Sampling and Estimation

Sean Hellingman ©

Introduction to Statistical Data Analysis (ADSC1000)

shellingman@tru.ca

Fall 2024

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- Now we are going to begin making inferences about the target population.
- **Estimation** involves assessing the value of an unknown population parameter.
  - mean, proportion, or variance.
- We use the corresponding sample estimates which are random variables characterized by some sampling distribution.
- We can then apply probability theory to our estimates.

Estimates

- Focused on two kinds of estimates:
  1. **Point estimates** are a single number used to estimate the value of a population parameter.
  2. **Confidence interval estimates** provide a range of values between which the value of the population parameter is believed to be.
- Due to sampling error, it is highly unlikely that our point estimate is equal to the population parameter.
- Confidence intervals also specify a probability that the interval correctly estimates the unknown parameter.
- The size of the confidence interval depends on the size of the sampling error.

Notation

Notation	Description
$\bar{x}$	Sample mean
μ	Population mean
s^2	Sample variance
σ^2	Population variance
s	Sample standard deviation
σ	Population standard deviation
$\hat{p}$	Sample proportion
π	Population proportion

Point Estimators

- The **sample mean** $\bar{x}$ of the observations x_1, x_2, ... x_n is given by

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

- The **sample variance** s^2 is calculated by:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}. \tag{2}$$

- The **sample standard deviation** s is calculated by:

$$s = \sqrt{s^2}. \tag{3}$$

- The **population variance** σ^2 is calculated by:

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}, \tag{4}$$

Unbiased Estimators

- A statistic is a **biased** estimator if it overestimates or underestimates the population parameter.
- The **Bias** of the estimator $\hat{\theta}$ is calculated by:

$$Bias(\hat{\theta}, \theta) = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta. \tag{5}$$

- A point estimator is said to be **unbiased** if the expected value of the estimator is the theoretical parameter θ ($Bias(\hat{\theta}, \theta) = 0$).
- The point estimators that we have examined have been shown to be unbiased.

Interval Estimates

- An **interval estimate** provides a range for a population parameter based on a sample.
- Intervals can be more informative than simple point estimates ("how plausible?")
- We generally assign a range of confidence out of 100%

Probability Interval

- A **probability interval** is any interval $[A, B]$ such that the probability of falling between A and B is $1 - \alpha$.
  - General notation: $100(1 - \alpha)\%$
- You can view α as the risk of incorrectly concluding that the confidence interval contains the true mean.

Confidence Intervals

- A **confidence interval** is an interval estimate that specifies the probability that the interval contains the true population parameter.
- This probability is called the **level of confidence**, denoted $1 - \alpha$.
- Usually expressed as a percentage:
  - Generally: 90%, 95%, or 99%
  - If $\alpha = 0.05$ then the level of confidence is 95%.
- This interval **still may not include** the actual population parameter.
- As the confidence level increases, so does the length of the confidence interval.

Confidence Interval Types

- Different formulas for estimating confidence intervals have been developed for different situations.

- The formulas may depend on:
  - Parameter we are trying to estimate.
  - Assumptions about the population distribution.
  - Any prior knowledge about the population variability.

Confidence Interval for the Mean with Known Standard Deviation I

- Simplest confidence interval for the mean of a population is with a known standard deviation.
- In practice the population standard deviation **will not** be known.
- **100(1-α) confidence interval for μ:**

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \tag{6}$$

- Determine $z_{\alpha/2}$ in R: `qnorm(p=`α`/2, lower.tail=FALSE)`

Confidence Interval for the Mean with Known Standard Deviation II

- $z_{\alpha/2}$ is a **quantile** of the standard normal distribution.
- 95% confidence interval for μ:

$$\left(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \quad \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \tag{7}$$

- Can also use the functionality of the `z.test()` function in R.

Example 1

- Calculate the mean and a 95% confidence interval for the mean of the *mpg* variable in the *mtcars* data. Assume a known population standard deviation: $\sigma = 6$.
- Calculate a 99% confidence interval for the mean.
- *Hint: Rework the existing example code.*

Finite Population Correction

- If our sample n is larger than 5% of the population size N we need to use a finite population correction factor (FPC).
- Otherwise, the confidence interval will be too large.
- 100(1-α) confidence interval for μ with FPC:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \tag{8}$$

Example 2

- Assume the *iris* data to be our population of flowers.
- Set your seed to be 123 and draw a random sample of $n = 20$ flowers.
- Use the population correction factor (FPC) and population standard deviation to obtain a 95% confidence interval for the *petal length* of the flowers.

