Chapter 4

4.1: Previously we looked at how to estimate the value of a parameter, using information from the sample statistic. In this section, we are going to look at how to use the statistic to answer a question about the location of a parameter.

A statistical test or hypothesis test uses data from a sample to assess a claim about a population. You can think of the test as asking a question about the parameter, and we use the statistic to help us answer the question. These tests have their own language.

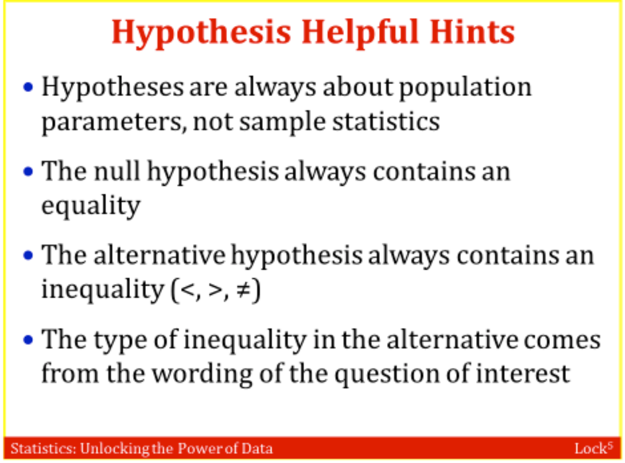
https://www.youtube.com/watch?v=VK-rnA3-41c (Intro to Hypothesis Testing in Statistics)

- **Hypothesis testing**
  is a procedure where claims about the value of a population parameter may be investigated using the sample evidence. (because it is usually impossible or impractical to gain access to the entire population)

How do I ask a question about the parameter using the language of a hypothesis test?
We set up a statistical test by first identifying 2 competing hypothesis.

- The **null hypothesis** denoted H_0. The null hypothesis is a statement of no change, no effect or no difference and is assumed true until evidence indicates otherwise.
- The **alternative hypothesis** denoted H_1 or H_a, also call the research hypothesis, involves the claim for which we seek evidence
- The alternative hypothesis is usually what we would like to prove. We observe evidence (data) that contradicts the null hypothesis and supports the alternative hypothesis.

- The major steps in hypothesis testing are
  - Formulate the appropriate null and alternative hypotheses
  - Calculate the test statistic
  - Determine the appropriate critical value(s)
  - Reach the reject / do not reject conclusions

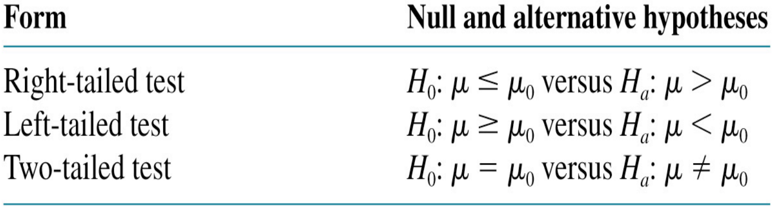


Hypothesis Helpful Hints

- Hypotheses are always about population parameters, not sample statistics
- The null hypothesis always contains an equality
- The alternative hypothesis always contains an inequality (<, >, ≠)
- The type of inequality in the alternative comes from the wording of the question of interest

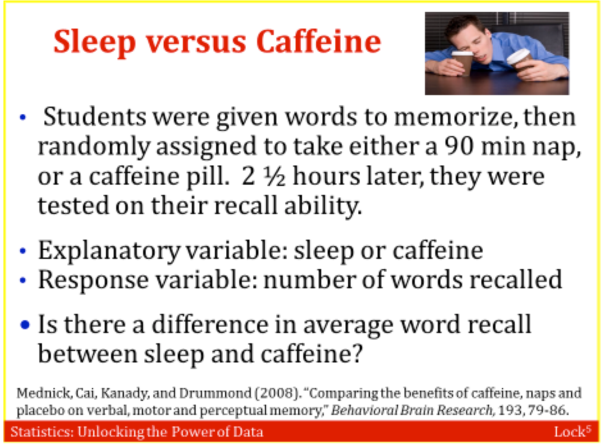
Statistics: Unlocking the Power of Data Lock5

