Test Performance and Multiple Testing

Grinnell College

November 11, 2026

Warm-Up

1. What is the definition of Type I and Type II error?
2. Can one decision about one null hypothesis be both types of error?
3. What impact does sample size have on power?
4. At fixed α, what impact does sample size have on the Type I error rate and interval width?
5. What impact does effect size have on power?

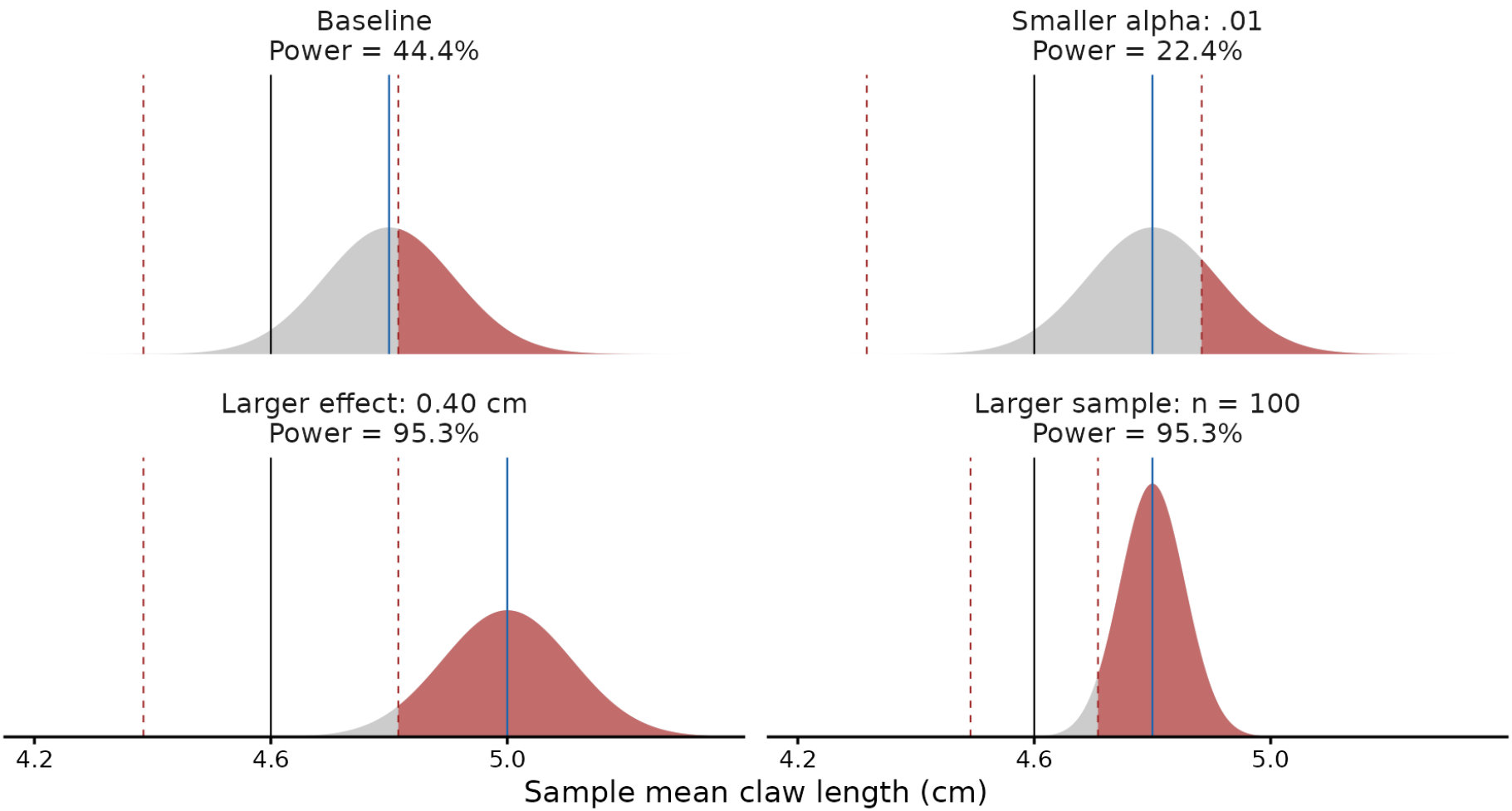
Warm-Up Solutions

1. Type I: reject a true null. Type II: fail to reject a false null.
2. No. The null is either true or false.
3. Larger n gives smaller SE and more power for a fixed nonzero effect.
4. The Type I error rate stays at α under our assumptions; larger n gives narrower intervals.
5. Larger effects are easier to detect, so power increases.

Today

1. A brief power recap
2. Test performance: true positives, false positives, and base rates
3. Multiple testing and family-wise error rates

Power: Four Comparisons



Baseline: $\mu_0 = 4.60$, $\mu = 4.80$, $n = 25$, $\sigma = .55$, $\alpha = .05$.

Red: power. Black:. null mean. Blue: true mean. Dashed: rejection boundaries.

Drawing Conclusions

For an individual statistical test, we can cross tabulate our decision possibilities with the following table

	H_0 true	H_0 false
Fail to reject	Correct: $1 - \alpha$	Type II error: β
Reject	Type I error: α	Correct: $1 - \beta$

For a diagnostic test, let H_0 mean the condition is absent. A positive result corresponds to rejecting H_0.

Confusion Matrix

In the context of classification, a **confusion matrix** compares the predicted category with the true category.

A positive result predicts that the condition is present; a negative result predicts that it is absent.

	Condition absent	Condition present
Predict negative	True negative (TN)	False negative (FN)
Predict positive	False positive (FP)	True positive (TP)

If H_0 means the condition is absent, a false positive is a Type I error and a false negative is a Type II error.

Two Rates, Two Denominators

In assessing the quality of a diagnostic test, we are typically interested in **sensitivity** and **specificity**

