      8      1.81          29.80      1.81          29.80
          1    311     70.20         100.00     70.20         100.00
       <NA>      0                                0.00         100.00
      Total    443    100.00         100.00    100.00         100.00
```

There are 295/421 positive diagnoses for the ultrasound group and 311/443 for the biopsy group.

```
ptest<- prop.test(x=c(295,311),
          n=c(443-22,443),
          conf.level=0.95,
          correct=TRUE)

ptest
```

```
    2-sample test for equality of proportions with continuity correction

data:  c(295, 311) out of c(443 - 22, 443)
X-squared = 1.3976e-29, df = 1, p-value = 1
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.06369117  0.06105315
sample estimates:
   prop 1    prop 2
0.7007126 0.7020316
```

The resulting 95% confidence interval is (-0.0637, 0.0611). 0 is part of the interval, which means it is a plausible value of the difference in detection rates between the two methods. Hence, it is plausible that the long-run detection rates of the two methods do not differ.

5.11 Exercise

In a randomized controlled trial, 23 out of 40 participants who received the treatment reported improvement while 14 out of 35 participants who received the placebo reported improvement.

5.11.1 Question

Calculate a 99% confidence interval **with and without the continuity correction** for the difference in improvement rate between the treatment and placebo groups.

5.11.2 Answer

Without continuity correction,

```
ptest<- prop.test(x=c(23,14),
          n=c(40,35),
          conf.level=0.99,
          correct=FALSE)

ptest
```

```
    2-sample test for equality of proportions without continuity correction

data:  c(23, 14) out of c(40, 35)
X-squared = 2.2871, df = 1, p-value = 0.1305
alternative hypothesis: two.sided
99 percent confidence interval:
 -0.1183113  0.4683113
sample estimates:
prop 1 prop 2
 0.575  0.400
```

With the continuity correction,

```
ptest<- prop.test(x=c(23,14),
          n=c(40,35),
          conf.level=0.99,
          correct=TRUE)

ptest
```

```
    2-sample test for equality of proportions with continuity correction

data:  c(23, 14) out of c(40, 35)
X-squared = 1.6405, df = 1, p-value = 0.2003
alternative hypothesis: two.sided
99 percent confidence interval:
 -0.145097  0.495097
sample estimates:
prop 1 prop 2
 0.575  0.400
```

The resulting 99% confidence interval with the continuity correction is is (-0.1451, 0.4951). 0 is part of the interval, which means it is a plausible value of the difference in improvement rates between the two groups. Hence, it is plausible that the long-run improvement rate due to the treatment does not differ from the placebo.