- The three possible forms for the hypotheses for a test for μ



Form	Null and alternative hypotheses
Right-tailed test	H_0: $\mu \leq \mu_0$ versus H_a: $\mu > \mu_0$
Left-tailed test	H_0: $\mu \geq \mu_0$ versus H_a: $\mu < \mu_0$
Two-tailed test	H_0: $\mu = \mu_0$ versus H_a: $\mu \neq \mu_0$

https://www.youtube.com/watch?v=_Qlxt0HmuOo (video for how to set up H_0 and H_a)

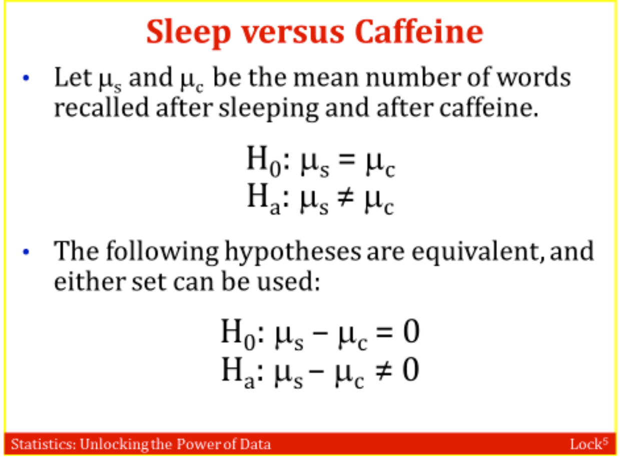


Sleep versus Caffeine

- Students were given words to memorize, then randomly assigned to take either a 90 min nap, or a caffeine pill. 2 ½ hours later, they were tested on their recall ability.
- Explanatory variable: sleep or caffeine
- Response variable: number of words recalled
- Is there a difference in average word recall between sleep and caffeine?

Mednick, Cai, Kanady, and Drummond (2008). "Comparing the benefits of caffeine, naps and placebo on verbal, motor and perceptual memory," *Behavioral Brain Research*, 193, 79-86.

Statistics: Unlocking the Power of Data Lock^5



Sleep versus Caffeine

- Let μ_s and μ_c be the mean number of words recalled after sleeping and after caffeine.

$$H_0: \mu_s = \mu_c$$
$$H_a: \mu_s \neq \mu_c$$

- The following hypotheses are equivalent, and either set can be used:

$$H_0: \mu_s - \mu_c = 0$$
$$H_a: \mu_s - \mu_c \neq 0$$

Statistics: Unlocking the Power of Data Lock^5

- The null hypothesis is a statement of *no difference* and always contains a statement of equality. The null hypothesis is assumed true until we have evidence to the contrary. We seek evidence that supports the statement in the alternative hypothesis
- Key words:
  Difference, change, differ $\neq$
  Less, more, decrease, increase $<$ or $>$

Ex: According to a study published in March, 2006 the mean length of a phone call on a cellular telephone was 3.25 minutes. A researcher believes that the mean length of a call has increased since then.

H_0: μ = 3.25. *H_a: $\mu > 3.25$, a right-tailed test.*

Write down the hypotheses for the test in each case below:

a). Does the proportion of people who support gun control differ between males and females?

p_f: proportion of females who support gun control

p_m: proportion of males who support gun control

H_0: $p_f = p_m$ H_a: $p_f \neq p_m$

b). Are the average hours of sleep per night for college students less than 7?

H_0: $\mu = 7$ where μ is the average number of hours of sleep at night for college students

H_a: $\mu < 7$ since we are looking for evidence that the mean is less than 7

Ex; State whether each set of hypotheses is valid for a statistical test.

(a) *H_0: $\mu_1 \neq \mu_2$ vs H_a: $\mu_1 = \mu_2$* *Invalid*

(b) H_0: $p = 0.5$ vs H_a: $p > 0.5$ *Valid*

(c) H_0: $\hat{p}_1 = \hat{p}_2$ vs H_a: $\hat{p}_1 < \hat{p}_2$ *Invalid*

(d) H_0: $p_1 < p_2$ vs H_a: $p_1 > p_2$ *Invalid*

4.2 : Measuring Evidence with p-values

In this section we are going to look at how to determine which outcome is appropriate. In other words, based on our observed statistic should we reject the null or fail to reject the null.

