Sampling and Experimental Design

STAT 332

Winter 2021 (1211)

LaTeXer: *Cameron Roopnarine*
Instructor: *Riley Metzger*

June 23, 2021

Contents

Chapter 1

Assignment 1

1.1 Lecture 1.00: PPDAC + Example

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

1. 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.
2. Plan: How exactly are you going to answer the problem you posed yourself?
   - **Study population** (SP): The set of units you *can* study. The study population is not necessarily a subset of the target population.

     EXAMPLE 1.1.2

     Does a drug reduce hair loss?
     - Target population: People.
     - Study population: Mice.

     Note that mice is not a subset of people, so the study population and target population are not subsets of one another.
   - **Sample**: A subset of the study population. In the prior example, it would be the set of mice you select from your study population that are of interest in the sample.
   - **Data**: Collect the data, according to the plan.
3. Analysis: We analyze the data.
4. Conclusion: Refers back to the problem. We also note some common *errors*.
   (a) **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.

(b) **Sample error**: The parameter differs from the sample statistic, sometimes called an 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.

(c) **Measurement error**: The difference between what *we want* to calculate and what *we do* calculate.

EXAMPLE 1.1.5

IQ is an interesting thing. You want to measure somebody's intelligence, and yet if you go and actually calculate it, they're using various statistical tests or various psychological tests that could have a lot of measurement error.

1.2 Lecture 2.00: Models, Model 1

DEFINITION 1.2.1: Model

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

DEFINITION 1.2.2: Model 1

Model 1 is defined as

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

where

- Y_j: random parameter that is the response of unit j.
- μ: study population parameter. In this case, it's the mean and is non-random. However, it is unknown.
- R_j: error term. It gives the distribution of responses about μ.

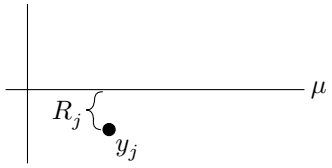


Figure 1.1: R_j diagram

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

We are interested in the average grade of STAT 101 students.

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

That would be a good place for us to use it because in our response the grade is related to the average grade of the class plus some error.

1.3 Lecture 3.00: Independent Groups

- Dependent: we randomly select one group and 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*.

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

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 $Y \sim$ 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 $y_1, \ldots, y_n$.

Process:

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

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

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** $\ell = \ln(L)$ 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

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

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 1860s. 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 not rigorous in this course.

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 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!

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

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 $\frac{\tilde{\mu} - \mu}{\tilde{\sigma}/\sqrt{n}}$?

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

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

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,

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

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

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.

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

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

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	d.f.
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	$Y_{ij} = \mu_i + R_{ij}$ $R_{ij} \sim \mathcal{N}(0,\sigma^2)$	$\bar{y}_{1+} \pm t^* \frac{s_1}{\sqrt{n_1}}$ $\bar{y}_{1+} - \bar{y}_{2+} \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$n_1 - 1$ $n_1 + n_2 - 2$
2B	$Y_{ij} = \mu_i + R_{ij}$ $R_{ij} \sim \mathcal{N}(0,\sigma_i^2)$	$\bar{y}_{1+} \pm t^* \frac{s_1}{\sqrt{n_1}}$ $\bar{y}_{1+} - \bar{y}_{2+} \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$n_1 - 1$ $n_1 + n_2 - 2$
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.

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

```
y <- c(65, 70, 80, 85, 75)
n <- length(y)
round(mean(y) + c(-1, 1) * qt(0.975, n - 1) * sd(y) / sqrt(n), 2)
```

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)
(y_b.bar - y_g.bar) + c(-1,1) * 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? Since $0 \in (-6.4, 3.1)$, 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. We select each girl so that she was born in the same month as a boy in the class. The response is their grade on the standardized test. 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)
y_d.bar + c(-1,1) * 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? Since $0 \notin (-2.9, -0.5)$, 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. 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)
(y_b.bar - y_g.bar) + c(-1, 1) * 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)
p.hat + c(-1, 1) * 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

We grow willow trees from cuttings. We grow these cuttings from 6 willow trees in two soils: high and low acidity. We assign two cuttings from each tree to the two levels of acidity. After 1 year the height, we measure the cuttings in centimetres.

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. 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

We plant a random assortment of pumpkin seeds and fertilize them 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. 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)
```

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

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

Chapter 2

Assignment 2

2.1 Lecture 18.00: Model 5, Estimates

DEFINITION 2.1.1: Completely randomized design, Model 5

The **completely randomized design** (CRD) is defined as

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

for $i = 1, 2, \dots, t$ (no. of treatments), $j = 1, 2, \dots, r$ (no. of replicates/treatment). The number of units is tr. In this course, this is **Model 5**.

- μ is the study population mean
- $\mu + \tau_i$ is the group mean
- τ_i is the treatment effect of group i
- R_{ij} is the distribution of values about the deterministic part of the model.

Constraint: $\tau_1 + \tau_2 + \cdots + \tau_t = 0$

EXAMPLE 2.1.2

Group 1	Group 2
60	70
65	75
70	80

- $\hat{\mu} = \frac{60+65+70+70+75+80}{6} = 70$
- $\hat{\mu} + \hat{\tau}_1 = \frac{60+65+70}{3} = 65$
- $\hat{\mu} + \hat{\tau}_2 = \frac{70+75+80}{3} = 75$
- $\hat{\tau}_1 = -5$
- $\hat{\tau}_2 = 5$

Note that $\hat{\tau}_1 + \hat{\tau}_2 = 5$.

EXAMPLE 2.1.3: LS for CRD

$$W = \sum_{ij} r_{ij}^2 + \lambda(\tau_1 + \cdots + \tau_t) = \sum_{ij} (y_{ij} - \mu - \tau_i)^2 + \lambda(\tau_1 + \tau_2 + \cdots + \tau_t)$$

Find $\frac{\partial W}{\partial \mu}$, $\frac{\partial W}{\partial \tau_1}, \dots, \frac{\partial W}{\partial \tau_t}$, and $\frac{\partial W}{\partial \lambda}$ and set to zero to solve.

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

$$\hat{\tau}_i = \bar{y}_{i+} - \bar{y}_{++}$$
$$\hat{\sigma}^2 = \frac{W}{n-q+c} = \frac{W}{(tr)-(t+1)+(1)}$$

- $n = tr$ since that is the number of parameters we have.
- $q = t+1$ since we have one μ and t τ's.
- $c = 1$ since we have one constraint $\tau_1 + \cdots + \tau_t = 0$.

2.2 Lecture 19.00: Model 5, Estimators

Suppose we have $i = 1, 2$ and $j = 1, 2, \ldots, r$. The number of units is $2r$. For the CRD model, the estimator is

$$\tilde{\mu} = \bar{Y}_{++}$$

Let's find the mean and variance of $\tilde{\mu}$ for $i = 1, 2$ and $j = 1, 2, \ldots, r$.

$$\begin{aligned}
\mathbb{E}[\bar{Y}_{++}] &= \mathbb{E}\left[\frac{\sum_{i=1}^{2}\sum_{j=1}^{r} Y_{ij}}{2r}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^{2}\sum_{j=1}^{r} (\mu + \tau_i + R_{ij})}{2r}\right] \\
&= \frac{\sum_{i=1}^{2}\sum_{j=1}^{r} \mathbb{E}[\mu] + \mathbb{E}[\tau_i] + \mathbb{E}[R_{ij}]}{2r} \\
&= \frac{\sum_{i=1}^{2}\sum_{j=1}^{r} \mu + \tau_i}{2r} \\
&= \frac{2r\mu + \sum_{j=1}^{r}(\tau_1 + \tau_2)}{2r} \\
&= \mu
\end{aligned}$$

Since $\mathbb{E}[\tilde{\mu}] = \mu$ we have an unbiased estimator.

$$\mathbb{V}(\bar{Y}_{++}) = \mathbb{V}\left(\frac{\sum_{i=1}^{2}\sum_{j=1}^{r} Y_{ij}}{2r}\right) = \frac{\sum_{i=1}^{2}\sum_{j=1}^{r} \mathbb{V}(Y_{ij})}{(2r)^2} = \frac{2r\sigma^2}{(2r)^2} = \frac{\sigma^2}{2r}$$

where the second equality used independence.

2.3 Lecture 20.00: Model 5, Estimators 2

An estimator for CRD is

$$\tilde{\tau}_1 = \bar{Y}_{1+} - \bar{Y}_{++}$$

Let's find the mean and variance of $\tilde{\tau}_1$ for $i = 1, 2$ and $j = 1, 2, \ldots, r$.

$$\mathbb{E}[\tilde{\tau}_1] = \mathbb{E}[\bar{Y}_{1+} - \bar{Y}_{++}] = \mathbb{E}[\bar{Y}_{1+}] - \mu = \mathbb{E}\left[\frac{\sum_{i=1}^{r} Y_{1j}}{r}\right] - \mu = \frac{\sum_{i=1}^{r}(\mu + \tau_1)}{r} - \mu = \frac{r\mu + r\tau_1}{r} - \mu = \tau_1$$

Working with the variance is slightly tricky.

$$\begin{aligned}
\mathbb{V}(\tilde{\tau}_1) &= \mathbb{V}(\bar{Y}_{1+} - \bar{Y}_{++}) \\
&= \mathbb{V}\left(\bar{Y}_{1+} - \left(\frac{\bar{Y}_{1+} + \bar{Y}_{2+}}{2}\right)\right) \\
&= \mathbb{V}\left(\frac{1}{2}\bar{Y}_{1+} - \frac{1}{2}\bar{Y}_{2+}\right) \\
&= \frac{1}{4}\,\mathbb{V}(\bar{Y}_{1+}) + \frac{1}{4}\,\mathbb{V}(\bar{Y}_{2+}) && \text{independence} \\
&= \frac{\sigma^2}{4r} + \frac{\sigma^2}{4r} \\
&= \frac{\sigma^2}{2r}
\end{aligned}$$

The confidence interval for τ_1 is given by

$$\tau_1 : \hat{\tau}_1 \pm c\sqrt{\frac{\hat{\sigma}^2}{2r}} \quad (c \sim t(n-q+c))$$

and the discrepancy is (obviously) given by

$$d = \frac{\hat{\tau}_1 - \tau_0}{\sqrt{\frac{\hat{\sigma}^2}{2r}}} \quad (c \sim t(n-q+c))$$

The confidence interval for μ is given by

$$\mu : \hat{\mu} \pm c\sqrt{\frac{\hat{\sigma}^2}{2r}} \quad (c \sim t(n-q+c))$$

Chapter 3

Assignment 3

3.1 Lecture 21.00: Model 5, Example 1

A study of intoxication measured two groups of students, one of which was drunk while the other was not as they drove a computer-simulated driving course with a max speed limit of 50 km/h. Of interest was the maximum speed of an individual doing the computer-simulated driving course. Group 1 was intoxicated, while Group 2 was not.

Is there a difference in speed between those that drive while intoxicated versus those that do not?

```
# Data frames
grp1 <- c(50, 53, 52, 58)
grp2 <- c(62, 55, 58, 60)
# Must run to get same results as textbook
options(contrasts = c("contr.sum", "contr.poly"))
Y <- c(grp1, grp2)
# Makes a discrete variable
x <- as.factor(c(rep(1, 4), rep(2, 4)))
# Builds the model
model <- lm(Y ~ x)
# Displays the output
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3.75  -1.75  -0.50   1.75   4.75
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   56.000      1.132  49.473 4.57e-09 ***
## x1            -2.750      1.132  -2.429   0.0512 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.202 on 6 degrees of freedom
## Multiple R-squared:  0.4959,Adjusted R-squared:  0.4119
## F-statistic: 5.902 on 1 and 6 DF,  p-value: 0.0512
```

(1) `Residuals`: Helps test our residuals.

(2) `Coefficients`: $\hat{\mu} = 56$, $\hat{\tau}_1 = -2.75$, $\hat{\tau}_2 = 2.75$.

(3) `Residual standard error`: $\hat{\sigma} = 3.202$ on 6 degrees of freedom.

(4) `Coefficients (Error to P(>|t|))`: This line tests H_0: $\mu = 0$ versus H_a: $\mu \neq 0$

$$d = \frac{56 - 0}{1.132} = 49.473$$

$$p\text{-value} = 2\,\mathbb{P}(D > 49.473) = 4.57 \times 10^{-9}$$

We have tons of evidence to reject H_0.

(5) H_0: $\tau_1 = 0$ versus H_a: $\tau_1 \neq 0$

$$d = \frac{-2.75 - 0}{1.132} = -2.429$$

$$p\text{-value} = 2\,\mathbb{P}(D > |-2.429|) = 2(1 - \mathbb{P}(D \leq 2.429)) = 0.0512$$

There is evidence to reject H_0.

What is the treatment effect for being inebriated?

$\hat{\tau}_1 = -2.75$.

Is there a difference between the treatment effect of group 1 and 2? Use a 95% CI.

$$\theta = \text{ave of grp1} - \text{ave of grp2} = (\mu + \tau_1) - (\mu + \tau_2) = \tau_1 - \tau_2$$

Estimator: $\tilde{\theta} = \tilde{\tau}_1 - \tilde{\tau}_2$ and is normal by Gauss.

$$\mathbb{E}[\tilde{\theta}] = \mathbb{E}[\tilde{\tau}_1 - \tilde{\tau}_2] = \mathbb{E}[\tilde{\tau}_1] - \mathbb{E}[\tilde{\tau}_2] = \tau_1 - \tau_2$$

since unbiased.

$$\mathbb{V}(\tilde{\theta}) = \mathbb{V}\big(\bar{Y}_{1+} - \bar{Y}_{++} - (\bar{Y}_{2+} - \bar{Y}_{++})\big) = \mathbb{V}(\bar{Y}_{1+} - \bar{Y}_{2+}) = \mathbb{V}(\bar{Y}_{1+}) + \mathbb{V}(\bar{Y}_{2+}) = \frac{\sigma^2}{4} + \frac{\sigma^2}{4} = \frac{\sigma^2}{2}$$

CI for θ:

$$\theta : \hat{\theta} \pm c\,\text{SE} = \hat{\tau}_1 - \hat{\tau}_2 \pm c\sqrt{\frac{\hat{\sigma}^2}{2}} \quad (c \sim t(n - q + c) = t(8 - 2 + 1) = t(6))$$

In our case,

$$\theta : (-2.75 - 2.75) \pm 2.447\sqrt{\frac{3.202^2}{2}} = (-11.04, 0.04)$$

0 is in the interval, so we conclude that there is no difference between the treatment effect of group 1 and 2. In R, we could do

```
-2.75 - 2.75 + c(-1, 1) * qt(0.975, 6) * sqrt(summary(model)$sigma^2/2)
## [1] -11.0394323   0.0394323
```

To obtain our CI $\theta : (-11.039, 0.039)$.

Is there a difference between the treatment effect of group 1 and 2? Use an HT.

H_0: $\tau_1 = \tau_2$ versus H_a: $\tau_1 \neq \tau_2$.

$$d = \frac{\hat{\tau}_1 - \hat{\tau}_2 - \tau_0}{\hat{\sigma}/\sqrt{2}} = \frac{(-2.75 - 2.75) - 0}{3.202/\sqrt{2}} = -2.489$$

$$p = 2\,\mathbb{P}(D \geq |d|) = (0.05, 0.10)$$

We have some evidence to reject H_0. In R, we could do