Confidence Interval for the Mean with Unknown Standard Deviation I

- In practice, the population standard deviation will not be known.
- In this case we now use the quantiles from a t-distribution.
- The t-distribution has a similar shape to the normal distribution but has an additional *degrees of freedom* (df) parameter.
- The t-distribution has a larger variability than the normal distribution.
    - Meaning that the confidence intervals are wider than if σ is known.

Confidence Interval for the Mean with Unknown Standard Deviation II

- Degrees of freedom $(df) = n$ - number of estimated parameters.
  - When estimating a confidence interval for the mean using a t-distribution: $df = n - 1$.
- In this case we now use the quantiles from a t-distribution.

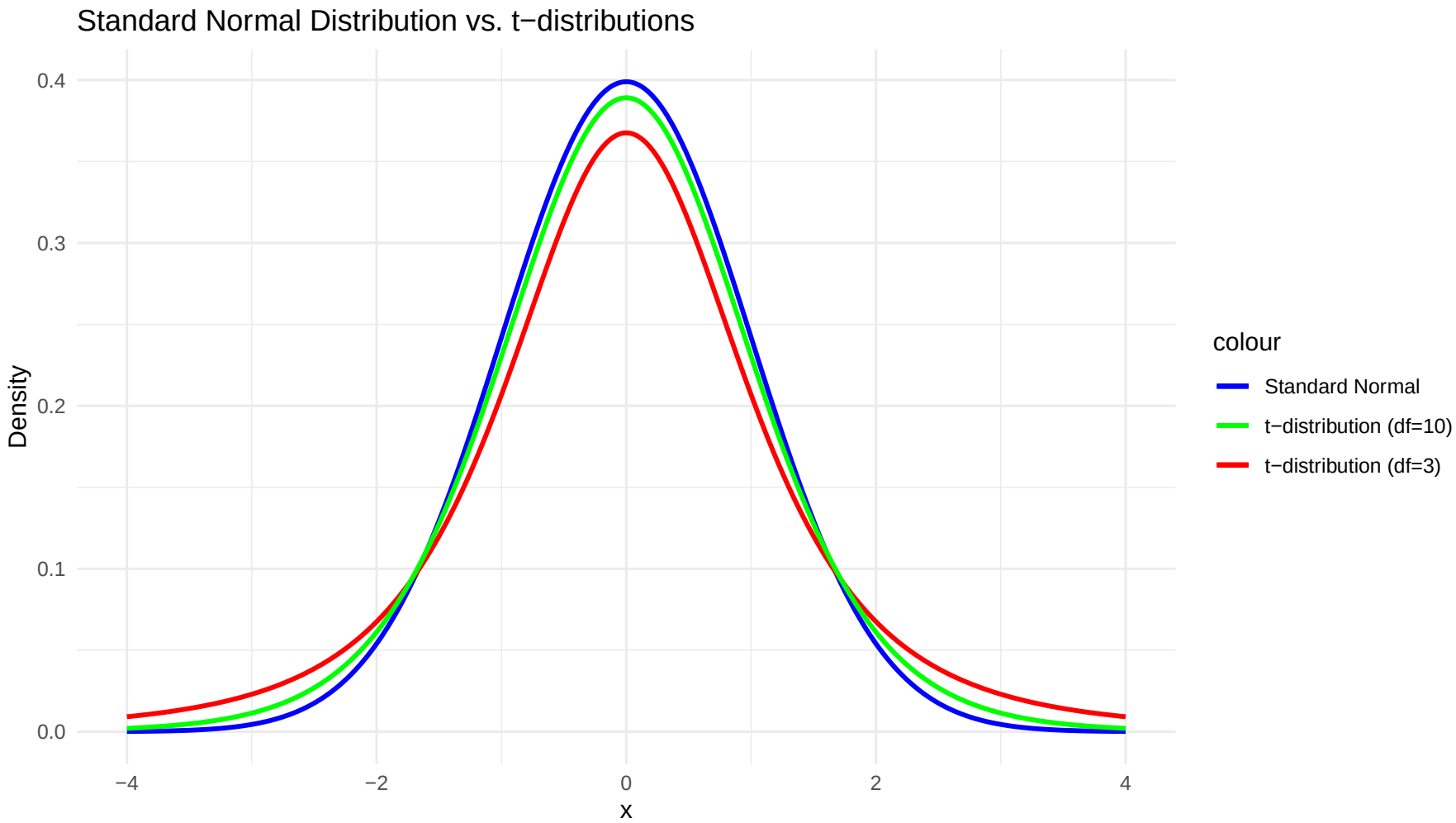
- **100(1-α) confidence interval for μ:**

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} \tag{9}$$

- To find $t_{\alpha/2,n-1}$ in R: `qt(p=`α`/2, df=n-1, lower.tail=FALSE)`.

