ANALYSIS OF VARIANCE

Analytics Primer

Two Parameters Instead of One?

- We have studied hypothesis tests and confidence intervals that have focused on one population parameter.
- However, sometimes we like to compare multiple parameters against each other.
- This is the foundation of a larger section of analysis called **analysis of variance (ANOVA)**.
- For now we will focus on only comparing two population means.

TWO-SAMPLE HYPOTHESIS TESTING

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

Two-Sample

$$H_0\colon \mu \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} \mu_0$$

$$H_a\colon \mu \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} \mu_0$$

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample **Two-Sample**

$$H_0\colon \mu \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} \mu_0$$

$$H_a\colon \mu \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} \mu_0$$

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

Two-Sample

$$H_0:\ \mu \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} \mu_0$$

$$H_a:\ \mu \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} \mu_0$$

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

Two-Sample

$$H_0\colon \mu \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} \mu_0$$

$$H_a\colon \mu \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} \mu_0$$

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample | **Two-Sample**

$$H_0: \mu \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} \mu_0$$

$$H_a: \mu \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} \mu_0$$

Parameter

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

Two-Sample

$$H_0: \mu \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} \mu_0$$

$$H_a: \mu \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} \mu_0$$

Parameter

Constant

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

$$H_0: \mu \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} \mu_0$$

$$H_a: \mu \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} \mu_0$$

Parameter Constant

Two-Sample

$$H_0: \mu_1 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} \mu_2$$

$$H_a: \mu_1 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} \mu_2$$

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

$$H_0: \mu \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} \mu_0$$

$$H_a: \mu \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} \mu_0$$

Parameter Constant

Two-Sample

$$H_0: \mu_1 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} \mu_2$$

$$H_a: \mu_1 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} \mu_2$$

Parameters

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

$$H_0: \mu \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} \mu_0$$

$$H_a: \mu \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} \mu_0$$

Parameter Constant

Two-Sample

$$H_0: \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0$$

$$H_a: \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

Parameters

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

$$H_0: \mu \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} \mu_0$$

$$H_a: \mu \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} \mu_0$$

Parameter Constant

Two-Sample

$$H_0: \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0$$

$$H_a: \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

Parameter Constant

Additional Assumptions

- How does our hypothesis testing structure change when we have two parameters instead of one?

$$H_0\colon \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0 \qquad H_a\colon \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

- Assume two samples are independent of each other – value selected in one sample has no bearing on value selected on other sample.

Additional Assumptions

- How does our hypothesis testing structure change when we have two parameters instead of one?

$$H_0\colon \ \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0 \qquad H_a\colon \ \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

- Assume two samples are independent of each other – value selected in one sample has no bearing on value selected on other sample.
- **Variance of the two samples becomes important.**

Equal or Unequal Variances?

- Depending on the relationship between the variances of the two samples, we have different hypothesis test structures.

$$\sigma_1^2 \begin{Bmatrix} = \\ \neq \end{Bmatrix} \sigma_2^2$$

Equal or Unequal Variances?

- Depending on the relationship between the variances of the two samples, we have different hypothesis test structures.

$$\sigma_1^2 \begin{Bmatrix} = \\ \neq \end{Bmatrix} \sigma_2^2$$

$=$

Two-Sample Hypothesis Test for Means with Equal Variances.

$\neq$

Two-Sample Hypothesis Test for Means with Unequal Variances.

TWO-SAMPLE HYPOTHESIS TESTING

Testing Differences in Means – Equal Variances

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)
2. Collect Data (Test Statistic)
3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0 \qquad H_a\colon \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

2. Collect Data (Test Statistic)
3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a\colon \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)
3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\!: \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a\!: \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{\text{Statistic} - \text{Null Value}}{\text{Standard Error}}$$

3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0:\ \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a:\ \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon\ \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a\colon\ \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Pooled Standard Deviation

3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Pooled Standard Deviation

- **Under the assumption of equal variances**, we have two estimates of the population variance, $\sigma^2 - s_1^2$ and s_2^2.
- The assumption of equal variances implies that both s_1^2 and s_2^2 are estimating the same population variance.
- Unreasonable to only use one estimate.
- Should combine both s_1^2 and s_2^2 to get our estimate.

Pooled Standard Deviation

- Should combine both s_1^2 and s_2^2 to get our estimate.

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

$$= \sqrt{\frac{\sum\left(x_{1,i} - \bar{x}_1\right)^2 + \sum\left(x_{2,i} - \bar{x}_2\right)^2}{n_1 + n_2 - 2}}$$

Pooled Standard Deviation

- Should combine both s_1^2 and s_2^2 to get our estimate.

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

$$= \sqrt{\frac{\sum\left(x_{1,i} - \bar{x}_1\right)^2 + \sum\left(x_{2,i} - \bar{x}_2\right)^2}{n_1 + n_2 - 2}}$$

"Pooled" Variance Calculation

Pooled Standard Deviation

- There are a couple of additional assumptions that goes along with the combined standard deviations.

1. Each population has an **approximate Normal distribution**.
   - Symmetric and Bell-Shaped
2. Variances of two groups are equal.

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon\ \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a\colon\ \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

3. What is probability this happens? (P-value)

   t-Distribution: $d.f. = n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$

4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0:\ \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a:\ \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

3. What is probability this happens? (P-value)

   t-Distribution: $d.f. = n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$

4. $\boxed{\text{Decision About Null Hypothesis}}$
5. $\boxed{\text{Summarize}}$

Exact Same

Confidence Interval

- If we can have hypothesis tests, then we can have confidence intervals as well:

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{\alpha/2} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example – Comparing Two Means

- A human resources manager of a large business firm is trying to determine if there exists gender bias in the pay scale of employees at the company. The manager assumes the variability of salaries between genders is the same, but wants to test if there is gender bias in salaries (males make significantly more than females). The manager samples 62 males and 77 females. The sample of males had an average salary of $87,547 with a s.d. of $5,910. The sample of females had an average salary of $78,289 with a s.d. of $6,276. Run a hypothesis test.

Example – Comparing Two Means

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon \mu_M - \mu_F \leq 0 \qquad H_a\colon \mu_M - \mu_F > 0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{(\bar{x}_M - \bar{x}_F) - D_0}{s_p\sqrt{\frac{1}{n_M} + \frac{1}{n_F}}} = \frac{(87547 - 78289) - 0}{6115.742\sqrt{\frac{1}{62} + \frac{1}{77}}} = 8.87$$

3. What is probability this happens? (P-value)

P-value < 0.001

4. Decision About Null Hypothesis

Reject H_0 → It does appear that males, on average, make significantly more than females!

TWO-SAMPLE HYPOTHESIS TESTING

Testing Differences in Means – Unequal Variances

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)
2. Collect Data (Test Statistic)
3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon\ \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0 \qquad H_a\colon\ \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

2. Collect Data (Test Statistic)
3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a\colon \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)
3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon\ \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a\colon\ \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{\text{Statistic } - \text{Null Value}}{\text{Standard Error}}$$

3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon\ \mu_1-\mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a\colon\ \mu_1-\mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{(\bar{x}_1-\bar{x}_2)-D_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\!: \ \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a\!: \ \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Standard Error Changes

3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Combination of Standard Deviations

- **Under the assumption of unequal variances**, we have two population variances, σ_1^2 and σ_2^2.
- Therefore, we need both estimates of each of these treated separately.
- Cannot “pool” them, just use them both in standard error calculation.

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon\ \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a\colon\ \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

3. What is probability this happens? (P-value)

   t-Distribution: $d.f. = \text{COMPLICATED!}$

4. Decision About Null Hypothesis
5. Summarize

t-Distribution Degrees of Freedom

- **Under the assumption of unequal variances**, we have two population variances, σ_1^2 and σ_2^2.
- This makes the degrees of freedom on the test a lot more complicated.

$$d.f. = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}\right)}$$

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0:\ \mu_1 - \mu_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a:\ \mu_1 - \mu_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

3. What is probability this happens? (P-value)

   t-Distribution: $d.f. = \text{COMPLICATED!}$

4. Decision About Null Hypothesis
5. Summarize

Exact Same

Confidence Interval

- If we can have hypothesis tests, then we can have confidence intervals as well:

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{\alpha/2} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Confidence Interval

- If we can have hypothesis tests, then we can have confidence intervals as well:

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{\alpha/2} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Standard Error Changes

Example – Comparing Two Means

- A human resources manager of a large business firm is trying to determine if there exists gender bias in the pay scale of employees at the company. The manager assumes the variability of salaries between genders is different, but wants to test if there is gender bias in salaries (males make significantly more than females). The manager samples 62 males and 77 females. The sample of males had an average salary of $87,547 with a s.d. of $5,910. The sample of females had an average salary of $78,289 with a s.d. of $6,276. Run a hypothesis test.

Example – Comparing Two Means

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon \mu_M - \mu_F \leq 0 \qquad H_a\colon \mu_M - \mu_F > 0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{(\bar{x}_M - \bar{x}_F) - D_0}{\sqrt{\frac{s_M^2}{n_M} + \frac{s_F^2}{n_F}}} = \frac{(87547 - 78289) - 0}{\sqrt{\frac{5910^2}{62} + \frac{6276^2}{77}}} = 8.93$$

3. What is probability this happens? (P-value)

P-value < 0.001

4. Decision About Null Hypothesis

Reject H_0 → It does appear that males, on average, make significantly more than females!

TWO-SAMPLE HYPOTHESIS TESTING

Testing Differences in Variances

Variances Instead of Means

- Previously, we introduced comparing two means through hypothesis testing.
- However, we needed to assume the relationship between two variances to conduct the test between two means.
- We can run a **hypothesis test comparing two variances** to tell us which means test to use.
- Although we commonly work with standard deviations, there is **no formal statistical test for comparing standard deviations** – only variances.

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

Two-Sample

$$H_0: \sigma_1^2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} \sigma_2^2$$

$$H_a: \sigma_1^2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} \sigma_2^2$$

Parameters

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

Two-Sample

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 1$$

$$H_a: \frac{\sigma_1^2}{\sigma_2^2} \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 1$$

Parameters

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have wo parameters instead of one?

One-Sample

Two-Sample

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 1$$

$$H_a: \frac{\sigma_1^2}{\sigma_2^2} \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 1$$

Parameter Constant

Additional Assumptions

- How does our hypothesis testing structure change when we have two parameters instead of one?

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 1 \qquad H_a: \frac{\sigma_1^2}{\sigma_2^2} \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 1$$

1. Assume two samples are independent of each other – value selected in one sample has no bearing on value selected on other sample.
2. Original population distributions are assumed to be approximately Normally distributed (symmetric, bell-shaped).

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)
2. Collect Data (Test Statistic)
3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad\qquad H_a\colon \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

2. Collect Data (Test Statistic)
3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad H_a : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

2. Collect Data (Test Statistic)

$$F = \frac{s_i^2}{s_j^2}$$

3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad H_a : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

2. Collect Data (Test Statistic)

$$F = \frac{s_i^2}{s_j^2}$$

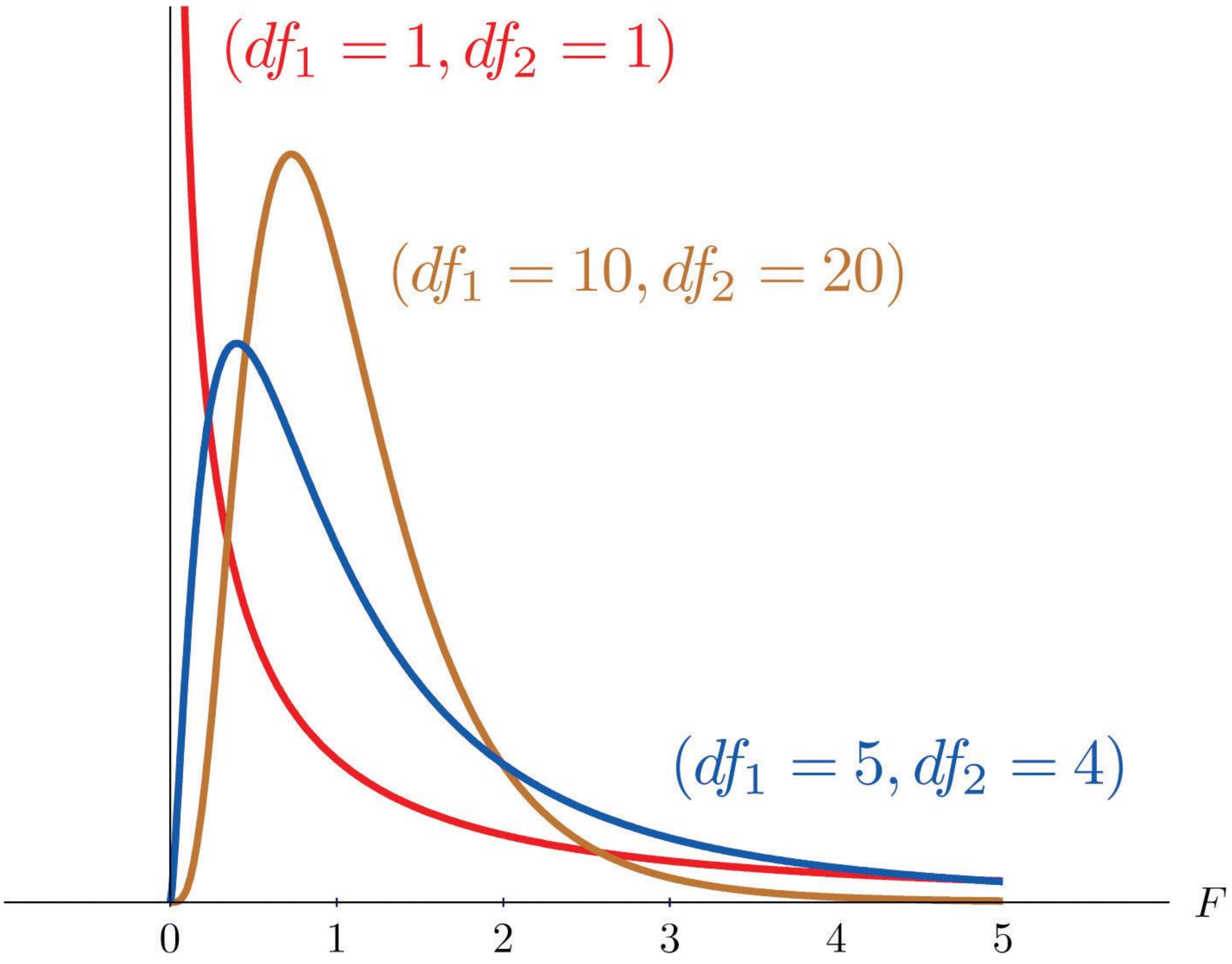
Larger standard deviation between two samples