```
d <- (-2.75 - 2.75)/(summary(model)$sigma/sqrt(2))
d
```

```
## [1] -2.429494
```

```
2 * (1 - pt(abs(d), 6))
```

```
## [1] 0.05119768
```

To obtain $d = -2.429$ and p-value $= 0.051$. There is some difference between the treatment effect of group 1 and 2.

3.2 Lecture 22.00: Model 5, Example 1 Cont.

We want to check our model assumptions of $R_j \sim \mathcal{N}(0, \sigma^2)$ independent. Four things to check:

- $\mathbb{E}[R_j] = 0$ (zero mean)
- $\mathbb{V}(R_j) = \sigma^2$ (constant variance)
- Normality
- Independence

To check these, we can

- plot residuals versus fitted values to check for both mean and variance assumption.

  `plot(model$residuals)`

- Q-Q plot to check for normality (straight line is normal).

  `qqnorm(model$residuals)`

- residuals plot to check for independence assumption.

  `plot(model$fitted.values, model$residuals)`

Example

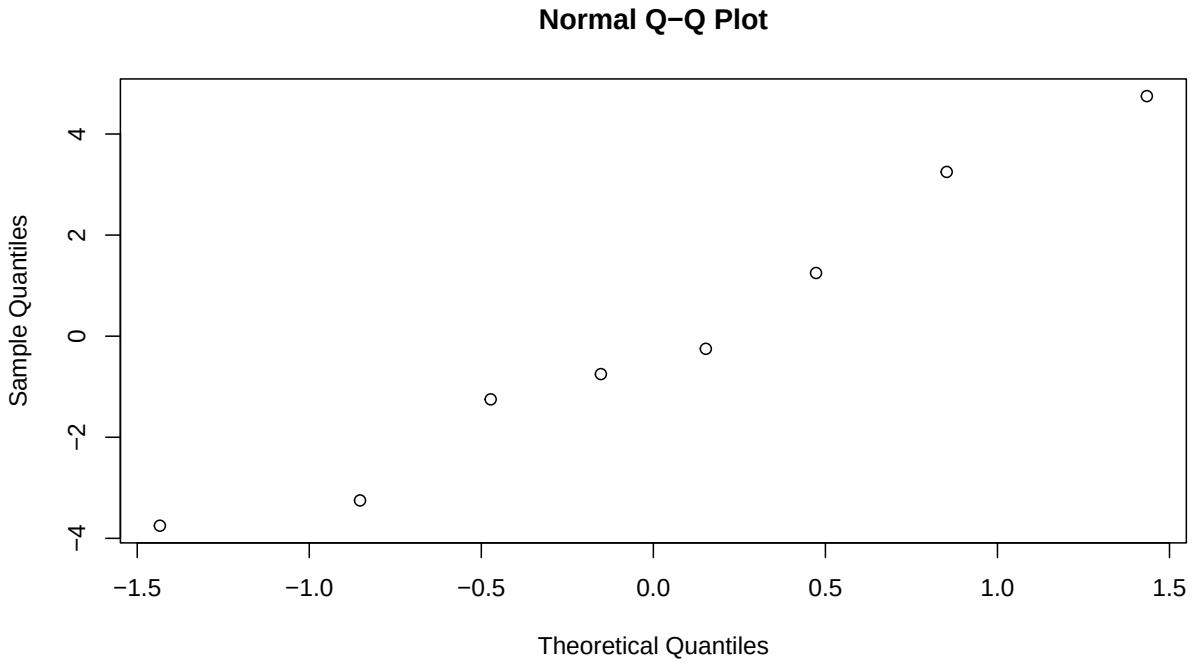
All the diagnostics for this example seem good.

```
# Data frames
grp1 <- c(50, 53, 52, 58)
grp2 <- c(62, 55, 58, 60)
# Must run to get same results as textbook
options(contrasts = c("contr.sum", "contr.poly"))
Y <- c(grp1, grp2)
# Makes a discrete variable
x <- as.factor(c(rep(1, 4), rep(2, 4)))
# Builds the model
model <- lm(Y ~ x)
```

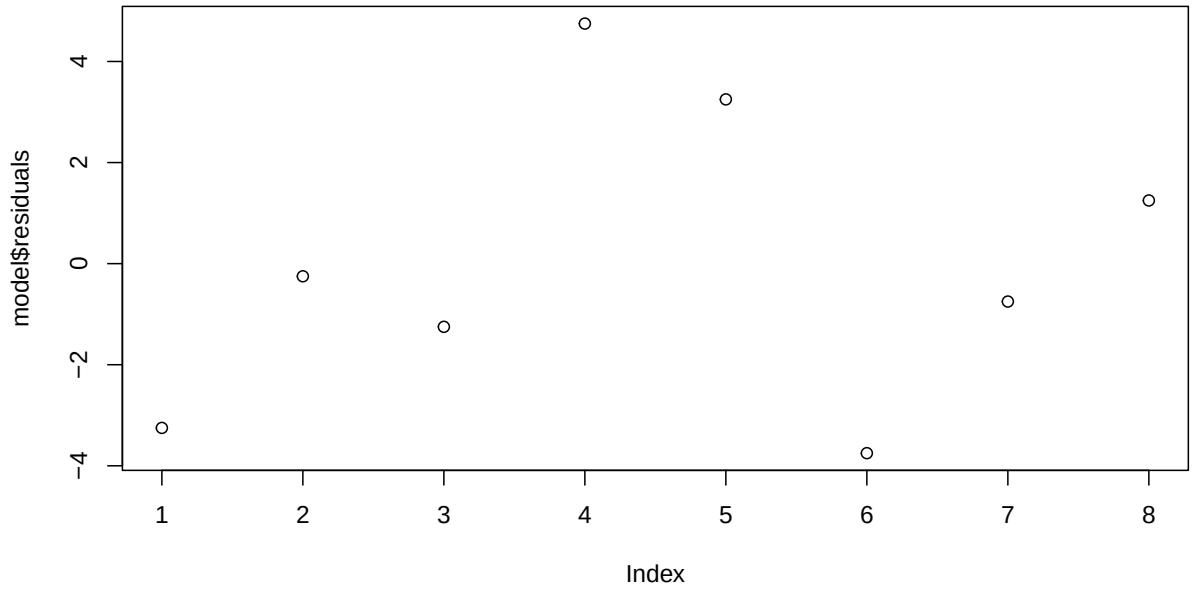
```
# Residuals
model$residuals
```

```
##     1     2     3     4     5     6     7     8
## -3.25 -0.25 -1.25  4.75  3.25 -3.75 -0.75  1.25
```

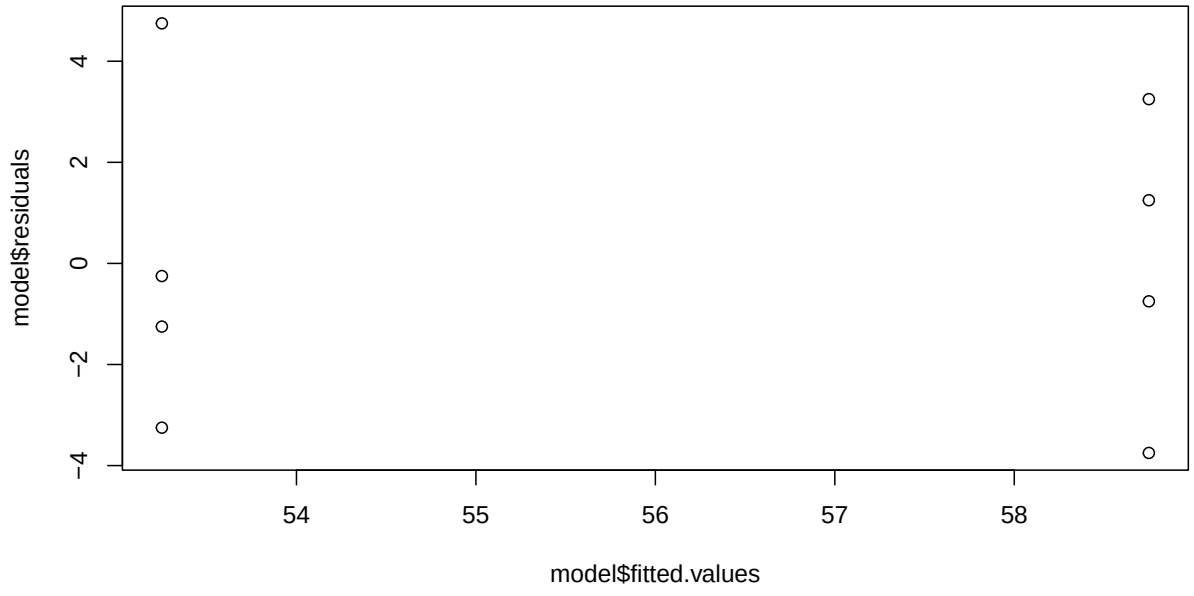
```
qqnorm(model$residuals)
```




```
plot(model$residuals)
```




```
plot(model$fitted.values, model$residuals)
```



3.3 Lecture 23.00: Model 5, Example 2

Suppose professors are coordinating 4 sections of the same course in a term. We want to look at the average mark for each section on the midterm. The treatment is the "instructor."

```r
options(contrasts = c("contr.sum", "contr.poly"))
Marks1 = c(55, 92, 48, 57, 66, 72)
Marks2 = c(62, 95, 84, 83, 66, 75)
Marks3 = c(89, 92, 94, 99, 87, 67)
Marks4 = c(25, 35, 71, 42, 44, 30)
Y = c(Marks1, Marks2, Marks3, Marks4)
x = as.factor(c(rep(1, 6), rep(2, 6), rep(3, 6), rep(4, 6)))
model = lm(Y ~ x)
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.0000 -10.2917   0.9167   6.1250  29.8333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67.917      2.861  23.741    4e-16 ***
## x1            -2.917      4.955  -0.589 0.562699
## x2             9.583      4.955   1.934 0.067381 .
## x3            20.083      4.955   4.053 0.000621 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.01 on 20 degrees of freedom
## Multiple R-squared:  0.6506,Adjusted R-squared:  0.5982
## F-statistic: 12.41 on 3 and 20 DF,  p-value: 8.281e-05
```

Note that

- $\hat{\tau}_4 = -(\hat{\tau}_1 + \hat{\tau}_2 + \hat{\tau}_3) = -26.749$.
- Degrees of freedom $= n - q + c = (24) - (4+1) + 1 = 20$.

Is there a difference between the treatment effect of group 1 and 2? Use a 95% CI.

$\tilde{\theta} = \tilde{\tau}_1 + \tilde{\tau}_2$ and by Gauss this is normal.

$$\mathbb{E}[\tilde{\theta}] = \mathbb{E}[\tilde{\tau}_1 - \tilde{\tau}_2] = \tau_1 + \tau_2$$

$$\mathbb{V}(\tilde{\theta}) = \mathbb{V}(\tilde{\tau}_1 - \tilde{\tau}_2) = \mathbb{V}(\bar{Y}_{1+} - \bar{Y}_{2+}) = \mathbb{V}(\bar{Y}_{1+}) + \mathbb{V}(\bar{Y}_{2+}) = \frac{\sigma^2}{6} + \frac{\sigma^2}{6} = \frac{\sigma^2}{3}$$

The 95% confidence interval for θ is

$$\hat{\tau}_1 - \hat{\tau}_2 \pm c\,\frac{\hat{\sigma}}{\sqrt{3}} = -2.917 - 9.583 \pm \frac{2.09(14.01)}{\sqrt{3}} = (-29.37, 4.37) \quad (c \sim t(20))$$

In R, we could do

```r
tau.1 <- summary(model)$coefficients[2]
tau.2 <- summary(model)$coefficients[3]
tau.3 <- summary(model)$coefficients[4]
```

```
tau.4 <- -1 * (tau.1 + tau.2 + tau.3)
tau.1 - tau.2 + c(-1, 1) * qt(0.975, 20) * summary(model)$sigma/sqrt(3)
```

```
## [1] -29.378554   4.378554
```

To get at 95% confidence interval θ: $(-29.38, 4.38)$. Since $0 \in (-29.38, 4.38)$, there is not a difference between the treatment effect of group 1 and 2.

Groups 2 and 3 were taught by the same instructor. Groups 1 and 4 are taught by another instructor. Is there a difference between the average treatment effect of instructor 1 to instructor 2? Use an HT.

$$\tilde{\theta} = \frac{\tilde{\tau}_1 + \tilde{\tau}_4}{2} - \left(\frac{\tilde{\tau}_2 + \tilde{\tau}_3}{2}\right)$$

$$\mathbb{E}[\tilde{\theta}] = \frac{\tau_1 + \tau_4}{2} - \left(\frac{\tau_2 + \tau_3}{2}\right)$$

$$\begin{aligned}
\mathbb{V}(\tilde{\theta}) &= \mathbb{V}\left(\frac{\tilde{\tau}_1 + \tilde{\tau}_4}{2} - \left(\frac{\tilde{\tau}_2 + \tilde{\tau}_3}{2}\right)\right) \\
&= \mathbb{V}\left(\frac{\bar{Y}_{1+} + \bar{Y}_{4+}}{2} - \left(\frac{\bar{Y}_{2+} + \bar{Y}_{3+}}{2}\right)\right) \\
&= \frac{1}{4}\,\mathbb{V}(Y_{1+}) + \cdots + \frac{1}{4}\,\mathbb{V}(Y_{4+}) \\
&= \frac{\sigma^2}{4(6)} + \cdots + \frac{\sigma^2}{4(6)} \\
&= \frac{\sigma^2}{6}
\end{aligned}$$

H_0: $\theta = 0$ versus H_a: $\theta \neq 0$.

$$d = \frac{\hat{\theta} - 0}{\hat{\sigma}/\sqrt{6}} = -5.19 \quad (D \sim t(20))$$

$$p = 2\,\mathbb{P}(D > |-5.19|) = (0, 0.001)$$

We have tons of evidence to reject H_0 in favour of the instructors having a different effect. In R, we could do

```
theta <- ((tau.1 + tau.4)/2) - ((tau.2 + tau.3)/2)
d <- (theta - 0)/(summary(model)$sigma/sqrt(6))
d
```

