Econ 103 – Statistics for Economists

Chapter 8: Hypothesis Testing

Mallick Hossain

University of Pennsylvania

The Lady Tasting Tea

An excerpt from *The Lady Tasting Tea* by David Salsburg

It was a summer afternoon in Cambridge, England, in the late 1920s. A group of university dons, their wives, and some guests were sitting around an outdoor table for afternoon tea. One of the women was insisting that tea tasted different depending upon whether the tea was poured into the milk or whether the milk was poured into the tea. The scientific minds among the men scoffed at this as sheer nonsense. What could be the difference? They could not conceive of any difference in the chemistry of the mixtures that could exist. A thin, short man, with thick glasses and a Vandyke beard beginning to turn gray, pounced on the problem. "Let us test the proposition" he said excitedly. He began to outline an experiment in which the lady who insisted there was a diference would be presented with a sequence of cups of tea, in some of which the milk had been poured into the tea and in others of which the tea had been poured into the milk.

Figure 1: The Orchard, Grantchester

Figure 2: What to have with your tea.

Figure 3: Why walk when you can punt?

Figure 4: What to wear.

Continued...

And so it was that summer afternoon in Cambridge. The man with the Vandyke beard was Ronald Aylmer Fisher, who was in his late thirties at the time. He would later be knighted Sir Ronald Fisher. In 1935, he wrote a book entitled The Design of Experiments, and he described the experiment of the lady tasting tea in the second chapter of that book. In his book, Fisher discusses the lady and her belief as a hypothetical problem. He considers the various ways in which an experiment might be designed to determine if she could tell the difference.

The Pepsi Challenge

The Pepsi Challenge

Our expert claims to be able to tell the difference between Coke and Pepsi. Let's put this to the test!

- Eight cups of soda
    - Four contain Coke
    - Four contain Pepsi
- The cups are randomly arranged
- How can we use this experiment to tell if our expert can *really* tell the difference?

# of Cokes Correctly Identified:

What do you think? Can our expert really tell the difference?

(a) Yes
(b) No

If you just guess randomly, what is the probability of identifying *all four cups of Coke correctly*?

- $\binom{8}{4} = 70$ ways to choose four of the eight cups.
- If guessing randomly, each of these is *equally likely*
- Only *one* of the 70 possibilities corresponds to correctly identifying all four cups of Coke.
- Thus, the probability is $1/70 \approx 0.014$

If you just guess randomly, what is the probability of identifying *all but one cup of Coke correctly*?

- $\binom{8}{4} = 70$ ways to choose four of the eight cups.
- If guessing randomly, each of these is *equally likely*
- There are 16 ways to mis-identify one Coke:
    - 4 choices of *which* Coke you call a Pepsi
    - 4 choices of *which* Pepsi you call a Coke
    - Total of $4 \times 4 = 16$ possibilities
- Thus, the probability is $16/70 \approx 0.23$

Probabilities if Guessing Randomly

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If you're just guessing, what is the probability of identifying *at least* three Cokes correctly?

- Probabilities of mutually exclusive events sum.
- P(all four correct) = 1/70
- P(exactly 3 correct)= 16/70
- $P(\text{at least three correct}) = 17/70 \approx 0.24$

- Even if you're just guessing randomly, the probability of correctly identifying three or more Cokes is around 24%
- In contrast, the probability of identifying *all four* Cokes correctly is only around 1.4% if you're guessing randomly.
- We should probably require the expert to get them all right.
- What if the expert gets them all wrong? This also has probability 1.4% if you're guessing randomly...

That was a Hypothesis Test!

We'll go through the details in a moment, but first an analogy...

Hypothesis Testing is Similar to a Criminal Trial

Criminal Trial

- The person on trial is either innocent or guilty (but not both!)
- "Innocent Until Proven Guilty"
- Only convict if evidence is "beyond a shadow of a doubt"
- *Not Guilty* rather than Innocent
    - Acquit $\neq$ Innocent
- Two Kinds of Errors:
    - Convict the innocent
    - Acquit the guilty
- Convicting the innocent is a worse error. Want this to be rare even if it means acquitting the guilty.

Hypothesis Testing

- Either the null hypothesis H_0 or the alternative H_1 hypothesis is true.
- Assume H_0 to start
- Only reject H_0 in favor of H_1 if there is strong evidence.
- *Fail to reject* rather than Accept H_0
    - (Fail to reject H_0) $\neq$ (H_0 True)
- Two Kinds of Errors:
    - Reject true H_0 (Type I)
    - Don't reject false H_0 (Type II)
- Type I errors (reject true H_0) are worse: make them rare even if that means more Type II errors.

Hypothesis Testing

How is the Pepsi Challenge a Hypothesis Test?

Null Hypothesis H_0
Can't tell the difference between Coke and Pepsi: just guessing.

Alternative Hypothesis H_1
Able to distinguish Coke from Pepsi.

Type I Error – Reject H_0 even though it's true
Decide expert can tell the difference when she's really just guessing.

Type II Error – Fail to reject H_0 even though it's false
Decide expert just guessing when she really can tell the difference.

How do we find evidence to reject H_0?

- Choose a significance level α maximum probability of Type I error that we are willing to tolerate.
  - Measures how often we will reject a true null, i.e. convict an innocent person
- Test Statistic T_n uses sample to measure plausibility of H_0
- Null Hypothesis $H_0 \Rightarrow$ Sampling Distribution for T_n
  - "Under the null" means "assuming the H_0 is true"
- Using α and the sampling distribution of T_n under the null, we construct a decision rule in terms of a critical value c_α
  - Reject H_0 if $T_n > c_\alpha$

We still have a random sampling model in mind!

Why does T_n have a sampling distribution?

- Random Sampling: new data $\Rightarrow$ different *realization* t of T_n
- Key point: T_n is a *random variable* with a particular distribution under the null hypothesis H_0

What do we mean by α?

- T_n is a RV $\Rightarrow$ outcome of hypothesis test is random!
- Sometimes we make mistake: either reject H_0 when it is true or fail to reject it when it is false.
- Repeated Sampling $\Rightarrow$ many different realizations of T_n $\Rightarrow$ many different outcomes of the test.
- Test is constructed so that, if H_0 is true, we will reject it no more than $100 \times \alpha\%$ of the time under repeated sampling.

Example: Pepsi Challenge

Test Statistic T_n

$T_n =$ Number of Cokes correctly identified

$H_0:$ **No skill, just guessing randomly**

Under this null hypothesis, the sampling distribution of T_n is:

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

Example: Pepsi Challenge

T_n: # of Cokes correctly identified. Sampling Dist. under H_0:

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If I choose a significance level of $\alpha = 0.05$, what critical value should I use?

(Remember that α is the probability of rejecting H_0 when it is actually true.)