3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

F-Distribution

- The F-test comes from the **F-distribution**.
- Characteristics of the F-distribution:
  1. Bounded Below By Zero
  2. Right Skewed
  3. Numerator **and** Denominator Degrees of Freedom

F-Distribution



Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad H_a: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

2. Collect Data (Test Statistic)

$$F = \frac{s_i^2}{s_j^2}$$

3. What is probability this happens? (P-value)

F-Dist.: Numerator $d.f. = n_i - 1$, Denominator $d.f. = n_j - 1$

4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad H_a\colon \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

2. Collect Data (Test Statistic)

$$F = \frac{s_i^2}{s_j^2}$$

3. What is probability this happens? (P-value)

F-Dist.: Numerator $d.f. = n_i - 1$, Denominator $d.f. = n_j - 1$

4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad\qquad H_a: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

2. Collect Data (Test Statistic)

$$F = \frac{s_i^2}{s_j^2}$$

3. What is probability this happens? (P-value)

If larger variance in numerator, then p-value is always 2*upper-tail probability!

4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad\qquad H_a: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

2. Collect Data (Test Statistic)

$$F = \frac{s_i^2}{s_j^2}$$

3. What is probability this happens? (P-value)

2*pf(F,numerator df, denominator df, lower.tail=F)

4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad\qquad H_a: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

2. Collect Data (Test Statistic)

$$F = \frac{s_i^2}{s_j^2}$$

3. What is probability this happens? (P-value)

F-Dist.: Numerator $d.f. = n_i - 1$, Denominator $d.f. = n_j - 1$

4. Decision About Null Hypothesis
5. Summarize

Exact Same

Example – Comparing Two Variances

- A human resources manager of a large business firm is trying to determine if there exists gender bias in the pay scale of employees at the company. The manager has no assumption about the variability of salaries between genders, but wants to test if makes have higher average salary than females. The manager samples 62 males and 77 females. The sample of males had an average salary of $87,547 with a s.d. of $5,910. The sample of females had an average salary of $78,289 with a s.d. of $6,276. Run a hypothesis test.
- **Need to first test if variances are equal or not before running test of means.**

Example – Comparing Two Variances

Develop your Hypothesis Statements (H_0 and H_a)

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad H_a: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

Collect Data (Test Statistic)

$$F = \frac{s_i^2}{s_j^2} = \frac{6276^2}{5910^2} = 1.13$$

What is probability this happens? (P-value)

2*(pf(1.13, 76, 61,lower.tail = F)) P-value > 0.05

Decision About Null Hypothesis

Do NOT Reject H_0 → Use Equal Variance Test

TWO-SAMPLE HYPOTHESIS TESTING

Matched / Paired Differences

Sources of Variation

- Generally comparing two population means works well in certain situations.
- However, there are some instances where a paired difference (**matched**) **sample** is used to control for **sources of variation** that might distort the conclusions.

Example – Sources of Variation

- You have SAT scores for both boys and girls from a local school.
- You believe that boys and girls have the same average test score, but want to test otherwise.
- Of the 39 females, 32 of them are part of the accelerated math and language arts program.
- Of the 39 males, 11 of them are part of the accelerated math and language arts program.

Example – Sources of Variation

- You have SAT scores for both boys and girls from a local school.
- You believe that boys and girls have the same average test score, but want to test otherwise.
- Of the 39 females, 32 of them are part of the accelerated math and language arts program.
- Of the 39 males, 11 of them are part of the accelerated math and language arts program.
- **Is this a fair test?**

Matched (Paired) Samples

- **Matched samples** are samples that are selected such that each data value from one sample is related (or matched / paired) with a corresponding data value from a second sample.
- Purposefully not guaranteeing independence.
- In our previous example, match boy and girls who were in the accelerated program and ones who were not.
- Compare these matched individuals instead.

Differences Become Important

- **Matched / Paired samples** are samples that are selected such that each data value from one sample is related (or matched / paired) with a corresponding data value from a second sample.
- Now our focus turns from the individual values in the populations and to the values of the **differences in the populations**.
- ALL assumptions and calculations are done on the differences, NOT the individual samples.

Additional Assumptions

- How does our hypothesis testing structure change when we have one parameter instead of two?

$$H_0: \mu_d \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a: \mu_d \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

- The assumptions for the matched pairs hypothesis test are the same as for regular hypothesis test for means.
- Large sample ($n > 50$) **of differences**.

OR

- Small samples with **differences** having Normal distribution.

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)
2. Collect Data (Test Statistic)
3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0 : \mu_d \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad\qquad H_a : \mu_d \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)
3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0 : \mu_d \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a : \mu_d \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)

$$t = \frac{\bar{x}_d - D_0}{\frac{s_d}{\sqrt{n_d}}}$$

3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0: \mu_d \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a: \mu_d \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)

$$t = \frac{\bar{x}_d - D_0}{\frac{s_d}{\sqrt{n_d}}}$$

Focus is on the differences!

3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0: \mu_d \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a: \mu_d \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)

$$t = \frac{\bar{x}_d - D_0}{\frac{s_d}{\sqrt{n_d}}}$$

3. What is probability this happens? (P-value)

   t-Distribution: $d.f. = n_d - 1$

4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon \mu_d \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D_0 \qquad H_a\colon \mu_d \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} D_0$$

2. Collect Data (Test Statistic)

$$t = \frac{\bar{x}_d - D_0}{\frac{s_d}{\sqrt{n_d}}}$$

3. What is probability this happens? (P-value)

   t-Distribution: $d.f. = n_d - 1$

4. Decision About Null Hypothesis
5. Summarize

Exact Same

Confidence Interval

- If we can have hypothesis tests, then we can have confidence intervals as well:

$$\bar{x}_d \pm t^*_{\alpha/2} \times \frac{s_d}{\sqrt{n_d}}$$

Example – Paired Samples

- A human resources manager of a large business firm is trying to determine if there exists gender bias in the pay scale of employees at the company. The manager samples 51 pairs of male and female employees where the pair has the same job title and experience at the company. The average difference in salaries is $2,131 with a s.d. of differences of $7,898. Run a hypothesis test.

Example – Comparing Two Means

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0: \mu_d \leq 0 \qquad\qquad H_a: \mu_d > 0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{\bar{x}_d - D_0}{\frac{s_d}{\sqrt{n_d}}} = \frac{2131 - 0}{\frac{7898}{\sqrt{51}}} = 1.93$$

3. What is probability this happens? (P-value)

P-value = (0.025, 0.05)

1-pt(1.93,50) = 0.03
OR
pt(1.93,50,lower.tail=F) = 0.03

4. Decision About Null Hypothesis

Reject H_0 $\rightarrow$ Males make more!

TWO-SAMPLE HYPOTHESIS TESTING

Testing Population Proportions

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

Two-Sample

$$H_0 : p \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} p_0$$

$$H_a : p \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} p_0$$

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample **Two-Sample**

$$H_0: p \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} p_0$$

$$H_a: p \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} p_0$$

Parameter

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample **Two-Sample**

$$H_0: p \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} p_0$$

$$H_a: p \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} p_0$$

Parameter Constant

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

$$H_0 \colon p \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} p_0$$

$$H_a \colon p \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} p_0$$

Parameter Constant

Two-Sample

$$H_0 \colon p_1 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} p_2$$

$$H_a \colon p_1 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} p_2$$

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

$$H_0: p \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} p_0$$

$$H_a: p \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} p_0$$

Parameter Constant

Two-Sample

$$H_0: p_1 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} p_2$$

$$H_a: p_1 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} p_2$$

Parameters

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

$$H_0: p \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} p_0$$

$$H_a: p \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} p_0$$

Parameter Constant

Two-Sample

$$H_0: p_1 - p_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0$$

$$H_a: p_1 - p_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

Parameters

Two Parameters Instead of One?

- How does our hypothesis testing structure change when we have two parameters instead of one?

One-Sample

$$H_0: p \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} p_0$$

$$H_a: p \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} p_0$$

Parameter Constant

Two-Sample

$$H_0: p_1 - p_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0$$

$$H_a: p_1 - p_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

Parameter Constant

Additional Assumptions

- How does our hypothesis testing structure change when we have two parameters instead of one?

$$H_0\!: p_1 - p_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0 \qquad H_a\!: p_1 - p_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

- Assume two samples are independent of each other – value selected in one sample has no bearing on value selected on other sample.

Additional Assumptions

- How does our hypothesis testing structure change when we have two parameters instead of one?

$$H_0\colon p_1 - p_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0 \qquad H_a\colon p_1 - p_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

- Assume two samples are independent of each other.
- **Sample size has to be large just like when we had a single proportion we were testing.**

$$n_1\hat{p}_1 \geq 5, \qquad n_2\hat{p}_2 \geq 5$$

$$n_1(1-\hat{p}_1) \geq 5, \quad n_2(1-\hat{p}_2) \geq 5$$

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)
2. Collect Data (Test Statistic)
3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon\ p_1 - p_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0 \qquad H_a\colon\ p_1 - p_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

2. Collect Data (Test Statistic)
3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon\ p_1 - p_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0 \qquad H_a\colon\ p_1 - p_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{\text{Statistic } - \text{Null Value}}{\text{Standard Error}}$$

3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon\ p_1 - p_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0 \qquad H_a\colon\ p_1 - p_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\bar{p}(1-\bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon\ p_1 - p_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0 \qquad H_a\colon\ p_1 - p_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

2. Collect Data (Test Statistic)

"Average" $\hat{p}$

$$\text{Test Statistic} = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\bar{p}(1-\bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

3. What is probability this happens? (P-value)
4. Decision About Null Hypothesis
5. Summarize

“Average” $\hat{p}$

- For the test statistic, the estimates of the proportion are combined together:

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon\ p_1 - p_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0 \qquad H_a\colon\ p_1 - p_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\bar{p}(1-\bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

3. What is probability this happens? (P-value)

   Derived from normal distribution!

4. Decision About Null Hypothesis
5. Summarize

Hypothesis Test Process

1. Develop your Hypothesis Statements (H_0 and H_a)

$$H_0\colon\ p_1 - p_2 \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} 0 \qquad H_a\colon\ p_1 - p_2 \begin{Bmatrix} < \\ \neq \\ > \end{Bmatrix} 0$$

2. Collect Data (Test Statistic)

$$\text{Test Statistic} = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\bar{p}(1-\bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

3. What is probability this happens? (P-value)

   Derived from normal distribution!

4. Decision About Null Hypothesis
5. Summarize

Exact Same

Confidence Interval

- If we can have hypothesis tests, then we can have confidence intervals as well:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Confidence Interval

- If we can have hypothesis tests, then we can have confidence intervals as well:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Standard Error Changes

TWO-SAMPLE HYPOTHESIS TESTING

Comprehensive Example

Example

- A researcher at a large university on the west coast is interested in comparing some factors between upperclassmen (juniors and seniors) and underclassmen (freshmen and sophomores) in the undergraduate school. The researcher believes that more experience in college may help students perform better in the classroom. The researcher is interested in testing if the average GPA of upperclassmen is greater than the average GPA of underclassmen. The researcher sampled 89 underclassmen with an average GPA of 2.75 with a s.d. of 0.91 and 102 upperclassmen with an average GPA of 3.07 and a s.d. of 1.02.

Example

1. The researcher did not use matched sampling. Do you agree with their decision? Explain.

Example

1. The researcher did not use matched sampling. Do you agree with their decision? Explain.
    - No! Other factors could influence GPA such as major. Would want to at least compare students in same major / college for equal comparison.