```
## [1] -5.185077
```

```
2 * (1 - pt(abs(d), 20))
```

```
## [1] 4.498007e-05
```

To obtain a p-value of 4.498007×10^{-5}.

EXAMPLE 3.3.1

An example of a *contrast* is

$$\theta = \frac{\tau_1 + \tau_4}{2} - \frac{(\tau_2 + \tau_3)}{2}$$

DEFINITION 3.3.2: Contrast

A **contrast** has the form

$$a_1\tau_1 + a_2\tau_2 + \cdots + a_n\tau_n$$

where $\sum_{i=1}^{n} a_i = 0$.

3.4 Lecture 24.00: ANOVA

Analysis of Variance

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

Recall:

$$\begin{aligned} W &= \sum_{ij} r_{ij}^2 \\ &= \sum_{ij} (y_{ij} - \hat{\mu} - \hat{\tau}_i)^2 \\ &= \sum_{ij} (y_{ij} - \hat{\mu})^2 + (-r) \sum_{i} (\bar{y}_{i+} - \bar{y}_{++})^2 \end{aligned}$$

Rearranging

$$\underbrace{\sum_{ij} (y_{ij} - \bar{y}_{++})^2}_{\text{SS(Tot)}} = \underbrace{r \sum_{i} (\bar{y}_{i+} - \bar{y}_{++})^2}_{\text{SS(Trt)}} + \underbrace{\sum_{ij} (y_{ij} - \hat{\mu} - \hat{\tau}_i)^2}_{\text{SS(Res)}}$$

$$\boxed{\text{SS(Tot)} = \text{SS(Trt)} + \text{SS(Res)}}$$

SS(Tot)

- Represents a measure of total variability in your data
- $s^2 = \dfrac{\text{SS(Tot)}}{n-1} = \text{MS(Tot)}$
- $\text{df} = n - 1$
- You get this by fitting Model 1; $Y_i = \mu + R_i$ where $R_i \sim \mathcal{N}(0, \sigma^2)$
- Recall $\hat{\sigma} = s$ in Model 1

SS(Res)

- The variability left over after you fit the model (unexplained)
- Synonymous with $\hat{\sigma}^2$
- $\hat{\sigma}^2 = \dfrac{W}{n-q+c} = \dfrac{\text{SS(Res)}}{\text{df}_{\text{Res}}} = \text{MS(Res)}$
- $\text{df} = n - q + c$

SS(Trt)

- Due to τ component
- $\text{MS(Trt)} = \dfrac{\text{SS(Trt)}}{\text{df}_{\text{Trt}}}$
- $\text{df}_{\text{Trt}} = t - 1$
- $\boxed{\text{df}_{\text{Tot}} = \text{df}_{\text{Trt}} + \text{df}_{\text{Res}}}$

- Variability explained by your model

$$\text{SS(Tot)} = \text{SS(Trt)} = \text{SS(Res)}$$

We want $\text{SS(Trt)} \gg \text{SS(Res)}$. We compare MS(Trt) to MS(Res) using the ratio

$$F = \frac{\text{MS(Trt)}}{\text{MS(Res)}}$$

3.5 Lecture 25.00: F Test

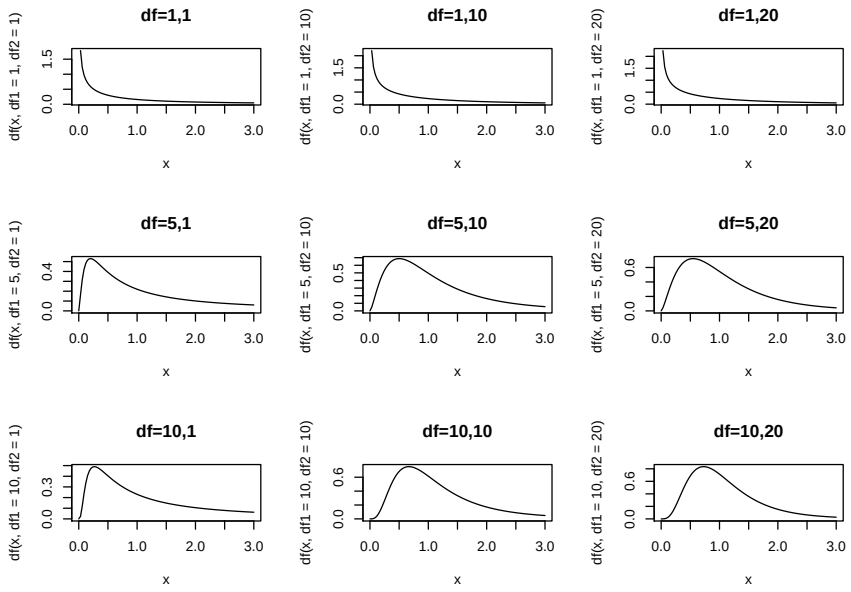


Figure 3.1: F distribution

THEOREM 3.5.1

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

$$\frac{X/m}{Y/n} \sim F(m,n)$$

THEOREM 3.5.2

Let $X \sim F(m,n)$ and $Y \sim 1/X$, then

$$Y \sim F(n,m)$$

EXAMPLE 3.5.3

$\alpha = \mathbb{P}(F(20,4) > 4) = (0.05, 0.1)$ since

α	0.1	0.05	0.01
critical value	3.84	5.80	14.0

In R, we can directly calculate α with `1-pf(4,20,4)=0.094`.

$$\tilde{F} = \frac{\widetilde{\text{MS(Trt)}}}{\widetilde{\text{MS(Res)}}}$$

Now, $\widetilde{\text{MS(Res)}} = \tilde{\sigma}^2$. We know

$$\frac{\tilde{\sigma}^2 \text{df}_{\text{Res}}}{\sigma^2} \sim \chi^2(\text{df}_{\text{Res}}) \tag{3.1}$$

Similarly,

$$\frac{\widetilde{\text{MS(Trt)}}\text{df}_{\text{Trt}}}{\sigma^2} \sim \chi^2(\text{df}_{\text{Trt}}) \tag{3.2}$$

Divide 3.2 by 3.1 to get

$$\frac{\widetilde{\text{MS(Trt)}}}{\widetilde{\text{MS(Res)}}} \sim F(n, d)$$

where $n = \text{df}_{\text{Trt}}$ and $d = \text{df}_{\text{Res}}$.

When is F large?

$$\mathbb{E}[\tilde{F}] = \mathbb{E}\left[\frac{\widetilde{\text{MS(Trt)}}}{\widetilde{\text{MS(Res)}}}\right] \approx \frac{\mathbb{E}[\widetilde{\text{MS(Trt)}}]}{\mathbb{E}[\widetilde{\text{MS(Res)}}]} = \frac{\sigma^2 + r\dfrac{\sum_{i=1}^{t} \tau_i^2}{t-1}}{\sigma^2}$$

$$\mathbb{E}[\tilde{F}] = 1 + \frac{r}{\sigma^2}\frac{\sum_{i=1}^{t} \tau_i^2}{t-1}$$

If $\tau_1 = \tau_2 = \cdots = \tau_t = 0$, then $\mathbb{E}[\tilde{F}] = 1$. However, if even one τ is not zero, then $\mathbb{E}[\tilde{F}] > 1$.

F Test

(1) H_0: $\tau_1 = \tau_2 = \cdots = \tau_t = 0$ versus H_a: at least one τ is not zero.

(2) $d = \dfrac{\text{MS(Trt)}}{\text{MS(Res)}}$ where $D \sim F(\text{df}_{\text{Trt}}, \text{df}_{\text{Res}})$

(3) p-value $= \mathbb{P}(D > d)$

(4) Conclusion.

3.6 Lecture 26.00: F Test, Example

See Section 3.3 for the data.

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          3 7315.5 2438.50  12.415 8.281e-05 ***
## Residuals 20 3928.3  196.42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- H_0: $\tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$
- H_a: At least one τ is not zero

$$d = \frac{\text{MS(Trt)}}{\text{MS(Res)}} = \frac{\text{SS(Trt)}/\text{df}_{\text{Trt}}}{\text{SS(Res)}/\text{df}_{\text{Res}}} = \frac{7315.5/3}{3928.3/20} = 12.415$$

Note that $D \sim F(3, 20)$, so

$$p = \mathbb{P}(D > 12.415) = 8.21 \times 10^{-5}$$

We have tons of evidence against H_0, so one of our treatment effects is not zero.

Chapter 4

Assignment 4

4.1 Lecture 27.00: Model 6

DEFINITION 4.1.1: Unbalanced CRD, Model 6

The **unbalanced completely randomized design** is

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

for $i = 1, 2, \dots, t$ (no. of treatments), $j = 1, 2, \dots, r_i$ (no. of replicates/treatment). In this course, this is **Model 6**.
Constraint: $\sum_{i=1}^{t} r_i \tau_i = 0$.

EXAMPLE 4.1.2: LS for Model 6

The LS for Model 6 is

$$W = \sum r_{ij}^2 + \lambda \left(\sum_{i=1}^{t} r_i \tau_i \right)$$

and results in

$$\hat{\mu} = \bar{y}_{++}$$
$$\hat{\tau}_i = \bar{y}_{i+} - \bar{y}_{++}$$
$$\hat{\sigma}^2 = \frac{W}{(r_1 + r_2 + \cdots + r_t) - (t+1) + 1}$$

4.1.1 Unbalanced CRD Example

Refer to Section 3.1, we remove the last element of group 2.

```
grp1 = c(50, 53, 52, 58)
grp2 = c(62, 55, 58)
Y = c(grp1, grp2)
x = as.factor(c(rep(1, 4), rep(2, 3)))
# Group Averages
grp_av = tapply(Y, x, mean, na.rm = T)
mu = mean(Y)
# Treatment Effects
tau1 = (grp_av - mean(Y))[1]
```

```
tau2 = (grp_av - mean(Y))[2]
# Estimated Sigma
sigma = summary(lm(Y ~ x))$sigma
# Values
sigma
```

```
## [1] 3.447221
```

```
tau1
```

```
##         1
## -2.178571
```

```
tau2
```

```
##        2
## 2.904762
```

```
mu
```

```
## [1] 55.42857
```

We obtain

- $\hat{\sigma} = 3.447221$
- $\hat{\tau}_1 = -2.178571$
- $\hat{\tau}_2 = 2.904762$
- $\hat{\mu} = 55.42857$
- Obviously, $4(\hat{\tau}_1) + 3(\hat{\tau}_2) = 0$

What is the treatment effect for being inebriated?

$\hat{\tau}_1 = -2.18$

Is there a difference between the treatment effect of group 1 and 2? Use a 95% CI.

$\theta = \tau_1 - \tau_2 \implies \tilde{\theta} = \tilde{\tau}_1 - \tilde{\tau}_2$.

$$\mathbb{E}[\tilde{\theta}] = \tau_1 - \tau_2$$

$$\mathbb{V}(\tilde{\theta}) = \mathbb{V}(\bar{Y}_{1+} - \bar{Y}_{2+}) = \frac{\sigma^2}{4} + \frac{\sigma^2}{3} = \frac{7\sigma^2}{12}$$

Confidence interval:

$$\hat{\tau}_1 - \hat{\tau}_2 \pm c\sqrt{\frac{7\hat{\sigma}^2}{12}} = (-11.85, 1.68)$$

In R,

```
tau1 - tau2 + c(-1, 1) * qt(0.975, 5) * sqrt((7 * sigma^2)/12)
```

```
## [1] -11.851312   1.684645
```

Is there a difference between the treatment effect of group 1 and 2? Use a HT.

```
anova(lm(Y ~ x))
```

```
## Analysis of Variance Table
##
## Response: Y
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## x          1 44.298  44.298  3.7277 0.1114
## Residuals  5 59.417  11.883
```

No evidence against H_0: $\tau_1 = \cdots = \tau_t = 0$, so this model is not great.

4.2 Lecture 28.00: Model 7

DEFINITION 4.2.1: Randomized block design, Model 7

The **randomized block design** (RBD) is defined as

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

where β_j is the j^{th} block (BIK) effect. Note that

- $i = 1, 2, \dots, t$
- $j = 1, 2, \dots, r$
- $\sum_{i=1}^{t} \tau_i = 0$
- $\sum_{j=1}^{r} \beta_j = 0$

EXAMPLE 4.2.2: LS for Model 7

The LS for Model 7 is

$$W = \sum_{ij} r_{ij} + \lambda_1 \left(\sum_{i=1}^{t} \tau_i \right) + \lambda_2 \left(\sum_{j=1}^{r} \beta_j \right)$$

Solving

$$\hat{\mu} = \bar{y}_{++}$$
$$\hat{\tau}_i = \bar{y}_{i+} - \bar{y}_{++}$$
$$\hat{\beta}_j = \bar{y}_{+j} - \bar{y}_{++}$$
$$\hat{\sigma}^2 = \frac{W}{(rt) - (t + r + 1) + 2}$$

4.3 Lecture 29.00: Model 7, Example

We grow willow trees from cuttings. We grow these cuttings from 6 willow trees in two soils: high and low acidity. We assign two cuttings from each tree to the two levels of acidity. After 1 year the height, we measure the cuttings in centimetres.

Is the growth in high and low acidity equal? Use an appropriate hypothesis test.

```
# Step 1 - Change the directory In R, select FILE, CHANGE
# DIR select the folder your data is located in.
# Step 2 - use read.table
Data = read.table("blocked.csv", sep = ",", header = T)
# Step 3 - Have a look at the data:
Data
```

```
##   Block Treatment Value
## 1     1      High    16
```

```
## 2      2      High    19
## 3      3      High    32
## 4      4      High    12
## 5      5      High     7
## 6      6      High    14
## 7      1       Low    17
## 8      2       Low    21
## 9      3       Low    33
## 10     4       Low    11
## 11     5       Low     8
## 12     6       Low    12
```

```
# To build a model we type:
options(contrasts = c("contr.sum", "contr.poly"))
attach(Data)
Treatment = as.factor(Treatment)
Block = as.factor(Block)
Model = lm(Value ~ Treatment + Block)
# To look at the output, we type:
summary(Model)
```

```
##
## Call:
## lm(formula = Value ~ Treatment + Block)
##
## Residuals:
##       1       2       3       4       5       6       7       8       9      10
## -0.3333 -0.8333 -0.3333  0.6667 -0.3333  1.1667  0.3333  0.8333  0.3333 -0.6667
##      11      12
##  0.3333 -1.1667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.8333     0.3073  54.775 3.84e-08 ***
## Treatment1   -0.1667     0.3073  -0.542 0.610881
## Block1       -0.3333     0.6872  -0.485 0.648131
## Block2        3.1667     0.6872   4.608 0.005797 **
## Block3       15.6667     0.6872  22.798 3.02e-06 ***
## Block4       -5.3333     0.6872  -7.761 0.000568 ***
## Block5       -9.3333     0.6872 -13.582 3.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.065 on 5 degrees of freedom
## Multiple R-squared:  0.9927,Adjusted R-squared:  0.984
## F-statistic: 113.5 on 6 and 5 DF,  p-value: 3.532e-05
```

```
anova(Model)
```

```
## Analysis of Variance Table
##
## Response: Value
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## Treatment  1   0.33   0.333   0.2941    0.6109
## Block      5 771.67 154.333 136.1765 2.446e-05 ***
```

```
## Residuals  5   5.67   1.133
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $\hat{\sigma} = 1.065$ on $n - q + c$ degrees of freedom. In this case, we have 6 blocks, 2 treatments, so 12 total values. One μ and two constraints. So $12 - 6 - 2 - 1 + 2 = 5$ degrees of freedom.
- $\tilde{\theta} = \tilde{\tau}_1 - \tilde{\tau}_2$.
- $\mathbb{E}[\tilde{\theta}] = \tau_1 - \tau_2$.
- $\mathbb{V}(\tilde{\theta}) = \mathbb{V}(\bar{Y}_{1+}) - \mathbb{V}(\bar{Y}_{2+}) = \frac{\sigma^2}{6} + \frac{\sigma^2}{6} = \frac{\sigma^2}{3}$.

The confidence interval for the difference in treatments is:

$$\hat{\tau}_1 - \hat{\tau}_2 \pm c\sqrt{\frac{\hat{\sigma}^2}{3}} = (-0.1667 - 0.1667) \pm 2.57\frac{1.065}{\sqrt{3}} = (-1.91, 1.25)$$

Suppose we ran a CRD instead.

```
Model = lm(Value ~ Treatment)
summary(Model)
```

```
##
## Call:
## lm(formula = Value ~ Treatment)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.667 -5.250 -1.667  2.750 16.000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.8333     2.5451   6.614 5.97e-05 ***
## Treatment1   -0.1667     2.5451  -0.065    0.949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.817 on 10 degrees of freedom
## Multiple R-squared:  0.0004286,Adjusted R-squared:  -0.09953
## F-statistic: 0.004288 on 1 and 10 DF,  p-value: 0.9491
```

```
anova(Model)
```

```
## Analysis of Variance Table
##
## Response: Value
##           Df Sum Sq Mean Sq F value Pr(>F)
## Treatment  1   0.33   0.333  0.0043 0.9491
## Residuals 10 777.33  77.733
```

$\hat{\sigma}$ has gone up since we are no longer accounting for the variability in the blocks.

$$\hat{\tau}_1 - \hat{\tau}_2 \pm c\frac{\hat{\sigma}}{\sqrt{3}} = (-11.7, 11.02)$$

which is much wider than the RBD. The ANOVA table for the CRD are just the sum of the Block and Residuals in the RBD. There was a benefit to using blocking. ANOVA gives us a simple test that we can use. In the RBD

ANOVA table, on the Block line we are testing

$$H_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_6 = 0$$

$$H_a\text{: At least one } \beta_j \text{ is not zero}$$

$$d = \frac{\text{MS(Block)}}{\text{MS(Res)}} = \frac{154.333}{1.133} = 136.1765$$

With p-value:

$$p = \mathbb{P}(D > 136.1765) = 2.44 \times 10^{-5}$$

We have tons of evidence to reject H_0 in favour of H_a. Since at least one β_j is not zero, RBD is a better model to use instead of CRD.