Our key question is …

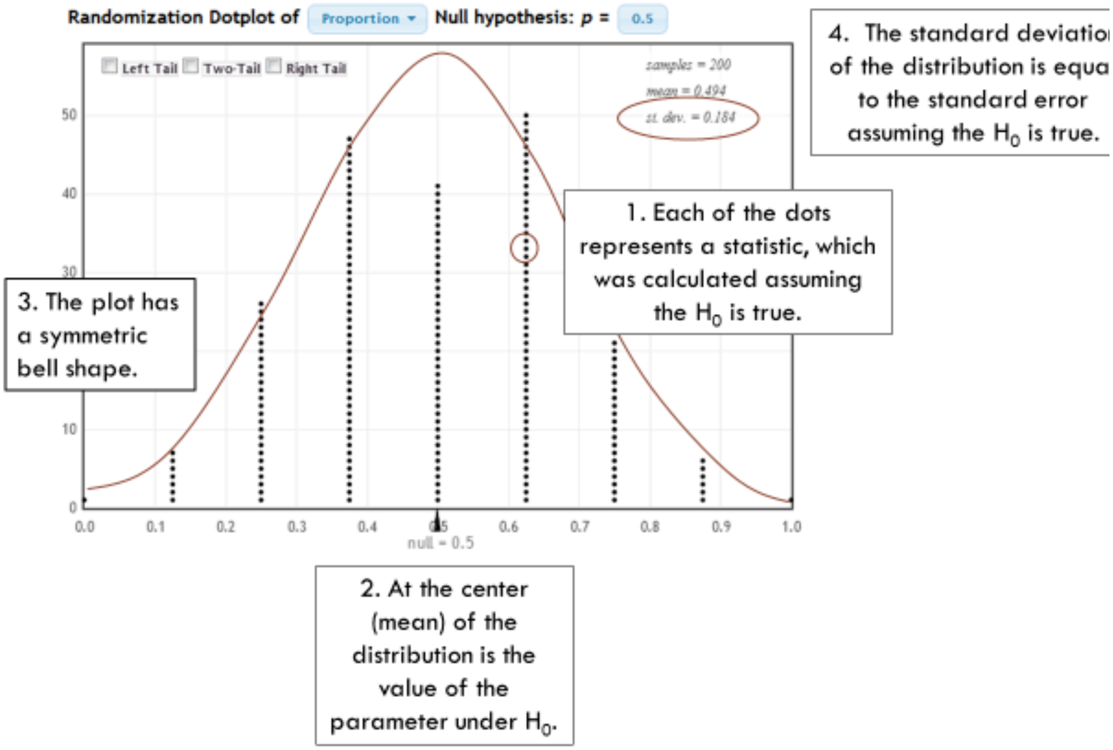
How unusual is it to see a sample statistic as extreme as that observed, if H_0 is true?

If it is very unusual, we have *statistically significant* evidence against the null hypothesis.

To see if a statistic provides evidence against $H_{0,}$ we need to see what kind of sample statistics we would observe, just by random chance, when the null hypothesis is true.

A randomization distribution is a collection of statistics from samples simulated assuming the null hypothesis is true. The randomization distribution shows what types of statistics would be observed, just by random chance, if the null hypothesis were true.

A randomization distribution simulates samples assuming the null hypothesis is true, so
A randomization distribution is centered at the value of the parameter given in the null hypothesis



4. The standard deviatio
of the distribution is equa
to the standard error
assuming the H_0 is true.

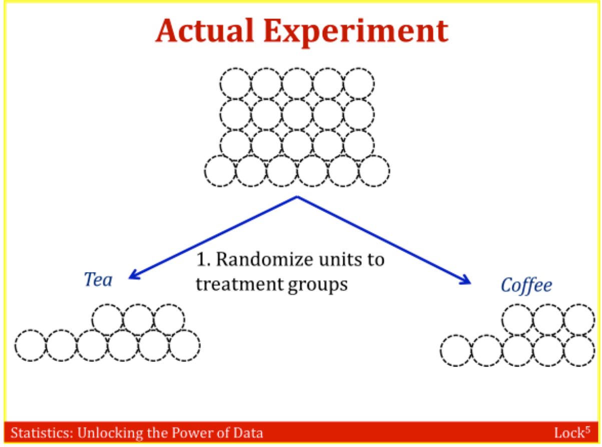
1. Each of the dots represents a statistic, which was calculated assuming the H_0 is true.