Want $P(\text{Reject } H_0 | H_0 \text{ True}) \leq 0.05$

$P(T_n \geq 3|\text{Just Guessing}) = 17/70 \approx 0.23 > 0.05$

$P(T_n \geq 4|\text{Just Guessing}) = 1/70 \approx 0.014 \leq 0.05$

Example: Pepsi Challenge

T_n: # of Cokes correctly identified. Sampling Dist. under H_0:

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If I choose a significance level of $\alpha = 0.25$, what critical value should I use?

Want $P(\text{Reject } H_0 | H_0 \text{ True}) \leq 0.25$
$P(T_n \geq 2 | \text{Just Guessing}) = 53/70 \approx 0.76 > 0.25$
$P(T_n \geq 3 | \text{Just Guessing}) = 17/70 \approx 0.23 \leq 0.25$

Example: Pepsi Challenge

H_0: Expert is just guessing randomly.

H_1: Expert can distinguish Coke from Pepsi.

T_n: # of Cokes correctly identified. Has following sampling under the null:

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If I choose $\alpha = 0.05$, what decision rule should I use?

(a) Reject H_0 if $T_n \geq 0$

(b) Reject H_0 if $T_n \geq 1$

(c) Reject H_0 if $T_n \geq 2$

(d) Reject H_0 if $T_n \geq 3$

(e) Reject H_0 if $T_n \geq 4$

Example: Pepsi Challenge

H_0 : Expert is just guessing randomly.

H_1 : Expert can distinguish Coke from Pepsi.

T_n : # of Cokes correctly identified. Has following sampling under the null:

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If I choose $\alpha = 0.05$, what decision rule should I use?

Need $P(\text{Reject } H_0 | H_0 \text{ True}) \leq \alpha = 0.05$

$$
\begin{aligned}
P(T_n \geq 3|\text{Just Guessing}) &= 17/70 \approx 0.23 > 0.05 \\
P(T_n \geq 4|\text{Just Guessing}) &= 1/70 \approx 0.014 \leq 0.05
\end{aligned}
$$

Critical value for $\alpha = 0.05$ is 4

Example: Pepsi Challenge

H_0 : Expert is just guessing randomly.

H_1 : Expert can distinguish Coke from Pepsi.

T_n : # of Cokes correctly identified. Has following sampling under the null:

# Correct	0	1	2	3	4
Prob.	1/70	16/70	36/70	16/70	1/70

If I choose $\alpha = 0.25$, what critical value should I use?

(a) 0
(b) 1
(c) 2
(d) 3
(e) 4

Hypothesis: Assertion about Population(s)

- A Big Mac contains, on average, 550 kcal: $\mu = 550$
- Midterm 2 was harder than Midterm 1: $\mu_1 > \mu_2$
- Equal proportions of Republicans and Democrats know that John Roberts is the chief justice of SCOTUS: $p = q$
- Google stock is riskier than IBM stock: $\sigma_X^2 > \sigma_Y^2$
- There is no correlation between height and income: $\rho = 0$

Hypothesis Testing: Try to Find Evidence *Against* H_0

Null Hypothesis: H_0

- Start off assuming H_0 is true – "innocent until proven guilty"
- "Under the Null" = Assuming the null is true
- $H_0 \Rightarrow$ know something about population, can calculate probs.

This Course: *Simple* Null Hypotheses

$H_0 : f(\text{Parameters}) = \text{Known Constant}$, for example

- $\mu_1 - \mu_2 = 0$
- $p = 0.5$
- $\mu = 0$
- $\sigma_X^2/\sigma_Y^2 = 1$

How do I know what my null hypothesis is?

There is no rule I can give you for this: it depends on the problem. Here are some guidelines:

- It will take the form $f(\text{Parameters}) = \text{Known Constant}$
- Nulls are typically things like "there is no effect," "these two groups are not different," i.e. the *status quo*.
- Nulls are *very specific*: we need to be able to do probability calculations under the null – c.f. the Pepsi Challenge.

Big Mac Example

Example: How many calories in a Big Mac?

- According to McDonald's: 550 kcal on average
- Measure calories in random sample of 9 Big Macs: $X_1, \ldots, X_9 \sim \text{iid } N(\mu, \sigma^2)$

If we wanted to test McDonald's claim, what would be H_0?

(a) $\sigma^2 = 1$

(b) $\mu = 0$

(c) $\mu > 550$

(d) $\mu = 550$

(e) $\mu \neq 550$

Example: How many calories in a Big Mac?

- According to McDonald's: 550 kcal on average
- Measure calories in random sample of 9 Big Macs: $X_1, \ldots, X_9 \sim \text{iid } N(\mu, \sigma^2)$

If McDonald's is telling the truth, approximately what value should we get for the sample mean caloric content of the 9 Big Macs?

Example: How many calories in a Big Mac?

- According to McDonald's: 550 kcal on average
- Measure calories in random sample of 9 Big Macs: $X_1, \ldots, X_9 \sim \text{iid } N(\mu, \sigma^2)$

If the sample mean does not equal 550, does this prove that McDonald's is lying?

(a) Yes

(b) No

(c) Not Sure

How to find evidence against H_0? Test Statistic!

Test Statistic: T_n

A statistic that gives us information about the parameter we are testing and has a *known* sampling distribution *under* H_0.

Example: How many calories in a Big Mac?

- Measure calories in random sample of n Big Macs: $X_1, \ldots, X_9 \sim \text{iid } N(\mu, \sigma^2)$
- $H_0: \mu = 550$

If McDonald's is telling the truth, i.e. under the null, what is *exact* sampling distribution of $(\bar{X} - 550)/(S/3)$?

(a) χ^2_9

(b) $N(550, 1)$

(c) $F(9, 1)$

(d) $N(0, 1)$

(e) t_8

What if the null is false?

Alternative hypothesis: H_1

The *negation* of the null hypothesis.

Examples:

1. - H_0: This parameter equals 5.
   - H_1: This parameter does *not* equal 5.
2. - H_0: There is no difference between these two groups.
   - H_1: There *is* a difference between these two groups.

Sometimes we only care about *certain kinds* of violations of H_0...

One-sided vs. Two-sided Alternative

Let θ be a population parameter and θ_0 be a specified constant.

Null Hypothesis

- $H_0 : \theta = \theta_0$

Two-sided Alternative

- $H_1 : \theta \neq \theta_0$

One-sided Alternative

Two possibilities, depending on the problem at hand:

- $H_1 : \theta > \theta_0$
- $H_1 : \theta < \theta_0$

Example: Suing McDonald's

A class action lawsuit claims that McDonald's has been understating the caloric content of the "Big Mac," misleading consumers into thinking the sandwich is healthier than it really is. McDonald's claims the sandwich contains 550 kcal on average.

Suppose you're the judge in this case. What is your alternative hypothesis?

(a) $H_1 \colon \mu \neq 550$ kcal
(b) $H_1 \colon \mu < 550$ kcal
(c) $H_1 \colon \mu > 550$ kcal
(d) $H_1 \colon \mu = 550$ kcal

Example: Quality Control at McDonald's

You are a senior manager at McDonald's and are concerned that franchises may be deviating from company policy on the calorie count of a Big Mac sandwich, which is supposed to be 550 kcal on average. Because intervening is costly, you will only take action is there is strong evidence of deviation from company policy.