Chapter 5

Assignment 5

5.1 Lecture 30.00: Factorial Designs

EXAMPLE 5.1.1: Cancer

- Chemo (high, low).
- Radiation (high, low).

A treatment might be high chemo and low radiation.

In factorial design, we're interested in the factorials individually as well as how they *interact*. Interaction means that the effects of the factors differ when used alone differs from when they are used together.

EXAMPLE 5.1.2: Interaction

- Radiation: kills 1/4 of cancer cells
- Chemo: kills 1/4 of cancer cells
- Together: kills 5/6 cancer cells

DEFINITION 5.1.3: Factorial CRD, Model 8

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

where

- $i = 1, 2, \dots, \ell_1$ (no. of levels of factor 1)
- $j = 1, 2, \dots, \ell_2$ (no. of levels of factor 2)
- $k = 1, 2, \dots, r$

Constraint: $\sum_{ij} \tau_{ij} = 0$

EXAMPLE 5.1.4: LS for Model 8

$$W = \sum_{ijk} r_{ijk}^2 + \lambda\left(\sum_{ij} \tau_{ij}\right)$$

Solving

$$\hat{\mu} = \bar{y}_{+++}$$
$$\hat{\tau}_{ij} = \bar{y}_{ij+} - \bar{y}_{+++}$$
$$\hat{\sigma}^2 = \frac{W}{(r\ell_1\ell_2) - (\ell_1\ell_2 + 1) + 1}$$

DEFINITION 5.1.5: Factorial RBD, Model 9

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

Constraints: $\sum_{ij} \tau_{ij} = 0$ and $\sum_k \beta_k = 0$

EXAMPLE 5.1.6: LS for Model 9

$$W = \sum_{ijk} r_{ijk}^2 + \lambda_1 \sum_{ij} \tau_{ij} + \lambda_2 \sum_k \beta_k$$

Solving

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

$$\hat{\tau}_{ij} = \bar{y}_{ij+} - \bar{y}_{+++}$$

$$\hat{\beta}_k = \bar{y}_{++k} - \bar{y}_{+++}$$

$$\hat{\sigma}^2 = \frac{W}{(r\ell_1\ell_2) - (\ell_1\ell_2 + r + 1) + 2}$$

5.2 Lecture 30.50: Factorial Designs, Example

An experiment was conducted by students at Ohio State to explore the nature of the relationship between a person's heart rate and the frequency at which that person stepped up and down on steps of various heights. The response, the difference in heart rate, was measured in beats per minute. There were two different step heights, coded as 0 and 1. There were two rates of stepping coded as 0 and 1. This resulted in four possible height/frequency combinations — treatments. Each subject performed the activity for three minutes, and were kept on pace by the beat of an electric metronome.

```
rm(list = ls())
options(contrasts = c("contr.sum", "contr.poly"))
data = read.table("stepping2.csv", header = T, sep = ",", as.is = T)
attach(data)
Y = HR - RestHR
Trt = 2 * Height + Frequency
Trt = as.factor(Trt)
Model = lm(Y ~ Trt)
summary(Model)
```

```
##
## Call:
## lm(formula = Y ~ Trt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22.20  -5.10  -0.90   5.85  16.80
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19.500      2.263   8.619 2.08e-07 ***
## Trt1         -11.700      3.919  -2.986  0.00874 **
## Trt2          -0.900      3.919  -0.230  0.82126
## Trt3           3.900      3.919   0.995  0.33444
## ---
```

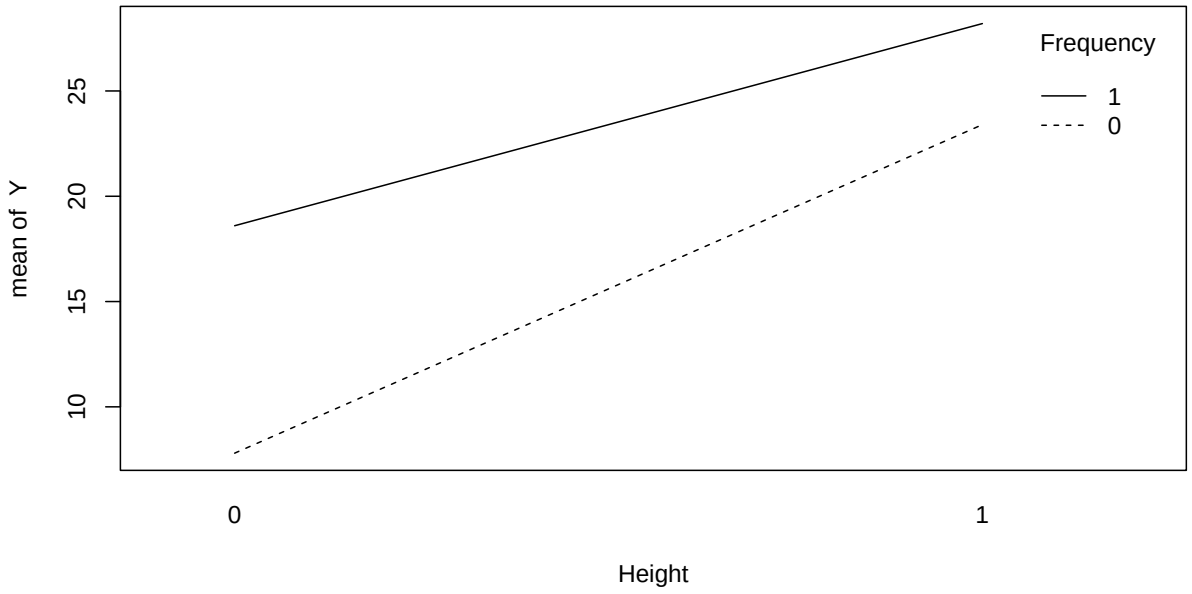


Figure 5.1: Interaction Plot

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.12 on 16 degrees of freedom
## Multiple R-squared:  0.411,Adjusted R-squared:  0.3006
## F-statistic: 3.722 on 3 and 16 DF,  p-value: 0.03332
```

5.2.1 Determining Interaction (Method 1)

Group Average	0	1
0	$19.5 - (11.7)$	$19.5 - 0.9$
1	$19.5 + 3.9$	$19.5 + 8.7$

We can only do this if we have 2 levels with 2 factors. We end up creating a contrast. Figure 5.1 shows that there is no interaction.

```
interaction.plot(Height, Frequency, Y)
```

5.2.2 Determining Interaction (Method 2)

- If $\Delta_1 = \Delta_2$, the lines are parallel; that is, there is no interaction.
- If $\Delta_1 \neq \Delta_2$, the lines are not parallel; that is, there is interaction.

$\Delta_1 - \Delta_2 = (\bar{Y}_{11+} - \bar{Y}_{01+}) - (\bar{Y}_{10+} - \bar{Y}_{00+})$. Add tildes to get $\tilde{\theta} = (\tilde{\tau}_{11} - \tilde{\tau}_{01}) - (\tilde{\tau}_{10} - \tilde{\tau}_{00})$

$$\mathbb{E}[\tilde{\theta}] = \tau_{11} - \tau_{01} - \tau_{10} - \tau_{00}$$

$$\mathbb{V}(\tilde{\theta}) = \frac{\sigma^2}{5} + \frac{\sigma^2}{5} + \frac{\sigma^2}{5} + \frac{\sigma^2}{5} = \frac{4\sigma^2}{5}$$

H_0: $\theta = 0$ (no interaction) versus H_a: $\theta \neq 0$ (interaction)

$$d = \frac{\hat{\tau}_{11} - \hat{\tau}_{01} - \hat{\tau}_{10} - \hat{\tau}_{00}}{\sigma\sqrt{\frac{4}{5}}} = -0.66 \quad (D \sim t(20 - 4 - 1 + 1) = t(16))$$

$$p = 2\,\mathbb{P}(D > 0.66) = (0.4, 0.6)$$

There is no evidence to reject H_0. Therefore, there is no interaction.

There is actually a *third* way to determine interaction, the ANOVA table.

5.2.3 Determining Interaction (Method 3)

```
rm(list = ls())
options(contrasts = c("contr.sum", "contr.poly"))
data = read.table("stepping2.csv", header = T, sep = ",", as.is = T)
attach(data)
Y = HR - RestHR
Height = as.factor(Height)
Freq = as.factor(Frequency)
Model = lm(Y ~ Freq + Height + Freq * Height)
anova(Model)
```

```
## Analysis of Variance Table
##
## Response: Y
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Freq         1  304.2  304.20  2.9714 0.10401
## Height       1  793.8  793.80  7.7538 0.01326 *
## Freq:Height  1   45.0   45.00  0.4396 0.51677
## Residuals   16 1638.0  102.37
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We calculate SS(Trt) with

$$\text{SS(Freq)} + \text{SS(Height)} + \text{SS(Interaction)} = \text{SS(Trt)}$$

In our case, SS(Trt) $= 1143.0$ on 3 degrees of freedom.

- H_0: no interaction
- H_a: interaction

$$d = \frac{\text{MS(Int)}}{\text{MS(Res)}} = \frac{45}{102.37} = 0.4396$$

With p-value,

$$p = 0.51667 \quad (D \sim F(1, 16))$$

$p > 0.1$, so there is no evidence to reject H_0, so it appears there is no interaction.

Chapter 6

Assignment 6

6.1 Lecture 31.00: Sampling

Let

- U be the **frame** (study population).
- $U = \{1, 2, \dots, N\}$.
- $\mathcal{S}$ be our sample. It has size $n \leq N$ and $\mathcal{S} \subset U$.

Also, let the **sampling protocol** refer the probability of selecting any particular sample.

- Define π_{ij} to be the **inclusion probability** for unit i and j. (note: $\pi_{ii} = \pi_i$)

SRSWOR

- Simple Random Sampling without Replacement.
- $U = \{1, 2, 3, 4\}$ we select samples of size $n = 2$.
- $\mathcal{S}_1 = \{1, 2\}$, $\mathcal{S}_2 = \{1, 3\}$, $\mathcal{S}_3 = \{1, 4\}$, $\mathcal{S}_4 = \{2, 3\}$, $\mathcal{S}_5 = \{2, 4\}$, $\mathcal{S}_6 = \{3, 4\}$
- What is the probability we select sample 1?

$$\mathbb{P}(\mathcal{S}_1) = \frac{1}{6}$$

- What is the probability that unit 1 is in our sample? That is, what is π_1? Well, $\pi_1 = 3/6 = 1/2$.

Let the **frame** be $\{1, 2, \dots, N\}$. We select SRSWOR a sample of size n.

$$\mathbb{P}(\mathcal{S}_1) = \frac{1}{\binom{N}{n}}$$

$$\pi_1 = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

6.2 Lecture 32.00: Model 1 Revisited

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

- Parameters: μ, and σ^2
- Estimates: $\hat{\mu} = \bar{y}_+$, and $\hat{\sigma}^2 = s^2$

- Estimators: $\tilde{\mu} = \bar{Y}_+$, and $\tilde{\sigma}^2 = S^2$, where

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

- CI: EST $\pm\, c$ SE. For μ:

$$\hat{\mu} \pm c\frac{\hat{\sigma}}{\sqrt{n}} \quad (c \sim t(n-1))$$

Use SRS (simple random sampling without replacement):

- Parameters:

$$\mu = \frac{\sum_{i=1}^{N} y_i}{N}$$

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

- Estimates:

$$\hat{\mu} = \frac{\sum_{i\in\mathcal{S}} y_i}{n}$$

$$\hat{\sigma}^2 = \frac{\sum_{i\in\mathcal{S}}(y_i - \hat{\mu})^2}{n-1}$$

- Estimator: y_i is not a realization of Y_i. Instead, y_i is a constant. What is random, is whether y_i is selected for the sample.

$$I_i = \begin{cases} 0 & \text{if } y_i \text{ is not in the sample} \\ 1 & \text{if } y_i \text{ is in the sample} \end{cases}$$

$$\tilde{\mu} = \frac{\sum_{i=1}^{N} I_i y_i}{n}$$

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{N} I_i (y_i - \hat{\mu})^2}{n-1}$$

$$\tilde{\mu} \sim \mathcal{N}\left(\mu, \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}\right)$$

  where
  - $\frac{n}{N}$ is the sampling fraction
  - $1 - \frac{n}{N}$ is the finite population correction.

Model 1	SRS
$\hat{\mu} \pm c\frac{\hat{\sigma}}{\sqrt{n}}$	$\hat{\mu} \pm c\sqrt{1 - \frac{n}{N}}\frac{\hat{\sigma}}{\sqrt{n}}$

Let's prove the distribution.

- $\mathbb{P}(I_i = 1) = \pi_i = \frac{n}{N}$
- $\mathbb{E}[I_i] = (0)\,\mathbb{P}(I_i = 0) + (1)\,\mathbb{P}(I_i = 1) = \frac{n}{N} = \mathbb{E}[I_i^2]$
- $\mathbb{V}(I_i) = \mathbb{E}[I_i^2] - \mathbb{E}[I_i]^2 = \frac{n}{N} - \left(\frac{n}{N}\right)^2 = \frac{n}{N}\left(1 - \frac{n}{N}\right)$

- $\mathbb{E}[I_i I_j] = (1)(1)\,\mathbb{P}(I_i = 1, I_j = 1) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$

Recall:

$$\tilde{\mu} = \frac{\sum_{i=1}^{N} I_i y_i}{n}$$

$$\mathbb{E}[\tilde{\mu}] = \frac{\sum_{i=1}^{N} \mathbb{E}[I_i] y_i}{n} = \frac{\sum_{i=1}^{N} (n/N) y_i}{n} = \mu$$

We note that I_i is not independent of I_j, so we must compute covariance in our variance calculation.

$$\mathbb{V}(\tilde{\mu}) = \mathbb{V}\left(\frac{\sum_{i=1}^{N} I_i y_i}{n}\right) = \frac{\sum_{i=1}^{N} y_i^2\, \mathbb{V}(I_i)}{n^2} + \frac{\sum_{ij} y_i y_j \mathsf{Cov}(I_i, I_j)}{n^2} = \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}$$

6.3 Lecture 33.00: Sample Size Calculation

Model 1: $\hat{\mu} \pm \frac{c\sigma}{\sqrt{n}}$ where σ is known.

Set $E = \frac{c\sigma}{\sqrt{n}}$ and solve for n. Therefore, $n = \frac{c^2\sigma^2}{E^2}$.

Process

- First, we take a small sample, then estimate σ.
- Find n.
- Perform a large study with n units.