3. The plot has a symmetric bell shape.

2. At the center (mean) of the distribution is the value of the parameter under H_0.

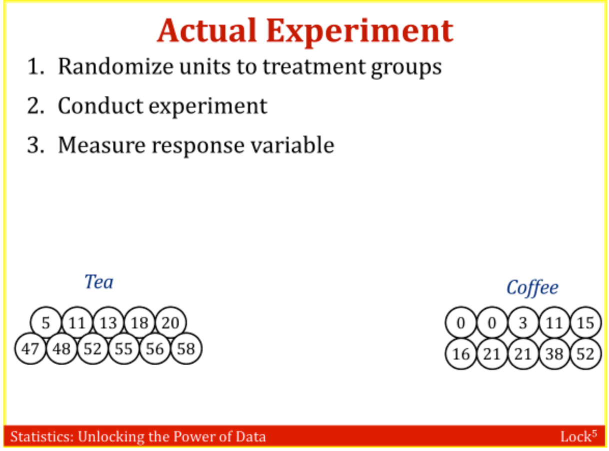
Does drinking tea boost your immune system?

- Explanatory variable: tea or coffee
- Response variable: immune system response



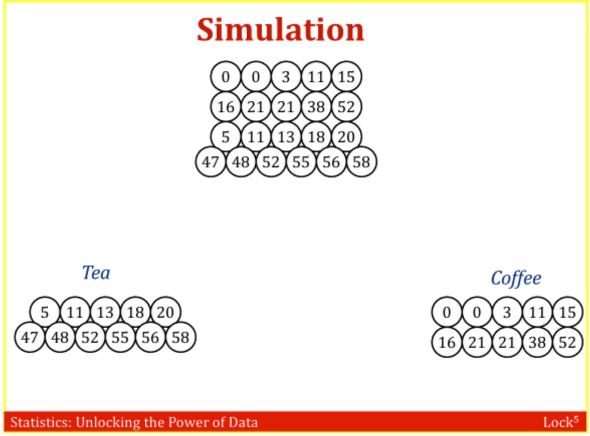
Actual Experiment

1. Randomize units to treatment groups

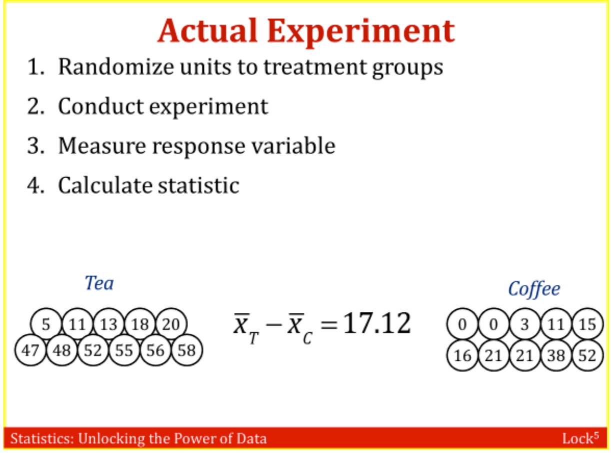


Actual Experiment

1. Randomize units to treatment groups
2. Conduct experiment
3. Measure response variable



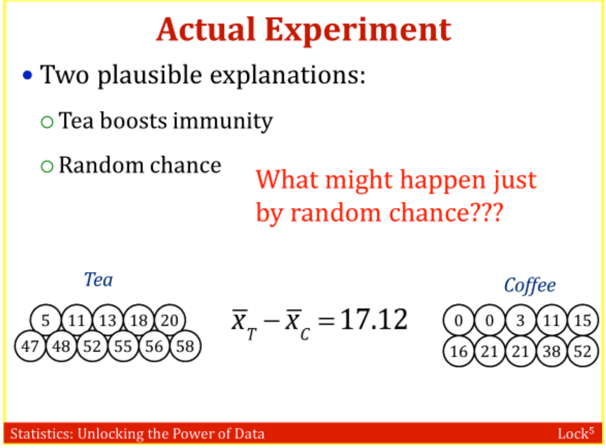
Simulation



Actual Experiment

1. Randomize units to treatment groups
2. Conduct experiment
3. Measure response variable
4. Calculate statistic

$\bar{x}_T - \bar{x}_C = 17.12$

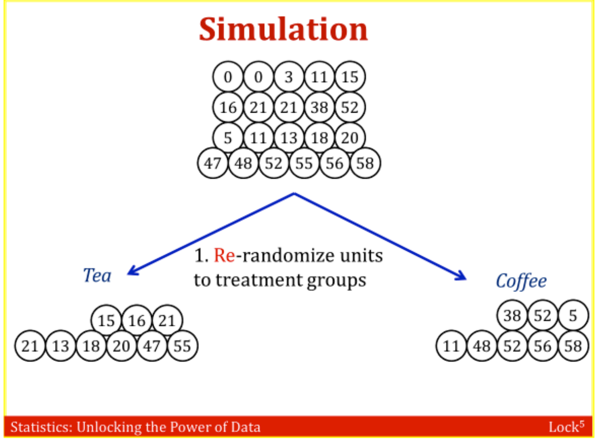


Actual Experiment

- Two plausible explanations:
  - Tea boosts immunity
  - Random chance

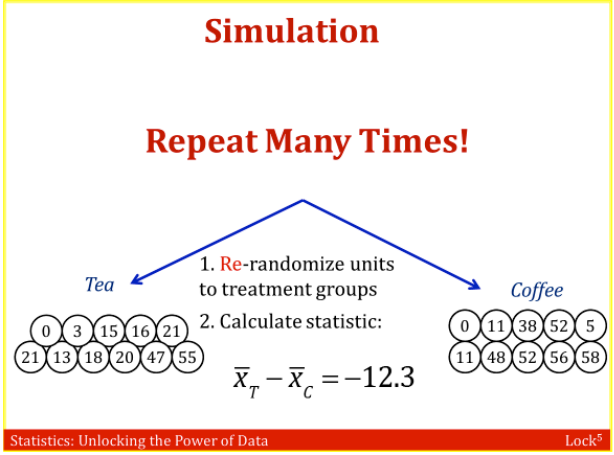
What might happen just by random chance???

$\bar{x}_T - \bar{x}_C = 17.12$



Simulation

1. Re-randomize units to treatment groups



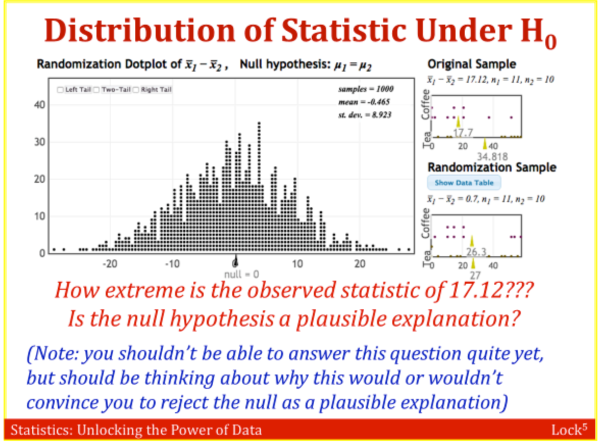
Simulation

Repeat Many Times!

1. Re-randomize units to treatment groups
2. Calculate statistic:

$$\bar{x}_T - \bar{x}_C = -12.3$$

Statistics: Unlocking the Power of Data Lock5



Distribution of Statistic Under H_0

How extreme is the observed statistic of 17.12???
Is the null hypothesis a plausible explanation?

(Note: you shouldn't be able to answer this question quite yet, but should be thinking about why this would or wouldn't convince you to reject the null as a plausible explanation)

Statistics: Unlocking the Power of Data Lock5

