STAT 332 - Sampling and Experimental Design

Cameron Roopnarine

Last updated: February 14, 2021

Contents

Chapter 1

Assignment 1

1.1 Lecture 1.00 - PPDAC + Example

PPDAC: Problem, Plan, Data, Analysis, Conclusion.

- Problem: Define the problem.
  - **Target population** (TP): The group of units referred to in the problem step.
  - **Response**: The answer provided by the TP to the problem.
  - **Attribute**: Statistic of the response.

    EXAMPLE 1.1.1

    What is the average grade of the students in STAT 101?
    - Target population: All STAT 101 students
    - Response: Grade of a STAT 101 student.
    - Attribute: Average grade.
- Plan:
  - **Study population** (SP): The set of units you *can* study

    EXAMPLE 1.1.2

    Does a drug reduce hair loss?
    - Target population: People.
    - Study population: Mice.
  - **Sample**: A subset of the study population.
- Analysis: We analyze the data.
- Conclusion: Refers back to the problem. We also note some common *errors*.
  - **Study error**: The attribute of the population the target population differs from the parameter of the study population.

    EXAMPLE 1.1.3

    Mathematically we can write it down as $a(\text{TP}) - \mu$, however this error is qualitative. Therefore, we cannot actually calculate it.

- **Sample error**: The parameter differs from the sample statistic (estimate).

EXAMPLE 1.1.4

Mathematically we can write it down as $\mu - \bar{x}$, however this error is qualitative. Therefore, we cannot actually calculate it.

- **Measurement error**: The difference between what *we want* to calculate and what *we do* calculate.

1.2 Lecture 2.00 - Models, Model 1

DEFINITION 1.2.1: Model

A **model** relates a parameter to a response.

DEFINITION 1.2.2: Model 1

Model 1 is defined as

$$Y_j = \mu + R_j \quad (R_j \sim \mathcal{N}(0, \sigma^2))$$

where

- Y_j: random parameter that is the response of unit j.
- μ: non-random unknown parameter that is the study population mean.
- R_j: the distribution of responses about μ.

REMARK 1.2.3

- R_j's are always independent.
- **Gauss**' Theorem: Any linear combination of normal random variables is normal.
- $Y_j \sim \mathcal{N}(\mu, \sigma^2)$ since

$$\mathbb{E}[Y_j] = \mathbb{E}[\mu + R_j] = \mathbb{E}[\mu] + \mathbb{E}[R_j] = \mu + 0 = \mu$$
$$\mathbb{V}(Y_j) = \mathbb{V}(\mu + R_j) = \mathbb{V}(R_j) = \sigma^2$$

EXAMPLE 1.2.4

Average grade of STAT 101 students.

$$Y_j = \mu + R_j \quad (R_j \sim \mathcal{N}(0, \sigma^2))$$

1.3 Lecture 3.00 - Independent Groups

- Dependent: we randomly select one group and we find a match, having the same explanatory variates, for each unit of the first group. For example, twins, reusing members of a group, or matching.
- Independent: are formed when we select units at random from mutually exclusive groups. For example, broken parts and non-broken parts.

1.4 Lecture 4.00 - Models 2A and 2B

DEFINITION 1.4.1: Model 2A

Model 2A is used when we assume the groups have the same standard deviation and is defined as

$$Y_{ij} = \mu_i + R_{ij} \quad (R_{ij} \sim \mathcal{N}(0, \sigma^2))$$

where

- Y_{ij}: response of unit j in group i.
- μ_i: mean for group i.
- R_{ij}: the distribution of responses about μ_i.

DEFINITION 1.4.2: Model 2B

Model 2B is used when $\sigma_1 \neq \sigma_2$ and is defined as

$$Y_{ij} = \mu_i + R_{ij} \quad (R_{ij} \sim \mathcal{N}(0, \sigma_i^2))$$

1.5 Lecture 5.00 - Model 3

We subtract Model 2A from Model 2B to model a difference between two groups, and we get *Model 3*.

$$\begin{array}{rcccccc} & Y_{1j} & = & \mu_1 & + & R_{1j} \\ - & Y_{2j} & = & \mu_2 & + & R_{2j} \\ \hline & Y_{1j} - Y_{2j} & = & \mu_1 - \mu_2 & + & R_{1j} - R_{2j} \end{array}$$

Let

- $Y_{1j} - Y_{2j} = Y_{dj}$
- $\mu_1 - \mu_2 = \mu_d$
- $R_{1j} - R_{2j} = R_{dj}$

DEFINITION 1.5.1: Model 3

Model 3 is defined as

$$Y_{dj} = \mu_d + R_{dj} \quad (R_{dj} \sim \mathcal{N}(0, \sigma_d^2))$$

EXAMPLE 1.5.2: Model 3

Heart Rate Before Exercise	Heart Rate After Exercise	d
70	80	10
80	100	20
90	90	0

We could use Model 3.