True positive rate (sensitivity):

$$P(\text{positive} \mid \text{condition present}) = 1 - \beta.$$

True negative rate (specificity):

$$P(\text{negative} \mid \text{condition absent}) = 1 - \alpha.$$

However, for an individual person taking a test, the more relevant question is: "Given a positive result, what is the probability is the condition present?"

Disease Testing

Suppose a disease affects 40% of a population. A diagnostic test has power (sensitivity) of 99% and a specificity of 95%.

- What is the null hypothesis?
- What are Type I and Type II errors? Which is more consequential here?
- Given a positive result, what is the probability the person has the disease?
- What if the disease affected only 1% of the population?

Disease Testing: Constructing the Table

Suppose a disease affects 40% of a population. A diagnostic test has sensitivity of 99% and a specificity of 95%.
Choose $N = 1000$ and construct the table:

Test Result	Disease Status D^-	D^+	Total
T^-	570	4	574
T^+	30	396	426
Total	600	400	1000

- What is the null hypothesis?
- What are Type I and Type II errors? Which is more consequential here?
- Given a positive result, what is the probability the person has the disease?
- What if the disease affected only 1% of the population?

Same Test, Different Base Rates

Consider 10,000 people. The sensitivity stays at 99% and specificity at 95%.

	40% have disease	1% have disease
Have disease	4,000	100
Do not have disease	6,000	9,900
True positives	3,960	99
False positives	300	495
All positive results	4,260	594

$$P(\text{disease} \mid +): \quad \frac{3960}{4260} = 93.0\% \qquad \frac{99}{594} = 16.7\%.$$

A Positive Result Is Not the Same as Power

The probability of disease *given a positive result* is called the **positive predictive value**.

It depends on the test's performance **and the base rate**: how common the condition is in the population being tested.

When the condition is rare, even a small false positive rate can produce many false positives because so many people do not have the condition.