Randomization Test:

1. **State hypotheses**
2. **Collect data**
3. **Calculate statistic:**
4. **Simulate statistics that could be observed, just by random chance, if the null hypothesis were true (create a randomization distribution)**
5. **How extreme is the observed statistic? Is the null hypothesis (random chance) a plausible explanation?**

The p-value is the chance of obtaining a sample statistic as extreme (or more extreme) than the observed sample statistic, if the null hypothesis is true.

When the p-value is low we reject H_0.

When the p-value is high we fail to reject H_0.

We measure the strength of evidence a sample shows against the null hypothesis with a p-value.

The p-value is the probability of obtaining a sample statistic as extreme as (or more extreme than) the observed sample statistic, when the null hypothesis is true
It describes how unusual the observed data would be if H_0 were true.

A small p-value means that the observed sample results would be unlikely to happen, when the null hypothesis is true, just by random chance.

The smaller the *P*-value, the stronger the evidence the data provide against the null hypothesis.
When making formal decisions based on the p-value, we use a pre-specified significance level, α.

• If p-value < α, we reject H0 and have statistically significant evidence for Ha.

• If p-value ≥ α, we do not reject H0, the test is inconclusive, and the results are not statistically significant.

One way to estimate a p-value is to construct a randomization distribution of sample statistics that we might see by random chance, if the null hypothesis were true.