For SRS for our mean we have:

$$\hat{\mu} \pm \underbrace{\frac{c\hat{\sigma}}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}}_{E}$$

$$n = \left(\frac{E^2}{c^2\hat{\sigma}^2} + \frac{1}{N}\right)^{-1}$$

EXAMPLE 6.3.1: SRSWOR Example 1

Part 1. Assume our class has 200 students in it. I draw a sample of 5 students to find the average on midterm 2 is 65% with a standard deviation of 3%. Build a 95% confidence interval for μ. Please assume that n is "large" enough to apply the normality assumption.
Solution.

$$\begin{aligned}
&\hat{\mu} \pm \frac{c\hat{\sigma}}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} \quad c \sim \mathcal{N}(0,1) \\
&= 65 \pm \frac{1.96(3)}{\sqrt{5}}\sqrt{1 - \frac{5}{200}} \\
&= (0.62, 0.68)
\end{aligned}$$

The width is $0.68 - 0.62 \approx 0.06$.
Part 2. If I want to be accurate to within 0.1, 19 times out of 20 how large should n be?

- $E = 0.1$.
- $\frac{19}{20} = 0.95 \implies 95\%$.
- $\sigma = 3$.
- $\mu = 65$.

SRS:

$$n = \left(\frac{E^2}{c^2\hat{\sigma}^2} + \frac{1}{N}\right)^{-1} = \left(\frac{0.1^2}{1.96^2 3^2} + \frac{1}{200}\right)^{-1} = 189.0634 = \uparrow 190$$

Model 1: Assumes $N \to \infty$.

$$n = \frac{c^2\hat{\sigma}^2}{E^2} = 3457.44 = \uparrow 3458$$

6.4 Lecture 34.00: Model 4 Revisited

Model 4: $\frac{Y_i}{n} \sim \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$ with confidence interval:

$$\pi : \hat{\pi} \pm c\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

SRS for π

- SP Parameter:

$$\pi = \frac{\sum_{i=1}^{N} y_i}{N} \quad (y_i = \{0, 1\})$$

- Statistic:

$$\hat{\pi} = \frac{\sum_{i\in\mathcal{S}} y_i}{n} = \bar{y}$$

$$\hat{\sigma}^2 = \frac{\sum_{i\in\mathcal{S}}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i\in\mathcal{S}}(y_i^2 + \bar{y}^2 - 2y_i\bar{y})}{n-1}$$

Recall that $y_i = \{0, 1\}$. Therefore,

$$\hat{\sigma}^2 = \frac{\sum_{i\in\mathcal{S}}(y_i + \bar{y}^2 - 2y_i\bar{y})}{n-1} = \frac{n\bar{y} + n\bar{y}^2 - 2\bar{y}n\bar{y}}{n-1} = \frac{n}{n-1}(\bar{y} - \bar{y}^2) = \frac{n}{n-1}[\hat{\pi}(1-\hat{\pi})]$$

since $\bar{y} = \hat{\pi}$. Now, assume that $n \to \infty$. Therefore,

$$\hat{\sigma}^2 = \hat{\pi}(1-\hat{\pi})$$

- Estimators:

$$\tilde{\pi} = \frac{\sum_{i=1}^{N} y_i I_i}{n}$$

EXERCISE 6.4.1

In SRS, clearly show that $\tilde{\pi}$ is an unbiased estimator for π.
Solution. In SRS, $\tilde{\pi}$ is defined by:

$$\tilde{\pi} = \frac{\sum_{i=1}^{N} y_i I_i}{n} \quad \text{where} \quad I_i = \begin{cases} 1 & \text{if } y_i \text{ is in the sample} \\ 0 & \text{if } y_i \text{ is not in the sample} \end{cases} \quad \text{and} \quad y_i = 0 \text{ or } 1.$$

In `Lec 32.00: Model 1 Revisited`, we derived that $\mathbb{E}[I_i] = n/N$. Hence,

$$\mathbb{E}[\tilde{\pi}] = \mathbb{E}\left[\frac{\sum_{i=1}^{N} y_i I_i}{n}\right] = \frac{\sum_{i=1}^{N} y_i\,\mathbb{E}[I_i]}{n} = \frac{\sum_{i=1}^{N} y_i(n/N)}{n} = \frac{\sum_{i=1}^{N} y_i}{N} = \pi$$

Therefore, in SRS $\tilde{\pi}$ is an unbiased estimator for π.

REMARK 6.4.2

$\tilde{\pi}$ is normal.

$$\mathbb{V}(\tilde{\pi}) = \mathbb{V}\left(\frac{\sum_{i=1}^{N} y_i I_i}{n}\right) = \cdots = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)$$

Replace σ by $\hat{\sigma}^2 = \hat{\pi}(1-\hat{\pi})$. Our confidence interval is:

$$\hat{\pi} \pm c\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}\left(1 - \frac{n}{N}\right)}$$

Model 4:

$$\hat{\pi} \pm c\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Sample Size Calculation

$$n = \left(\frac{E^2}{\sigma^2 c^2} + \frac{1}{N}\right)^{-1}$$

However, $\hat{\sigma}^2 = \hat{\pi}(1-\hat{\pi})$. Often, we replace $\hat{\sigma}^2$ by $1/4$ which is the maximum.

6.5 Lecture 35.00: SRS Examples

EXAMPLE 6.5.1: SRSWOR Example 2

According to a new poll conducted by Ipsos Reid on behalf of Postmedia News and Global Television, 42%, 'approve' of the performance of the Conservative government under the leadership of Stephen Harper. For this survey, a sample of 1053 Canadians, from Ipsos' Canadian online panel was interviewed online. This result is 3% lower than last years' results.
Is the difference between this year and last years' results significant; that is, is there a difference? Use a confidence interval with a 95% level of confidence to answer the question.
Solution.

$$\hat{\pi} \pm c\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.42 \pm 1.96\sqrt{\frac{0.42(1-0.42)}{1053}} = (0.39, 0.45)$$

Now, 0.45 is in the interval (it's at the edge which is fine for our purposes), so there is no difference between this year and last years' results; that is, they are the same.

EXAMPLE 6.5.2: SRSWOR Example 3

Jeff Henry, a counsellor for Waterloo wants to know how many people he should poll so that with 95% confidence his poll ("will you vote for me?"), is accurate with a margin of error of 1%. There are $150\,000$ people in his Waterloo riding.
Solution.

- $c = 1.96$.
- $E = 0.01$.
- $N = 150\,000$.
- We should assume the worst-case scenario for the proportion; that is, $\hat{\pi} = 1/2$.
- $\hat{\sigma}^2 = \hat{\pi}(1-\hat{\pi})$.

$$n = \left(\frac{E^2}{\hat{\sigma}^2 c^2} + \frac{1}{N}\right)^{-1} = 9026.09 = \uparrow 9027$$

EXAMPLE 6.5.3: SRSWOR Example 4

Sheila is an auditor. She has taken a sample of 15 accounts across a large company to see whether the company is being compliant (i.e., following accounting laws). In these she has found that the amount of misstated account values is, on average, $\$143.95$. The variance of her sample values is $\$81.09$. If there are a total of 200 accounts to look at, and her auditing company allows a level of non-compliance up to $\$25\,000$ dollars, then is this company being compliant? Make your decision at a 90% level of confidence.

Solution.

- $c = 1.645$.
- $\hat{\mu} = 143.95$.
- $\hat{\sigma}^2 = 81.09$.
- $N = 200$.
- $n = 15$.

$$\hat{\mu} \pm \frac{c\hat{\sigma}}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} = (140.27, 147.63)$$

On average the discrepancy is $(140.27, 147.63)$ with 90% confidence. However,

$$200(140.27, 147.63) = (28\,054, 29\,526) > 25\,000$$

We can say that the company is not being compliant.

Chapter 7

Assignment 7

7.1 Lecture 36.00: Regression Sampling

We want our parameter to be

$$\mu_y = \frac{\sum_{i=1}^{N} y_i}{N}$$

which is our population average.

We use

$$\hat{\mu}_y = \frac{\sum_{i \in \mathcal{S}} y_i}{n} = \bar{y}$$

which is our sample average.

Suppose Y_i is linearly related to a *continuous* explanatory variate called x_i. If that's the case, x_i has its own population average

$$\mu_x = \frac{\sum_{i=1}^{N} x_i}{N}$$

with sample mean

$$\hat{\mu}_x = \frac{\sum_{i \in \mathcal{S}} x_i}{n} = \bar{x}$$

Suppose we have a linear relationship of the form $Y_i = \alpha + \beta(x_i - \bar{x}) + R_i$ where $R_i \sim \mathcal{N}(0, \sigma^2)$.

We use least squares,

$$W = \sum_i r_i^2 = \sum_i \left[y_i - \alpha - \beta(x_i - \bar{x}) \right]^2$$

We find $\frac{\partial W}{\partial \alpha}$ and $\frac{\partial W}{\partial \beta}$, and you can show this for homework:

$$\hat{\alpha} = \bar{y}$$
$$\hat{\beta} = \frac{\sum_i y_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = \frac{s_{xy}}{s_x^2}$$

where

$$S_{xy} = \sum_i y_i(x_i - \bar{x}) = \sum_i (y_i - \bar{y})(x_i - \bar{x})$$
$$s_{xy} = \frac{S_{xy}}{n-1}$$
$$S_{xx} = \sum_i (x_i - \bar{x})^2$$
$$s_x^2 = \frac{S_{xx}}{n-1}$$

We had $y_i = \alpha + \beta(x_i - \bar{x}) + R_i$. We used least squares to estimate α and β to obtain (ignoring the R_i term)

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}(x_i - \bar{x})$$

- If $x_i = \bar{x}$, then $\hat{y}_i = \hat{\alpha} = \bar{y}$.
- If $x_i = \mu_x$, then $\hat{y}_i = \hat{\alpha} + \hat{\beta}(\mu_x - \bar{x}) = \hat{\mu}_{\text{reg}}$.

$$\boxed{\hat{\mu}_{\text{reg}} = \hat{\alpha} + \hat{\beta}(\mu_x - \bar{x})}$$

Estimators

The $\alpha, \beta, \mu_x, \mu_y$ estimators are all unbiased. However,

$$\begin{aligned}
\hat{\mu}_{\text{reg}} &= \hat{\alpha} + \hat{\beta}(\mu_x - \bar{x}) \\
&= \hat{\alpha} - \hat{\beta}(\bar{x} - \mu_x) \\
&= \bar{y} - \hat{\beta}(\bar{x} - \mu_x) \\
&= \frac{\sum_{i \in \mathcal{S}} y_i}{n} - \hat{\beta}\left(\frac{\sum_{i \in \mathcal{S}} x_i}{n} - \frac{n\mu_x}{n}\right) \\
&= \frac{\sum_{i \in \mathcal{S}} [y_i - \hat{\beta}(x_i - \mu_x)]}{n} \\
&= \frac{\sum_{i \in \mathcal{S}} r_i}{n}
\end{aligned}$$

$$\tilde{\mu}_{\text{reg}} = \frac{\sum_{i=1}^N I_i r_i}{n}$$

We're interested in three things for $\tilde{\mu}_{\text{reg}}$:

- Distribution. We're not going into the details, but we get that $\tilde{\mu}_{\text{reg}}$ is normally distributed.
- Expected Value.
- Variance.

Expected Value and Variance of $\tilde{\mu}_{\text{reg}}$

$$\begin{aligned}
\mathbb{E}\big[\tilde{\mu}_{\text{reg}}\big] &= \mathbb{E}\big[\tilde{\alpha} + \tilde{\beta}(\tilde{\mu}_x - \mu_x)\big] \\
&= \mathbb{E}\big[\tilde{\mu}_y + \tilde{\beta}(\tilde{\mu}_x - \mu_x)\big] \\
&= \mu_y + \underbrace{\mathbb{E}\big[\tilde{\beta}(\tilde{\mu}_x - \mu_x)\big]}_{\text{small}}
\end{aligned}$$

Therefore, $\tilde{\mu}_{\text{reg}}$ is a biased estimator for μ_y.

$$\begin{aligned}\mathbb{V}(\tilde{\mu}_{\text{reg}}) &= \mathbb{V}\left(\frac{\sum_{i=1}^{N} I_i r_i}{n}\right) \\ &= \left(1 - \frac{n}{N}\right)\frac{\sigma_r^2}{n}\end{aligned}$$

We estimate σ_r^2 by

$$\hat{\sigma}_r^2 = \frac{\sum_{i \in \mathcal{S}}(r_i - \bar{r})^2}{n-1} = \cdots = \frac{W}{n-1}$$

The confidence interval is:

$$\hat{\mu}_{\text{reg}} \pm c\sqrt{1 - \frac{n}{N}}\frac{\hat{\sigma}_r}{\sqrt{n}}$$

7.2 Lecture 37.00: Regression Sampling, Example

- In R the data set is `women`. Simply type `women` to see the data.
- We assume this is our population and that we want to know the mean height μ_{height}.
- When you do regression sampling you need to have a y and an x.
- y: height.
- x: weight.
- Now when we talk about our x being weight we have to assume that we know the mean weight μ_{weight}; that is, you need to know the population value for your weight. You don't know your population value for your height that's what you're trying to build the interval about.

```
attach(women)
mean(height)
```

```
## [1] 65
```

```
mean(weight)
```

```
## [1] 136.7333
```

- $\mu_{\text{height}} = 65$ is unknown!
- $\mu_{\text{weight}} \approx 136.7333$ is known, and must be known to do regression sampling.

We're almost there where we have everything we need. Once we get everything we need, we can build an SRS confidence interval. We need one more thing, and that's getting a sample. The sample command below grabs five `heights` from the set of heights that are there. So it grabs five of them, and then we can get the mean of the sample height and the standard deviation of the sample heights, so this would be sigma hat for simple random sampling.

Using SRSWOR, we take a sample of size 5 and use this as our estimate for the height.

```
set.seed(45376)
sample_heights = sample(height, 5)
mean(sample_heights)
```

```
## [1] 63.4
```

```
sd(sample_heights)
```

```
## [1] 3.209361
```

- $\hat{\mu}_{\text{height}} = 63.4$.
- $\hat{\sigma}_{\text{SRS}} \approx 3.209361$.

Now we have enough information that we can actually build a confidence interval.

```
N <- nrow(women)
print(N)
```

```
## [1] 15
```

```
n <- 5
c <- qnorm(0.975)
round(mean(sample_heights) + c(-1, 1) * ((c * sd(sample_heights))/sqrt(n)) *
  sqrt(1 - n/N), 1)
```

```
## [1] 61.1 65.7
```

- $N = 15$.
- $n = 5$.
- $c \approx 1.96$.

$$\text{SRS: } \hat{\mu}_{\text{height}} \pm \frac{c\hat{\sigma}_{\text{SRS}}}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} = 63.4 \pm \frac{1.96(3.209361)}{\sqrt{5}}\sqrt{1 - \frac{5}{15}} = (61.1, 65.7)$$

which has a width of 4.6.