High sensitivity and specificity can coexist with a low positive predictive value.

EXAM-ple

We evaluate software that sends suspected spam to a junk folder and other email to the inbox.

- What is the null hypothesis for an incoming email?
- What are Type I and Type II errors? Which is more consequential here?
- The filter has a Type I error rate of 2% and a Type II error rate of 20%. Of 1,000 monthly emails, 5% are spam. Given that an email is flagged as spam, what is the probability it is legitimate?

Which Emails Reach the Junk Folder?

	Legitimate	Spam	Total
Sent to inbox	931	10	941
Flagged as spam	19	40	59
Total	950	50	1,000

We can use the table to construct the probability we're interested in

$$P(\text{legitimate} \mid \text{flagged}) = \frac{19}{59} = 32.2\%.$$

About one third of flagged emails are legitimate, even though the filter incorrectly flags only 2% (19/950) of legitimate emails

Transition slide

Multiple Comparisons

Suppose we test several hypotheses as part of one study.

Each test can have a small Type I error rate, while the chance of **at least one false positive across the study** is much larger.

A **family** is the collection of tests whose errors we want to control together. Define it from the study questions, before inspecting the p-values.

Two Tests

Suppose two null hypotheses are both true, the tests are independent, and each uses $\alpha = .05$.

1. What is the probability of no Type I error on either test?
2. What is the probability of at least one Type I error?

$$P(\text{no errors}) = .95^2 = .9025$$

$$P(\text{at least one error}) = 1 - .9025 = .0975.$$

Multiple Testing

xkcd: Significant

Family-Wise Error Rate (FWER)

The **family-wise error rate** is the probability of making at least one Type I error in a family of tests.

If all m null hypotheses are true and the tests are independent, each with Type I error rate α, then

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

Only tests of true null hypotheses can contribute Type I errors.

Eight Regions, Eight Tests

We test whether cost and average faculty salary are correlated within each of the eight regions of our college dataset.

Region	p-value
Far West	.7667
Great Lakes	.0085
Mid East	.0001
New England	.0061
Plains	.9487
Rocky Mountains	.7394
South East	.0143
South West	.0344

Which tests reject using a separate .05 cutoff for each region?

How Often Would We Find Something?

If all eight regional null hypotheses were true and the tests were independent:

$$P(\text{no Type I errors}) = .95^8 = .6634$$

$$\text{FWER} = 1 - .95^8 = .3366.$$

About one third of repeated studies would produce at least one false positive.

Bonferroni Adjustment

To control the family-wise error rate at .05 across m planned tests, use the p-value cutoff

$$\alpha^* = \frac{.05}{m}.$$

Then for independent tests, the FWER is approximately

$$FWER = (1 - \alpha/m)^m \approx \alpha$$

Reject each null only when its p-value is below this cutoff.

For eight tests: $\alpha^* = .05/8 = .00625$.

Which Conclusions Change?

Region	p-value	$p < .05$	$p < .00625$
Far West	.7667	No	No
Great Lakes	.0085	Yes	No
Mid East	.0001	Yes	Yes
New England	.0061	Yes	Yes
Plains	.9487	No	No
Rocky Mountains	.7394	No	No
South East	.0143	Yes	No
South West	.0344	Yes	No

If all eight nulls are true and the tests independent, the adjusted FWER is

$$1 - (1 - .00625)^8 \approx .0489.$$

The Cost of Controlling More Errors

Bonferroni requires stronger evidence for each rejection.

- It reduces the chance of false positives across the family.
- At fixed sample size and true effect, it also reduces power for each test.
- A result that no longer rejects has not established that the effect is zero.

Plan the comparisons that answer the study question, and collect enough data to detect effects worth finding.

Review

1. Power and the false positive rate condition on the true state.
2. Predictive values condition on the observed test result, and depend on base rates.
3. Multiple tests create more opportunities for false positives.
4. Bonferroni controls family-wise error, with a cost in power.