1.6 Lecture 6.00 - Model 4

Suppose $Y \sim \text{Binomial}(n, p)$; that is, we have n outcomes where each outcome is binary.

$$\mathbb{E}[Y] = np$$

$$\mathbb{V}(Y) = np(1-p)$$

By the Central Limit Theorem, $Y \dot{\sim} \mathcal{N}(np, np(1-p))$. The proportion is

$$\frac{Y}{n} \dot{\sim} \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Let's find the expected value and variance of Y/n.

$$\mathbb{E}\left[\frac{Y}{n}\right] = \frac{\mathbb{E}[Y]}{n} = \frac{np}{n} = p$$

$$\mathbb{V}\left(\frac{Y}{n}\right) = \frac{\mathbb{V}(Y)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

DEFINITION 1.6.1: Model 4

Model 4 is defined as

$$\frac{Y}{n} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

1.7 Lecture 7.00 - MLE

- What is MLE? It connects the population parameter θ to your sample statistic $\hat{\theta}$.
- How? It chooses the most probable value of θ given our data $y_1, \dots, y_n$.

Process:

(1) Define the **likelihood function**.

$$L = f(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n)$$

We assume $Y_i \perp Y_j$ for all $i \neq j$. Therefore,

$$L = f(Y_1 = y_1) f(Y_2 = y_2) \cdots f(Y_n = y_n)$$

(2) Define the **log-likelihood function** and use log rules to clean it up!

(3) Find $\frac{\partial \ell}{\partial \theta}$.

(4) Set $\frac{\partial \ell}{\partial \theta} = 0$, put hat on all θ's.

(5) Solve for $\hat{\theta}$.

EXAMPLE 1.7.1

Let $Y_{ij} = \mu_i + R_{ij}$ where $R_{ij} \sim \mathcal{N}(0, \sigma^2)$.

$$\begin{aligned} L &= f(Y_{11} = y_{11}, \dots, Y_{2n_2} = y_{2n_2}) \\ &= \prod_{j=1}^{n_1} f(y_{1j}) \prod_{j=1}^{n_2} f(y_{2j}) \\ &= \prod_{j=1}^{n_1} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_{1j}-\mu_1)^2}{2\sigma^2}\right\} \prod_{j=1}^{n_2} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_{2j}-\mu_2)^2}{2\sigma^2}\right\} \end{aligned}$$

Let $n_1 + n_2 = n$, then

$$L = (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{\sum_{j=1}^{n_1}(y_{1j}-\mu_1)^2}{2\sigma^2}\right\} \exp\left\{-\frac{\sum_{j=1}^{n_2}(y_{2j}-\mu_2)^2}{2\sigma^2}\right\}$$

The log-likelihood is given by

$$\ell = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{\sum_{j=1}^{n_1}(y_{1j}-\mu_1)^2}{2\sigma^2} - -\frac{\sum_{j=1}^{n_2}(y_{2j}-\mu_2)^2}{2\sigma^2}$$

Now,

$$\frac{\partial \ell}{\partial \hat{\mu}_1} = 0 + 0 - \frac{\sum_{j=1}^{n_1} 2(y_{1j}-\hat{\mu})(-1)}{2\hat{\sigma}^2} + 0 = 0$$

Hence,

$$0 = \sum_{j=1}^{n_1}(y_{1j}-\hat{\mu}) \implies \sum_{j=1}^{n_1} y_{1j} = \sum_{j=1}^{n_1}\hat{\mu}$$

Note that

$$\sum_{j=1}^{n_1} y_{1j} = \frac{n_1}{n_1}\sum_{j=1}^{n_1} y_{1j} = n_1\bar{y}_{1+}$$

Therefore,

$$n_1\bar{y}_{1+} = n_1\hat{\mu} \implies \bar{y}_{1+} = \hat{\mu}_1$$

By symmetry,

$$\bar{y}_{2+} = \hat{\mu}_2$$

The second partial is given by

$$\frac{\partial \ell}{\partial \sigma} = 0 + \frac{(-n)}{\hat{\sigma}} - \frac{\sum_{j=1}^{n_1}(y_{1j}-\hat{\mu}_1)^2}{2}(-2\hat{\sigma}^{-3}) - -\frac{\sum_{j=1}^{n_2}(y_{2j}-\hat{\mu}_2)^2}{2}(-2\hat{\sigma}^{-3})$$