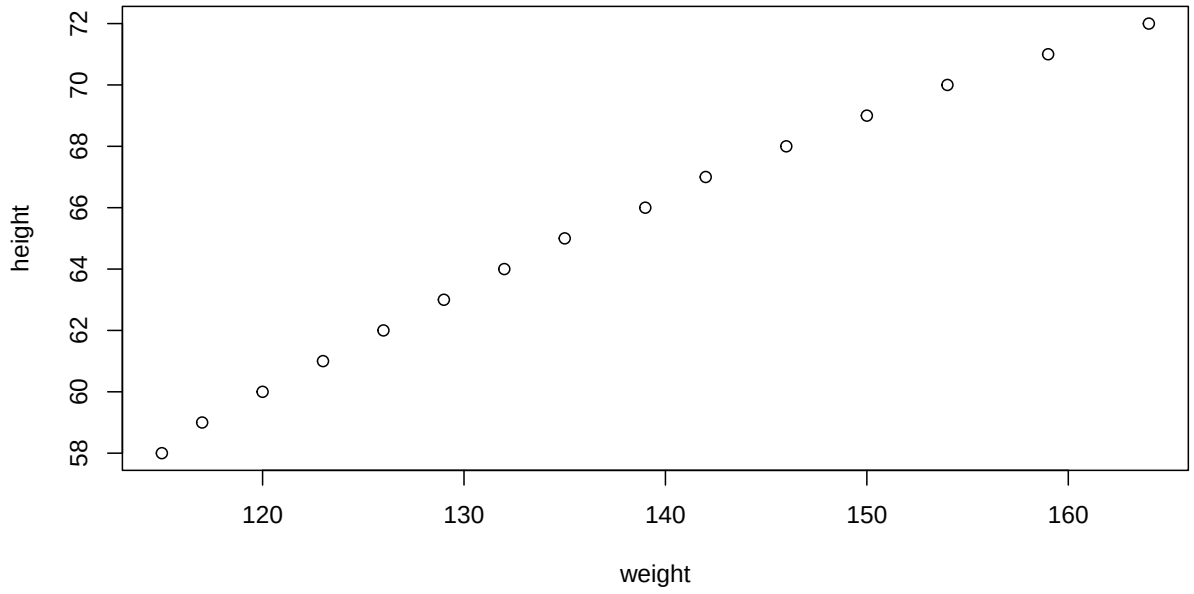
```
sample_weights = c(123, 129, 135, 146, 120)
mean(sample_weights)
```

```
## [1] 130.6
```

- $\hat{\mu}_{\text{weight}} = 130.6$.

We are wrong by $\mu_y - \hat{\mu}_y = 65 - 63.4 = 1.6$ units. We note that there is a linear relationship between height and weight.

```
plot(weight, height)
```



Thus, we decide to use Regression Sampling.

```
sample_weights = sample_weights - mean(sample_weights)  # x_i - bar(x)
summary(lm(sample_heights ~ sample_weights))  # Y_i ~ (x_i - bar(x))

##
## Call:
## lm(formula = sample_heights ~ sample_weights)
##
## Residuals:
##       1       2       3       4       5
## -0.04846  0.09506  0.23858 -0.16496 -0.12022
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    63.400000   0.085624  740.45 5.43e-09 ***
## sample_weights  0.309413   0.009242   33.48 5.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1915 on 3 degrees of freedom
## Multiple R-squared:  0.9973,Adjusted R-squared:  0.9964
## F-statistic:  1121 on 1 and 3 DF,  p-value: 5.858e-05
```

We didn't use a factor because this is not a discrete variable. We consider this factor to be continuous, so we consider our weights to be a continuous value.

Therefore,

- $\hat{\alpha} = \hat{\mu}_y = 63.4$.
- $\hat{\beta} = 0.309413$.

Right now, the degrees of freedom is $n-2=5-2=3$, but we want to multiply by $(n-2)/(n-1)$ as the degrees of freedom should be $n-1=5-1=4$, so

```
sigma_r <- summary(lm(formula = sample_heights ~ sample_weights))$sigma
print(sigma_r)
```

```
## [1] 0.1914611
```

```
sigma_r_sq <- sigma_r^2 * (n - 2)/(n - 1)
print(sigma_r_sq)
```

```
## [1] 0.02749301
```

$$\hat{\sigma}_r^2 = \hat{\sigma}_{\text{r}}^2 \frac{3}{4} = 0.1915^2(3/4) = 0.02749$$

$$\hat{\mu}_{\text{reg}} = \hat{\mu}_{\text{height}} = \hat{\alpha} + \hat{\beta}(x_i - \bar{x}) = 63.4 - 0.31(x_i - 130.6)$$

```
alpha_hat <- summary(lm(formula = sample_heights ~ sample_weights))$coefficients[1]
beta_hat <- summary(lm(formula = sample_heights ~ sample_weights))$coefficients[2]
reg <- mean(sample_heights) + beta_hat * (mean(weight) - mean(c(123,
  129, 135, 146, 120)))
print(reg)
```

```
## [1] 65.29773
```

The regression estimate is:

$$\hat{\mu}_{\text{reg}} = \hat{\mu}_{\text{height}}(\mu_{\text{weight}}) = 63.4 + 0.31(136.7333 - 130.6) = 65.3$$

```
round(reg + c(-1, 1) * c * sqrt(sigma_r_sq)/sqrt(5) * (1 - n/N),
  1)
```

```
## [1] 65.2 65.4
```

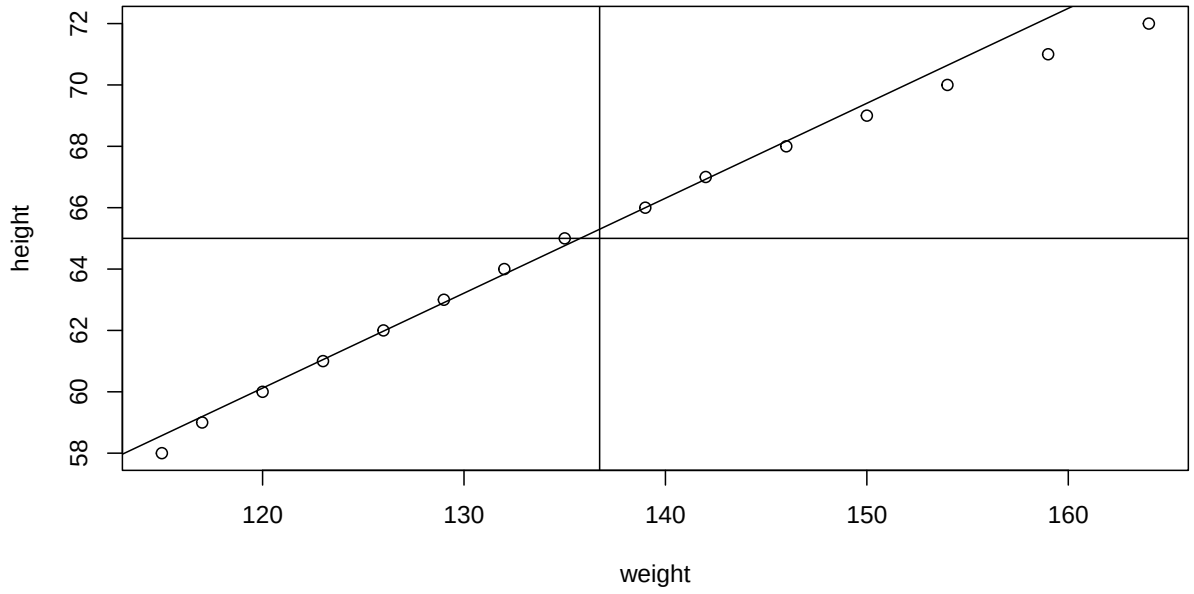
The confidence interval is:

$$\hat{\mu}_{\text{reg}} \pm \frac{c\hat{\sigma}_r}{\sqrt{n}}\sqrt{1-\frac{n}{N}} = 65.3 \pm \frac{1.96\sqrt{0.02749}}{\sqrt{5}}\sqrt{1-\frac{5}{15}} = (65.2, 65.4)$$

In this case, you are only 0.3 from the true mean.

- The width of this interval is much narrower than that of the SRS. In fact, for the SRS if we go back in time it had a width of 4.6.
- Note that your interval does not actually contain the population mean height. The population mean height is 65, but it's not in your interval and that's because of the bias that comes from a regression interval. So, the bias of a regression interval means that we may not always contain the actual value of interest. You won't be far off from it because of the regression line, but your interval might not contain it.

```
plot(weight, height)
abline(h = mean(height))
abline(v = mean(weight))
abline(alpha_hat - beta_hat * mean(c(123, 129, 135, 146, 120)),
  beta_hat)
```



7.3 Lecture 38.00: Regression Sampling, Example 2

A student was curious about whether they performed better with more sleep. To test this hypothesis, she decided to write various tests on a certain number of hours (x) of sleep. The grade on their test was considered to be the response (y). In total, she has written 94 tests since coming to UW. On average, she slept for 5.1 hours during those 94 tests. We consider the 9 below to be a random sample.

- **Model**: $Y_i = \alpha + \beta(x_i - \bar{x}) + R_i$ where $R_i \sim \mathcal{N}(0, \sigma^2)$.

Questions:

- Assume the explanatory variate was not present. Build a 95interval for her mean grade using SRSWOR.
- Use Regression sampling to build a 95
- Compare your SRSWOR results to your regression results, what do you notice?

```
x <- c(4, 6, 2, 7, 5, 9, 2, 1, 8)
y <- c(75, 78, 69, 80, 77, 82, 65, 55, 85)
n <- 9
N <- 94
mu_x <- 5.1
s_xy <- sum(y * (x - mean(x)))/(n - 1)
print(s_xy)
```

```
## [1] 24.75
```

```
s_xsq <- var(x)
print(s_xsq)
```

```
## [1] 8.111111
```

```
s_ysq <- var(y)
print(s_ysq)
```

```
## [1] 89.25
```

```
xbar <- mean(x)
print(xbar)
```

```
## [1] 4.888889
```

```
ybar <- mean(y)
print(ybar)
```

```
## [1] 74
```

```
r <- (y - ybar - (x - xbar) * sum((y - ybar) * (x - xbar))/sum((x -
  xbar)^2))
sigma_rsq <- sum(r^2)/(n - 1)
print(sigma_rsq)
```

```
## [1] 13.7286
```

```
sqrt(sigma_rsq)
```

```
## [1] 3.705212
```

- $s_{xy} = 24.75$.
- $s_x^2 = 8.11$.
- $\hat{\sigma}_y^2 = s_y^2 = 89.25$.
- $\bar{x} = 4.89$.
- $\bar{y} = 74$.
- $\hat{\sigma}_{\text{reg}}^2 = 13.72$.
- $N = 94$.
- $n = 9$.
- $\mu_x = 5.1$ (given).

```
alpha_hat <- ybar
print(alpha_hat)
```

```
## [1] 74
```

```
beta_hat <- s_xy/s_xsq
print(beta_hat)
```

```
## [1] 3.05137
```

```
reg <- alpha_hat + beta_hat * (mu_x - xbar)
print(reg)
```

```
## [1] 74.64418
```

$$\hat{\alpha} = \bar{y} = \hat{\mu}_y = 74$$

$$\hat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{24.75}{8.11} = 3.0514$$

$$\hat{\mu}_{\text{reg}} = \hat{\alpha} + \hat{\beta}(\mu_x - \bar{x}) = 74 + 3(5.1 - 4.89) = 74.63$$

```
c <- qnorm(0.975)
round(alpha_hat + c(-1, 1) * c * sqrt(s_ysq/n) * sqrt(1 - n/N),
  1)
```

```
## [1] 68.1 79.9
```

SRS:

$$\hat{\mu}_y \pm \frac{c\hat{\sigma}_y}{\sqrt{n}}\sqrt{1-\frac{n}{N}} = 74 \pm 1.96\sqrt{\frac{89.25}{9}}\sqrt{1-\frac{9}{94}} = (68.1, 79.9)$$

Roughly a width of 12.

```
round(reg + c(-1, 1) * c * sqrt(sigma_rsq/n) * sqrt(1 - n/N),
  1)
```

```
## [1] 72.3 76.9
```

Reg:

$$\hat{\mu}_{\text{reg}} \pm \frac{c\hat{\sigma}_r}{\sqrt{n}}\sqrt{1-\frac{n}{N}} = 74.63 \pm 1.96\sqrt{\frac{13.72}{9}}\sqrt{1-\frac{9}{94}} = (72.3, 76.9)$$

Roughly a width of 5.

As you will notice, the big difference between the two intervals is that the width of the regression interval is narrower than the width of the SRS interval.

Chapter 8

Assignment 8

8.1 Lecture 39.00: Ratio Estimation (Ave.)

Model:

$$Y_i = \beta x_i + R_i \qquad \text{where } R_i \sim \mathcal{N}(0, x_i\sigma^2)$$

Divide by $\sqrt{x_i}$ $\implies$

$$\frac{Y_i}{\sqrt{x_i}} = \frac{\beta x_i}{\sqrt{x_i}} + \frac{R_i}{\sqrt{x_i}} \qquad \text{where } \frac{R_i}{\sqrt{x_i}} \sim \mathcal{N}(0, \sigma^2)$$

- It goes through the origin.
- Variance that increases with x_i. In fact, there is a funnel effect. To fix the funnel effect, we divide by $\sqrt{x_i}$ as done above.

Therefore, $\mathbb{E}\left[\frac{R_i}{\sqrt{x_i}}\right] = 0$ since $\mathbb{E}[R_i] = 0$ and x_i is a constant. Also,

$$\mathbb{V}\left(\frac{R_i}{\sqrt{x_i}}\right) = \frac{\mathbb{V}(R_i)}{x_i} = \frac{\sigma^2 x_i}{x_i} = \sigma^2$$

Let $Y_i' = \frac{Y_i}{\sqrt{x_i}}$, $x_i' = \frac{x_i}{\sqrt{x_i}}$, and $R_i' = \frac{R_i}{\sqrt{x_i}}$. So, our new model is

$$Y_i' = \beta x_i' + R_i' \qquad \text{where } R_i' \sim \mathcal{N}(0, \sigma^2)$$

We use LS to estimate our parameters. We find

$$\hat{\beta} = \frac{\bar{y}}{\bar{x}}$$
$$\hat{\sigma}^2_{\text{ratio}} = \frac{W}{n-1}$$

Our prediction is:

$$\hat{y}_i = \hat{\beta} x_i' = \left(\frac{\bar{y}}{\bar{x}}\right) x_i'$$

- If $x_i' = \bar{x}$, then $\hat{y}_i = \bar{y}$. Then, if $x_i' = \mu_x$, then $\hat{\mu}_{\text{ratio}} = \left(\frac{\bar{y}}{\bar{x}}\right)\mu_x$.

Now, we're going out, and we're going to build a confidence interval for this. And when we build a confidence interval for this, we basically get the same logic that we got for regression sampling. All the same mathematics

basically kicks in, there's no real difference between the mathematics, so I'm simply going to state it. A confidence interval for μ_y is:

$$\hat{\mu}_{\text{ratio}} \pm \frac{c\hat{\sigma}_{\text{ratio}}}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

8.2 Lecture 40.00: Ratio Estimation (Ave.), Example

In R, the data set is `women`. Simply type `women` to see the data. We assume this is our population, and that we want to know the mean height μ_{height}. We also assume we know the mean weight μ_{weight}. In fact, directly from the data we have:

```
attach(women)
mu_height <- mean(height)
print(mu_height)
```

```
## [1] 65
```

```
mu_weight <- mean(weight)
print(mu_weight)
```

```
## [1] 136.7333
```

- $\mu_{\text{height}} = 65$
- $\mu_{\text{weight}} = 136.7333$
- y: `height`.
- x: `weight`.