Example

2. Conduct a hypothesis test on the variances to see if they are equal.

Example

2. Conduct a hypothesis test on the variances to see if they are equal.

$$H_0\colon \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad H_a\colon \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

$$F = \frac{s_i^2}{s_j^2} = \frac{1.02^2}{0.91^2} = 1.256$$

P-value > 0.05

2*pf(1.256,101,88,lower.tail=F)
0.2741516

Do NOT Reject H_0 → Use Equal Variance Test

Example

3. Conduct the appropriate hypothesis test on the means to see if they are equal.

Example

3. Conduct the appropriate hypothesis test on the means to see if they are equal.

$$H_0\text{: } \mu_U - \mu_L \leq 0 \qquad H_a\text{: } \mu_U - \mu_L > 0$$

$$\text{Test Statistic} = \frac{(\bar{x}_U - \bar{x}_L) - D_0}{s_p\sqrt{\frac{1}{n_U} + \frac{1}{n_L}}} = \frac{(3.07 - 2.75) - 0}{0.97\sqrt{\frac{1}{102} + \frac{1}{89}}} = 2.27$$

P-value = (0.01, 0.025)

df= 102+89-2=189
1-pt(2.27,189)=0.012
OR
pt(2.27,189,lower.tail=F)=0.012

Reject H_0 → Upperclassmen have higher GPA

Example

- Same researcher as before also believes that a higher proportion of upperclassmen live off campus compared to the proportion of underclassmen. While sampling the students in the previous sample, the researcher also asked whether the student lived off campus. Of the 89 underclassmen sampled, 27 lived off campus. Of the 102 upperclassmen sampled, 65 lived off campus.

Example

4. Construct a 95% confidence interval for the difference between the proportion of upperclassmen living off campus to the proportion of underclassmen living off campus.

Example

4. Construct a 95% confidence interval for the difference between the proportion of upperclassmen living off campus to the proportion of underclassmen living off campus.

$$\hat{p}_U = \frac{65}{102} = 0.637 \qquad \hat{p}_L = \frac{27}{89} = 0.303$$

$$(0.637 - 0.303) \pm 1.96 \times \sqrt{\frac{0.637(1-0.637)}{102} + \frac{0.303(1-0.303)}{89}}$$

$$\mathbf{0.334 \pm 0.134}$$

(0.2, 0.468)

Example

5. Conduct the appropriate hypothesis test to test the researcher's claim.

Example

5. Conduct the appropriate hypothesis test to test the researcher's claim.

$$H_0: p_U - p_L \leq 0 \qquad H_a: p_U - p_L > 0$$

$$\text{Test Statistic} = \frac{(0.637 - 0.303) - 0}{\sqrt{0.482(1 - 0.482) \times \left(\frac{1}{102} + \frac{1}{89}\right)}} = 4.61$$

P-value < 0.0005

1-pnorm(4.61) = 2.013e-06
OR
pnorm(4.61,lower.tail=F)=2.013e-06

Reject H_0 → Upperclassmen live off campus more!

Example

6. Can you compare the confidence interval and the hypothesis test?

Example

6. Can you compare the confidence interval and the hypothesis test?
   - No! Hypothesis test was one sided, while confidence interval is two-sided.

ANOVA

More than Two Parameters

- Sometimes we like to compare multiple parameters against each other.
- Now we will focus on comparing *means for more than two groups*.
- These tests are called **analysis of variance (ANOVA)**.

ANOVA

One-Way ANOVA

One-Way ANOVA

- The simplest form of ANOVA is the one-way model.

- **One-Way ANOVA** – design in which independent samples are obtained from k levels of a single factor, then testing whether the k levels have equal means.

One-Way ANOVA

- The simplest form of ANOVA is the one-way model.
- **One-Way ANOVA** – design in which independent samples are obtained from k levels of a single **factor**, then testing whether the k levels have equal means.

One-Way ANOVA

- The simplest form of ANOVA is the one-way model.

- **One-Way ANOVA** – design in which independent samples are obtained from k levels of a single **explanatory variable**, then testing whether the k levels have equal means.

One-Way ANOVA

- The simplest form of ANOVA is the one-way model.

- **One-Way ANOVA** – design in which independent samples are obtained from k **levels** of a single **explanatory variable**, then testing whether the k **levels** have equal means.

One-Way ANOVA

- The simplest form of ANOVA is the one-way model.

- **One-Way ANOVA** – design in which independent samples are obtained from k **categories** of a single **explanatory variable**, then testing whether the k **categories** have equal means.

One-Way ANOVA

- The simplest form of ANOVA is the one-way model.
- **One-Way ANOVA** – design in which independent samples are obtained from k categories of a single explanatory variable, then testing whether the k categories have equal means.
- Similar to regression analysis where we have one categorical variable predicting a continuous response.

Hypotheses

- **One-Way ANOVA** – design in which independent samples are obtained from k categories of a single explanatory variable, then **testing** whether the k categories **have equal means**.

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_a: \text{At least one mean different than another}$$

Assumptions

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_a: \text{At least one mean different than another}$$

- To test the above hypotheses, there are some assumptions that are made:
  1. Normally distributed categories
  2. Equality of variances between categories
  3. Independence

Assumptions

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_a: \text{At least one mean different than another}$$

- To test the above hypotheses, there are some assumptions that are made:
  1. Normally distributed categories
  2. Equality of variances between categories
  3. Independence

Slightly different than with two categories

Equality of Variance for More than 2

- Testing the equality of variance for more than two variances is slightly different than when we tested only two variances.
- Hypotheses: $H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$

  H_a: At least one variance different than another

- Test Statistic:

$$F = \frac{s_{max}^2}{s_{min}^2}$$

- P-value: Calculated from Hartley's F-max distribution which won't be covered in this course.

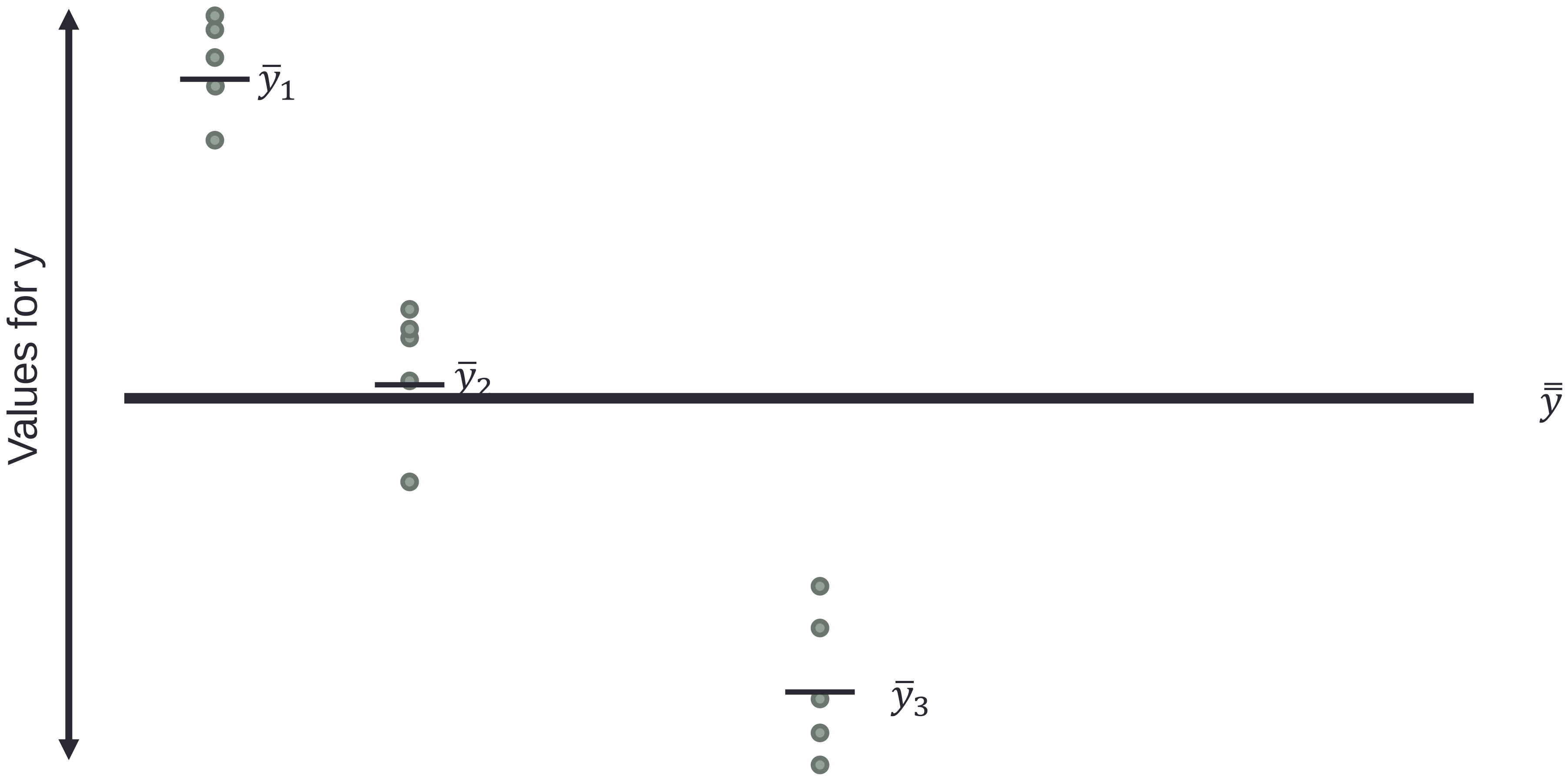
Sources of Variation

- Variation in an ANOVA can come from two places – within a category / level and across different categories / levels.

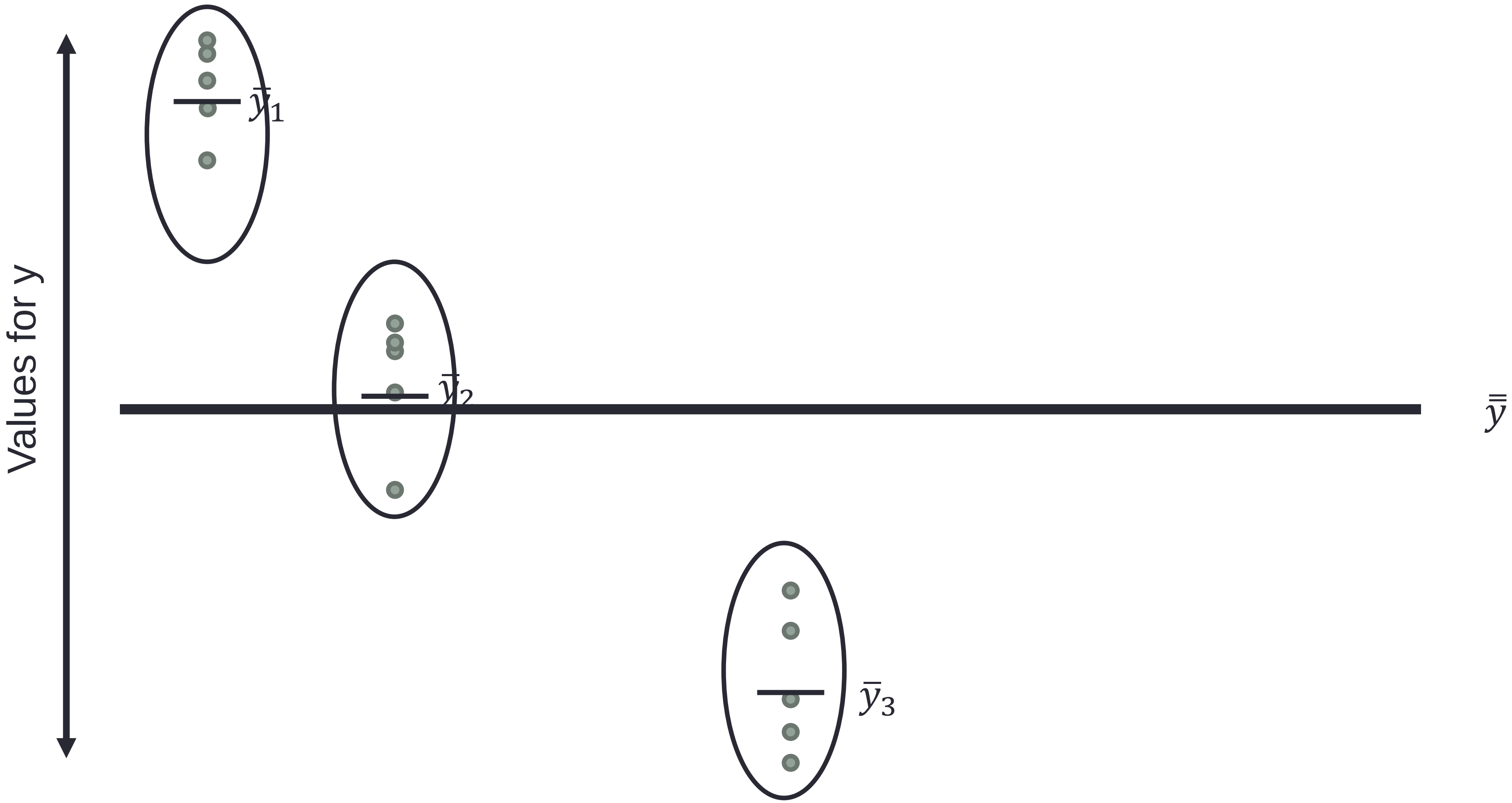
- **Within-Sample Variability** – variability in the response that exists within a category of a variable