The p-value is the proportion of randomization statistics that are as extreme as the observed sample statistic.

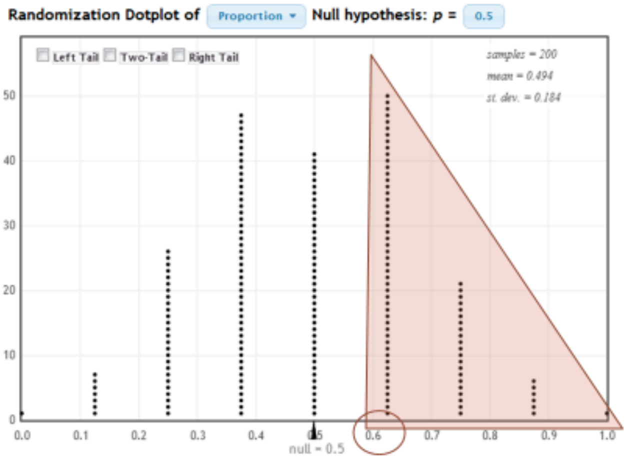
Remember that the randomization distribution shows us what statistics look like when H_0 is true.

Suppose we are testing:

- H_0: p = .5
- H_a: p > .5

And we observe a statistics equal to .6

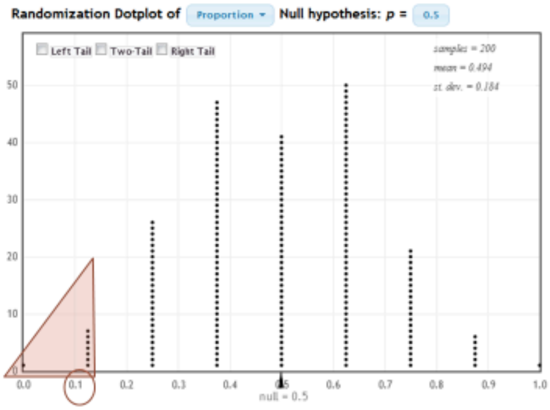
All of these statistics are as or more extreme than the one we observed



Randomization Dotplot of Proportion Null hypothesis: p = 0.5

samples = 200
mean = 0.494
st. dev. = 0.184

78 of the 200 statistics fall in the triangle, therefore:
p-value = 78/200 = .39



Randomization Dotplot of Proportion Null hypothesis: p = 0.5

samples = 200
mean = 0.494
st. dev. = 0.184

8 of the 200 statistics fall in the triangle, therefore:
p-value = 8/200 = .04

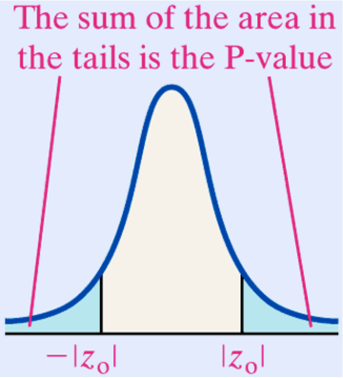
Remember that the randomization distribution shows us what statistics look like when H_0 is true.

Suppose we are testing:

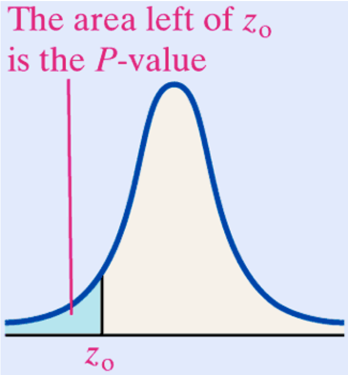
- **H_0: p = .5**
- **H_a: p < .5**

And we observe a statistics equal to .1

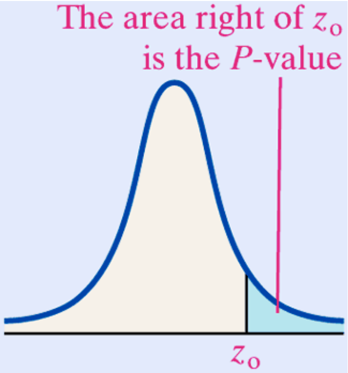
All of these statistics are as or more extreme than the one we observed



