Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

Lecture 23

Experiment

- Weigh a known 10 gram mass 16 times on the same scale.
- Scale makes normally distributed measurement errors:
  $X_1, \ldots, X_{16} \sim$ iid $N(\mu, \sigma^2 = 4)$

Measurement Errors?

Weigh same object repeatedly $\Rightarrow$ slightly different result each time.

Average deviation from mean ≈ 2 grams.

Two Kinds of Scales

Unbiased Correct on average: $\mu = 10$ grams

Biased *Too high* on average: $\mu = 11$ grams

An Idea for Deciding if a Scale is Biased

1. Test $H_0 : \mu = 10$ against $H_1 : \mu > 10$ with $\alpha = 0.025$.
2. Decide based on the outcome of test:
   - Reject $H_0 \Rightarrow$ decide scale is biased, throw it away.
   - Fail to reject $H_0 \Rightarrow$ decide scale is unbiased, keep it.

Testing Whether a Scale is Biased

$X_1, \ldots, X_{16} \sim$ iid $N(\mu, \sigma^2)$ where we *know* $\sigma^2 = 4$

Suppose I want to test $H_0 : \mu = 10$. What is my test statistic?

(a) $4\bar{X}/S$

(b) $4(\bar{X} - 10)/S$

(c) $(\bar{X} - \mu)/(S/\sqrt{n})$

(d) $2\bar{X}$

(e) $2(\bar{X} - 10)$

$$T_n = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{X} - 10}{2/\sqrt{16}} = 2(\bar{X} - 10)$$

Testing Whether a Scale is Biased

$X_1, \ldots, X_{16} \sim$ iid $N(\mu, \sigma^2)$ where we *know* $\sigma^2 = 4$

What is the sampling distribution of $2(\bar{X} - 10)$ under $H_0 \colon \mu = 10$?

(a) $N(\mu, 4)$

(b) $N(0, 4)$

(c) $t(15)$

(d) $\chi^2(15)$

(e) $N(0, 1)$

$$H_0 \colon \mu = 0 \implies T_n = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = 2(\bar{X} - 10) \sim N(0, 1)$$

Testing Whether a Scale is Biased

$X_1, \ldots, X_{16} \sim$ iid $N(\mu, \sigma^2)$ where we *know* $\sigma^2 = 4$

Suppose I want to test $H_0 : \mu = 10$ against the *one-sided* alternative $\mu > 10$ with $\alpha = 0.025$. What is my decision rule?

(a) Reject H_0 if $2(\bar{X} - 10) > 1$

(b) Reject H_0 if $2(\bar{X} - 10) < 1$

(c) Reject H_0 if $2(\bar{X} - 10) > 2$

(d) Reject H_0 if $2(\bar{X} - 10) < 2$

(e) Reject H_0 if $|2(\bar{X} - 10)| > 2$

Reject H_0 if $T_n = 2(\bar{X} - 10) >$ `qnorm(1 - 0.025)` ≈ 2

Testing an *Unbiased* Scale

Unbeknowst to me the scale I am testing is in fact *unbiased*. What is the probability that I will decide, based on the outcome of my test, to throw it away?

This is simply a Type I Error! Hence the probability is $\alpha = 0.025$

Testing a *Biased* Scale

Unbeknowst to me the scale I am testing is in fact *biased*. What is the probability that I will decide, based on the outcome of my test, to throw it away?

This is the *opposite* of a Type II error...

What is the probability of throwing away a *biased scale?*

Decision Rule

Decide scale is biased if $2(\bar{X} - 10) > 2$ or *equivalently* if $\bar{X} > 11$

Biased Scale

$\mu = 11 \quad \Longrightarrow \quad X_1, \ldots, X_{16} \sim \text{iid } N(11, \sigma^2 = 4)$

Which implies...

Testing a *Biased* Scale

Suppose $X_1, \ldots, X_{16} \sim N(11, \sigma^2 = 4)$. What is the sampling distribution of $\bar{X}$?