- **Between-Sample Variability** – variability in the response that exists between categories of a variable

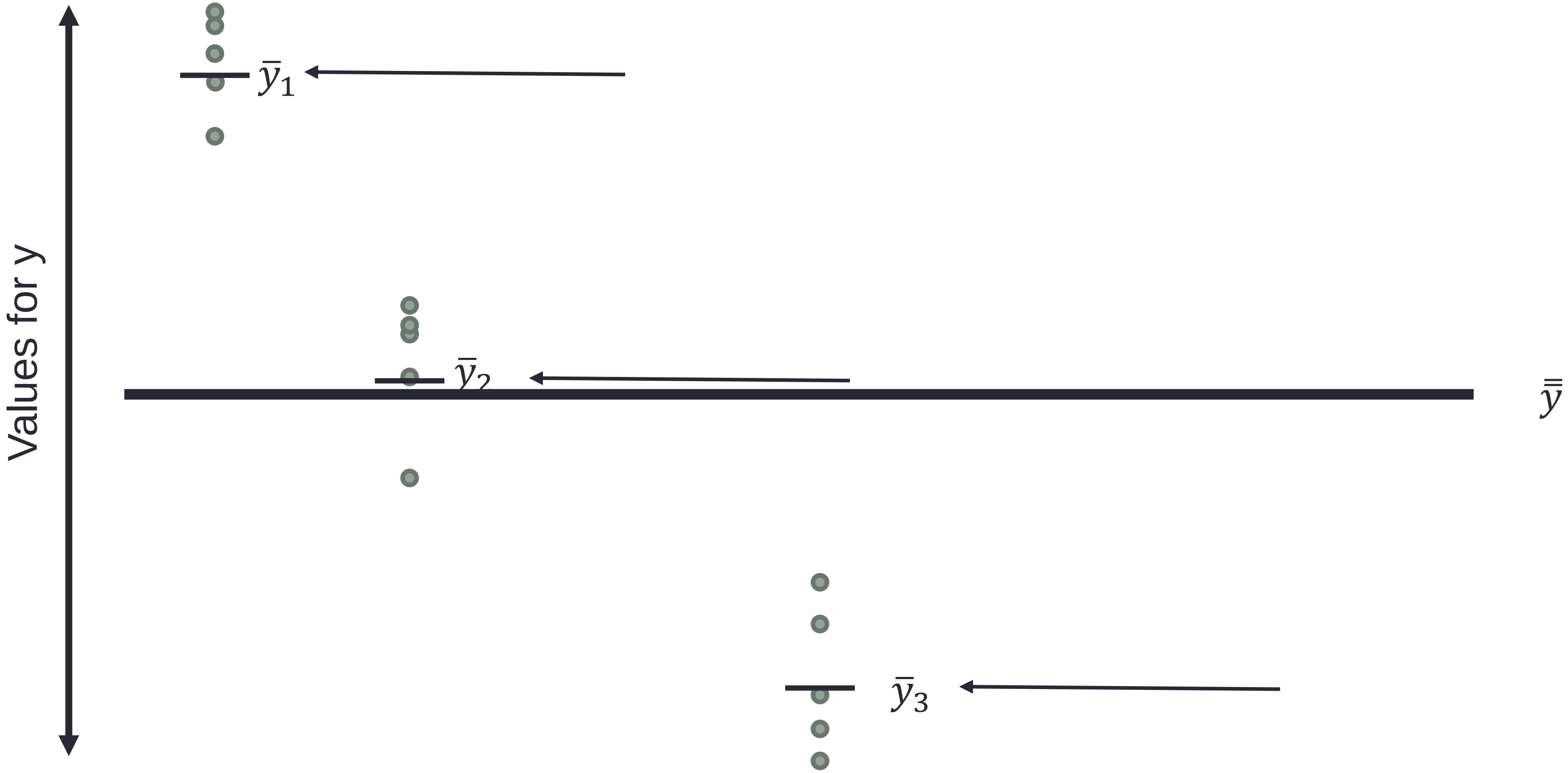
Within versus Between variability



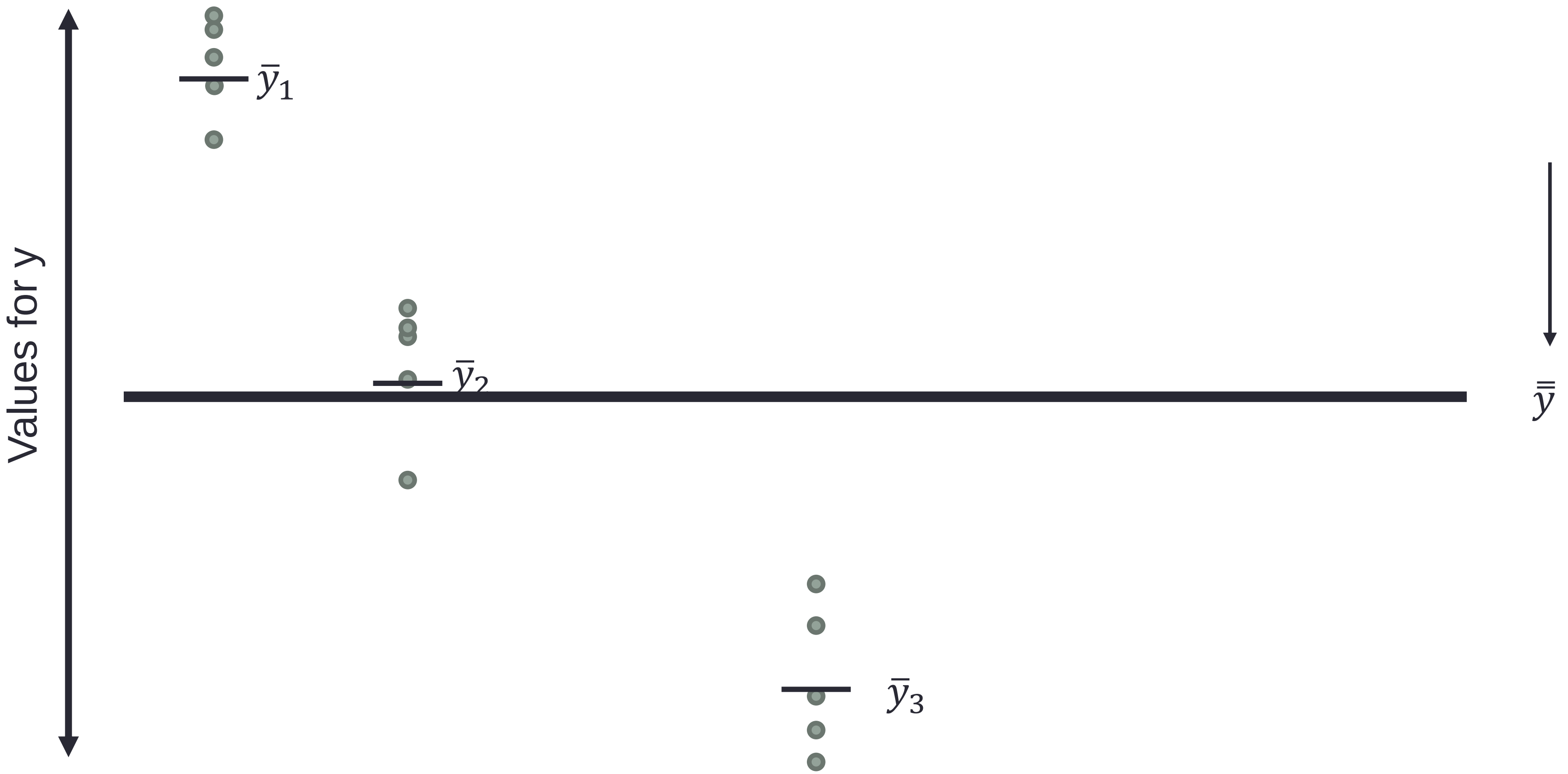
Within versus Between variability



Within versus Between variability

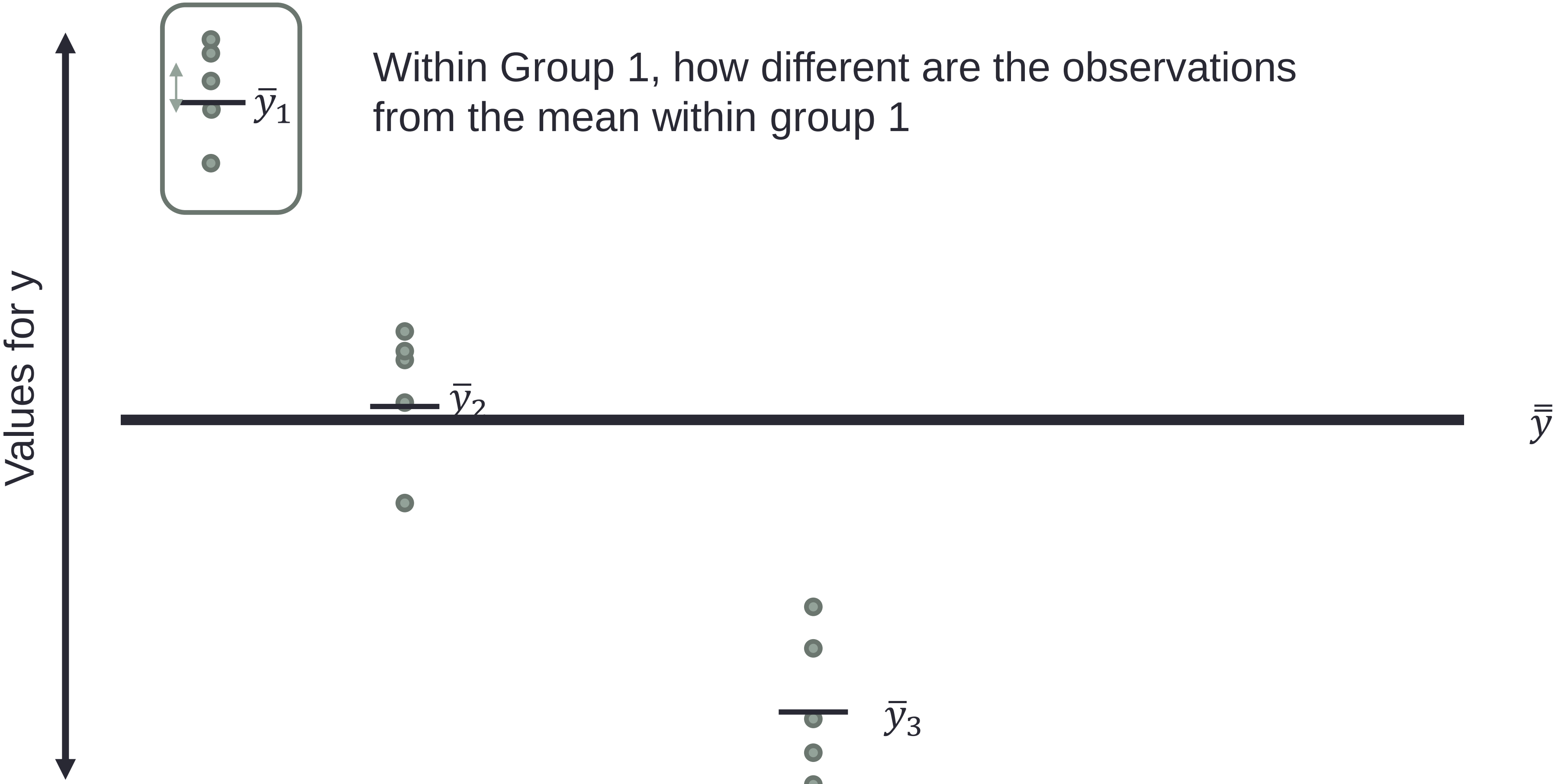


Within versus Between variability

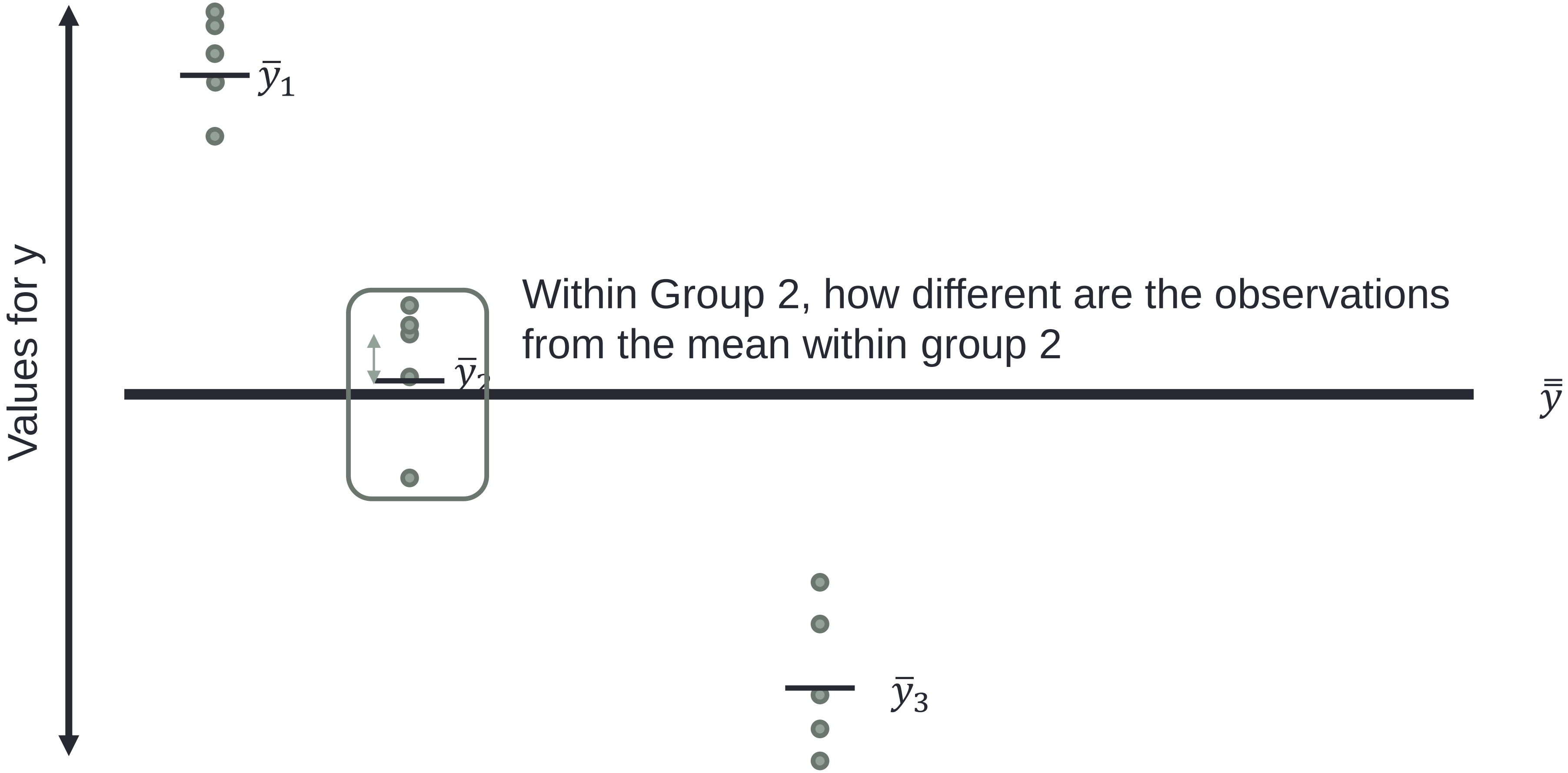


Within versus Between variability