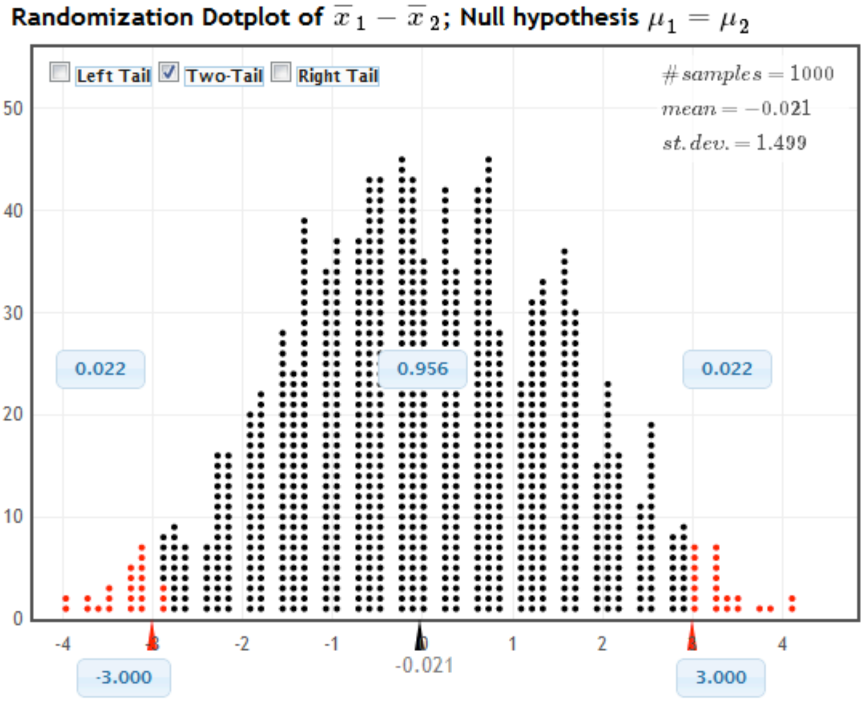
Two-tailed test



left tailed test



Right tailed test



A randomization distribution for difference in mean memory recall between sleep and caffeine groups for data in SleepCaffeine is shown. Each dot is a difference in means that might be observed just by random assignment to treatment groups, if there were no difference in terms of mean (memory) response. The hypotheses are:

H_0: $\mu_s = \mu_c$ vs H_a: $\mu_s \neq \mu_c$.

The sample statistic is $\bar{x}_s - \bar{x}_c = 3.0$.

Use the randomization distribution to state the p-value.

We see that 0.022*2= 0.044 (because two tailed test)of the simulated statistics are as extreme as the observed statistic ($\bar{x}_S - \bar{x}_C = 3$), so the p-value is 0.044. This p-value is less than 0.05, so the results are statistically significant at $\alpha = 0.05$, giving moderately strong evidence that sleeping is better than drinking caffeine for memory.

Example: Support for the Death Penalty

In 1980 and again in 2010, a Gallup poll asked a random sample of 1000 US citizens" Are you in favor of the death penalty for a person convicted of murder?" In 1980, the proportion saying yes was 0.66. In 2010, it was 0.64. Does this data provide evidence that the proportion of US citizens favoring the death penalty was higher in 1980 than it was in 2010? Using p1 for the proportion in 1980 and p2 for the proportion in 2010:

(a) State the null and alternative hypotheses:

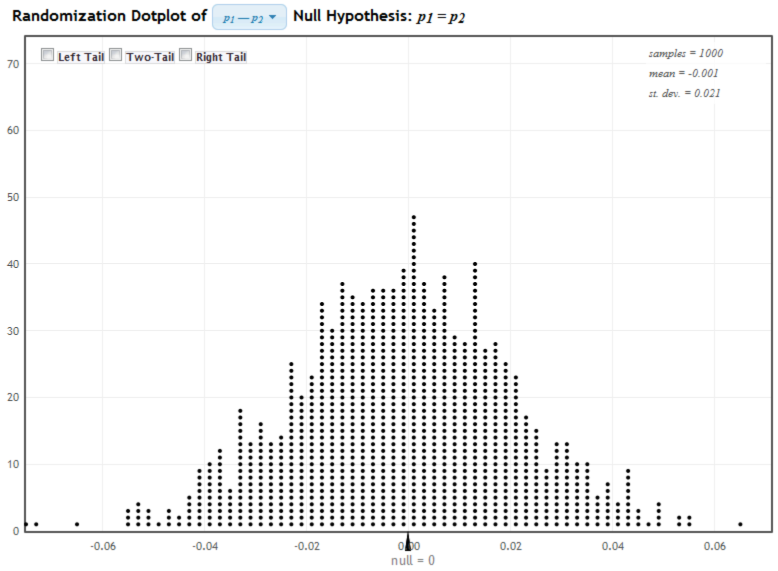
This is a difference in proportions test, with hypotheses $H_0 : p_1 = p_2$ vs $H_a : p_1 > p_2$.