(a) $N(11, 1)$

(b) $N(0, 1)$

(c) $t(15)$

(d) $N(11, 1/4)$

(e) $N(10, 1/4)$

$$\bar{X}_n \sim N(\mu, \sigma^2/n) = N(11, 1/4)$$

What is the probability of throwing away a *biased scale?*

Decision Rule

Decide scale is biased if $2(\bar{X} - 10) > 2$ or *equivalently* if $\bar{X} > 11$

Biased Scale

$$\mu = 11 \quad \Longrightarrow \quad X_1, \ldots, X_{16} \sim \text{iid } N(11, \sigma^2 = 4)$$

Which implies

$$\bar{X} \sim N(11, 1/4) \quad \Longrightarrow \quad P(\bar{X} > 11) = 1/2$$

The *power* of this test is 50%

Recall:

Type I Error

Rejecting H_0 when it is true: $P(\text{Type I Error}) = \alpha$

Type II Error

Failing to reject H_0 when it is false: $P(\text{Type II Error}) = \beta$

Statistical Power

The probability of rejecting H_0 when it is false: $\text{Power} = 1 - \beta$

i.e. the probability of *convicting* a guilty person.

Hypothesis tests designed to control Type I error rate (α). But we also care about Type II errors. What can learn about these?

Recall: Normal Population Known Variance

Sampling Model

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ where σ^2 is known

Sampling Distribution

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Under $H_0 : \mu = 0$

$$T_n = \frac{\bar{X}_n}{\sigma/\sqrt{n}} \sim N(0,1)$$

What happens if $\mu \neq 0$?

Key Point #1

- Test Statistic $T_n = \sqrt{n}(\bar{X}_n/\sigma)$
- Unless $\mu = 0$, test statistic is *not* standard normal!
- When $\mu \neq 0$, distribution of T_n *depends on* μ!

Key Point #2

Regardless of the value of μ,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

since the population is normally distributed!

Distribution of T_n *Under the Alternative*

$$
\begin{aligned}
T_n &= \frac{\bar{X}_n}{\sigma/\sqrt{n}} \\
&= \frac{\bar{X}_n}{\sigma/\sqrt{n}} - \frac{\mu}{\sigma/\sqrt{n}} + \frac{\mu}{\sigma/\sqrt{n}} \\
&= \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) + \frac{\mu}{\sigma/\sqrt{n}} \\
&= Z + \sqrt{n}(\mu/\sigma) \sim N\left(\sqrt{n}(\mu/\sigma), 1\right)
\end{aligned}
$$

Where $Z \sim N(0,1)$

Power of One-Sided Test

Under the Alternative

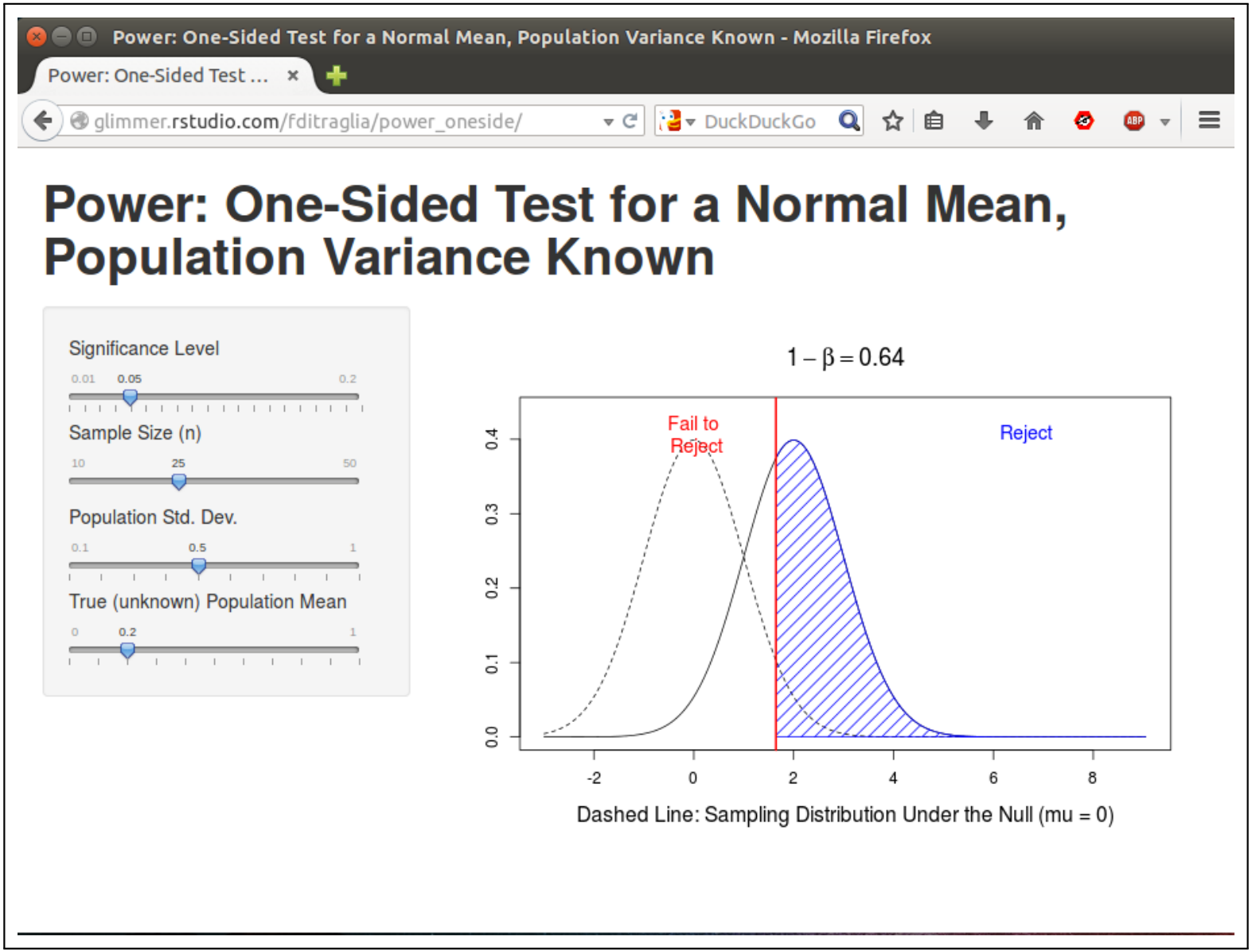
$T_n = \sqrt{n}(\bar{X}_n/\sigma) \sim N\left(\sqrt{n}(\mu/\sigma), 1\right)$