Multiply both sizes by $\hat{\sigma}^3$, yields

$$0 = -n\hat{\sigma}^2 + \sum_{j=1}^{n_1}(y_{1j}-\hat{\mu}_1)^2 + \sum_{j=1}^{n_2}(y_{2j}-\hat{\mu}_2)^2$$

Divide both sizes by n and rearrange to get

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{n_1}(y_{1j}-\hat{\mu}_1)^2 + \sum_{j=1}^{n_2}(y_{2j}-\hat{\mu}_2)}{n}$$

Recall that

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

$$s_1^2 = \sum_{j=1}^{n_1}\frac{(y_{1j}-\bar{y}_{1+})^2}{n_1-1}$$

$$s_2^2 = \sum_{j=1}^{n_2}\frac{(y_{2j}-\bar{y}_{2+})^2}{n_2-1}$$

Therefore,

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

1.8 Lecture 8.00 - LS

- What is LS? Another technique to find $\hat{\theta}$.
- How? It minimizes the "residuals."
- Models:

$$\text{Response} = \text{Deterministic Part} + \text{Random Part}$$

$$Y = f(\theta) + R$$

Let $y_1, y_2, \ldots, y_n$ be realizations of Y. Let $\hat{y}_i = f(\hat{\theta})$, where $f(\hat{\theta})$ is simply $f(\theta)$ with θ replaced by $\hat{\theta}$. We call $\hat{y}_i$ our "prediction."

DEFINITION 1.8.1: Residual

A **residual** is

$$r_i = y_i - f(\hat{\theta}) = y_i - \hat{y}_i$$

Process:

(1) Define the W function, $W = \sum r^2$.

(2) Calculate $\frac{\partial W}{\partial \theta}$ for all non-σ parameters

(3) Set $\frac{\partial W}{\partial \theta} = 0$ and replace θ by $\hat{\theta}$.

(4) Solve for $\hat{\theta}$.

1.9 Lecture 9.00 - LS Example

Let's determine the LS of Model 2A.

$$Y_{ij} = \mu_i + R_{ij}$$

Also, let $n = n_1 + n_2$.

$$\begin{aligned}
W = \sum_{ij} r_{ij}^2 &= \sum_{ij} (y_{ij} - \hat{\mu}_i)^2 \\
&= \sum_{j=1}^{n} \sum_{i=1}^{2} (y_{ij} - \hat{\mu}_i)^2 \\
&= \sum_{j=1}^{n_1} (y_{1j} - \hat{\mu}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \hat{\mu}_2)^2
\end{aligned}$$

$$\begin{aligned}
0 &= \frac{\partial W}{\partial \hat{\mu}_1} \\
&= \sum_{j=1}^{n_1} (y_{1j} - \hat{\mu}_1)(-2) \\
&= \frac{n_1}{n_1} \sum_{j=1}^{n_1} y_{ij} - \sum_{j=1}^{n_1} \hat{\mu}_1 \\
&= n_1 \bar{y}_{1+} - n\hat{\mu}_1
\end{aligned}$$

Therefore, $\hat{\mu}_1 = \bar{y}_{1+}$ and by symmetry $\hat{\mu}_2 = \bar{y}_{2+}$.

REMARK 1.9.1

For LS, $\hat{\sigma}^2$ is always of the form

$$\hat{\sigma}^2 = \frac{W}{n-q+c}$$

where

- $n =$ number of units
- $q =$ number of non-σ parameters
- $c =$ number of constraints

Note that $\hat{\sigma}^2 = s_p^2$.

REMARK 1.9.2: MLE versus LS

- LS is from 1860's. Unbiased provided R_j is normal.
- MLE is a recent technique and it is much more flexible since it does not require R_j to be normal.
- Minimum? You need to calculate the second derivative, but we're too lazy and unrigorous in this course. No thanks.

1.10 Lecture 10.00 - Estimators

Our sample data is $y_1, \dots, y_n$. It is non-random and is a realization of a random variable $Y_1, \dots, Y_n$. A statistic is a function of the sample data; $\hat{\theta}$. It is non-random, but if $y_1, \dots, y_n$ changes, then so does $\hat{\theta}$. For that reason, you can think of $\hat{\theta}$ as the realization of a random variable $\tilde{\theta}$, called an estimator. To move from $\hat{\theta}$ to $\tilde{\theta}$ we capitalize our Y's.