Using SRSWOR, we take a sample of size 5 and use this for our estimate for the height:

```
set.seed(45376)
sample_heights = sample(height, 5)
muhat_height <- mean(sample_heights)
print(muhat_height)
```

```
## [1] 63.4
```

```
sd(sample_heights)
```

```
## [1] 3.209361
```

```
sample_weights = c(123, 129, 135, 146, 120)
muhat_weight <- mean(sample_weights)
print(muhat_weight)
```

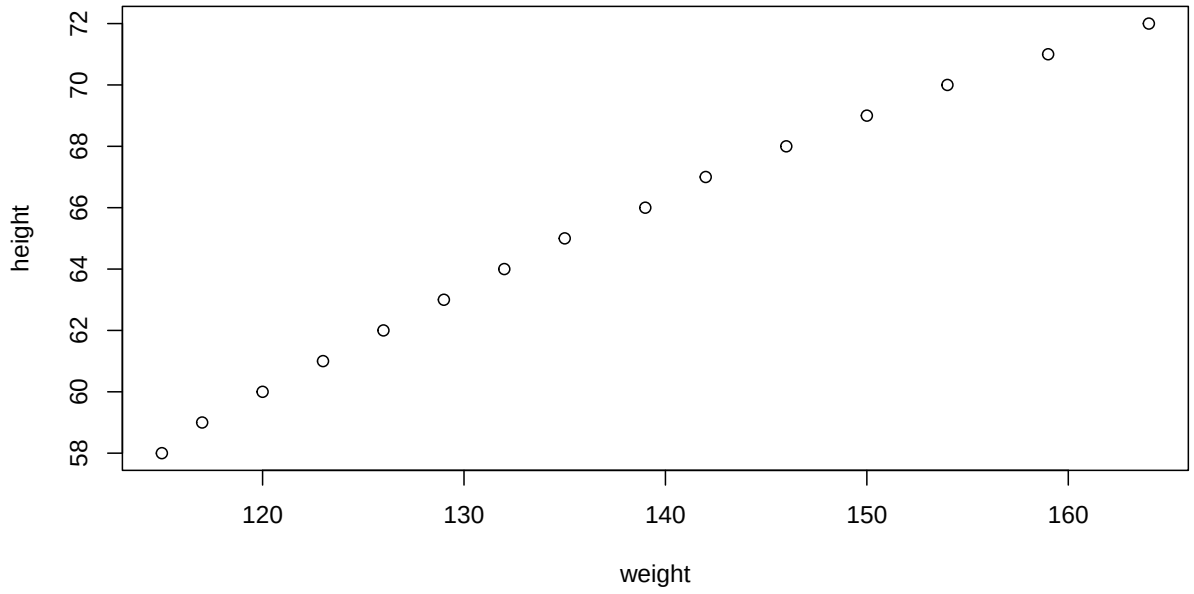
```
## [1] 130.6
```

- $\hat{\mu}_{\text{height}} = 63.4$ which is our SRS estimate for μ_y.
- $\hat{\sigma}_y = 3.209361$.
- $\bar{x} = \hat{\mu}_{\text{weight}} = 130.6$.

We found out that we were wrong by 1.4 units. Going back a long time ago, when we used SRS, we ended up with a confidence interval which was $(60.6, 66.2)$. We noticed how wide it was.

When we deal with ratio estimation, the first thing we want is a linear relationship between height and weight.

```
plot(weight, height)
```



Thus, we decide to use Ratio Sampling.

```
Sqrt_weights = sqrt(sample_weights)
sample_weights = sample_weights/Sqrt_weights
sample_heights = sample_heights/Sqrt_weights
```

- `Sqrt_weights = sqrt(sample_weights)` $= \sqrt{x_i}$.
- `sample_weights = sample_weights / Sqrt_weights` $= x_i/\sqrt{x_i}$.
- `sample_heights = sample_heights / Sqrt_weights` $= y_i/\sqrt{x_i}$.

In order to remove the intercept, we need to use the `-1` in the following code.

```
sum <- summary(lm(sample_heights ~ sample_weights - 1))
print(sum)
```

```
##
## Call:
## lm(formula = sample_heights ~ sample_weights - 1)
##
## Residuals:
##        1        2        3        4        5
##  0.11626  0.03317 -0.04613 -0.23802  0.15937
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## sample_weights  0.48545    0.00615   78.93 1.54e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1572 on 4 degrees of freedom
## Multiple R-squared:  0.9994,Adjusted R-squared:  0.9992
```

```
## F-statistic:  6231 on 1 and 4 DF,  p-value: 1.544e-07
```

- $\hat{\beta} = 0.48545$
- $\hat{\sigma}_{\text{ratio}} = 0.1572$

$$\hat{\mu}_{\text{height}} = \hat{\beta} x_i = \frac{\bar{y}}{\bar{x}} = \frac{63.4}{130.6} x_i = 0.48545 x_i$$

which is our line of best fit.

The ratio estimate is

$$\hat{\mu}_{\text{ratio}} = \hat{\mu}_{\text{height}}(\mu_{\text{weight}}) = 0.48545(136.7333) = 66.4$$

Well the real answer, was 65, so we are 1.4 units away from the real answer. However, that was closer than SRS which was 1.6 units away from the real answer.

```
beta <- sum$coefficients[1]
mu_ratio <- beta * mu_weight
sigma_ratio <- sum$sigma
n <- 5
N <- 15
c <- qnorm(0.975)
round(mu_ratio + c(-1, 1) * ((c * sigma_ratio)/sqrt(n)) * sqrt(1 -
  n/N), 1)
```

```
## [1] 66.3 66.5
```

A 95% confidence interval for μ_y is:

$$\hat{\mu}_{\text{ratio}} \pm \frac{c\hat{\sigma}_{\text{ratio}}}{\sqrt{n}} \sqrt{1 - \frac{n}{N}} = 66.4 \pm \frac{1.96(0.1572)}{\sqrt{5}} \sqrt{1 - \frac{5}{15}} = (66.3, 66.5)$$

Width: 0.2

Note:

- Width of CI using Ratios is narrower than SRS.
- There is bias in ratio estimation. Notice that the interval doesn't contain 65 which is the real answer.

Requirements:

- Regression and Ratio require highly correlated Y_i and x_i.
- Ratio requires an intercept of zero.
- Both Regression and Ratio are narrower than SRS, but Regression and Ratio are both biased.

Technique	Estimate	CI
SRS	$\hat{\mu}_y$	$\hat{\mu}_y \pm \frac{c\hat{\sigma}_y}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$
Reg	$\hat{\mu}_{\text{reg}} = \bar{y} + \hat{\beta}(\mu_x - \bar{x})$	$\hat{\mu}_{\text{reg}} \pm \frac{c\hat{\sigma}_{\text{reg}}}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$
Ratio	$\hat{\mu}_{\text{ratio}} = \frac{\bar{y}}{\bar{x}} \mu_x$	$\hat{\mu}_{\text{ratio}} \pm \frac{c\hat{\sigma}_{\text{ratio}}}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$

8.3 Lecture 41.00: Ratio Estimation

Suppose we had six students in our class. We selected them out at random using SRS. We record their grade in Calculus 1 and their gender.

Gender	Grade
M	70
F	70
M	85
F	85
M	90
F	90

The males on average: $\bar{x}_M = (70 + 85 + 90)/3$. Let

- y_i be the grade, and
- $z_i = \begin{cases} 1 & \text{if male} \\ 0 & \text{otherwise} \end{cases}$

Our estimate is $\hat{\theta}$,

$$\hat{\theta} = \frac{\sum_{i \in \mathcal{S}} y_i z_i}{\sum_{i \in \mathcal{S}} z_i}$$

- $\sum_{i \in \mathcal{S}} y_i z_i$ counts the number of grades of those males.
- $\sum_{i \in \mathcal{S}} z_i$ counts the number of males you happen to have.

Our parameter is θ,

$$\theta = \frac{\sum_{i=1}^{N} \frac{y_i z_i}{N}}{\sum_{i=1}^{N} \frac{z_i}{N}} = \frac{\mu}{\pi}$$

- μ is the average of the male grades.
- π is the proportion of the people that are male.

Therefore, we can write our estimate as $\hat{\theta} = \hat{\mu}/\hat{\pi}$.

Estimator

Our estimator asks "what's random?" What's random is whether you're in the sample. We define an indicator variable I_i which will be 1 if it's in our population.

$$\tilde{\theta} = \frac{\frac{\sum_{i=1}^{N} I_i y_i z_i}{n}}{\frac{\sum_{i=1}^{N} I_i z_i}{n}} = \frac{\tilde{\mu}}{\tilde{\pi}}$$

Now there's a bit of math involved in this because unfortunately we have never looked at the ratio of two random variables, and it's very difficult to do. Instead, we're going to use *Taylor's Approximation* (Lecture 41.50 goes more in detail).

Taylor's Approximation gives:

$$\frac{\tilde{\mu}}{\tilde{\pi}} \approx \frac{\mu}{\pi} + \frac{1}{\pi}(\tilde{\mu} - \mu) - \frac{\mu}{\pi^2}(\tilde{\pi} - \pi)$$

where $\tilde{\mu}$ and $\tilde{\pi}$ are both obtained by SRS. So we obtain the proportion and average from simple random sampling.

- The approximation is approximately normal (there's a Gaussian extension that allows this to be true).

Expectation and Variance

$$\begin{aligned}\mathbb{E}\left[\frac{\tilde{\mu}}{\tilde{\pi}}\right] &\approx \mathbb{E}\left[\frac{\mu}{\pi} + \frac{1}{\pi}(\tilde{\mu}-\mu) - \frac{\mu}{\pi^2}(\tilde{\pi}-\pi)\right] \\ &= \frac{\mu}{\pi} + \frac{1}{\pi}\Big(\mathbb{E}[\tilde{\mu}]-\mu\Big) - \frac{\mu}{\pi^2}\Big(\mathbb{E}[\tilde{\pi}]-\pi\Big) \\ &= \frac{\mu}{\pi}\end{aligned}$$

since $\mathbb{E}[\tilde{\mu}] = \mu$ and $\mathbb{E}[\tilde{\pi}] = \pi$ (by SRS, unbiased).

$$\begin{aligned}\mathbb{V}\left(\frac{\tilde{\mu}}{\tilde{\pi}}\right) &\approx \mathbb{V}\Big(\frac{\mu}{\pi} + \frac{1}{\pi}(\tilde{\mu}-\mu) - \frac{\mu}{\pi^2}(\tilde{\pi}-\pi)\Big) \\ &= \frac{1}{\pi^2}\,\mathbb{V}(\tilde{\mu}-\mu-\frac{\mu\tilde{\pi}}{\pi}+\mu) \\ &= \frac{1}{\pi^2}\,\mathbb{V}\left(\tilde{\mu}-\frac{\mu\tilde{\pi}}{\pi}\right)\end{aligned}$$

This is an average. Therefore, we end up with an estimated variance.

$$\mathbb{V}\left(\frac{\tilde{\mu}}{\tilde{\pi}}\right) \approx \frac{1}{\pi^2}\frac{\sigma^2_{\text{ratio}}}{n}\left(1-\frac{n}{N}\right)$$

A confidence interval is:

$$\text{EST} \pm c\,\text{SE} = \hat{\theta} \pm c\,\frac{1}{\hat{\pi}}\sqrt{1-\frac{n}{N}}\,\frac{\hat{\sigma}_{\text{ratio}}}{\sqrt{n}}$$

where

$$\hat{\sigma}^2_{\text{ratio}} = \frac{\sum_{i\in\mathcal{S}}(y_i - \hat{\theta}z_i)^2}{n-1}$$

8.4 Lecture 41.50: Taylor's Approximation

Calculus 1: $f(x) \approx f(x_0) + f'(x_0)(x-x_0)$.

EXAMPLE 8.4.1: Calculus 1 Taylor's Approximation

Approximate $f(1.1) = \ln(1.1)$ about $x_0 = 1$.
Solution.

$$f(1.1) \approx f(1) + f'(1)(1.1-1) = \ln(1) + \frac{1}{x}\bigg|_{x=1}(1.1-1) = 0 + 1(0.1) = 0.1$$

Calculus 3:

$$f(x,y) \approx f(x_0,y_0) + \frac{\partial f(x_0,y_0)}{\partial x}(x-x_0) + \frac{\partial f(x_0,y_0)}{\partial y}(y-y_0)$$

EXAMPLE 8.4.2: Calculus 3 Taylor's Approximation

Approximate $f(1.1, 1.1) = \ln(1.1 \times 1.1)$ about the point $(1,1)$.

Solution.

$$\begin{aligned} f(1.1, 1.1) &\approx f(1,1) + \frac{\partial f(1,1)}{\partial x}(x - x_0) + \frac{\partial f(1,1)}{\partial y}(y - y_0) \\ &= \ln(1) + \frac{1}{x}\bigg|_{x=1,y=1}(1.1 - 1) + \frac{1}{y}\bigg|_{x=1,y=1}(1.1 - 1) \\ &= 0 + 0.1 + 0.1 \\ &= 0.2 \end{aligned}$$

Approximate: $f(x, y) = x/y$ about the point (x_0, y_0).

$$\begin{aligned} f(x, y) &\approx f(x_0, y_0) + \frac{1}{y_0}(x - x_0) + \left(-\frac{x_0}{y_0^2}\right)(y - y_0) \\ &= \frac{x_0}{y_0} + \frac{1}{y_0}(x - x_0) - \frac{x_0}{y_0^2}(y - y_0) \end{aligned}$$

Therefore, approximating $\tilde{\mu}/\tilde{\pi}$ about (μ, π):

$$\frac{\tilde{\mu}}{\tilde{\pi}} \approx \frac{\mu}{\pi} + \frac{1}{\pi}(\tilde{\mu} - \mu) - \frac{\mu}{\pi^2}(\tilde{\pi} - \pi)$$

8.5 Lecture 42.00: Ratio Estimation, Example

EXAMPLE 8.5.1

The number of people in the Kitchener riding is $89\,422$. Stephen Harper wants to know the average age of people in the riding who would vote for him. Using SRSWOR, he selects 80 people, and finds that the average age of those who vote for him is 67. 42 of those poled would vote for him. If the estimated variance is 5.42, build a 95% confidence interval for the average age of those who vote for Stephen Harper.

Solution.

- The proportion of people that would vote for Stephen Harper is $\hat{\pi} = 42/80$.
- The average age of those that would vote for him is $\hat{\theta} = \hat{\mu}/\hat{\pi} = 67$.

Therefore, a 95% confidence interval for the average age of those who vote for Stephen Harper is:

$$\begin{aligned} \hat{\theta} \pm c\,\frac{1}{\hat{\pi}}\frac{\hat{\sigma}_{\text{ratio}}}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} &= 67 \pm 1.96\left(\frac{1}{42/80}\right)\sqrt{\frac{5.42}{80}}\sqrt{1 - \frac{80}{89422}} \\ &= (66.03, 67.97) \end{aligned}$$

Chapter 9

Assignment 9

9.1 Lecture 43.00: Stratified Sampling

Suppose that you're interested in the population as a whole, but at the same time you're interested in some sub-population of the population. For example suppose you're interested in University of Waterloo students, but we're also interested in Math faculty students. You would use a *stratified* sample. In this case, you might perform SRS over a sub-population and then combine that together to get the entire population so what would that notationally look like?

Suppose we have a frame U, and we divide U into sub-frames $U_1, U_2, \dots, U_H$ where

(1) $U_1 \cup U_2 \cup U_3 \cup \cdots \cup U_H = U$.

(2) For any $i \neq j$, $U_i \cap U_j = \emptyset$.

(3) Define $|U_h| = N_h$ and $|U| = N$.

(4) $N_1 + N_2 + \cdots + N_h = N$.

Parameter

$$\begin{aligned}\mu &= \frac{N_1\mu_1 + N_2\mu_2 + \cdots + N_H\mu_H}{N} \\ &= \frac{N_1}{N}\mu_1 + \frac{N_2}{N}\mu_2 + \cdots + \frac{N_h}{N}\mu_H \\ &= w_1\mu_1 + \cdots + w_H\mu_H \\ &= \sum_{i=1}^{H} w_i\mu_i\end{aligned}$$

Estimate

We would use SRS to estimate the strata's average.

$$\hat{\mu} = \sum_{i=1}^{H} w_i\hat{\mu}_i$$

When we get the strata's average, and we multiply by the weight we add over the strata's to get the population average.