Decision Rule

Reject $H_0 : \mu = 0$ if $T_n > \texttt{qnorm}(1-\alpha)$

$$
\begin{aligned}
1-\beta &= P(\text{Reject } H_0 | H_0 \text{ false}) = P(T_n > \texttt{qnorm}(1-\alpha)) \\
&= P\left(Z + \sqrt{n}(\mu/\sigma) > \texttt{qnorm}(1-\alpha)\right) \\
&= P\left(Z > \texttt{qnorm}(1-\alpha) - \sqrt{n}(\mu/\sigma)\right) \\
&= 1 - P\left(Z \leq \texttt{qnorm}(1-\alpha) - \sqrt{n}(\mu/\sigma)\right) \\
&= 1 - \texttt{pnorm}\left(\texttt{qnorm}(1-\alpha) - \sqrt{n}(\mu/\sigma)\right)
\end{aligned}
$$

https://fditraglia.shinyapps.io/power_oneside



Power of Two-Sided Test

Under the Alternative

$T_n = \sqrt{n}(\bar{X}_n/\sigma) \sim N\left(\sqrt{n}(\mu/\sigma), 1\right)$

Decision Rule

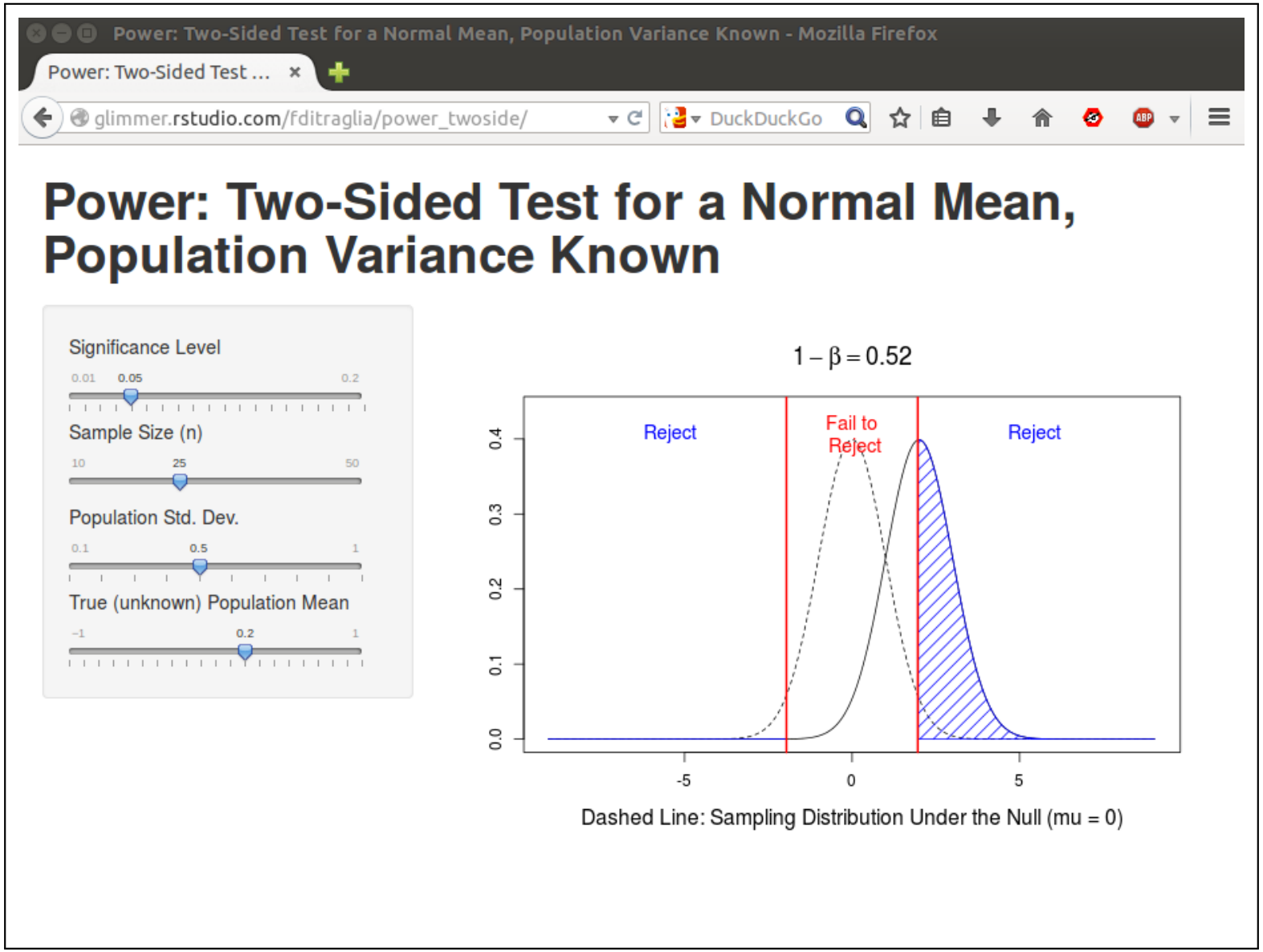
Reject $H_0 : \mu = 0$ if $|T_n| > \texttt{qnorm}(1-\alpha)$

$$
\begin{aligned}
1-\beta &= P(\text{Reject } H_0 | H_0 \text{ false}) = P(|T_n| > \texttt{qnorm}(1-\alpha/2)) \\
&= \underbrace{P(T_n < -\texttt{qnorm}(1-\alpha/2)}_{\text{Lower}} + \underbrace{P(T_n > \texttt{qnorm}(1-\alpha/2)}_{\text{Upper}}
\end{aligned}
$$

$$
\begin{aligned}
\text{Upper} &= (\text{Power of One-Sided Test with } \alpha/2 \text{ instead of } \alpha) \\