EXAMPLE 1.10.1

Model 2A: $\underbrace{\hat{\mu}_1 = \bar{y}_{1+}}_{\text{STATISTIC}} \to \underbrace{\tilde{\mu}_1 = \bar{Y}_{1+}}_{\text{ESTIMATOR}}$

THEOREM 1.10.2: Gauss' Theorem

Any linear combination of normal random variables is still normal.

EXAMPLE 1.10.3

Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ be independent random variables and $a, b, c \in \mathbf{R}$, then

$$L = aX + bY + c \sim \mathcal{N}(\mathbb{E}[L], \mathbb{V}(L))$$

THEOREM 1.10.4: Central Limit Theorem (CLT)

Let $Y_1, \dots, Y_n$ be a i.i.d. random variables with $\mathbb{E}[Y_i] = \mu$, $\mathbb{V}(Y_i) = \sigma^2 < \infty$, then

$$\bar{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

1.11 Lecture 11.00 - Estimators Example

EXAMPLE 1.11.1

Model 2A: $Y_{ij} = \mu_i + R_{ij}$ where $R_{ij} \sim \mathcal{N}(0, \sigma^2)$. What is the distribution of $\tilde{\mu}$?
Solution. Using LS or MLE we obtain

$$\hat{\mu} = \bar{y}_{1+}$$

Or corresponding estimator is

$$\tilde{\mu}_1 = \bar{Y}_{1+} = \frac{\sum_{j=1}^{n_1} Y_{1j}}{n_1}$$

and by Gauss it is normal!

$$\mathbb{E}[\tilde{\mu}_1] = \mathbb{E}\left[\frac{\sum_{j=1}^{n_1} Y_{1j}}{n_1}\right] = \frac{\sum_{j=1}^{n_1} \mathbb{E}[Y_{1j}]}{n_1} = \frac{\sum_{j=1}^{n_1} \mathbb{E}[\mu + R_{1j}]}{n_1} = \frac{\sum_{j=1}^{n_1} \mu + \mathbb{E}[R_{1j}]}{n_1} = \mu_1$$

DEFINITION 1.11.2: Unbiased estimator

If $\mathbb{E}[\tilde{\theta}] = \theta$, we say $\tilde{\theta}$ is an **unbiased estimator** of θ.

$$\begin{aligned}
\mathbb{V}(\tilde{\mu}_1) &= \mathbb{V}(\bar{Y}_{1+}) \\
&= \mathbb{V}\left(\frac{\sum_{j=1}^{n_1} Y_{1j}}{n_1}\right) \\
&= \frac{1}{n_1^2}\mathbb{V}\left(\sum_{j=1}^{n_1} Y_{1j}\right) \\
&= \frac{1}{n_1^2}\sum_{j=1}^{n_1} \mathbb{V}(Y_{ij}) && \text{since } Y_{1j} \perp Y_{1i} \\
&= \frac{1}{n_1^2}\sum_{j=1}^{n_1} \mathbb{V}(\mu_1 + R_{1j}) \\
&= \frac{1}{n_1^2}\sum_{j=1}^{n_1} \mathbb{V}(Y_{1j}) \\
&= \frac{1}{n_1^2}(n_1\sigma^2) \\
&= \frac{\sigma^2}{n_1}
\end{aligned}$$

Therefore,

$$\tilde{\mu}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma^2}{n_1}\right)$$

and by symmetry

$$\tilde{\mu}_2 \sim \mathcal{N}\left(\mu_2, \frac{\sigma^2}{n_2}\right)$$

1.12 Lecture 12.00 - Sigma

THEOREM 1.12.1

Let $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi^2(1)$

THEOREM 1.12.2

Let $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$ be independent, then

$$X + Y \sim \chi^2(n+m)$$

THEOREM 1.12.3

Let $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2(m)$, then

$$\frac{Z}{\sqrt{X/m}} \sim t(m)$$

THEOREM 1.12.4

Let $Y = \dfrac{(n-q+c)\tilde{\sigma}^2}{\sigma^2}$, then $Y \sim \chi^2(n-q+c)$.

1.13 Lecture 13.00 - Sigma Example

EXAMPLE 1.13.1

Model 1: $Y_j = \mu + R_j$ where $R_j \sim \mathcal{N}(0, \sigma^2)$. What is the distribution of $\dfrac{\tilde{\mu} - \mu}{\tilde{\sigma}/\sqrt{n}}$?