Estimator

$$\tilde{\mu} = \sum_{i=1}^{H} w_i \tilde{\mu}_i$$

where $\tilde{\mu}_i$ is the SRS random variable, hence it is unbiased and normally distributed. Therefore, the weighted average of the estimator will be normally distributed. Now, we need to find the expectation and variance of $\tilde{\mu}$.

Expectation of $\tilde{\mu}$

$$\mathbb{E}[\tilde{\mu}] = \mathbb{E}\left[\sum_{i=1}^{H} w_i \tilde{\mu}_i\right] = \sum_{i=1}^{H} w_i \, \mathbb{E}[\tilde{\mu}_i] = \sum_{i=1}^{H} w_i \mu_i \quad \text{since SRS is unbiased.}$$

Variance of $\tilde{\mu}$

At this point in the course the formulas become ugly.

$$\mathbb{V}(\tilde{\mu}) = \mathbb{V}\left(\sum_{i=1}^{H} w_i \tilde{\mu}_i\right) = \sum_{i=1}^{H} w_i^2 \, \mathbb{V}(\tilde{\mu}_i)$$

where $\tilde{\mu}_i \perp\!\!\!\perp \tilde{\mu}_j$ since no unit is in both groups. Therefore,

$$\mathbb{V}(\tilde{\mu}) = \sum_{i=1}^{H} w_i^2 \frac{\sigma_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$$

Confidence Interval for μ

$$\hat{\mu} \pm c\sqrt{\sum_{i=1}^{H} w_i^2 \frac{\sigma_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)} \quad \text{where } c \sim \mathcal{N}(0,1).$$

Proportion

For a proportion, we want to estimate π.

Parameter

$$\pi = \sum_{i=1}^{H} w_i \pi_i$$

Estimate

$$\hat{\pi} = \sum_{i=1}^{H} w_i \hat{\pi}_i \quad \text{which uses SRS.}$$

Estimator

$$\tilde{\pi} = \sum_{i=1}^{H} w_i \tilde{\pi}_i$$

Confidence Interval for π

$$\hat{\pi} \pm c\sqrt{\sum_{i=1}^{H} w_i^2 \frac{\sigma_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)} \quad \text{where } c \sim \mathcal{N}(0,1).$$

You'll remember that we have worked out before that $\sigma_i^2 = \pi_i(1-\pi_i)$.

9.2 Lecture 44.00: Stratified, Allocation

Today we're going to talk about something called *Allocation*. For example, allocation is when you have a sample of 100 units, and you have four strata. How should you spend those 100 units? Should three quarters of them go to one stratum, and the remaining quarter be split among the last three strata? We define how you divide them up to be stratification. To make that decision there are two types that we're going to talk about today.

(1) Proportional.

(2) Neyman or optimal.

Proportional

We sample based on the size of the strata. In other words, the bigger the strata size, the bigger the sample size.

$$n_h = w_h n$$

EXAMPLE 9.2.1

Provinces	Population (millions)
ON	10
QUE	5
BC	3
ALB	2
Total	20

If we have $n = 100$ units to sample, ON should get $n_{\text{ON}} = w_{\text{ON}}(n) = 1/2(100) = 50$ units.

Neyman

In Neyman allocation, we select our sample size and values that minimize the stratified variance.

$$\mathbb{V}(\tilde{\mu}) = \sum_{i=1}^{H} w_i^2 \frac{\sigma_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$$

subject to $n = n_1 + n_2 + \cdots + n_H$. This ends up being a Lagrange multiplication problem. So minimize

$$W(\tilde{\mu}) = \sum_{i=1}^{H} w_i^2 \frac{\sigma_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right) + \lambda(n - n_1 - \cdots - n_H)$$

Find $\frac{\partial W}{\partial \lambda}, \frac{\partial W}{\partial n_i}$ and set to zero to get:

$$n_i = \frac{n\sigma_i w_i}{\sum_{j=1}^{H} \sigma_j w_j}$$

REMARK 9.2.2

- $n_i \propto \sigma_i$. So, if you have more variability in your stratum, then you're going to want a larger sample size. That should make a lot of sense because if you have a lot of variability, you can reduce the variability by having a larger sample size, remember, the variability is the average of the variance divided by n. So, the larger your sample size the smaller the variance ends up being.
- $n_i \propto w_i$. The larger the strata, the more units you will want allocated to it.
- If $\sigma_1 = \sigma_2 = \cdots = \sigma_H$, then
$$n_i = \frac{nw_i}{\sum_{i=1}^{H} w_i} = nw_i$$
which is proportional allocation since $\sum_{i=1}^{H} w_i = 1$.

Just like when we did with the small sample size, we take a small sample, and you use the small sample to estimate these unknown σ's. Once you've estimated these unknown σ's, you'd use them to determine how you should allocate your larger sample size to the actual strata of interest.

9.3 Lecture 45.00: Stratified Example

9.3.1 Stratified 1

I am interested in the average tuition paid by students at the University of Waterloo. Additionally, I want to know how much each faculty student is paying on average. Hence, I decide to stratify by Faculty (assume students belong to a single faculty).

Faculty	N	$\hat{\mu}$	n	W	$\hat{\sigma}$
Math	6600	4500	15	0.22	400
Arts	9000	3000	10	0.30	200
Science	5400	4500	25	0.18	300
AHS	1500	3200	35	0.05	100
Engineer	6000	7000	15	0.20	100
EVS	1500	3500	20	0.05	200
Total	30000		120		

Build a 95% confidence interval for the mean tuition in Math.

```
ci <- 4500 + c(-1, 1) * qnorm(0.975) * 400/sqrt(15) * (1 - 15/6600)
round(ci)
```

```
## [1] 4298 4702
```

$$\hat{\mu}_{\text{math}} \pm \frac{c\hat{\sigma}_{\text{math}}}{\sqrt{n_{\text{math}}}}\sqrt{1 - \frac{n_{\text{math}}}{N_{\text{math}}}} = 4500 \pm \frac{1.96(400)}{\sqrt{15}}\sqrt{1 - \frac{15}{6600}} = (4298, 4702)$$

Build a 95% confidence interval for the mean tuition at UW.

Since we've used SRS in each of the strata, we have to use stratified sampling.

```
N_i <- c(6600, 9000, 5400, 1500, 6000, 1500)
N <- sum(N_i)
w_i <- N_i/N
n_i <- c(15, 10, 25, 35, 15, 20)
```

```
mu_i <- c(4500, 3000, 4500, 3200, 7000, 3500)
sigma_i <- c(400, 200, 300, 100, 100, 200)
mu <- sum(w_i * mu_i)
variance <- sum(w_i^2 * sigma_i^2/n_i * (1 - n_i/N_i))
ci <- mu + c(-1, 1) * qnorm(0.975) * sqrt(variance)
mu
```

```
## [1] 4435
```

```
round(variance, 3)
```

```
## [1] 1023.024
```

```
round(ci)
```

```
## [1] 4372 4498
```

$$\hat{\mu} = \sum_{i=1}^{H} w_i \hat{\mu}_i = 4435$$

$$\widehat{\mathbb{V}(\tilde{\mu})} = \sum_{i=1}^{H} \frac{w_i^2 \hat{\sigma}_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) = 1023.024$$

Therefore, a 95% confidence interval for μ is:

$$\hat{\mu} \pm c\sqrt{\widehat{\mathbb{V}(\tilde{\mu})}} = 4435 \pm 1.96\sqrt{1023.024} = (4372, 4498)$$

9.3.2 Stratified 2

We continue from Section 9.3.1.

A proportional allocation of our sample values to each stratum.

```
n <- 120
round(n * w_i)
```

```
## [1] 26 36 22  6 24  6
```

- $n_{\text{math}} = nw_{\text{math}} = 120(0.22) = 26.$
- $n_{\text{arts}} = 36.$
- $n_{\text{science}} = 22.$
- $n_{\text{ahs}} = 6.$
- $n_{\text{eng}} = 24.$
- $n_{\text{evs}} = 6.$

An optimal allocation of our sample values to each stratum.

```
round(n * sigma_i * w_i/sum(w_i * sigma_i))
```

```
## [1] 45 30 27  3 10  5
```

- $n_{\text{math}} = \frac{n\sigma_{\text{math}} w_{\text{math}}}{\sum_{j=1}^{6} w_j \sigma_j} = \frac{120(400)(0.22)}{237} = 45.$
- $n_{\text{arts}} = 30.$

- $n_{\text{science}} = 27$.
- $n_{\text{ahs}} = 3$.
- $n_{\text{eng}} = 10$.
- $n_{\text{evs}} = 120 - 45 - 30 - 27 - 3 - 10 = 5$.

9.3.3 Stratified 3

A course has 3 sections all taught by one instructor. There are 205, 212, and 253 people in each of the sections 1, 2, and 3 respectively. At the end of the term the instructor is curious about how well the students performed. The administration takes a simple random sampling of 15, 12, and 14 people respectively from each section. The averages for each section are 75, 70, and 72 respectively with standard deviations of 10, 15, and 5. Build a 95confidence interval for the mean grade of the instructors course.

```
N_i <- c(205, 212, 253)
N <- sum(N_i)
w_i <- N_i/N
n_i <- c(15, 12, 14)
mu_i <- c(75, 70, 72)
sigma_i <- c(10, 15, 5)
mu <- sum(w_i * mu_i)
variance <- sum(w_i^2 * sigma_i^2/n_i * (1 - n_i/N_i))
ci <- mu + c(-1, 1) * qnorm(0.975) * sqrt(variance)
mu
```

```
## [1] 72.28507
```

```
variance
```

```
## [1] 2.589983
```

```
round(ci, 3)
```

```
## [1] 69.131 75.439
```

$$\hat{\mu} = \sum_{i=1}^{H} w_i \hat{\mu}_i = 72.28507$$

$$\widehat{\mathbb{V}(\tilde{\mu})} = \sum_{i=1}^{H} \frac{w_i^2 \hat{\sigma}_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) = 2.589983$$

Therefore, a 95% confidence interval for μ is:

$$\hat{\mu} \pm c\sqrt{\widehat{\mathbb{V}(\tilde{\mu})}} = 72.28507 \pm 1.96\sqrt{2.589983} = (69.131, 75.439)$$

9.4 Lecture 46.00: Post Stratification

Until now what you had some strata, performed SRS on each stratum, then combined them, and looked at the population value μ. See Figure 9.1. In post stratification, the idea is that you have done an SRS of some large population, and then decided afterwards that you wanted to stratify. At this point, you then break it into three. Notice that the SRS is done at the *start* as opposed to at the *end*. See Figure 9.2.

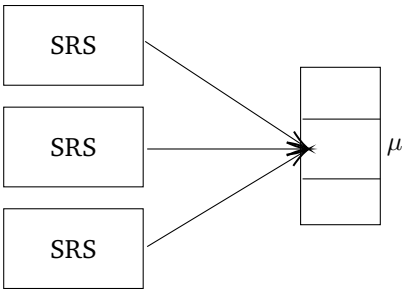


Figure 9.1: Regular Stratification

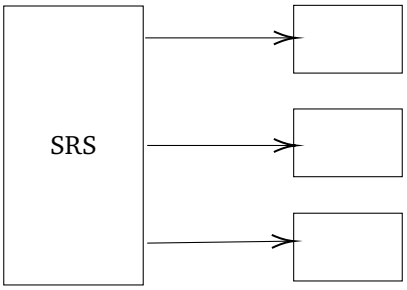


Figure 9.2: Post Stratification

Mathematically, your sample sizes end up being random, which has an influence on how we calculate things. Although, it doesn't actually have an influence on the actual mathematics at the end of the day, so the resultants are the same. For example, the post stratification estimate is very similar to that of stratified:

$$\hat{\mu}_{\text{post}} = w_1\mu_1 + \cdots + w_H\mu_H$$

The estimated variance for post stratification is:

$$\widehat{\mathbb{V}(\tilde{\mu}_{\text{post}})} = \sum_{i=1}^{H} w_i^2 \left(1 - \frac{n_i}{N_i}\right)\frac{\hat{\sigma}_i^2}{n_i}$$

A confidence interval for μ_{post} is given by:

$$\hat{\mu}_{\text{post}} \pm c\sqrt{\widehat{\mathbb{V}(\tilde{\mu}_{\text{post}})}} \quad \text{where } c \sim \mathcal{N}(0,1)$$

9.5 Lecture 47.00: Non-Response

Non-response means that someone didn't respond to our survey. All the surveys of human respondents have non-response. Non-response causes bias, it is a form of error that can skew our results. The response rate (or the proportion of people that respond) is hard to define and your text goes to some length to define it. To correct non-response, we can do something called *two-phase sampling*:

- Phase 1: Typical SRS with sample size n.
- Phase 2: Sub-sample m non-responders from Phase 1.

This is a stratified design with responders and non-responders as your strata.

The estimate is

$$\hat{\mu} = \frac{n_R}{n}\hat{\mu}_R + \frac{n_m}{n}\hat{\mu}_m$$

- $\hat{\mu} =$ population estimate.

- $n_R =$ number of responders.
- $n =$ number of people you ask in general.
- $\hat{\mu}_R =$ response average.
- $n_m =$ number of missing people.
- $\hat{\mu}_m =$ average of the missing people (the non-responders).

There's a similar one for the proportion.

$$\hat{\pi} = \frac{n_R}{n}\hat{\pi}_R + \frac{n_m}{n}\hat{\pi}_m$$

Those are the estimates that you would use. The variance is very ugly, so we'll ignore it today.

Chapter 10

Appendix

Normal:

$$f(x) = \mathbb{P}(X = x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- $F(x) = \mathbb{P}(X \leq x)$ can be obtained with `pnorm(`x`,`μ`,`σ`)` and gives the value of p.
- $F^{-1}(p)$ can be obtained with `qnorm(`p`,`μ`,`σ`)` and gives the value of x.
- $f(x) = \mathbb{P}(X = x)$ can be obtained with `dnorm(`x`,`μ`,`σ`)`. Note that this is <u>not</u> a probability.

10.1 Tables

10.1.1 $\mathcal{N}(0,1)$ Cumulative Distribution Function

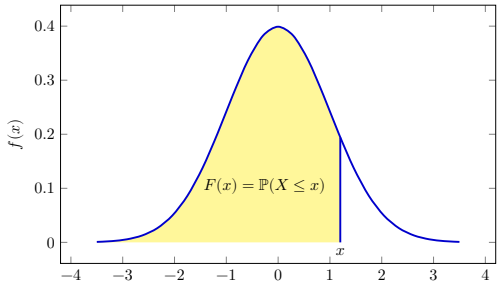


Table 10.1: $F(x) = \mathbb{P}(X \leq x)$ where $X \sim \mathcal{N}(0,1)$ and $x \geq 0$

x	**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

10.1.2 t Quantiles

Table 10.2: Quantiles of the t distribution with n degrees of freedom

n / p	**0.6**	**0.7**	**0.8**	**0.9**	**0.95**	**0.975**	**0.99**	**0.995**	**0.9995**
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657	636.62
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925	31.599
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	12.924
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	8.610
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	6.869
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	5.959
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	5.408
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355	5.041
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.781
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.587
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	4.437
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	4.318
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	4.221
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977	4.140
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	4.073
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921	4.015
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898	3.965
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878	3.922
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861	3.883
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845	3.850
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831	3.819
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819	3.792
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807	3.768
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797	3.745
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787	3.725
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779	3.707
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771	3.690
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763	3.674
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756	3.659
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750	3.646
$\mathcal{N}(0,1)$	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576	3.291