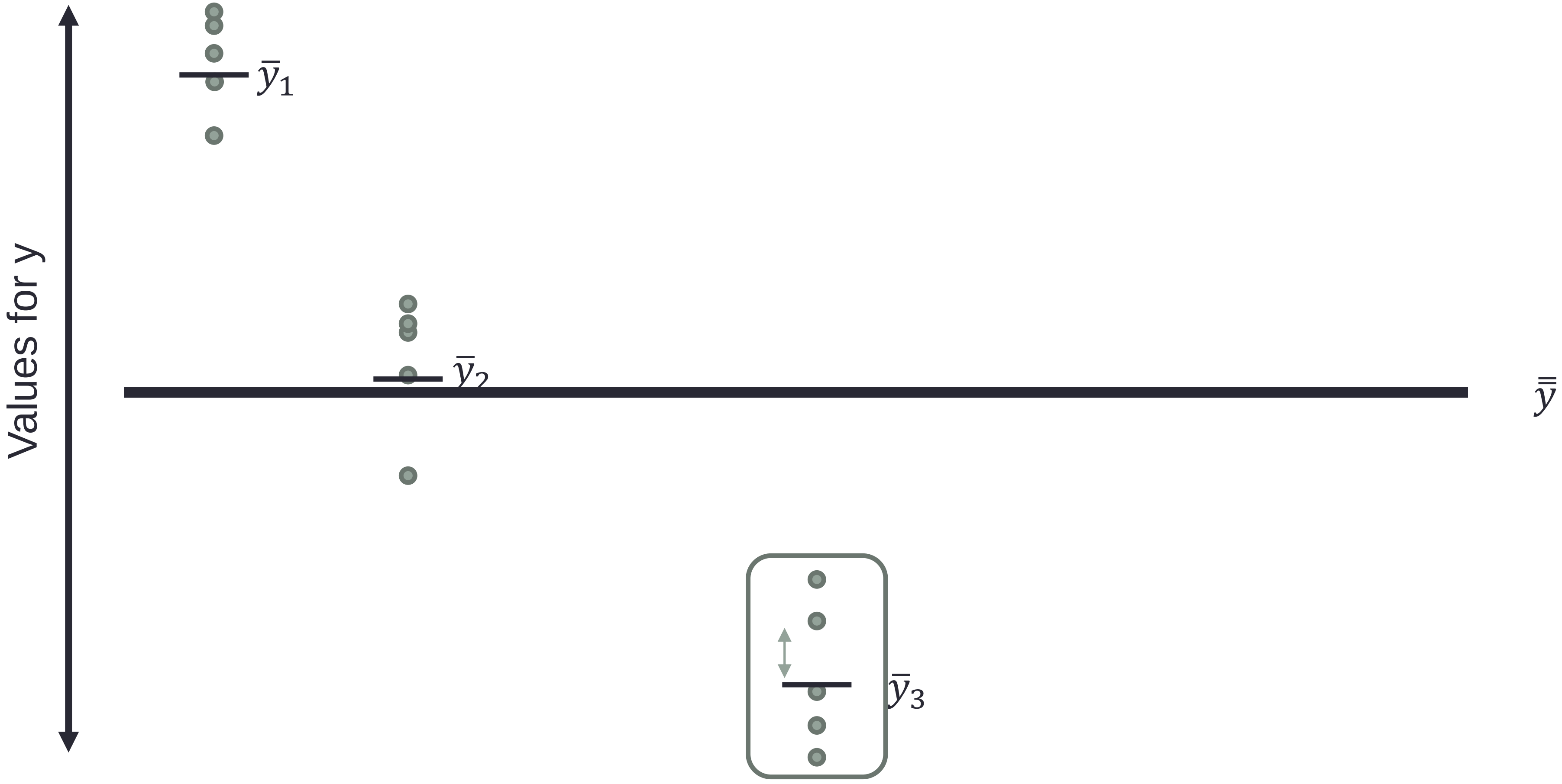
Within Group 1, how different are the observations from the mean within group 1



Within versus Between variability

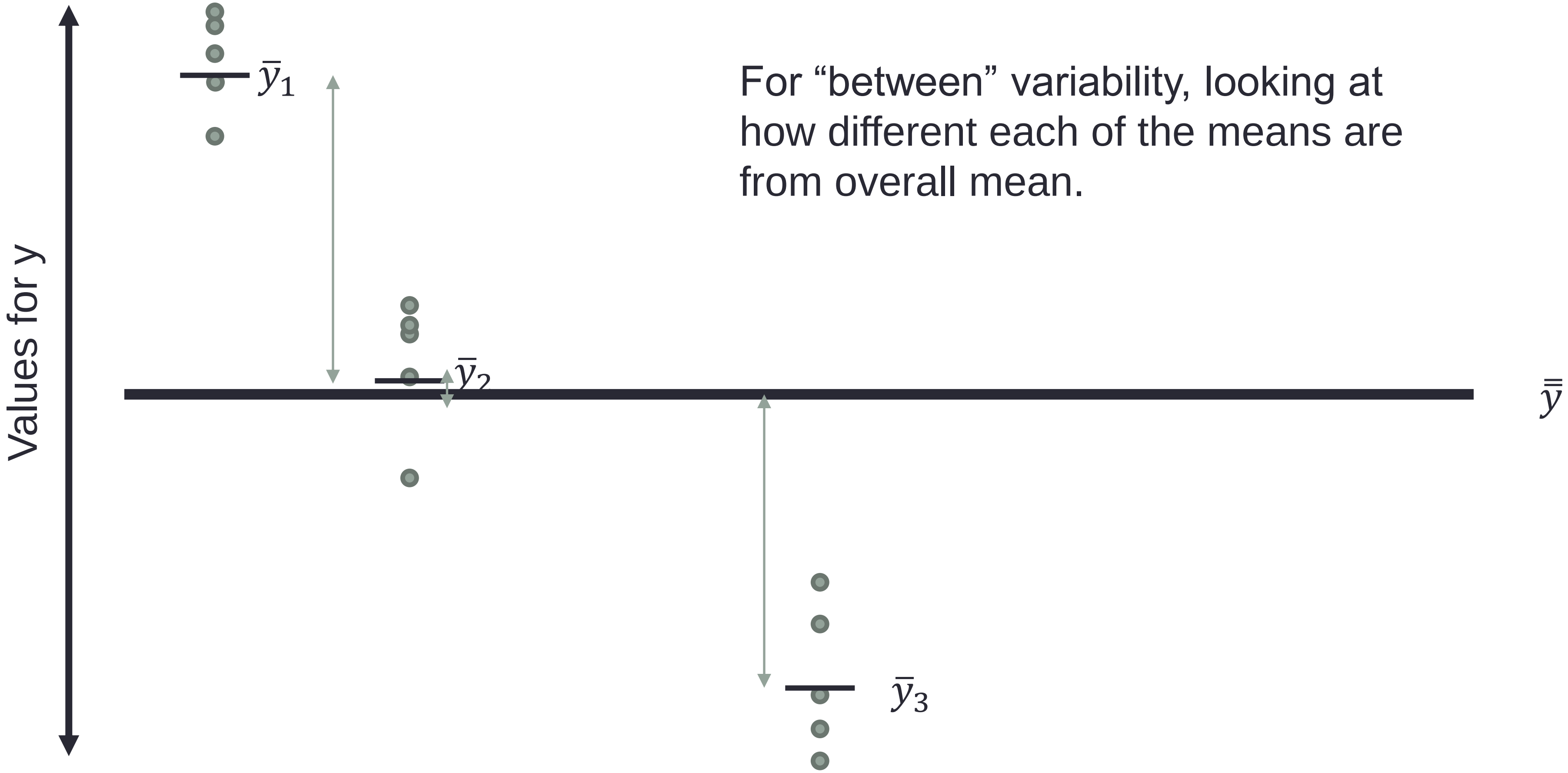


Within versus Between variability



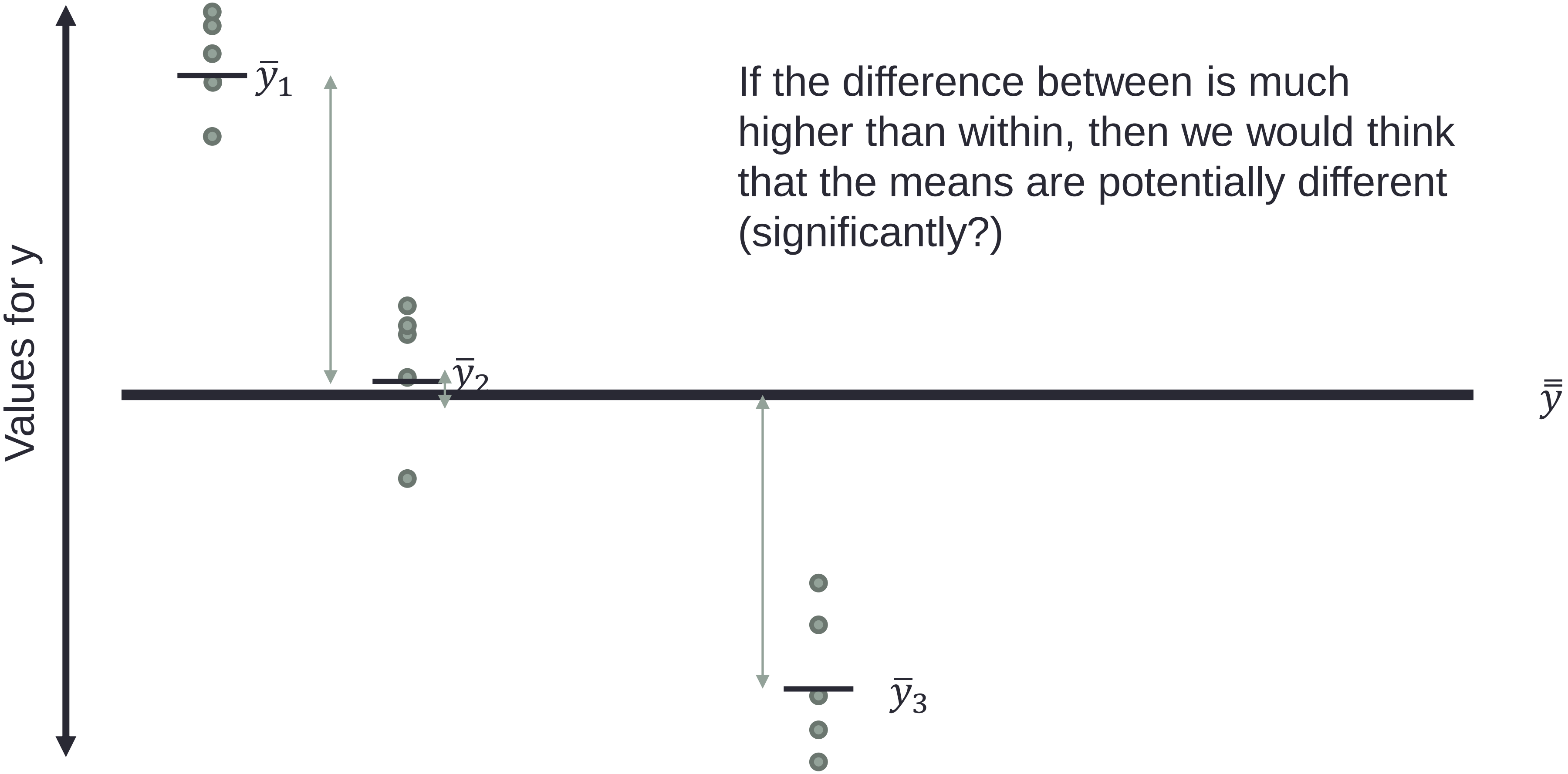
Within Group 3, how different are the observations from the mean within group 3

Within versus Between variability



For "between" variability, looking at how different each of the means are from overall mean.