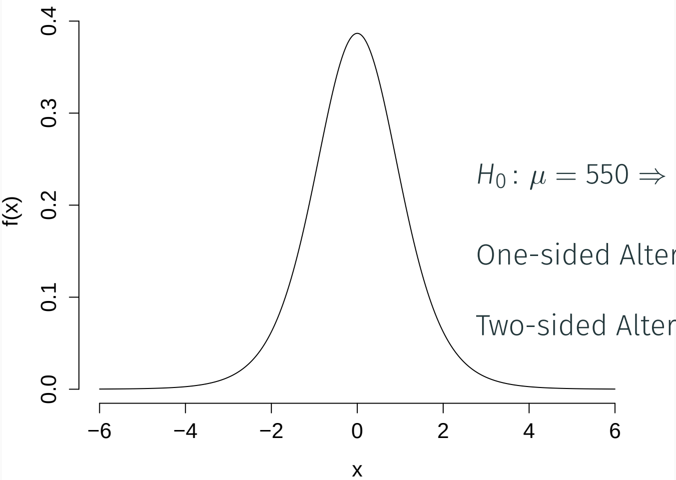
What is your alternative hypothesis?

(a) $H_1 \colon \mu \neq 550$ kcal
(b) $H_1 \colon \mu < 550$ kcal
(c) $H_1 \colon \mu > 550$ kcal
(d) $H_1 \colon \mu = 550$ kcal

Decision Rule: When should we reject H_0?

- Test statistic: RV with known sampling distribution under H_0
- McDonald's Example: $T_n = 3(\bar{X} - 550)/S$
- *Random* since $\bar{X}$ and S are RVs under random sampling: functions of $X_1, \ldots, X_9$.
- Observed dataset: *realizations* $x_1, \ldots, x_9$ of RVs $X_1, \ldots, X_9$
- Plug in observed data to get estimates (constants) $\bar{x}$ and s.
- Plug these into the formula for the test statistic to get a *number* – this is a *realization* of T_n
- Depending on this number, decide whether to reject H_0.

What Form Should the Decision Rule Take?



$$H_0\colon \mu = 550 \Rightarrow \frac{\bar{X} - 550}{S/3} \sim t(8)$$

One-sided Alternative $H_1\colon \mu > 550$

Two-sided Alternative $H_1\colon \mu \neq 550$

Example: Suing McDonald's

The plaintiffs allege that McDonald's has *understated* the true caloric content of a Big Mac: it's actually *greater* than 550 kcal. Suppose the plaintiffs are right. Then what sort of value should we expect the test statistic $3(\bar{X} - 550)/S$ to take on?

(a) A value *less* than zero.

(b) A value close to zero.

(c) A value *greater* than zero.

Example: Quality Control at McDonald's

The senior manager is worried that franchises are deviating from company policy that Big Macs should contain approximately 550 kcal. If the franchises *are* deviating, what sort of of value should we expect the test statistic $3(\bar{X} - 550)/S$ to take on?

(a) A value *less* than zero.

(b) A value close to zero.

(c) A value *greater* than zero.

(d) A value different from zero but we can't tell whether it will be positive or negative.

What Form Should the Decision Rule Take?

$X_1, \ldots, X_n \sim \text{iid } N(\mu, \sigma^2)$

Common Null Hypothesis $H_0 \colon \mu = 550$

Under H_0, $T_n = \sqrt{n}(\bar{X}_n - 550)/S \sim t(n-1)$

One-sided Alternative $H_1 \colon \mu > 550$

Reject H_0 if T_n is "too big"

Two-sided Alternative $H_1 \colon \mu \neq 550$

Reject H_0 if T_n is "too big" or "too small"

But how big of a discrepancy is "big enough" to reject?

Two Kinds of Mistakes in Hypothesis Testing

Type I Error

- Rejecting the null when it's actually true.
- $P(\text{Type I Error}) = \alpha$ $\boxed{\alpha = \text{“Significance Level” of Test}}$

Type II Error

- Failing to reject the null when it's false.
- $P(\text{Type II Error}) = \beta$ $\boxed{1 - \beta = \text{“Power” of Test}}$

Important!

Hypothesis testing *controls* probability of a Type I error since this is assumed to be the *worse* kind of mistake: convicting the innocent.

Construct a Decision Rule to *Fix* α at User-Chosen Level

Critical Value c_α

- Threshold for rejecting H_0
- Chosen so that $P(\text{Reject } H_0 | H_0 \text{ is True}) = \alpha$
- Depends on *both* α *and* the alternative hypothesis.

One-Sided Alternative

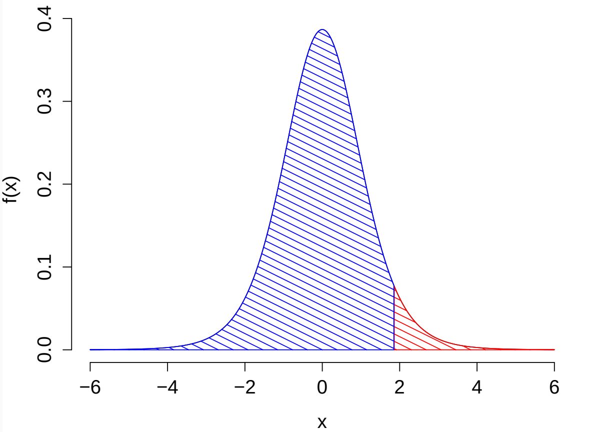
Reject H_0 if $T_n >$ Critical Value

Two-Sided Alternative

Reject H_0 if $|T_n| >$ Critical Value

Example: One-sided Alternative H_1: $\mu > 550$

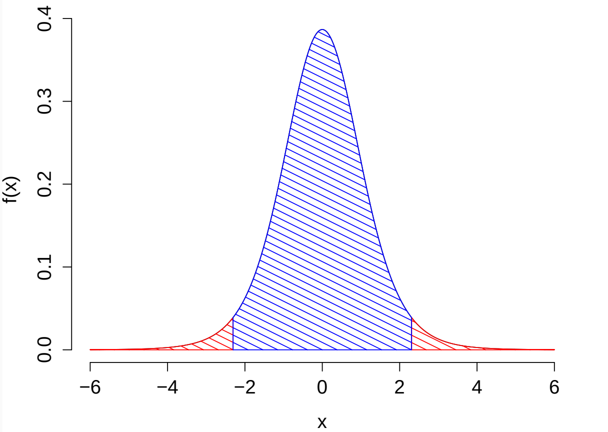
The critical value is chosen to reflect both the alternative hypothesis and the significance level.



One-sided Critical Value: `qt(`$1-\alpha$`, df = `$n-1$`)`

Example: Two-sided Alternative $H_1: \mu \neq 550$

The critical value is chosen to reflect both the alternative hypothesis and the significance level.