Solution. We know by LS or MLE that $\hat{\mu} = \bar{y}_+$, therefore $\tilde{\mu} = \bar{Y}_+$. We know $\tilde{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. We standardize

$$Z = \frac{\tilde{\mu} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

By Theorem 1.12.4, we know

$$X = \frac{(n-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2(n-1)$$

By Theorem 1.12.3,

$$\frac{Z}{\sqrt{X/(n-1)}} = \frac{\frac{\tilde{\mu}-\mu}{\sigma/\sqrt{n}}}{\frac{(n-1)\tilde{\sigma}^2}{\sigma^2}} = \frac{\tilde{\mu} - \mu}{\tilde{\sigma}/\sqrt{n}} \sim t(n-1)$$

REMARK 1.13.2

Recall that

$$\frac{\tilde{\mu} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

By replacing σ by $\tilde{\sigma}$, we end up using a t-distribution instead of a normal distribution.

1.14 Lecture 14.00 - CI

We assume our estimator is

$$\tilde{\theta} \sim \mathcal{N}(0, \mathbb{V}(\tilde{\theta}))$$

The CI:

$$\theta : \text{EST} \pm c\,\text{SE} = \hat{\theta} \pm c\sqrt{\mathbb{V}(\tilde{\theta})}$$

If we don't know σ, we replace it by $\hat{\sigma}$ and obtain

$$\theta : \hat{\theta} \pm c\sqrt{\widehat{\mathbb{V}(\tilde{\theta})}}$$

EXAMPLE 1.14.1

Model 1: $Y_j = \mu + R_j$ where $R_j \sim \mathcal{N}(0, \sigma^2)$. By LS we know $\hat{\mu} = \bar{y}_+$. The estimator is $\tilde{\mu} = \bar{Y}_+$ with distribution

$$\tilde{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Our CI:

$$\mu : \text{EST} \pm c\,\text{SE} = \hat{\mu} \pm c\frac{\sigma}{\sqrt{n}} = \bar{y}_+ \pm c\frac{\sigma}{\sqrt{n}} \quad (c \sim \mathcal{N}(0,1))$$

$$\mu : \bar{y}_+ \pm c\frac{s}{\sqrt{n}} \sim t(n-1)$$

Recall: $s = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$.

EXAMPLE 1.14.2

Model 2A: $Y_{ij} = \mu_i + R_{ij}$ where $R_{ij} \sim \mathcal{N}(0, \sigma^2)$. By LS, $\hat{\mu}_1 = \bar{y}_{1+}$ and $\hat{\mu}_2 = \bar{y}_{2+}$. The estimators $\tilde{\mu}_1 = \bar{Y}_{1+}$ and $\tilde{\mu}_2 = \bar{Y}_{2+}$. The distributions are

$$\tilde{\mu}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma^2}{n_1}\right)$$

$$\tilde{\mu}_2 \sim \mathcal{N}\left(\mu_2, \frac{\sigma^2}{n_2}\right)$$

$$\tilde{\mu}_1 - \tilde{\mu}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Our CI:

$$\mu_1 - \mu_2 : \text{EST} \pm c\,\text{SE} = \hat{\mu}_1 - \hat{\mu}_2 \pm c\,\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (c \sim \mathcal{N}(0,1))$$

$$\mu_1 - \mu_2 : \text{EST} \pm c\,\text{SE} = \hat{\mu}_1 - \hat{\mu}_2 \pm c\,s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (c \sim t(n_1 + n_2 - 2))$$

EXAMPLE 1.14.3

Model 2B: $Y_{ij} = \mu_i = R_{ij}$ where $R_{ij} \sim \mathcal{N}(0, \sigma_i^2)$.

$$\tilde{\mu}_1 - \tilde{\mu}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Our CI:

$$\hat{\mu}_1 - \hat{\mu}_2 \pm c\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (c \sim \mathcal{N}(0,1))$$

$$\hat{\mu}_1 - \hat{\mu}_2 \pm c\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (c \sim t(n_1 + n_2 - 2))$$

EXAMPLE 1.14.4

Model 3: $Y_{dj} = \mu_d + R_{dj}$ where $R_{dj} \sim \mathcal{N}(0, \sigma_d^2)$, which is the same as Model 1.

$$\mu_d : \bar{y}_{d+} \pm c\frac{\sigma_d}{\sqrt{n_d}} \quad (c \sim \mathcal{N}(0,1))$$

$$\mu_d : \bar{y}_{d+} \pm c\frac{s_d}{\sqrt{n_d}} \sim t(n_d - 1)$$

EXAMPLE 1.14.5

Model 4:

$$\tilde{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Our CI:

$$\hat{p} \pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (c \sim \mathcal{N}(0,1))$$