Within versus Between variability



If the difference between is much higher than within, then we would think that the means are potentially different (significantly?)

Sources of Variation

- Variation in an ANOVA can come from two places – within a category / level and across different categories / levels.

- **Within-Sample Variability** – variability in the response that exists within a category of a variable
  - What your categories CANNOT explain (similar to SSE also referred to SSW)

- **Between-Sample Variability** – variability in the response that exists between categories of a variable
  - What your categories CAN explain (similar to SSR also referred to as SSB)

Sum of Squares Within (SSW or SSE)

- Within sample variability is the variability that you cannot explain by just knowing which category your observation falls into.
- For example, red cars may be more expensive than blue cars, but that doesn't mean all red cars have the same price – price differences within red cars are SSW!

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{i,j} - \bar{y}_i\right)^2$$

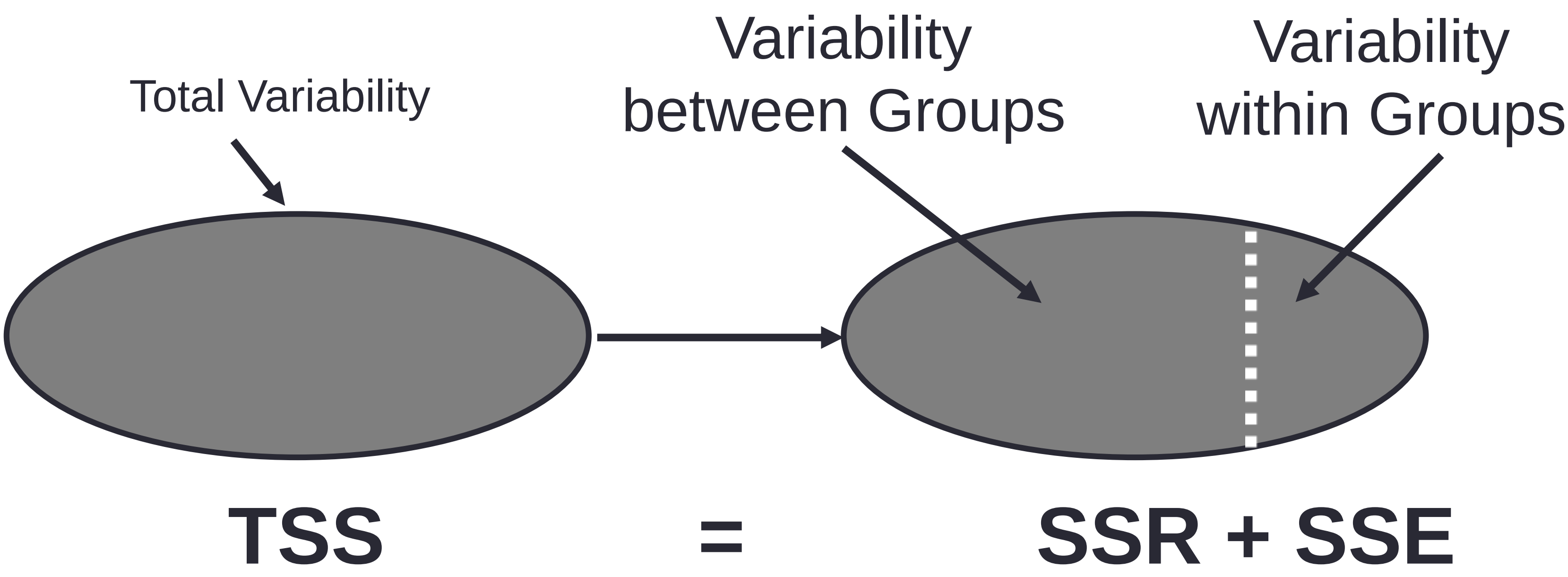
i defines which group observation comes from and j defines the observation

Sum of Squares Between (SSB or SSR)

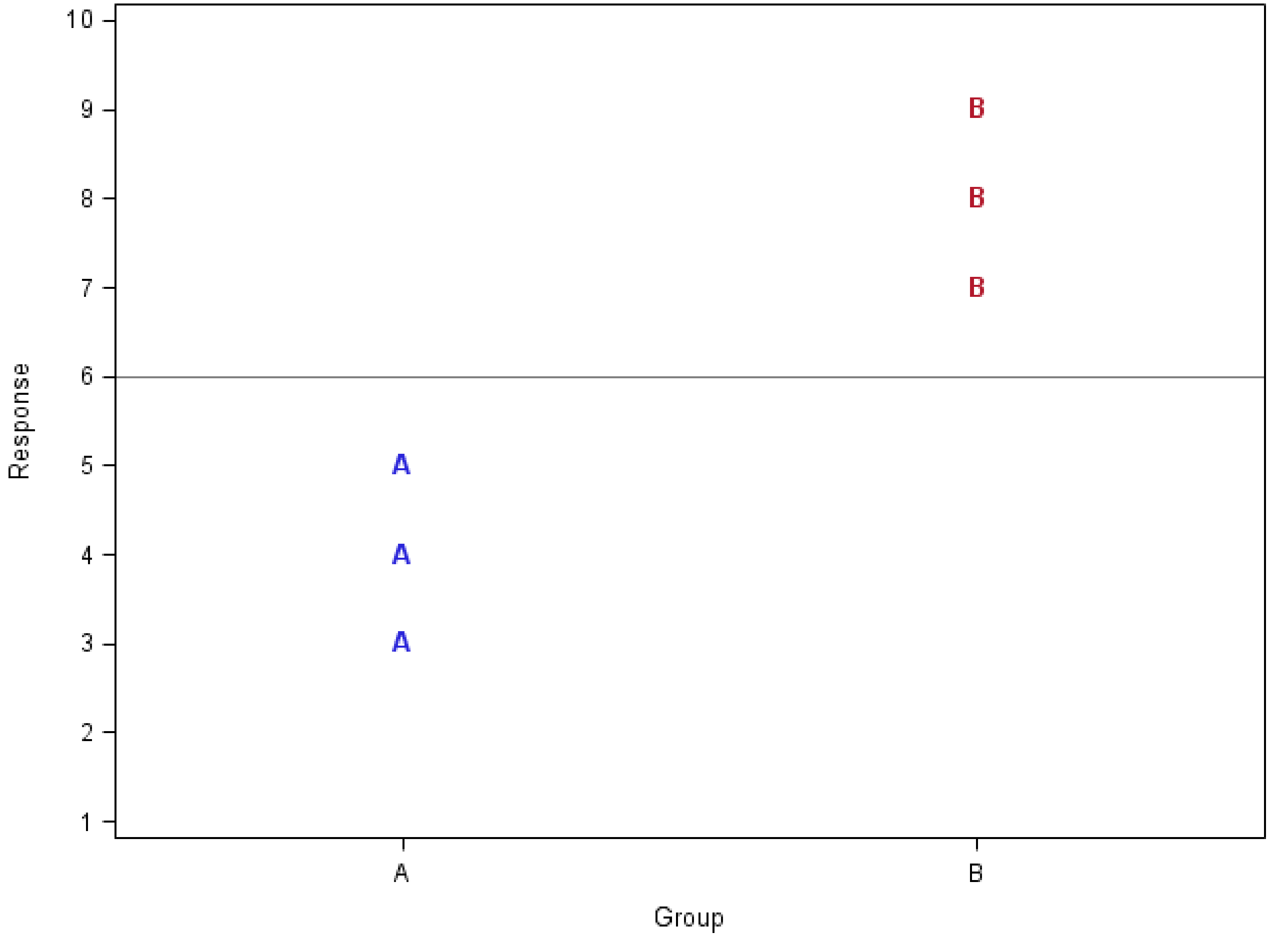
- Between sample variability is the variability that you can explain by just knowing which category your observation falls into.
- For example, red cars may be more expensive than blue cars – price differences between red and blue cars are SSB!

