Decision Error

Grinnell College

November 10, 2025

Strength of Evidence

So far, our process has been as follows:

1. Being with a null hypothesis, $H_0 : \mu = \mu_0$
2. Collected data and comptute statistic, i.e, $\overline{x}$
3. Compare our statistic against the null distribution, i.e., $t = \frac{\overline{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
4. Derive a p-value based on the statistic and the distribution

We found that we could use our p-value to quantify the strength of evidence against our null: the smaller the p-value, the less likely our observed data if the null were true

Decision Making

Based on the evidence we have collected, we must ultimately decide between one of two decisions:

1. There is sufficient evidence to reject H_0
2. There is *not* sufficient evidence to reject H_0

Decision Making

Just as our confidence intervals were correct or incorrect, so to may be our decision regarding H_0. In this case, however, there are two distinct ways in which our decision can be incorrect:

1. H_0 is *TRUE* (i.e., there is no effect), yet we reject anyway
2. H_0 is *FALSE* (i.e., there is an effect), yet we fail to reject it

Decision Making

These two types of errors are known as Type I and Type II errors, respectively:

1. H_0 is *TRUE* (i.e., there is no effect), yet we reject anyway
    - Type I error
    - "False positive"
    - Evidence leads to wrong conclusion
2. H_0 is *FALSE* (i.e., there is an effect), yet we fail to reject it
    - Type II error
    - "False negative"
    - Not enough evidence to conclude

Decision Making

Test Result	True State of Nature	
	H_0 True	H_0 False
Fail to reject H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

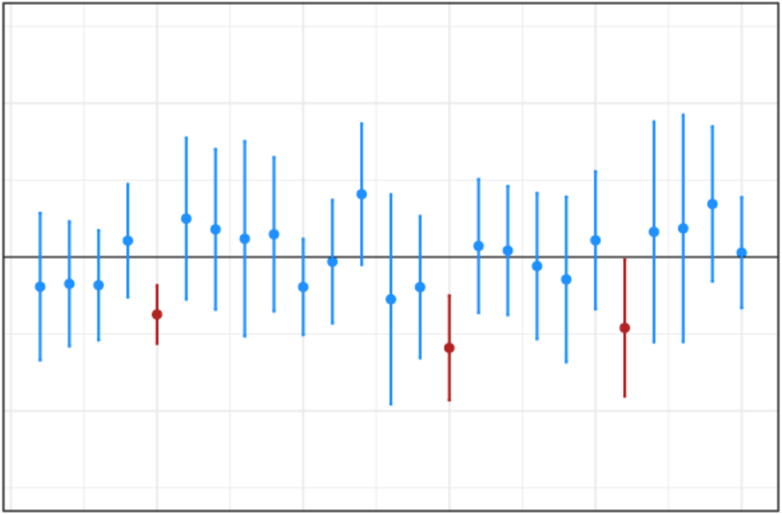
Type I Errors

A Type I error describes a situation in which we incorrectly identify a null effect:

- Conclude that an intervention works when it does not
- Conclude that there is a relationship between two variables when there are not

A Type I error will occur, for example, when our constructed confidence does not contain μ_0 when $\mu_0 = \mu$

Type I Errors



Type I Error Rate

We can control the rate at which we commit Type I errors with adjusting the *level of significance*, denoted α.

This is also called the *Type I error rate*

The Type I error rate has a *one-to-one* correspondence with our confidence intervals: a 95% confidence interval will permit a Type I error 5% of the time, corresponding to $\alpha = 0.05$

We *reject* our null hypothesis when p-value $< \alpha$

Type II Errors

A Type II error describes a situation in which the null hypothesis is false, yet based on the evidence gathered we fail to reject it:

- An intervention has a clinical effect, but it is not detected
- An email is considered spam, but the filter does not detect it

Typically, a Type II error is the result of one or more factors:

- Too few observations in our sample
- The population has large variability
- The effect size is small

Effect Size

One important concept is identifying statistical differences is that of **effect size**, a value that measures the strength of association between two variables:

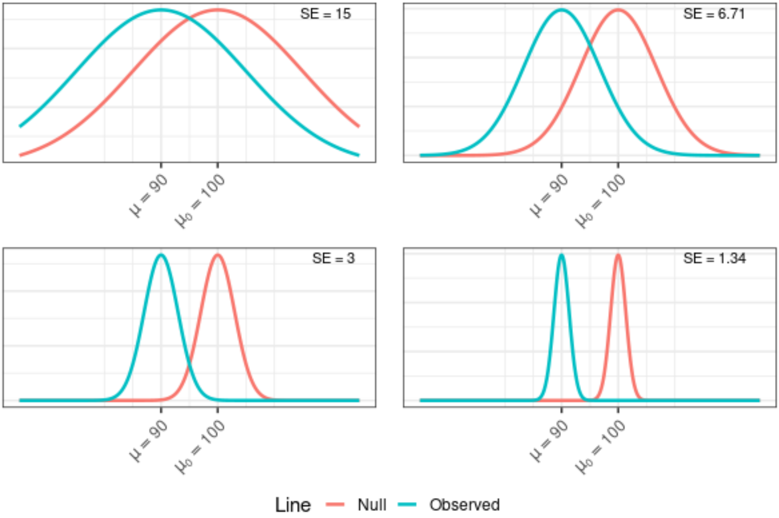
- Absolute difference between sample and hypothesis $(\overline{X} - \mu)$
- Standardized value of difference (t-statistic, z-score)
- Odds ratio, correlation, etc.,

Large effect sizes are much easier to detect, accommodating larger variances or smaller sample sizes. When the true effect size is small, more observations need to be collected to detect a difference

Practical vs Statistical Significance

Suppose I have a coin and hypothesize that the probability of landing on heads is $H_0 : p = 0.5$. A collect a sample and find that $\hat{p} = 0.51$. Is this considered significant?

$\hat{p}$	n	t	p-value
0.51	100	0.400	0.6899
0.51	500	0.895	0.3713
0.51	1000	1.265	0.2060
0.51	10000	4.002	0.0001


SE = 15
SE = 6.71
SE = 3
SE = 1.34
$\mu = 90$
$\mu_0 = 100$
Line
Null
Observed

Type II Error Rate

The Type II error rate is typically denoted β

More frequently, we consider the rate at which Type II errors do not occur $(1 - \beta)$, a term we refer to as **power**

A study that is unable to detect a true effect is said to be **underpowered**

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

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

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

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?

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

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

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 α, the FWER is defined as

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

FWER Correction

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

For m hypothesis tests controlled at level α, 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 α

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 $p < \alpha$

	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

Testing $p < \alpha^*$

	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

Review

Based on the evidence observed, we will ultimately make one of two decisions:

1. Reject H_0
2. Fail to reject H_0

Depending on the true state of H_0, we can be incorrect in two ways:

1. Type I Error (α): H_0 is true, yet we reject anyway
2. Type II Error (β): H_0 is false, yet we fail to reject it

Finally, there is the issue of *multiple comparisons*

1. Family-wise error rate
2. Bonferonni correction