Table 1.1: Confidence Intervals

#	Model	CI	df
1	$Y_i = \mu + R_i$ $R_i \sim \mathcal{N}(0,\sigma^2)$	$\bar{y} \pm t^* \frac{s}{\sqrt{n}}$	$n-1$
2A	$Y_{ij} = \mu_i + R_{ij}$ $R_{ij} \sim \mathcal{N}(0,\sigma^2)$	$\bar{y}_{1+} \pm t^* \frac{s_1}{\sqrt{n_1}}$	$n_1 - 1$
		$\bar{y}_{1+} - \bar{y}_{2+} \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$n_1 + n_2 - 2$
2B	$Y_{ij} = \mu_i + R_{ij}$ $R_{ij} \sim \mathcal{N}(0,\sigma_i^2)$	$\bar{y}_{1+} \pm t^* \frac{s_1}{\sqrt{n_1}}$	$n_1 - 1$
		$\bar{y}_{1+} - \bar{y}_{2+} \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$n_1 + n_2 - 2$
3	$Y_{dj} = \mu_d + R_{dj}$ $R_{dj} \sim \mathcal{N}(0,\sigma_d^2)$	$\bar{y}_d \pm t^* \frac{s_d}{\sqrt{n_d}}$	$n_d - 1$
4	$\frac{Y}{n} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$	$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\mathcal{N}(0,1)$

1.15 Lecture 15.00 - CI Examples

EXAMPLE 1.15.1: Model 1

- **Problem**: What is the mean calculus grade of students in STAT 332?
- **Plan**: We randomly select 5 students from the class.
- **Data**: 65, 70, 80, 85, 75
- **Analysis**: Build a 95% confidence interval for the mean grade.

$$\mu : \bar{y} \pm t^* \frac{s}{\sqrt{n}}$$

```
dat <- c(65, 70, 80, 85, 75)
y.bar <- mean(dat)
s <- sd(dat)
n <- length(dat)
df <- length(dat) - 1
t <- qt(0.975, df)
left <- y.bar - t * s / sqrt(n)
right <- y.bar + t * s / sqrt(n)
```

The 95% confidence interval is: $(65.18, 84.82)$. We are 95% confident that the mean grade is in the interval. What we mean is that if we drew 100 samples, and built 100 confidence intervals for these samples, then we would expect to find μ in 95 of these intervals that we created. This is not a probability because at the end of the day, you estimated your data for y.

EXAMPLE 1.15.2: Model 2A

- **Problem**: In grade 9, there is a standardized test in Ontario. We wish to compare the mean performance of girls to boys.
- **Plan**: They collect data from a class of 30 students; 15 boys and girls. Their response is their grade on the standardized test. If necessary, assume the variances of the two groups are the same.
- **Data**:
  - Boys: 39, 42, 47, 50, 52, 52, 54, 55, 55, 56, 56, 56, 58, 60, 62
  - Girls: 44, 45, 48, 50, 51, 52, 53, 53, 57, 58, 59, 60, 62, 63, 64
- **Analysis**: Build a 95% confidence interval for the mean difference in grades.

```
boys <- c(39, 42, 47, 50, 52, 52, 54, 55, 55, 56, 56, 56, 58, 60, 62)
girls <- c(44, 45, 48, 50, 51, 52, 53, 53, 57, 58, 59, 60, 62, 63, 64)
y_b.bar <- mean(boys)
y_g.bar <- mean(girls)
s_b.sq <- var(boys)
s_g.sq <- var(girls)
n_b <- length(boys)
n_g <- length(girls)
s_p.sq <-
((n_g - 1) * s_g.sq + (n_b - 1) * s_b.sq) / (n_g + n_b - 2)
df <- n_g + n_b - 2
t <- qt(0.975, df)
left <- (y_b.bar - y_g.bar) - t * sqrt(s_p.sq * (1 / n_g + 1 / n_b))
right <- (y_b.bar - y_g.bar) + t * sqrt(s_p.sq * (1 / n_g + 1 / n_b))
```

- $\bar{y}_{b+} = 52.9$
- $\bar{y}_{g+} = 54.6$
- $s_b^2 = 39.6$
- $s_g^2 = 41$
- $s_p^2 = 40.3$

- $t^* = 2.048$

The 95% confidence interval for the mean difference grade is $(-6.4, 3.1)$. Is there a difference between male and female grades? 0 is in this interval, so we conclude there is no difference between male and female grades.