$$SSB = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{\bar{y}})^2$$

Partitioning Variability in ANOVA

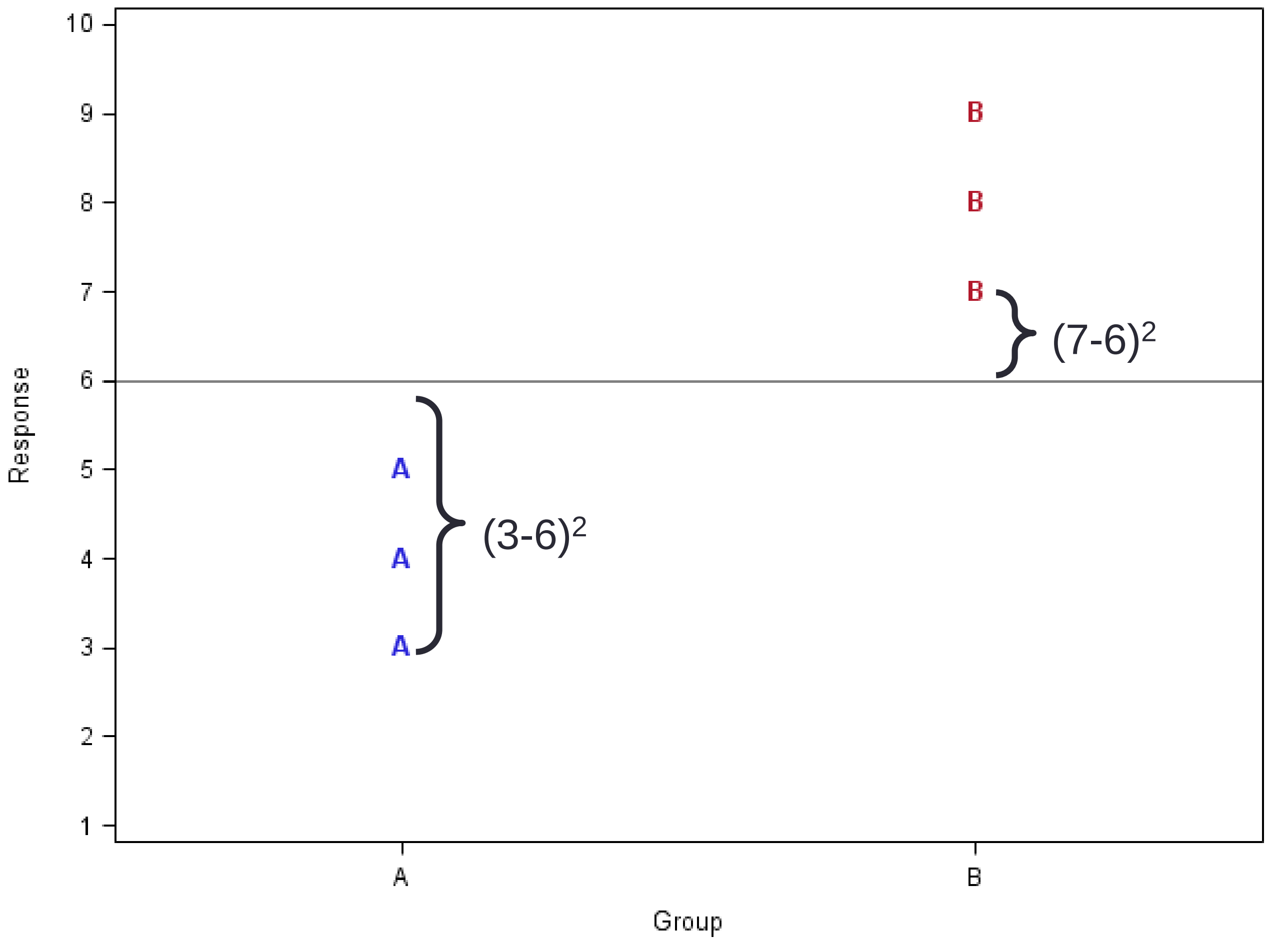


Sums of Squares



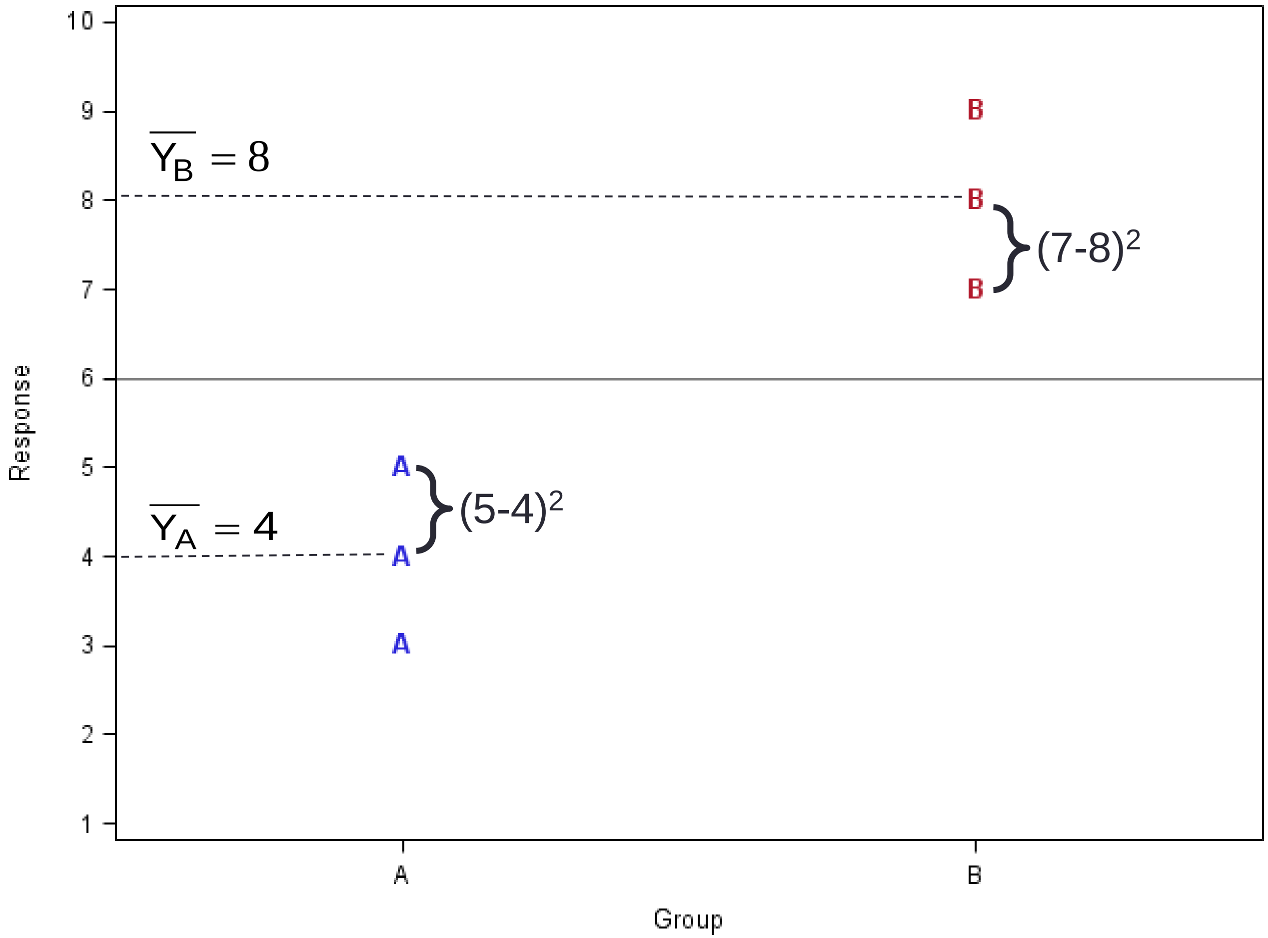
$$\text{Overall mean} = \bar{\bar{y}} = \frac{3+4+5+7+8+9}{6} = \mathbf{6}$$

Total Sum of Squares



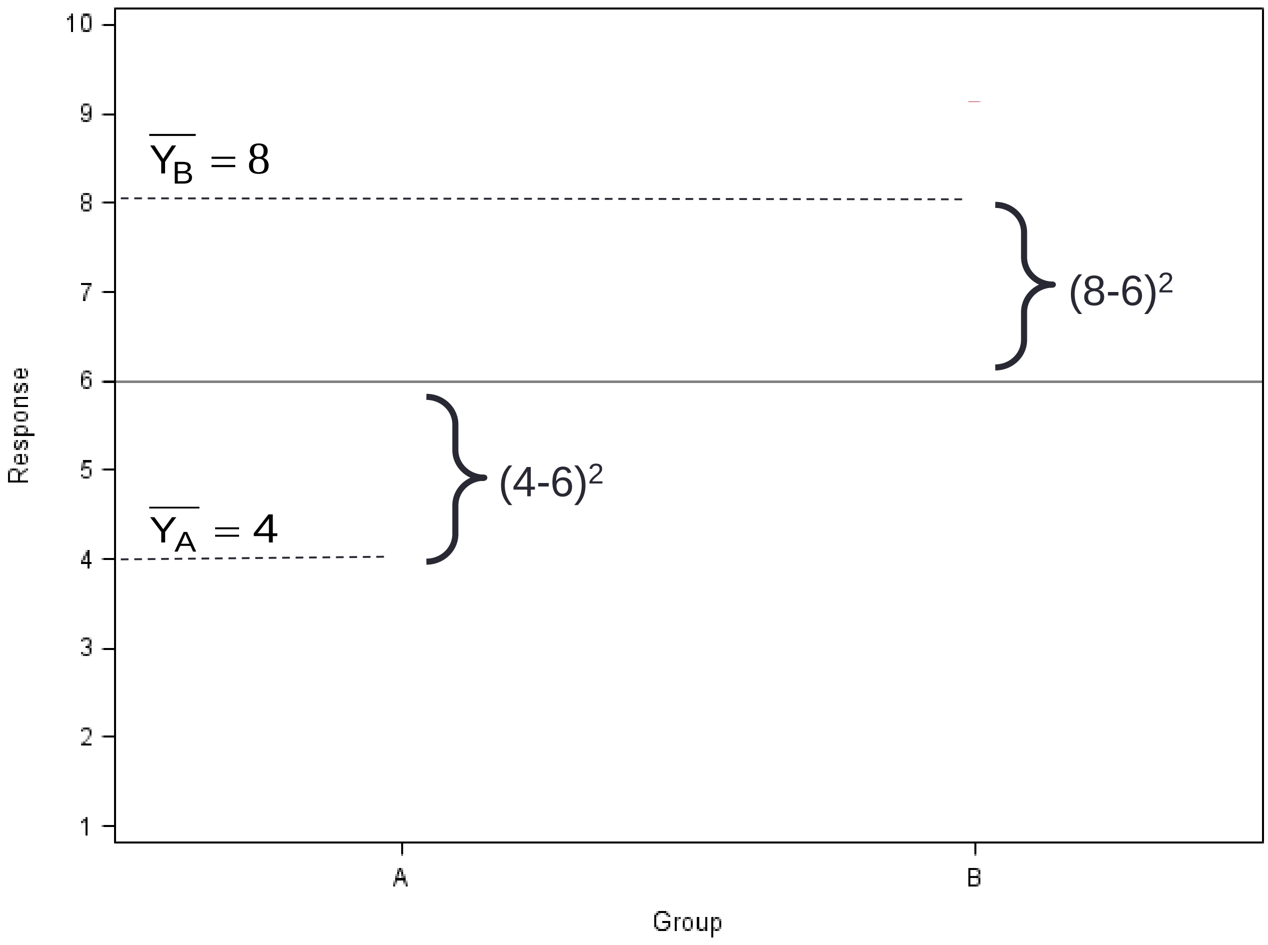
TSS= $(3\text{-}6)^2 + (4\text{-}6)^2 + (5\text{-}6)^2 + (7\text{-}6)^2 + (8\text{-}6)^2 + (9\text{-}6)^2 =$ **28**

Error Sum of Squares



SSE = $(3-4)^2 + (4-4)^2 + (5-4)^2 + (7-8)^2 + (8-8)^2 + (9-8)^2 =$ **4**

Model Sum of Squares



$\textbf{SSR} = 3*(4-6)^2 + 3*(8-6)^2 = \textbf{24}$

ANOVA F-test

- The ANOVA F-test is similar to the global F-test in multiple linear regression – trying to test multiple things at once.

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_a: \text{At least one mean different than another}$$

ANOVA F-test

- The ANOVA F-test is similar to the global F-test in multiple linear regression – trying to test multiple things at once.

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_a: \text{At least one mean different than another}$$

- The test statistic follows an *F*-distribution and is calculated as follows:

$$F = \frac{\left(\frac{SSB}{k-1}\right)}{\left(\frac{SSW}{N-k}\right)}$$

ANOVA F-test

- The ANOVA F-test is similar to the global F-test in multiple linear regression – trying to test multiple things at once.

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_a: \text{At least one mean different than another}$$

- The test statistic follows an *F*-distribution and is calculated as follows:

$$F = \frac{\left(\frac{SSB}{k-1}\right)}{\left(\frac{SSW}{N-k}\right)}$$

K categories would be k-1 variables in a regression model

ANOVA F-test

- The ANOVA F-test is similar to the global F-test in multiple linear regression – trying to test multiple things at once.

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_a: \text{At least one mean different than another}$$

- The test statistic follows an *F*-distribution and is calculated as follows:

$$F = \frac{\left(\frac{SSB}{k-1}\right)}{\left(\frac{SSW}{N-k}\right)}$$

Total sample size across all categories

ANOVA F-test

- The ANOVA F-test is similar to the global F-test in multiple linear regression – trying to test multiple things at once.

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_a: \text{At least one mean different than another}$$

- The test statistic follows an *F*-distribution and is calculated as follows:

$$F = \frac{\left(\frac{SSB}{k-1}\right)}{\left(\frac{SSW}{N-k}\right)} = \frac{MSB}{MSW}$$

One-Way ANOVA Table

Source	DF	SS	MS	F-Value	P-Value
Between	$k-1$	SSB	$\frac{SSB}{k-1}$	$\frac{MSB}{MSW}$	…
Within	$N-k$	SSW	$\frac{SSW}{N-k}$		
Total	$N-1$	TSS			

P-value is calculated by P(F>F-value)

Using R, this would be:

```
1-pf(F-value, numerator df, denominator df)   OR
  pf(F-value, numerator df, denominator df, lower.tail=F)
```

Example

- A marketing analyst is interested in testing the effectiveness of 4 different commercials describing their company’s new product. The marketing analyst randomly assigns a commercial to each of 32 cities across the country and measures the average increase in sales of their new product at their stores. The marketing analyst wants to test if there is a difference in sales between the commercials.

Example

1. Fill in the blanks on the ANOVA table.

Source	DF	SS	MS	F-Value	P-Value
Between		2.3236			
Within		0.9587			
Total		3.2823			

Example

1. Fill in the blanks on the ANOVA table.

Source	DF	SS	MS	F-Value	P-Value
Between	**3**	2.3236	**0.775**	**22.79**	**< 0.05**
Within	**28**	0.9587	**0.034**		
Total	**31**	3.2823			

pf(22.79,3,28,lower.tail=F) = 1.130282e-07

ANOVA

Multiple Comparisons

Next Steps

- If you reject the null hypothesis on the F-test what does that mean?

Next Steps

- If you reject the null hypothesis on the F-test what does that mean? **Evidence shows at least one category is different**.

- But which category?!?!?!?!

Next Steps

- If you reject the null hypothesis on the F-test what does that mean? **Evidence shows at least one category is different**.

- Once a difference is detected, must test each individual pair of categories to find where all the differences are – a process called **multiple comparisons** or **ad-hoc testing**.

Multiple Comparisons Problem

- You have a coin which lands on heads 50% of the time when flipped.
- What is the probability of flipping a head on your first flip?
- What is the probability of flipping a head on your second flip?
- What is the probability of flipping *at least* one head in two flips?

Multiple Comparisons Problem

- You have a coin which lands on heads 50% of the time when flipped.
- What is the probability of flipping a head on your first flip?
  - 50%
- What is the probability of flipping a head on your second flip?
  - 50%
- What is the probability of flipping *at least* one head in two flips?
  - 75%

Multiple Comparisons Problem

- You have a test which makes an error 5% of the time when performed.
- What is the probability of making an error on your first test?

- What is the probability of making an error on your second test?

- What is the probability of making *at least* one error in two tests?

Multiple Comparisons Problem

- You have a test which makes an error 5% of the time when performed.
- What is the probability of making an error on your first test?
  - 5%
- What is the probability of making an error on your second test?
  - 5%
- What is the probability of making *at least* one error in two tests?
  - 9.75%

Multiple Comparisons Problem

- You have a test which makes an error 5% of the time when performed.
- What is the probability of making an error on your first test?
  - 5%

Comparison-wise Error

- What is the probability of making an error on your second test?
  - 5%
- What is the probability of making *at least* one error in two tests?
  - 9.75%

Multiple Comparisons Problem

- You have a test which makes an error 5% of the time when performed.
- What is the probability of making an error on your first test?
  - 5%
- What is the probability of making an error on your second test?
  - 5%
- What is the probability of making *at least* one error in two tests?
  - 9.75%

Experiment-wise Error

Different Types of Errors

- **Comparison-wise error rate** is the error rate for each individual test or comparison.
- **Experiment-wise error rate** is the error rate across all comparisons – proportion of experiments/comparisons in which at least one error occurs.

- Tests and confidence intervals typically control for comparison-wise error rates, α, but ideally we want to control for experiment-wise error.