&= 1 - \texttt{pnorm}\left(\texttt{qnorm}(1-\alpha/2) - \sqrt{n}(\mu/\sigma)\right)
\end{aligned}
$$

$$
\text{Lower} = \texttt{pnorm}\left(-\texttt{qnorm}(1-\alpha/2) - \sqrt{n}(\mu/\sigma)\right)
$$

https://fditraglia.shinyapps.io/power_twoside



What Determines Power?

Power $= 1 - P(\text{Type II Error})$

Chance of detecting an effect given that one exists.

Depends On:

1. Magnitude of Effect: *true* value of μ
   - Easier to detect large deviations from $H_0 : \mu = 0$
2. Amount of variability in the population: σ
   - Lower $\sigma \Rightarrow$ easier to detect effect of given magnitude
3. Sample Size: n
   - Larger sample size $\Rightarrow$ easier to detect effect of given magnitude
4. Significance Level: α
   - Fewer Type I errors $\Rightarrow$ more Type II errors

Study Tip

Compare determinants of *width* of $(1-\alpha)\times 100\%$ CI to determinants of *power* of corresponding two-sided test.

Some Final Thoughts on Hypothesis Testing and Confidence Intervals

Terminology I Have Intentionally Avoided Until Now

Statistical Significance

Suppose we carry out a hypothesis test at the $\alpha\%$ level and, based on our data, reject the null. You will often see this situation described as "statistical significance."