EXAMPLE 1.15.3: Model 3

- **Problem**: In grade 9 there is a standardized test in Ontario. We wish to compare the mean performance of girls to boys.
- **Plan**: They collect data from a class of 30 students; 15 boys and 15 girls. Each girl is selected so that she was born in the same month as a boy in the class. The response is their grade on the standardized test. If necessary assume the variances of the two groups are different.
- **Data**:
  - Boys: 39, 42, 47, 50, 52, 52, 54, 55, 55, 56, 56, 56, 58, 60, 62
  - Girls: 44, 45, 48, 50, 51, 52, 53, 53, 57, 58, 59, 60, 62, 63, 64
- **Analysis**: Build a 95% confidence interval for the mean difference in grades.

By matching, they have created a dependent group. Paired data implies we use Model 3.

```
boys <- c(39, 42, 47, 50, 52, 52, 54, 55, 55, 56, 56, 56, 58, 60, 62)
girls <- c(44, 45, 48, 50, 51, 52, 53, 53, 57, 58, 59, 60, 62, 63, 64)
diff <- boys - girls
y_d.bar <- mean(diff)
s_d <- sd(diff)
n_d <- length(diff)
df <- length(diff) - 1
t <- qt(0.975, df)
left <- y_d.bar - t * s_d / sqrt(n_d)
right <- y_d.bar + t * s_d / sqrt(n_d)
```

- $\bar{y}_{d+} = 1.7$
- $s_d = 2.1$
- $n_d = 15$
- $t^* = 2.145$

The 95% confidence interval for the mean difference grade is $(-2.9, -0.5)$. Is there a difference between male and female grades? 0 is not in this interval, so we conclude there is a difference between male and female grades. In fact, we may argue that the boys are doing worse than the girls.

EXAMPLE 1.15.4: Model 2B

- **Problem**: In grade 9 there is a standardized test in Ontario. We wish to compare the mean performance of girls to boys.
- **Plan**: They collect data from a class of 30 students; 15 boys and 15 girls. The response is their grade on the standardized test. If necessary assume the variances of the two groups are different.
- **Data**:
  - Boys: 39, 42, 47, 50, 52, 52, 54, 55, 55, 56, 56, 56, 58, 60, 62
  - Girls: 44, 45, 48, 50, 51, 52, 53, 53, 57, 58, 59, 60, 62, 63, 64
- **Analysis**: Build a 95% confidence interval for the mean difference in grades.

```
boys <- c(39, 42, 47, 50, 52, 52, 54, 55, 55, 56, 56, 56, 58, 60, 62)
girls <- c(44, 45, 48, 50, 51, 52, 53, 53, 57, 58, 59, 60, 62, 63, 64)
y_b.bar <- mean(boys)
y_g.bar <- mean(girls)
s_b.sq <- var(boys)
s_g.sq <- var(girls)
n_b <- length(boys)
n_g <- length(girls)
```

```
df <- n_g + n_b - 2
t <- qt(0.975, df)
left <- (y_b.bar - y_g.bar) - t * sqrt(s_b.sq / n_g + s_g.sq / n_b)
right <-
(y_b.bar - y_g.bar) + t * sqrt(s_b.sq / n_g + s_g.sq / n_b)
```

The 95% confidence interval for the mean difference grade is $(-6.41, 3.08)$.

EXAMPLE 1.15.5: Model 4

- **Problem**: In October there will be a federal election. Prior to the election pollsters will gauge the popularity of the candidates. One party of interest will be the Liberals.
- **Plan**: They ask 430 randomly selected people whether they would vote liberal.
- **Data**: 267 people would be willing to vote Liberal.
- **Analysis**: Build a 95% confidence interval for the proportion of people willing to vote Liberal.

```
n <- 430
voters <- 267
p.hat <- 267 / 430
z <- qnorm(0.975)
left <- p.hat - z * sqrt((p.hat * (1 - p.hat)) / n)
right <-  p.hat + z * sqrt((p.hat * (1 - p.hat)) / n)
```

- $n = 430$
- $\hat{p} = 267/430$
- $z^* = 1.96$

The 95% confidence interval for the proportion of people willing to vote Liberal is $(0.575, 0.667)$.