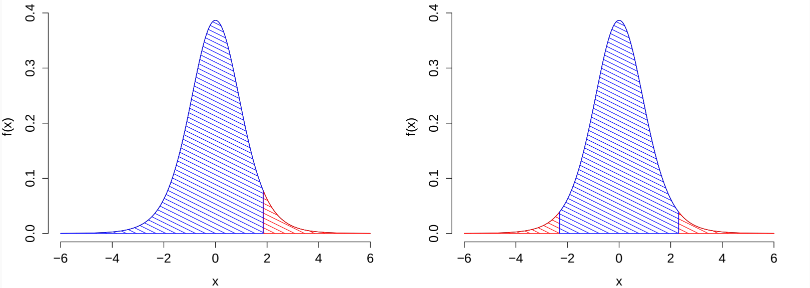


Two-sided Critical Value: `qt(`$1-\alpha/2$`, df = `$n-1$`)`

Suppose, for example, $\alpha = 0.05$, $n = 9$

$$\texttt{qt(0.95, df = 8)} \approx 1.86$$

$$\texttt{qt(0.975, df = 8)} \approx 2.3$$



One-sided Alternative: Reject H_0 if $3(\bar{X}_n - 550)/S \geq 1.86$

Two-sided Alternative: Reject H_0 if $\left|3(\bar{X}_n - 550)/S\right| \geq 2.3$

McDonald's Example

Suppose $n = 9$, $\bar{x} = 563$, $s = 34$. What is the value of our test statistic?

$$\frac{563 - 550}{34/\sqrt{9}} = \frac{13}{34/3} \approx 1.14$$

McDonald's Example: $\alpha = 0.05$

Recall that:

$$\texttt{qt(0.95, df = 8)} \approx 1.86$$
$$\texttt{qt(0.975, df = 8)} \approx 2.3$$

Based on an observed test statistic of 1.14, would we reject H_0 against the one-sided alternative at the 5% significance level?

(a) Yes
(b) No
(c) Not Sure

McDonald's Example: $\alpha = 0.05$

Recall that:

$$\texttt{qt(0.95, df = 8)} \approx 1.86$$
$$\texttt{qt(0.975, df = 8)} \approx 2.3$$

Based on an observed test statistic of 1.14, would we reject H_0 against the two-sided alternative at the 5% significance level?

(a) Yes

(b) No

(c) Not Sure

Reporting the Results of a Hypothesis Test

Lawsuit Example

The judge *failed to reject* the null hypothesis that $\mu = 550$ against the one-sided alternative $\mu > 550$ at the 5% significance level.

Quality Control Example

The senior manager *failed to reject* the null hypothesis that $\mu = 550$ against the two-sided alternative at the 5% significance level.

Interpretation

In each of these two cases, there was *insufficient evidence* the initial assumption that $\mu = 550$ *given the significance level used*.

But what if we have used a *different* significance level?

P-Values

The P-Value of a Hypothesis Test

Two Equivalent Definitions:

1. Given the value we calculated for our test statistic, what is the *smallest* α at which we would have rejected the null?
2. Under the null, what is the probability of observing a test statistic *at least as extreme* as the one we *actually* observed?

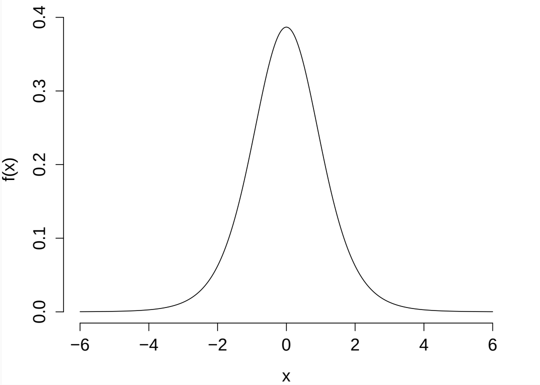
Why Report P-Values?

- More informative than reporting α and Reject/Fail to Reject
- E.g. a p-value of 0.03 means we would have rejected the null for any $\alpha \geq 0.03$ and failed to reject it for any $\alpha < 0.03$

P-Value Depends on Which Alternative We Have Specified!

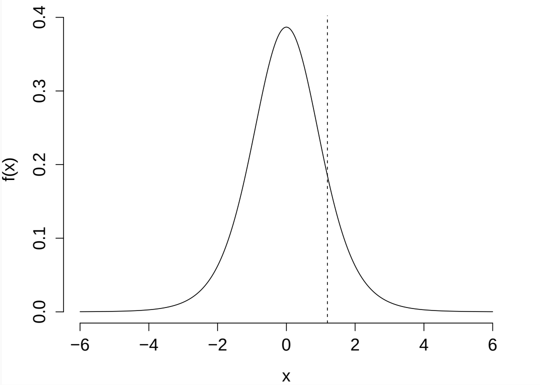
What is the p-value? (One-sided Test)

Recall: p-value is *smallest significance level* at which our observed test statistic would cause us to reject H_0. Test statistic is 1.14. What is the one-sided p-value?



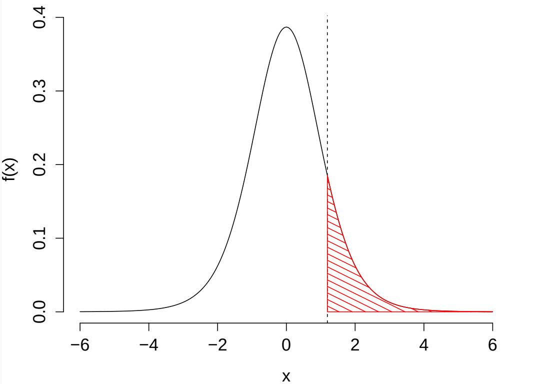
What is the p-value? (One-sided Test)

Recall: p-value is *smallest significance level* at which our observed test statistic would cause us to reject H_0. Test statistic is 1.14. What is the one-sided p-value?



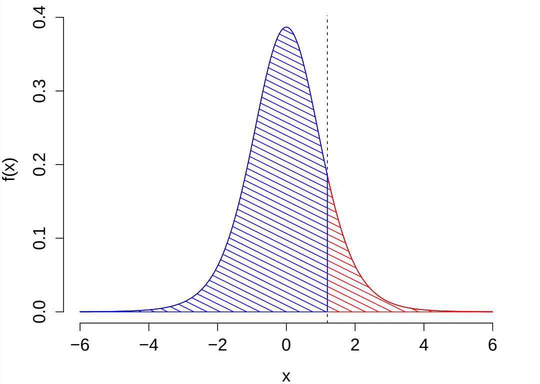
What is the p-value? (One-sided Test)

Recall: p-value is *smallest significance level* at which our observed test statistic would cause us to reject H_0. Test statistic is 1.14. What is the one-sided p-value?