In Other Words...

When people say "statistically significant" what they really mean is that they rejected the null hypothesis.

Some Examples

- We found a difference between the "Hi" and "Lo" groups in the anchoring experiment that was statistically significant at the 5% level based on data from a past semester.
- Our 95% CI for the proportion of US voters who know who John Roberts is did not include 0.5. Viewed as a two-sided test, we found that the difference between the true population proportion and 0.5 was statistically significant at the 5% level.

Why Did I Avoid this Terminology?

Statistical Significance $\neq$ Practical Importance

- You need to understand the term "statistically significant" since it is widely used. A better term for the idea, however, would be "statistically discernible"
- Unfortunately, many people are confuse "significance" in the narrow, technical sense with the everyday English word meaning "important"
- Statistically Significant Does Not Mean Important!"
  - A difference can be practically unimportant but statistically significant.
  - A difference can be practically important but statistically insignificant.

P-value Measures Strength of Evidence Against H_0

Not The Size of an Effect!

Statistically Significant but Not Practically Important

I flipped a coin 10 million times (in R) and got 4990615 heads.

Test of $H_0: p = 0.5$ against $H_1: p \neq 0.5$

$$T = \frac{\widehat{p} - 0.5}{\sqrt{0.5(1-0.5)/n}} \approx -5.9 \implies \text{p-value} \approx 0.000000003$$

Approximate 95% Confidence Interval

$$\widehat{p} \pm \texttt{qnorm}(1 - 0.05/2)\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \implies (0.4988, 0.4994)$$

(Such a huge sample size that refined vs. textbook CI makes no difference)

Actual p was 0.499

Practically Important But Not Statistically Significant

Vickers: "What is a P-value Anyway?" (p. 62)

Just before I started writing this book, a study was published reporting about a 10% lower rate of breast cancer in women who were advised to eat less fat. If this indeed the true difference, low fat diets could reduce the incidence of breast cancer by tens of thousands of women each year – astonishing health benefit for something as simple and inexpensive as cutting down on fatty foods. The p-value for the difference in cancer rates was 0.07 and here is the key point: this was widely misinterpreted as indicating that low fat diets don't work. For example, the New York Times editorial page trumpeted that "low fat diets flub a test" and claimed that the study provided "strong evidence that the war against all fats was mostly in vain." However failure to prove that a treatment is effective is not the same as proving it ineffective.

Do Students with 4-Letter Surnames Do Better?

Based on Data from Midterm 1 Last Semester

4-Letter Surname

$\bar{x} = 88.9$

$s_x = 10.4$

$n_x = 12$

Other Surnames

$\bar{y} = 74.4$

$s_y = 20.7$

$n_y = 92$

Difference of Means

$\bar{x} - \bar{y} = 14.5$

Standard Error

$SE = \sqrt{s_x^2/n_x + s_y^2/n_y} \approx 3.7$

Test Statistic

$T = 14.5/3.7 \approx 3.9$

What is the p-value for the two-sided test?

Test Statistic ≈ 3.9

(a) $p < 0.01$

(b) $0.01 \leq p < 0.05$

(c) $0.05 \leq p < 0.1$

(d) $p > 0.1$

(e) Not Sure

What do these results mean?

Evaulate this statement in light of our hypothesis test:

> *Students with four-letter long surnames do better, on average, on the first midterm of Econ 103 at UPenn.*

(a) Strong evidence in favor

(b) Moderate evidence in favor

(c) No evidence either way

(d) Moderate evidence against

(e) Strong evidence against

I just did 134 Hypothesis Tests...

... and 11 of them were significant at the 5% level.

	group	sign	p.value	x.bar	N.x	s.x	y.bar	N.y	s.y
26	first1 = P	1	0.000	93.8	3	2.9	75.5	101	20.4
70	id2 = 7	1	0.000	94.6	5	3.3	75.1	99	20.4
134	id8 = 0	1	0.000	92.6	7	4.9	74.8	97	20.5
5	Nlast = 4	1	0.001	88.9	12	10.4	74.4	92	20.7
90	id4 = 8	1	0.003	87.7	9	9.0	74.9	95	20.7
105	id6 = 8	1	0.003	88.1	5	5.8	75.4	99	20.6
109	id6 = 2	1	0.007	88.9	8	10.7	75.0	96	20.6
9	Nlast = 2	1	0.016	90.4	5	9.3	75.3	99	20.5
49	last1 = P	-1	0.036	65.2	6	9.9	76.7	98	20.6
65	id2 = 1	1	0.038	84.3	9	10.1	75.3	95	20.9
117	id7 = 8	1	0.041	83.4	13	11.6	75.0	91	21.1