1.16 Lecture 16.00 - HT

(1) Define the hypothesis

Table 1.2: Hypotheses

H_0	H_a
$\theta = \theta_0$	$\theta \neq \theta_0$
$\theta \geq \theta_0$	$\theta < \theta_0$
$\theta \leq \theta_0$	$\theta > \theta_0$

(2) Discrepancy

$$d = \frac{\text{EST} - H_0 \text{ value}}{\text{SE}} = \frac{\hat{\theta} - \theta_0}{\sqrt{\mathbb{V}(\tilde{\theta})}}$$

(3) Given $\tilde{\theta} \sim \mathcal{N}(\theta, \mathbb{V}(\tilde{\theta}))$ where $D \sim \mathcal{N}(0, 1)$ when σ is known or $D \sim t(n - q + c)$ when σ is known.

(4) p-value

Table 1.3: p-value

H_a	p-value
$\theta \neq \theta_0$	$2\mathbb{P}(D > \lvert d \rvert)$
$\theta < \theta_0$	$\mathbb{P}(D < d)$
$\theta > \theta_0$	$\mathbb{P}(D > d)$

(5) Conclusion

Table 1.4: Guidelines for interpreting p-values

p-value	Interpretation
$p > 0.1$	No evidence to reject H_0.
$0.05 < p \leq 0.10$	Weak evidence against H_0.
$0.01 < p < 0.05$	Evidence against H_0.
$p < 0.01$	Tons of evidence against H_0.

1.17 Lecture 17.00 - HT Examples

EXAMPLE 1.17.1

Willow trees are grown from cuttings. These cuttings from 6 willow trees were subjected to two soils: high and low acidity. 2 cuttings from each tree are assigned to the two levels of acidity. After 1 year the height, in cm, of the cuttings was measured.

Cutting	1	2	3	4	5	6
High	11	19	32	12	7	14
Low	17	21	14	11	18	9

Is the growth in high and low acidity equal? Use an appropriate hypothesis test. If necessary assume group variances are the same.

Solution.

- H_0: $\mu_d = 0$
- H_a: $\mu_d \neq 0$

```
high <- c(11, 19, 32, 12, 7, 14)
low <- c(17, 21 , 14 , 11 , 18 , 9)
diff <- high - low
y_d.bar <- mean(diff)
s_d <- sd(diff)
n_d <- length(diff)
df <- length(diff) - 1
d <- (y_d.bar - 0) / (s_d / sqrt(n_d))
pval <- 2 * (1 - pt(d, df))
```

- $\bar{y}_{d+} = 0.83$
- $s_d = 10.07$
- $d = 0.2$
- $p = 2\mathbb{P}(D > |d|) = 2[1 - \mathbb{P}(D \leq 0.2)] =$ `2 * (1 - pt(d, df))` ≈ 0.84.

We obtain a p-value of 0.84. There is no evidence to reject H_0. In other words, we can argue in favour of saying that they have the same growth in different acidic soils.

EXAMPLE 1.17.2

A random assortment of pumpkin seeds were planted and fertilized using two types of plant feed, coke and water. After 4 weeks the plant heights, in cm, were measured.

- Coke: 8, 7, 18, 42, 21
- Water: 5, 11, 21, 9, 14

Is coke a better fertilizer for pumpkin's than water? Use an appropriate hypothesis test. If necessary assume group variances are the same.

- H_0: $\mu_c = \mu_w$
- H_a: $\mu_c - \mu_w > 0$

```
coke <- c(8, 7, 18, 42, 21)
water <- c(5, 11, 21, 9, 14)
y_c.bar <- mean(coke)
y_w.bar <- mean(water)
s_c.sq <- var(coke)
s_w.sq <- var(water)
n_c <- length(coke)
n_w <- length(water)
s_p.sq <-
((n_w - 1) * s_w.sq + (n_c - 1) * s_c.sq) / (n_w + n_c - 2)
df <- n_c + n_w - 2
t <- qt(0.975, df)
d <- (y_c.bar - y_w.bar) / sqrt(s_p.sq * (1 / n_w + 1 / n_c))
pval <- 1 - pt(d, df)
```

- $\bar{y}_{c+} = 19.2$
- $\bar{y}_{w+} = 12$
- $n_w = n_c = 5$
- $s_c^2 = 199.7$
- $s_w^2 = 36$
- $s_p^2 = 117.85$
- $d = \frac{\bar{y}_{c+} - \bar{y}_{d+} - 0}{s_p\sqrt{\frac{1}{n_w} + \frac{1}{n_c}}} = 1.049$
- $p = \mathbb{P}(D > d) = 1 - \mathbb{P}(D \leq 1.049) =$ `1 - pt(d, df)` ≈ 0.162.

There is no evidence to reject H_0. Therefore, coke is not a better fertilizer than water.