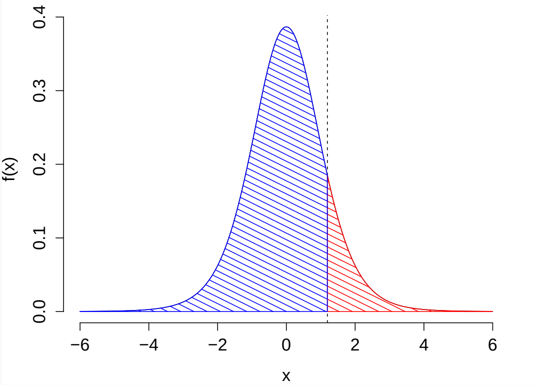
What is the p-value? (One-sided Test)

Recall: p-value is *smallest significance level* at which our observed test statistic would cause us to reject H_0. Test statistic is 1.14. What is the one-sided p-value?



What is the p-value? (One-sided Test)

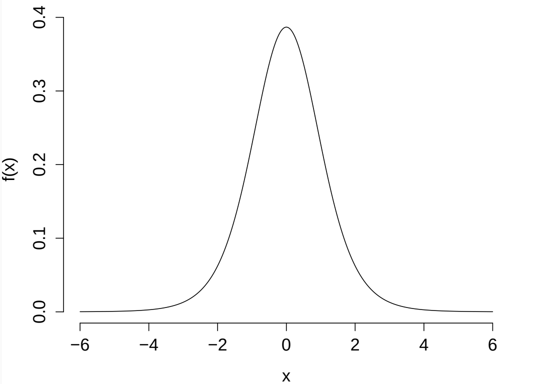
Recall: p-value is *smallest significance level* at which our observed test statistic would cause us to reject H_0. Test statistic is 1.14. What is the one-sided p-value?



`1 - pt(1.14, df = 8)` ≈ 0.14

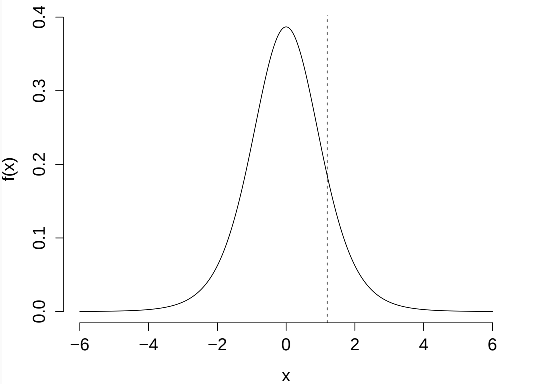
What is the p-value? (Two-sided Test)

Recall: p-value is *smallest significance level* at which our observed test statistic would cause us to reject H_0. Test statistic is 1.14. What is the two-sided p-value?



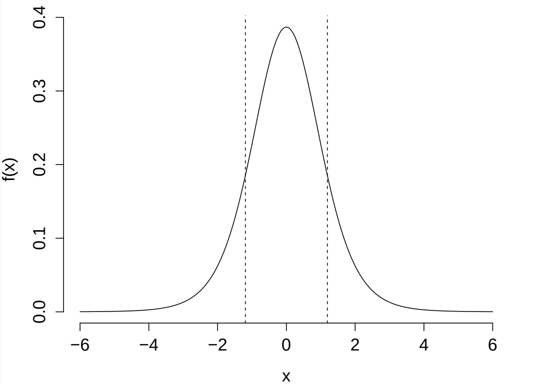
What is the p-value? (Two-sided Test)

Recall: p-value is *smallest significance level* at which our observed test statistic would cause us to reject H_0. Test statistic is 1.14. What is the two-sided p-value?



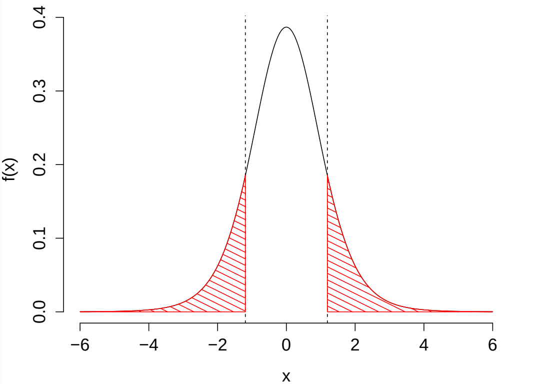
What is the p-value? (Two-sided Test)

Recall: p-value is *smallest significance level* at which our observed test statistic would cause us to reject H_0. Test statistic is 1.14. What is the two-sided p-value?



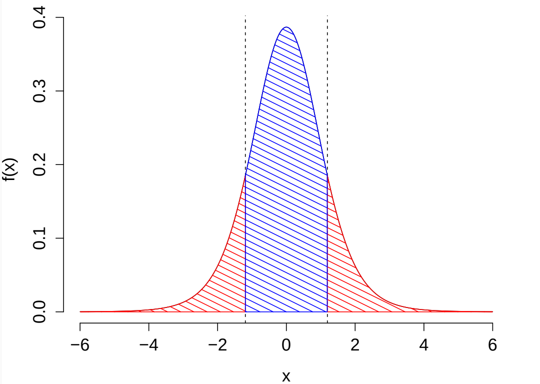
What is the p-value? (Two-sided Test)

Recall: p-value is *smallest significance level* at which our observed test statistic would cause us to reject H_0. Test statistic is 1.14. What is the two-sided p-value?



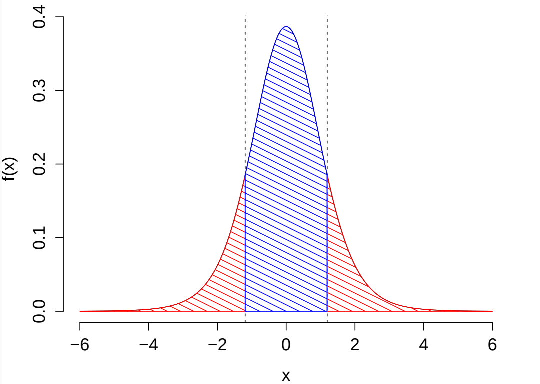
What is the p-value? (Two-sided Test)

Recall: p-value is *smallest significance level* at which our observed test statistic would cause us to reject H_0. Test statistic is 1.14. What is the two-sided p-value?



What is the p-value? (Two-sided Test)

Recall: p-value is *smallest significance level* at which our observed test statistic would cause us to reject H_0. Test statistic is 1.14. What is the two-sided p-value?



`2 * pt(-1.14, df = 8)`≈ 0.28 This is twice the one-sided p-value!

Two-sided Test is More Stringent

P-value measures strength of evidence against H_0

Lower p-value means stronger evidence.

(Two-sided p-value) $= 2 \times$ **(one-sided p-value)**

Reject H_0 based on two-sided test $\implies$ Reject H_0 based on appropriate one-sided test. The converse is *false*.

Steps in Hypothesis Testing

1. Specify Null and Alternative Hypotheses
2. Identify a Test Statistic: a function of the data that has a known sampling distribution under the null.
3. Specify a Decision Rule and a Critical Value so the Type I Error Rate equals α.