Data-Dredging

- Suppose you have a long list of null hypotheses and assume, for the sake of argument that all of them are true.
  - E.g. there's no difference in grades between students with different 4th digits of their student id number.
- We'll still reject about 5% of the null hypotheses.
- Academic journals tend only to publish results in which a null hypothesis is rejected at the 5% level or lower.
- We end up with the bizarre result that "most published studies are false."

I posted a reading about this on Piazza: "The Economist - Trouble in the Lab." To learn even more, see Ioannidis (2005)

Green Jelly Beans Cause Acne!

xkcd #882

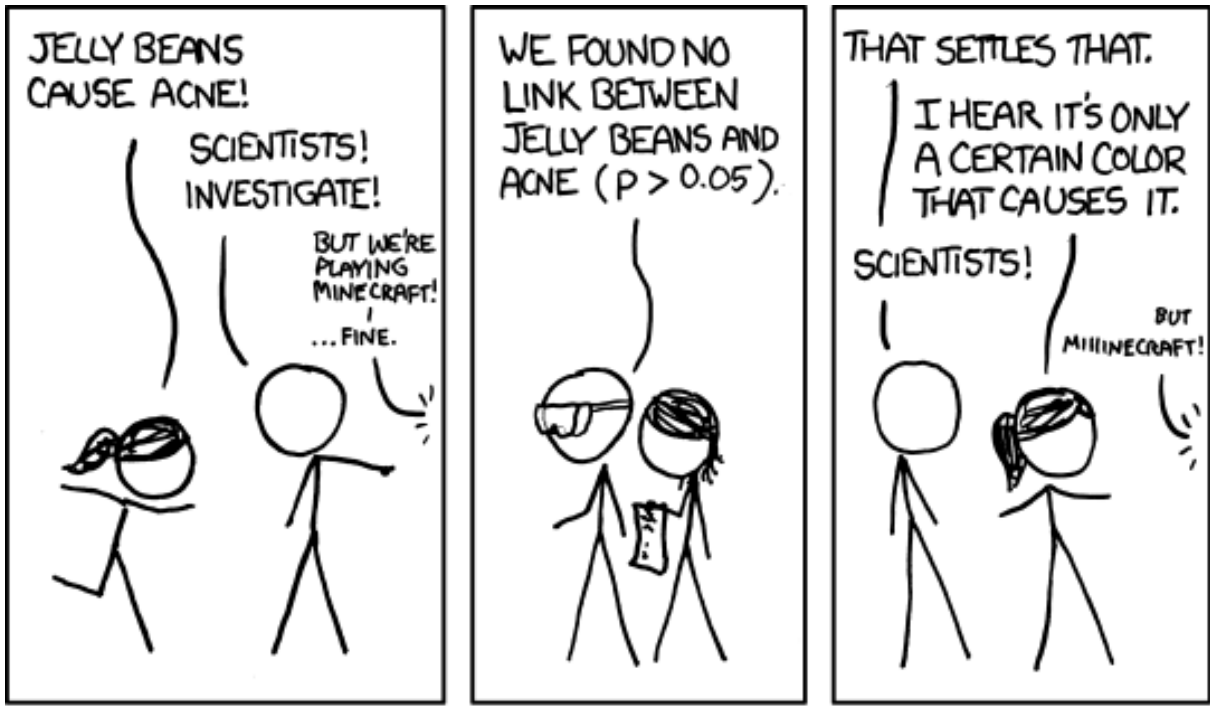


Figure: Go and read this comic strip: before today's lecture you wouldn't have gotten the joke!

Some Final Thoughts

- Failing to reject H_0 does not mean H_0 is true.
- Rejecting H_0 does not mean H_1 is true.
- P-values are always more informative than simply reporting "Reject" vs. "Fail To Reject" at a given significance level.
- Confidence intervals are more informative that hypothesis tests, since they give an idea of the size of an effect.
- If H_0 is actually plausible a priori (this is rarer than you may think), reporting a p-value can be a good complement to a CI.
- To avoid data-dredging be honest about the tests you have carried out: report *all of them*, not just the ones where you rejected the null.