Multiple Comparison Methods

Number of Groups Compared	Number of Comparisons	Experimentwise Error Rate (α=0.05)
2	1	.05
3	3	.14
4	6	.26
5	10	.40

Comparisonwise Error Rate = α = 0.05

EER $\leq 1 - (1 - \alpha)^{nc}$ where *nc* = number of comparisons

Multiple Comparison Methods

Control
Comparisonwise
Error Rate

→

Pairwise t-tests

Control
Experimentwise
Error Rate

→

Compare All Pairs
Tukey

Tukey's HSD Test

- HSD = Honest Significant Difference
- This method is appropriate when you consider pairwise comparisons.
- The experimentwise error rate is
  - equal to alpha when ***all*** pairwise comparisons are considered
  - less than alpha when ***fewer*** than all pairwise comparisons are considered.
- Also known as the **Tukey-Kramer Test.**
- This can be done with hypothesis tests or confidence intervals.

Tukey-Kramer Critical Range

- This approach essentially replaces the margin of error calculation for a typical confidence interval for a difference in means with an adjusted margin of error:

$$\text{Critical Range (Margin of Error)} = q_\alpha \times \sqrt{\frac{MSW}{2} \times \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Tukey-Kramer Critical Range

- This approach essentially replaces the margin of error calculation for a typical confidence interval for a difference in means with an adjusted margin of error:

$$\text{Critical Range (Margin of Error)} = q_\alpha \times \sqrt{\frac{MSW}{2} \times \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

From studentized range distribution – not covered in detail here

Example

- The same marketing analyst as before is interested in testing the effectiveness of 4 different commercials describing their company's new product. The marketing analyst randomly assigns a commercial to each of 32 cities across the country and measures the average increase in sales of their new product at their stores. The four commercial average sales were $1.2M for commercial A, $1.8M for B, $0.76M for C, and $1.3M for D. Where are the differences in sales?

Example

$$CR = q_\alpha \times \sqrt{\frac{MSW}{2} \times \left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = 3.90 \times \sqrt{\frac{0.034}{2} \times \left(\frac{1}{8} + \frac{1}{8}\right)}$$

$$CR = 0.254$$

Commercial Pair	Absolute Difference	Larger than CR (Diff. Exists)?
A, B	0.6	**YES**
A, C	0.44	**YES**
A, D	0.1	**NO**
B, C	0.5	**YES**
B, D	1.04	**YES**
C, D	0.54	**YES**

ANOVA

Fixed vs. Random Effects

Inference Considerations

- The inference drawn from an ANOVA procedure depends on whether the factor levels in the procedure are selected on purpose or randomly.

Inference Considerations

- The inference drawn from an ANOVA procedure depends on whether the factor levels in the procedure are selected on purpose or randomly.

- **Fixed Effects –** inferences extend only to the factor levels being analyzed because the levels were **purposefully** chosen as the levels of interest (ONLY interested in levels chosen!).

Inference Considerations

- The inference drawn from an ANOVA procedure depends on whether the factor levels in the procedure are selected on purpose or randomly.

- **Fixed Effects –** inferences extend only to the factor levels being analyzed because the levels were **purposefully** chosen as the levels of interest.

- **Random Effects –** inferences extend beyond just the factor levels being tested because the levels were **randomly** selected from a larger group of levels.

ANOVA

Randomized Blocking

Recall – Sources of Variation

- Generally comparing two population means works well in certain situations.
- However, there are some instances where a paired difference (**matched**) **sample** is used to control for **sources of variation** that might distort the conclusions.

Sources of Variation – ANOVA

- Generally comparing many population means works well in certain situations.
- However, there are some instances where **blocking** is used to control for **sources of variation** that might distort the conclusions.

Example

- The same marketing analyst as before is interested in testing the effectiveness of 4 different commercials describing their company’s new product. The marketing analyst randomly assigns a commercial to each of 32 cities across the country and measures the average increase in sales of their new product at their stores. The four commercial average sales were $1.2M for commercial A, $1.8M for B, **$0.76M for C**, and $1.3M for D. Where are the differences in sales?

What if the new product is a warm coat and a majority of the cities seeing C were warm weather cities?

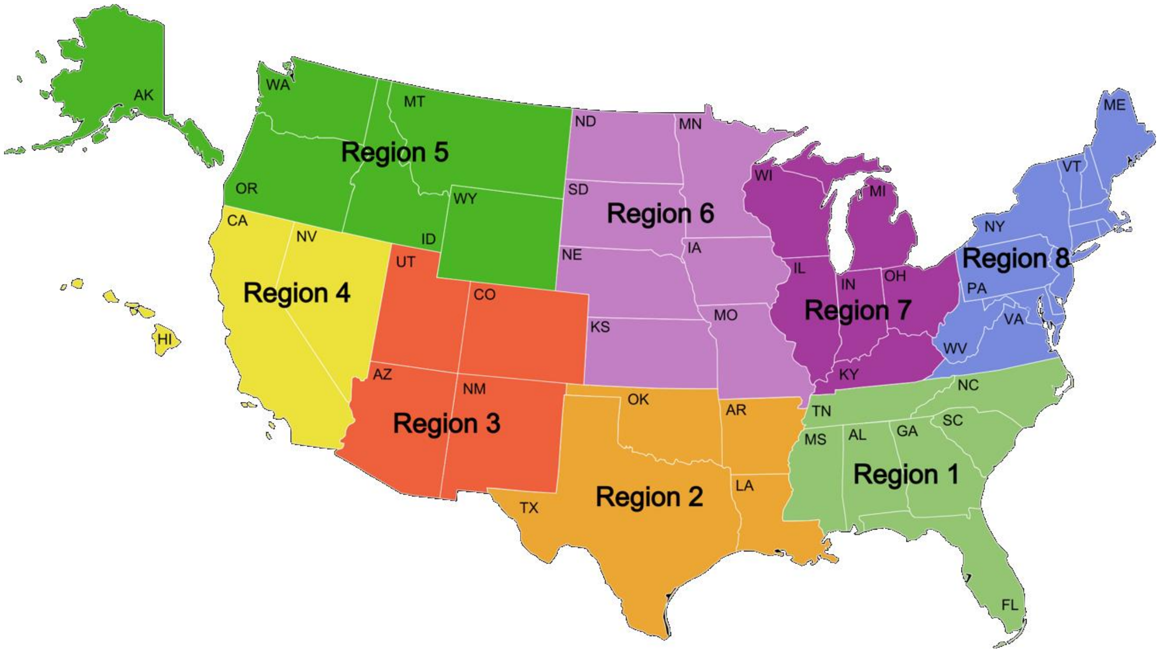
Example

- The same marketing analyst as before is interested in testing the effectiveness of 4 different commercials describing their company's new product. The marketing analyst randomly assigns a commercial to each of 32 cities across the country and measures the average increase in sales of their new product at their stores. The four commercial average sales were $1.2M for commercial A, $1.8M for B, **$0.76M for C**, and $1.3M for D. Where are the differences in sales?

What if the new product is a warm coat and a majority of the cities seeing C were warm weather cities? **FAIR?**

Example

- The same marketing analyst as before is interested in testing the effectiveness of 4 different commercials describing their company’s new product. Split (**block**) country into 8 regions. Show each commercial to one city in each region. Sample size still 32.



Assumptions

- Same as with One-Way ANOVA:
  1. Normally distributed categories
  2. Equality of variances between categories
  3. Independence

Sources of Variation – ANOVA

- Blocking not only comes in with the collection of the data, but also in the analysis of the data as a variable being added to the model.
- With a new variable being added to the model, comes a new source of variation – **the sum of squares of blocking**.

Sum of Squares Blocks

- The sum of squares blocks is the following:

$$SSBL = \sum_{j=1}^{b} k\left(\bar{x}_j - \bar{\bar{x}}\right)^2$$

- This sum of squares comes out of the error sum of squares and gets brought into the model – essentially shrinking the SSW (SSE) even more.

Sum of Squares Blocks

- The sum of squares blocks is the following:

$$SSBL = \sum_{j=1}^{b} k\left(\bar{x}_j - \bar{\bar{x}}\right)^2$$

- This sum of squares comes out of the error sum of squares and gets brought into the model – essentially shrinking the SSW (SSE) even more.

Sum of Squares Blocks

- The sum of squares blocks is the following:

$$SSBL = \sum_{j=1}^{b} k\left(\bar{x}_j - \bar{\bar{x}}\right)^2$$

- This sum of squares comes out of the error sum of squares and gets brought into the model – essentially shrinking the SSW (SSE) even more.

Blocking ANOVA Table

Source	DF	SS	MS	F-Value	P-Value
Blocking	$b - 1$	$SSBL$	$\frac{SSBL}{b-1}$	$\frac{MSBL}{MSW}$	...
Between	$k - 1$	SSB	$\frac{SSB}{k-1}$	$\frac{MSB}{MSW}$	...
Within	$(k-1)(b-1)$	SSW	$\frac{SSW}{(k-1)(b-1)}$		
Total	$N - 1$	TSS			

Blocking ANOVA Table

Source	DF	SS	MS	F-Value	P-Value
Blocking	$b-1$	$SSBL$	$\frac{SSBL}{b-1}$	$\frac{MSBL}{MSW}$	...
Between	$k-1$	SSB	$\frac{SSB}{k-1}$	$\frac{MSB}{MSW}$	...
Within	$(k-1)(b-1)$	SSW	$\frac{SSW}{(k-1)(b-1)}$		
Total	$N-1$	TSS			

Mean Square
Blocking

Blocking ANOVA Table

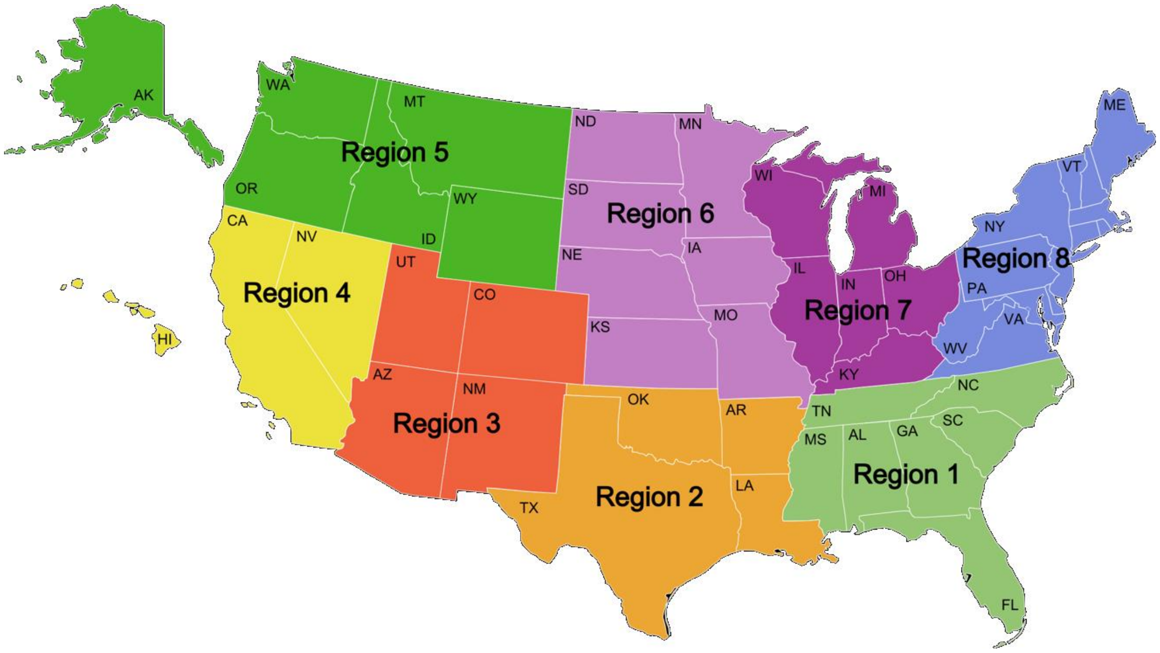
Source	DF	SS	MS	F-Value	P-Value
Blocking	$b-1$	$SSBL$	$\frac{SSBL}{b-1}$	$\frac{MSBL}{MSW}$	...
Between	$k-1$	SSB	$\frac{SSB}{k-1}$	$\frac{MSB}{MSW}$	...
Within	$(k-1)(b-1)$	SSW	$\frac{SSW}{(k-1)(b-1)}$		
Total	$N-1$	TSS			

$H_0: \mu_{b_1} = \cdots = \mu_{b_b}$

H_a: at least one block mean not equal

Example

- The same marketing analyst as before is interested in testing the effectiveness of 4 different commercials describing their company's new product. Split (**block**) country into 8 regions. Show each commercial to one city in each region. Sample size still 32. Fill out new ANOVA table.



Example

Source	DF	SS	MS	F-Value	P-Value
Blocking		1.587			
Between		2.923			
Within		0.470			
Total		4.980			

Example

Source	DF	SS	MS	F-Value	P-Value
Blocking	7	1.587	0.227	10.32	< 0.05
Between	3	2.923	0.974	44.27	< 0.05
Within	21	0.470	0.022		
Total	31	4.980			

Post-hoc Analysis for Blocking

- Tukey-Kramer ANOVA comparisons do not work for blocking designs.
- Instead use **Fisher's Least Significant Difference**.
- Similar to Tukey's critical range, the Fisher's LSD is a recalculation of the margin of error for the difference in means confidence interval:

$$LSD = t^* \times \sqrt{MSW} \times \sqrt{\frac{2}{b}}$$