Alternative to Step 3

Calculate P-Value: the minimum significance level (α) at which we would reject H_0 given the observed data.

How to Handle Other Examples?

You already know lots of sampling distributions! Testing is very similar to constructing confidence intervals in that the steps are always the same, and the only thing that differs is *which* sampling distribution we work with.

Hypothesis Testing and Confidence Intervals

Relationship between CI and Two-Sided Test

- There is a *very close* relationship between CIs and hypothesis tests against a two-sided alternative.
- I'll illustrate this using a generic version of the example from last class but the relationship holds *in general*.

Relationship between CI and Two-sided Test

Suppose $X_1, \ldots, X_n \sim \text{iid } N(\mu, \sigma^2)$

Test $H_0\colon \mu = \mu_0$ vs. $H_1\colon \mu \neq \mu_0$ at significance level α

- Test Statistic: $T_n = \sqrt{n}(\bar{X}_n - \mu_0)/S \sim t(n-1)$ under H_0
- Decision Rule: Reject H_0 if $|T_n| > \texttt{qt}(1-\alpha/2, \texttt{df} = n-1)$

$100 \times (1-\alpha)\%$ CI for μ

$$\bar{X}_n \pm \texttt{qt}(1-\alpha/2, \texttt{df} = n-1)\frac{S}{\sqrt{n}}$$

Relationship between CI and Two-sided Test

$c = \texttt{qt}(1-\alpha/2, \texttt{df} = n-1)$

Decision Rule: Reject H_0 if

$$\left|\frac{\bar{X}_n - \mu_0}{S/\sqrt{n}}\right| > c \quad \Longleftrightarrow \quad \left(\frac{\bar{X}_n - \mu_0}{S/\sqrt{n}} > c \quad \text{OR} \quad \frac{\bar{X}_n - \mu_0}{S/\sqrt{n}} < -c\right)$$

Equivalent to: *Don't Reject* H_0 provided

$$-c \leq \frac{\bar{X}_n - \mu_0}{S/\sqrt{n}} \leq c$$

$$\bar{X}_n - c \times \frac{S}{\sqrt{n}} \leq \mu_0 \leq \bar{X}_n + c \times \frac{S}{\sqrt{n}}$$

What does this mean?

Two-sided Test $\iff$ Checking if $\mu_0 \in$ CI

A two-sided test of $H_0\colon \mu = \mu_0$ against $H_1\colon \mu \neq \mu_0$ at significance level α is equivalent to checking whether μ_0 lies inside the corresponding $100 \times (1-\alpha)\%$ confidence interval for μ.

"Inverting" Two-sided Test to get a CI

Collect all the values μ_0 such that we cannot reject $H_0\colon \mu = \mu_0$ against the two-sided alternative. The result is *precisely* a $100 \times (1-\alpha)\%$ CI for μ.

The Anchoring Effect

The Anchoring Experiment

Shown a “random” number and then asked what proportion of UN member states are located in Africa.

“Hi” Group – Shown 65 ($n_{Hi} = 46$)
Sample Mean: 30.7, Sample Variance: 253

“Lo” Group – Shown 10 ($n_{Lo} = 43$)
Sample Mean: 17.1, Sample Variance: 86

Fairly large samples here, so we’ll proceed via the CLT…

In words, what is our null hypothesis?

(a) There is a *positive* anchoring effect: seeing a higher random number makes people report a higher answer.

(b) There is a *negative* anchoring effect: seeing a lower random number makes people report a lower answer.

(c) There *is* an anchoring effect: it could be positive or negative.

(d) There is *no* anchoring effect: people aren't influenced by seeing a random number before answering.

In symbols, what is our null hypothesis?

(a) $\mu_{Lo} < \mu_{Hi}$

(b) $\mu_{Lo} = \mu_{Hi}$

(c) $\mu_{Lo} > \mu_{Hi}$

(d) $\mu_{Lo} \neq \mu_{Hi}$

$\mu_{Lo} = \mu_{Hi}$ is *equivalent to* $\mu_{Hi} - \mu_{Lo} = 0$!

Under the null, what should we expect to be true about the values taken on by $\bar{X}_{Lo}$ and $\bar{X}_{Hi}$?

(a) They should be similar in value.
(b) $\bar{X}_{Lo}$ should be the smaller of the two.
(c) $\bar{X}_{Hi}$ should be the smaller of the two.
(d) They should be different. We don't know which will be larger.

What is our Test Statistic?

Sampling Distribution

$$\frac{(\bar{X}_{Hi} - \bar{X}_{Lo}) - (\mu_{Hi} - \mu_{Lo})}{\sqrt{\frac{S^2_{Hi}}{n_{Hi}} + \frac{S^2_{Lo}}{n_{Lo}}}} \approx N(0,1)$$

Test Statistic: Impose the Null

Under $H_0: \mu_{Lo} = \mu_{Hi}$

$$T_n = \frac{\bar{X}_{Hi} - \bar{X}_{Lo}}{\sqrt{\frac{S^2_{Hi}}{n_{Hi}} + \frac{S^2_{Lo}}{n_{Lo}}}} \approx N(0,1)$$

What is our Test Statistic?

$\bar{X}_{Hi} = 30.7$, $s_{Hi}^2 = 253$, $n_{Hi} = 46$
$\bar{X}_{Lo} = 17.1$, $s_{Lo}^2 = 86$, $n_{Lo} = 43$

Under $H_0 : \mu_{Lo} = \mu_{Hi}$

$$T_n = \frac{\bar{X}_{Hi} - \bar{X}_{Lo}}{\sqrt{\frac{S_{Hi}^2}{n_{Hi}} + \frac{S_{Lo}^2}{n_{Lo}}}} \approx N(0, 1)$$

Plugging in Our Data

$$T_n = \frac{\bar{X}_{Hi} - \bar{X}_{Lo}}{\sqrt{\frac{S_{Hi}^2}{n_{Hi}} + \frac{S_{Lo}^2}{n_{Lo}}}} \approx 5$$

Anchoring Experiment Example

Approximately what critical value should we use to test $H_0 : \mu_{Lo} = \mu_{Hi}$ against the two-sided alternative at the 5% significance level?

α	0.10	0.05	0.01
`qnorm(`$1-\alpha$`)`	1.28	1.64	2.33
`qnorm(`$1-\alpha/2$`)`	1.64	1.96	2.58

… Approximately 2

Which of these commands would give us the p-value of our test of $H_0\colon \mu_{Lo} = \mu_{Hi}$ against $H_1\colon \mu_{Lo} < \mu_{Hi}$ at significance level α?

(a) `qnorm(`$1-\alpha$`)`
(b) `qnorm(`$1-\alpha/2$`)`
(c) `1 - pnorm(5)`
(d) `2 * (1 - pnorm(5))`

P-values for $H_0: \mu_{Lo} = \mu_{Hi}$

We plug in the value of the test statistic that we observed: 5

Against $H_1: \mu_{Lo} < \mu_{Hi}$

`1 - pnorm(5)` < 0.0000

Against $H_1: \mu_{Lo} \neq \mu_{Hi}$

`2 * (1 - pnorm(5))` < 0.0000

If the null is true (the two population means are equal) it would be extremely unlikely to observe a test statistic as large as this!