- May also use the `t.test(x,conf.level = 0.95)` function in R.

Distribution Plots



Example 3

- Assuming a sufficiently large population, obtain a 95% confidence interval for the petal length of the flowers in the *iris* data.
    - *Hint: We do not know the population standard deviation.*

Sample Proportions

- Sample proportions are a natural way to examine categorical data.
- The numerical summaries account for the individual frequencies and the relative frequencies.
  - What brand of laptops students own.
  - Own a laptop or not.
- A sample proportion reflects how many observations fall into a certain category relative to the other possible categories.
- An unbiased estimator of a population proportion π is the statistic $\hat{p} = x/n$ (sample proportion).

Confidence Intervals for Sample Proportions

- We will use quantiles of the standard normal distribution $z_{\alpha/2}$.
- **100(1-α) confidence interval for π:**

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{10}$$

- Because $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the standard error of the sampling distribution of the proportion.
- May use the `prop.test(x,n,conf.level=0.95)` function but the results will be slightly different.

Example 4

- Assume we have a sample from a sufficiently large population of the favourite colour of twelve ($n = 12$) students: red, red, blue, green, green, orange, red, orange, yellow, green, red, orange.
    - Calculate a confidence 95% confidence interval for the proportion of students that like red.

Measures of Variability

- Understanding variability is very important in the implementation of statistical theory to decision-making process.
- We have point estimates for variability (standard deviation & variance).
- It might be useful to calculate confidence intervals for the variability.
- For example: The estimated variance in food prices.

Chi-square (χ^2) Distribution

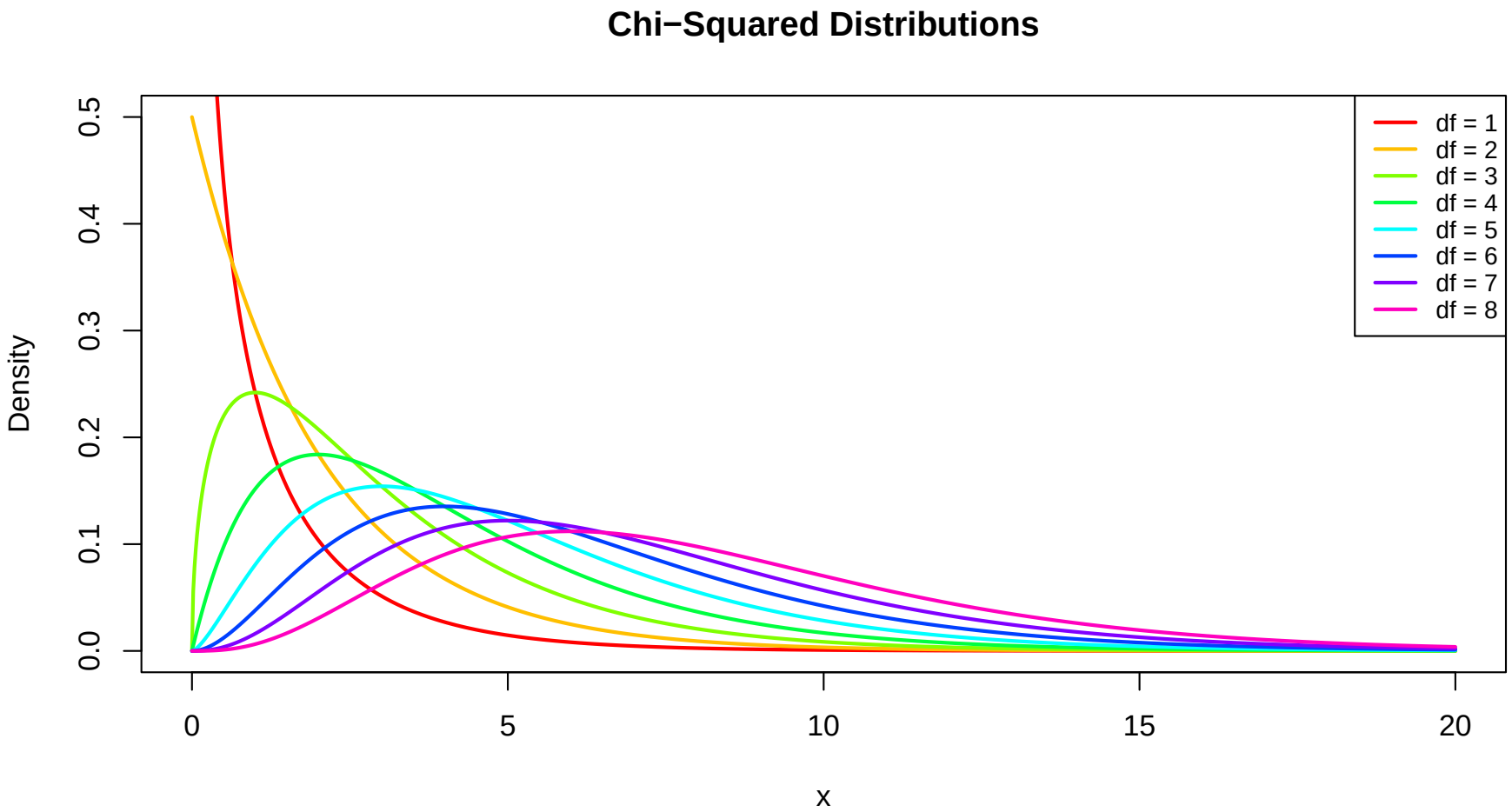
- The sampling distribution of s^2 is **not** normally distributed.
- Instead, the chi-square (χ^2) distribution is used.
- The chi-square (χ^2) distribution also relies on degrees of freedom (df) and it is **not** symmetric.
- 100(1-α) confidence interval for the variance σ^2:

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right] \tag{11}$$

Chi-square (χ^2) Distribution in R

- Use the `pchisq(q, df, ncp = 0, lower.tail = TRUE)` function to get the chi-square quantiles.
    - Not symmetric so you need to change to `lower.tail = FALSE` for the upper limit.
- We can also use the `VarCI(x,method = "classic",conf.level = 0.95)` function from the *DescTools* package.

Chi-square Plots



Example 5

- Assuming a sufficiently large population, obtain a 95% confidence interval for the **variance** of the petal length of the flowers in the *iris* data.

Applications of Confidence Intervals

- Confidence intervals may be used to help with the decision-making process.

- Example: Government regulator mandates that water tanks must have 64 litres.
  - Take a sample of $n = 30$ tanks and they have an average fill of $\bar{x} = 63.82$
  - Is the equipment under filling the tanks?
  - 95% confidence interval: $(63.43, 64.21)$

- With 95% certainty we can say that the equipment fills are not different than 64 litres.

Sample Sizes for Mean Confidence Intervals

- We can see using algebra that we can determine the size of n for any given half interval length E.

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2 \tag{12}$$

- *Be sure to round up to the nearest whole number.*

- We can use R to obtain n using the *EnvStats* package:
  - `ciNormN(half.width, sigma.hat = 1, conf.level = 0.95)`

- **This only works with random samples, we need other methods for more robust experimental designs.**

Example 6

- How many observations n, drawn from a random sample will we need to obtain a half confidence interval of 0.25 from a sample with $\sigma = 3.3$?

Differences Between Means, Independent Samples I

- A confidence interval for independent samples with **unequal** variances is:

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

- where:

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

Differences Between Means, Independent Samples II

- A confidence interval for independent samples with **equal** variances is:

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

- where:

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

Differences Between Means, Paired Samples

- A confidence interval for the difference in paired samples:

$$\overline{D} \pm \left(t_{n-1,\alpha/2}\right) s_D / \sqrt{n}$$

- where:

$$s_D = \sqrt{\frac{\sum_{i=1}^{n} \left(D_i - \overline{D}\right)^2}{n \ - \ 1}}$$

- $\bar{D}$ is a point estimate for the mean difference between the populations.

