# Power and FWER

Grinnell College

November 12, 2025

# Warm-Up

1. Is it possible to simultaneously commit both a Type I and Type II error? Why or why not?
2. What impact does our sample size have on power?
3. What impact does our sample size have on confidence/error?
4. What impact does effect size have on both power?

# Today

1. Tension between confidence and power
2. Tables, accuracy, and error (oh my!)
3. Family-wise error rates

# Drawing Conclusions

As we never truly know whether $H_0$ is correct or not, we must simultaneously be prepared to combat both types of error

| | True State of Nature | |
|---|---|---|
| Test Result | $H_0$ True | $H_0$ False |
| Fail to reject $H_0$ | Correct $(1 - \alpha)$ | Type II Error $(\beta)$ |
| Reject $H_0$ | Type I Error $(\alpha)$ | Correct $(1 - \beta)$ |

▶ Type I error $= P(\text{Reject } H_0 | H_0 \text{ true}) = $ false alarm
▶ Type II error $= P(\text{Fail to reject } H_0 | H_A \text{ true}) = $ missed opportunity

# Tension

Board-work

# Rare Disease Testing

Suppose we have created a diagnostic test for a rare disease that affects approximately 1% of the population. The test has power of 90% and a Type I error rate of 90%

- ▶ What is the null hypothesis for this test?
- ▶ What is a Type I and Type II error? Which is more important?
- ▶ Suppose somebody tests positive for the disease. What is the probability that they actually have it? (*True positive rate*)

# EXAM-ple

We are asked to evaluate a piece of diagnostic software that is used to sort through and categorize email as being spam or not spam. Email that is not spam is sent to your inbox, while email that is spam is sent to a junk folder.

- ▶ What is the null hypothesis for incoming email?

- ▶ What is a Type I and Type II error? Which is more important here?

- ▶ Suppose that the diagnostics of the spam software filter has a Type I error rate of 2% and a Type II error rate of 20%. In a typical month, the average user will receive 1,000 emails, with approximately 5% of them being spam. *Supposing that an email is marked as spam*, what is the probability that a given email was not spam? (*False positive rate*)

# Multiple Comparisons

One prevalent issue in hypothesis testing is that of **multiple comparisons** whereby several hypothesis tests are conducted simultaneously

As the number of hypothesis tests conducted grows in number, so to does the probability of one of those tests being decided in error

# Multiple Comparisons

Consider conducting 2 hypothesis tests, each with a Type I error rate of 5%

For any given test, the probability of *not* making an error is

$$P(\text{No type I error}) = 0.95$$

1. What is the probability that neither test has a Type I error?
2. What is the probability that *at least* one test has a Type I error?

## Example

Suppose that I am interested in testing if there is a non-zero correlation between cost and average faculty salary in each of the 8 regions of our college dataset

Suppose further we are testing for significance at the level $\alpha = 0.05$

|   | Region | $p$-value |
|---|--------|-----------|
| 1 | Far West | 0.7667 |
| 2 | Great Lakes | 0.0085 |
| 3 | Mid East | 0.0001 |
| 4 | New England | 0.0061 |
| 5 | Plains | 0.9487 |
| 6 | Rocky Mountains | 0.7394 |
| 7 | South East | 0.0143 |
| 8 | South West | 0.0344 |

# Family-wise error rates (FWER)

For a collection of independent hypothesis tests, the **family-wise error rate (FWER)** describes the probability of making one or more Type I errors

For $m$ independent tests with a Type I error rate of $\alpha$, the FWER is defined as

$$\text{FWER} = 1 - (1 - \alpha)^m$$

# Example

Suppose that I am interested in testing if there is a non-zero correlation between cost and average faculty salary in each of the 8 regions of our college dataset

If my Type I error rate for each test is 5%, what is the probability that I make at least one Type I error?

$$P(\text{At least one Type I error}) = 1 - P(\text{Probability of no Type I errors})$$
$$= 1 - (1 - 0.05)^8$$
$$= 33.6\%$$

That is, instead of making a Type I error 1 in 20 times, we are now making it 1 in 3 times

# FWER Correction

Just as we control the Type I error rate of a single hypothesis test with $\alpha$, we also have an interest in controlling the FWER

For $m$ hypothesis tests controlled at level $\alpha$, the correction $\alpha^* = \alpha/m$ is known as the **Bonferonni Adjustment**

If instead for a series of $m$ tests we reject the null hypothesis when $p < \alpha^*$, we will control the FWER at level $\alpha$

Assuming the 8 regions of our hypothesis test are independent, our Bonferonni adjustment for $\alpha = 0.05$ should be

$$\alpha^* = 0.05/8 = 0.00625$$

| | Testing $p < \alpha$ | | | | Testing $p < \alpha^*$ | |
|---|---|---|---|---|---|---|
| | Region | $p$-value | | | Region | $p$-value |
| 1 | Far West | 0.7667 | | 1 | Far West | 0.7667 |
| 2 | Great Lakes | 0.0085 | | 2 | Great Lakes | 0.0085 |
| 3 | Mid East | 0.0001 | | 3 | Mid East | 0.0001 |
| 4 | New England | 0.0061 | | 4 | New England | 0.0061 |
| 5 | Plains | 0.9487 | | 5 | Plains | 0.9487 |
| 6 | Rocky Mountains | 0.7394 | | 6 | Rocky Mountains | 0.7394 |
| 7 | South East | 0.0143 | | 7 | South East | 0.0143 |
| 8 | South West | 0.0344 | | 8 | South West | 0.0344 |

# Review

Tension between Type I and Type II errors

1. Requiring more evidence makes Type I error less likely

2. Requiring more evidence also reduces power

We are often interested in the behavior of diagnostic tests done on a sample of the population

1. Important to be able to construct tables to categorize outcomes

2. Priorities will depend on problem at hand

Finally, there is the issue of *multiple comparisons*

1. Family-wise error rate

2. Bonferonni correction