What should we conclude?

Exam Difficulty

Which Exam is Harder?

Student	Exam 1	Exam 2	Difference
1	57.1	60.7	3.6
$\vdots$	$\vdots$	$\vdots$	$\vdots$
71	78.6	82.9	4.3
Sample Mean:	79.6	81.4	1.8
Sample Var.	117	151	124
Sample Corr.	0.54		

Again, large sample size here so we'll use CLT.

One-Sample Hypothesis Test Using Differences

$\boxed{\text{Let } D_i = X_i - Y_i \text{ be (Midterm 2 Score - Midterm 1 Score) for student } i}$

Null Hypothesis

$H_0 \colon \mu_1 = \mu_2$, i.e. both exams were of the same difficulty

Two-Sided Alternative

$H_1 \colon \mu_1 \neq \mu_2$, i.e. one exam was harder than the other

One-Sided Alternative

$H_1 \colon \mu_2 > \mu_1$, i.e. the second exam was easier

Decision Rules

$\boxed{\text{Let } D_i = X_i - Y_i \text{ be (Midterm 2 Score - Midterm 1 Score) for student } i}$

Test Statistic

$$\frac{\bar{D}_n}{\widehat{SE}(\bar{D}_n)} = \frac{1.8}{\sqrt{124/71}} \approx 1.36$$

Two-Sided Alternative

Reject $H_0: \mu_1 = \mu_2$ in favor of $H_1: \mu_1 \neq \mu_2$ if $|\bar{D}_n|$ is sufficiently large.

One-Sided Alternative

Reject $H_0: \mu_1 = \mu_2$ in favor of $H_1: \mu_2 > \mu_1$ if $\bar{D}_n$ is sufficiently large.

Reject against *Two-sided* Alternative with $\alpha = 0.1$?

$$\boxed{\frac{\bar{D}_n}{\widehat{SE}(\bar{D}_n)} = \frac{1.8}{\sqrt{124/71}} \approx 1.36}$$

α	0.10	0.05	0.01
`qnorm(`$1-\alpha$`)`	1.28	1.64	2.33
`qnorm(`$1-\alpha/2$`)`	1.64	1.96	2.58

(a) Reject

(b) Fail to Reject

(c) Not Sure

Reject against *One-sided* Alternative with $\alpha = 0.1$?

$$\boxed{\frac{\bar{D}_n}{\widehat{SE}(\bar{D}_n)} = \frac{1.8}{\sqrt{124/71}} \approx 1.36}$$

α	0.10	0.05	0.01
`qnorm(`$1-\alpha$`)`	1.28	1.64	2.33
`qnorm(`$1-\alpha/2$`)`	1.64	1.96	2.58

(a) Reject

(b) Fail to Reject

(c) Not Sure

P-Values for the Test of $H_0: \mu_1 = \mu_2$

$$\boxed{\frac{\bar{D}_n}{\widehat{SE}(\bar{D}_n)} = \frac{1.8}{\sqrt{124/71}} \approx 1.36}$$

One-Sided $H_1: \mu_2 > \mu_1$

`1 - pnorm(1.36)` $=$ `pnorm(-1.36)` ≈ 0.09

Two-Sided $H_1: \mu_1 \neq \mu_2$

`2 * (1 - pnorm(1.36))` $=$ `2 * pnorm(-1.36)` ≈ 0.18

Basic Idea

The population *can't be* normal (it's Bernoulli) so we use the CLT to get approximate sampling distributions.

But there's a small twist!

Bernoulli RV only has a *single* unknown parameter $\implies$ we know *more* about the population under H_0 in a proportions problem than in the other testing examples we've examined...

For best results, always *fully* impose the null.

2012 Voter Polls

Tests for Proportions: One-Sample Example

From Pew Polling Data

54% of a random sample of 771 registered voters correctly identified 2012 presidential candidate Mitt Romney as Pro-Life.

Sampling Model

$X_1, \ldots, X_n \sim$ iid Bernoulli(p)

Sample Statistic

Sample Proportion: $\widehat{p} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$

$\boxed{\text{Suppose I wanted to test } H_0 : p = 0.5}$

Tests for Proportions: One Sample Example

Under $H_0: p = 0.5$ what is the standard error of $\widehat{p}$?

(a) 1

(b) $\sqrt{\widehat{p}(1-\widehat{p})/n}$

(c) $\sigma/\sqrt{n}$

(d) $1/(2\sqrt{n})$

(e) $p(1-p)$

$p = 0.5 \implies \sqrt{0.5(1-0.5)/n} = 1/(2\sqrt{n})$

Under the null we know the SE! Don't have to estimate it!

Sampling Model

$X_1, \ldots, X_n \sim$ iid Bernoulli(p)

Null Hypothesis

$H_0\colon\, p =$ Known Constant p_0

Test Statistic

$$T_n = \frac{\widehat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \approx N(0, 1) \text{ under } H_0 \text{ provided } n \text{ is large}$$

One-Sample Example $H_0: p = 0.5$

54% of a random sample of 771 registered voters knew Mitt Romney is Pro-Life.

$$
\begin{aligned}
T_n &= \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = 2\sqrt{771}(0.54 - 0.5) \\
&= 0.08 \times \sqrt{771} \approx 2.2
\end{aligned}
$$

One-Sided p-value

`1 - pnorm(2.2)` ≈ 0.014

Two-Sided p-value

`2 * (1 - pnorm(2.2))` ≈ 0.028

Tests for Proportions: Two-Sample Example

From Pew Polling Data

53% of a random sample of 238 Democrats correctly identified Mitt Romney as Pro-Life versus 61% of 239 Republicans.

Sampling Model

Republicans: $X_1, \ldots, X_n \sim$ iid Bernoulli(p) independent of
Democrats: $Y_1, \ldots, Y_m \sim$ iid Bernoulli(q)

Sample Statistics

Sample Proportions: $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{q} = \frac{1}{m}\sum_{i=1}^{m} Y_i$

$\boxed{\text{Suppose I wanted to test } H_0: p = q}$

A More Efficient Estimator of the SE Under H_0

Don't Forget!

Standard Error (SE) means "std. dev. of sampling distribution" so you should know how to prove that that:

$$SE(\widehat{p} - \widehat{q}) = \sqrt{\frac{p(1-p)}{n} + \frac{q(1-q)}{m}}$$

Under $H_0: p = q$

Don't know values of p and q: only that they are equal.

A More Efficient Estimator of the SE Under H_0

One Possible Estimate

$$\widehat{SE} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{\widehat{q}(1-\widehat{q})}{m}}$$

A *Better* Estimate Under H_0

$$\widehat{SE}_{Pooled} = \sqrt{\widehat{\pi}(1-\widehat{\pi})\left(\frac{1}{n}+\frac{1}{m}\right)} \quad \text{where} \quad \widehat{\pi} = \frac{n\widehat{p}+m\widehat{q}}{n+m}$$

Why Pool?