Differences Between Proportions

- A confidence interval for differences between proportions of two populations:

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

Summary of Confidence Interval Formulas

Type of Confidence Interval	Formula
Mean, standard deviation known	$\bar{x} \pm z_{\alpha/2}(\alpha/\sqrt{n})$
Mean, standard deviation unknown	$\bar{x} \pm t_{\alpha/2,n-1}(s/\sqrt{n})$
Proportion	$\hat{p} \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}\,(1-\hat{p})}{n}}$
Population total	$N\bar{x} \pm t_{\alpha/2,\,n-1}N\dfrac{s}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$
Difference between means, independent samples, equal variances	$\bar{x}_1 - \bar{x}_2 \pm (t_{\alpha/2,\,n_1+n_2-2})s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}$
	$s_p = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$
Difference between means, independent samples, unequal variances	$\bar{x}_1 - \bar{x}_2 \pm (t_{\alpha/2,\,df^*})\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}$
	$df^* = \dfrac{\left[\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}\right]^2}{\left[\dfrac{(s_1^2/n_1)^2}{n_1-1}\right]+\left[\dfrac{(s_2^2/n_2)^2}{n_2-1}\right]}$
Difference between means, paired samples	$\bar{D} \pm (t_{n-1,\,\alpha/2})s_D/\sqrt{n}$
Differences between proportions	$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
Variance	$\left[\dfrac{(n-1)s^2}{\chi^2_{n-1,\,\alpha/2}}, \dfrac{(n-1)s^2}{\chi^2_{n-1,\,1-\alpha/2}}\right]$

Figure: *Source:* (1)

Using R

- Confidence interval for the difference of two means (*DescTools*) package:
  - `MeanDiffCI(x, y, method = "classic", conf.level = 0.95, paired = FALSE)`
  - *Set* `paired = TRUE` *for paired samples*

- Confidence interval for the difference of two proportions using the functionality of the prop.test() function:
  - `prop.test(x=c(`x_1`,`x_2`), n=c(`n_1`,`n_2`), correct=FALSE)`

Example 7

- Calculate a 90% confidence interval for the differences in petal lengths between the *setosa* and *versicolor* flowers from the *iris* dataset.

Example 8

- In Kamloops, 25% of ($n_1 = 52$) people interviewed said they prefer winter. In Kelowna, 35% of ($n_2 = 60$) people interviewed said they prefer winter. Construct a 95% confidence interval for the differences in proportions.

Exercise 1

- Given the following situations, which method should you select to obtain a confidence interval:
  - Confidence interval for the variance of annual incomes in Canada.
  - Confidence interval for the mean of a sample of bird weights with a population standard deviation $\sigma = 8.23$.
  - Confidence interval for the proportion of students that have Apple laptops.
  - Confidence interval for the mean generated by a Sample with a mean of $\bar{x} = 3.4$ and sample standard deviation $s = 4.4$.
  - Confidence interval for the mean heights of ADSC1000 students obtained from a sample $n = 10$. *Hint:* $N = 35$.

Exercise 2

- Assuming a sufficiently large population, obtain a 95% confidence interval for the sepal length of the flowers in the iris data.
- Assuming a sufficiently large population, obtain a 95% confidence interval for the **variance** of the sepal length of the flowers in the *iris* data.

Exercise 3

- Assume we have a sample from a sufficiently large population of the favourite colour of twelve ($n = 12$) students: red, red, blue, green, green, orange, red, orange, yellow, green, red, orange.
    - Calculate a confidence 95% confidence interval for the proportion of students that like orange.

Exercise 4

- How many observations n, drawn from a random sample will we need to obtain a half confidence interval of 0.66 from a sample with $\sigma = 2.35$?

Exercise 4

- Before taking additional lessons, students' mathematics grades were as follows: 65, 55, 78, 80, 55, 45, 65, 66, 76, 45.
- After additional training the **same** students' mathematics grades were as follows: 75, 64, 81, 90, 62, 56, 71, 70, 79, 55.
- Construct a 95% confidence interval for the differences in mathematics scores.

References & Resources

1. Evans, J. R., Olson, D. L., & Olson, D. L. (2007). *Statistics, data analysis, and decision modeling.* Upper Saddle River, NJ: Pearson/Prentice Hall.

2. Devore, J. L., Berk, K. N., & Carlton, M. A. (2012). *Modern mathematical statistics with applications (Second Edition).* New York: Springer.

- BSDA
- VarCI()
- Chi-squared Distribution
- The Student t Distribution
- ciNormN()