If $p = q$, the two populations *are the same*. This means we can get a *more precise* estimate of the *common* population proportion by pooling. More data = Lower Variance $\implies$ better estimated SE.

Two-Sample Test for Proportions

Sampling Model

$X_1, \ldots, X_n \sim$ iid Bernoulli(p) indep. of $Y_1, \ldots, Y_m \sim$ iid Bernoulli(q)

Sample Statistics

Sample Proportions: $\widehat{p} = \dfrac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{q} = \dfrac{1}{m}\sum_{i=1}^{m} Y_i$

Null Hypothesis

$H_0\colon p = q \quad \Leftarrow \boxed{\text{i.e. } p - q = 0}$

Pooled Estimator of SE under H_0

$$\widehat{\pi} = \frac{n\widehat{p} + m\widehat{q}}{n+m}, \quad \widehat{SE}_{Pooled} = \sqrt{\widehat{\pi}(1-\widehat{\pi})\,(1/n + 1/m)}$$

Test Statistic

$T_n = \dfrac{\widehat{p} - \widehat{q}}{\widehat{SE}_{Pooled}} \approx N(0,1)$ under H_0 provided n and m are large

Two-Sample Example $H_0: p = q$

53% of 238 Democrats knew Romney is Pro-Life vs. 61% of 239 Republicans

$$\widehat{\pi} = \frac{n\widehat{p} + m\widehat{q}}{n+m} = \frac{239 \times 0.61 + 238 \times 0.53}{239 + 238} \approx 0.57$$

$$\begin{aligned} \widehat{SE}_{Pooled} &= \sqrt{\widehat{\pi}(1-\widehat{\pi})\,(1/n + 1/m)} = \sqrt{0.57 \times 0.43(1/239 + 1/238)} \\ &\approx 0.045 \end{aligned}$$

$$T_n = \frac{\widehat{p} - \widehat{q}}{\widehat{SE}_{Pooled}} = \frac{0.61 - 0.53}{0.045} \approx 1.78$$

One-Sided P-Value

`1 - pnorm(1.78)` ≈ 0.04

Two-Sided P-Value

`2 * (1 - pnorm(1.78))` ≈ 0.08

Biased Scales

Experiment

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

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

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

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

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

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

(a) $N(\mu, 4)$
(b) $N(0, 4)$
(c) $t(15)$
(d) $\chi^2(15)$
(e) $N(0, 1)$

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

Testing Whether a Scale is Biased

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

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

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

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 $H_0\colon \mu = 0$

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

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

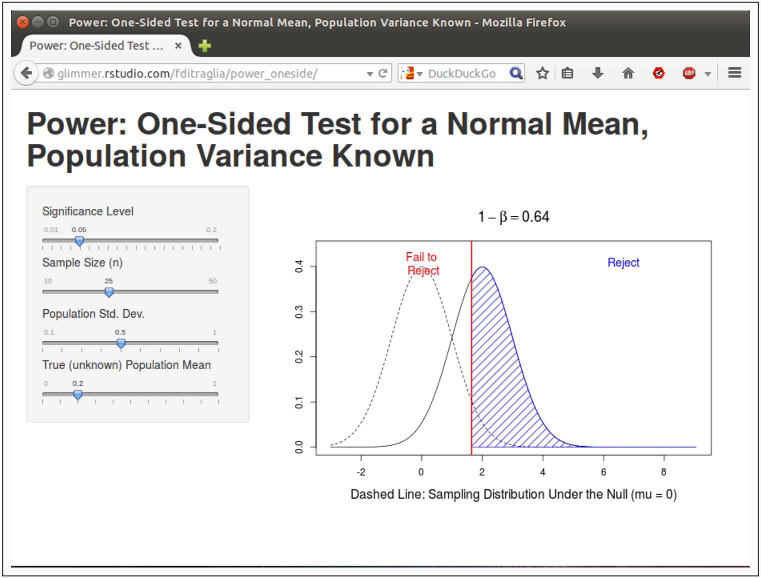
$T_n = \sqrt{n}(\bar{X}_n/\sigma) \sim N\left(\sqrt{n}(\mu/\sigma), 1\right)$

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

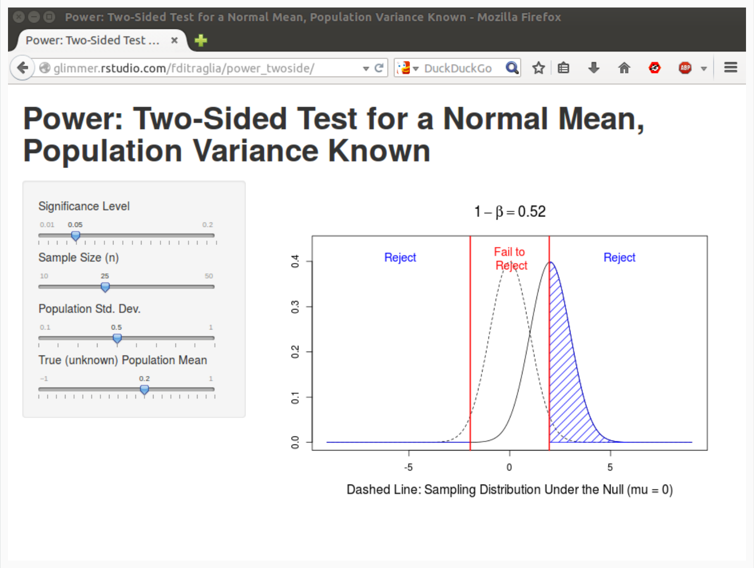
Reject $H_0: \mu = 0$ if $|T_n| > \texttt{qnorm}(1-\alpha)$

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

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

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

https://fditraglia.shinyapps.io/power_twoside



What Determines Power?

Power $= 1 - P$(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

Statistical Significance

Suppose we carry out a hypothesis test at the α% 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)

$\boxed{\text{Actual } p \text{ was } 0.499}$

Practically Important But Not Statistically Significant

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?

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?

$$\boxed{\text{Test Statistic} \approx 3.9}$$

(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)

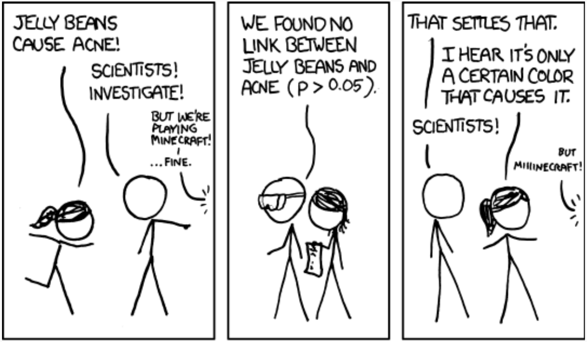


Figure 5: Go and read this comic strip: before today's lecture you wouldn't have gotten the joke!

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.