(b) What is the sample statistic?

The sample statistic is the difference in sample statistics: $\hat{p}_1 - \hat{p}_2 = 0.66 - 0.64 = 0.02$

(c) To create the randomization distribution, what do we have to assume?
To create the simulated statistics, we assume the proportions are equal, as stated in the null hypothesis.

(d) (Show a randomization distribution on StatKey or a slide and ask:) Which of the following is closest to the p-value? 0.001, 0.05, 0.20, 0.5



Randomization Dotplot of $p_1 - p_2$ ▾ Null Hypothesis: $p_1 = p_2$

□ Left Tail □ Two-Tail □ Right Tail

samples = 1000
mean = -0.001
st. dev. = 0.021

The p-value is the proportion of dots in the area indicated(right tailed test), which is closest to 0.20

Quick Self-Quiz: P-values from Randomization Distributions

To test H_0: $\mu = 50$ vs H_a: $\mu < 50$ using sample data with $\bar{x} = 43.7$:

Where will the randomization distribution be centered? Why?

At 50, since we must assume the null hypothesis is true when we create the randomization distribution.

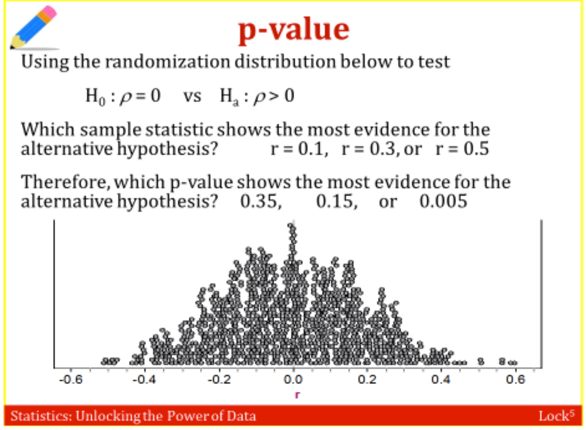
Is this a left-tail test, a right-tail test, or a two-tail test?

It is a left-tail test, since the alternative hypothesis is $\mu < 50$.

How can we find the p-value once we have the randomization distribution?

We see how extreme the sample statistic of 43.7 is in the left tail of the randomization distribution.

4.3. **Determining Statistical Significance**



p-value

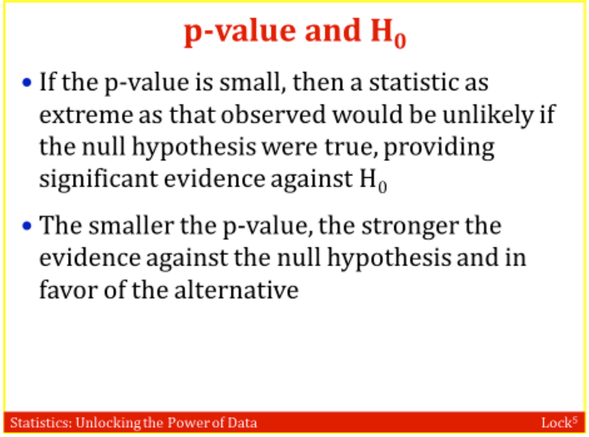
Using the randomization distribution below to test

$H_0: \rho = 0$ vs $H_a: \rho > 0$

Which sample statistic shows the most evidence for the alternative hypothesis? $r = 0.1$, $r = 0.3$, or $r = 0.5$

Therefore, which p-value shows the most evidence for the alternative hypothesis? 0.35, 0.15, or 0.005

Statistics: Unlocking the Power of Data Lock5



p-value and H_0

- If the p-value is small, then a statistic as extreme as that observed would be unlikely if the null hypothesis were true, providing significant evidence against H_0
- The smaller the p-value, the stronger the evidence against the null hypothesis and in favor of the alternative

Statistics: Unlocking the Power of Data Lock5

$r = 0.5$ shows the most evidence and $p = 0.005$ shows the most evidence. Idea to get across: Sample statistics far out in the tail show the most evidence against the null, so small p-values show the most evidence.

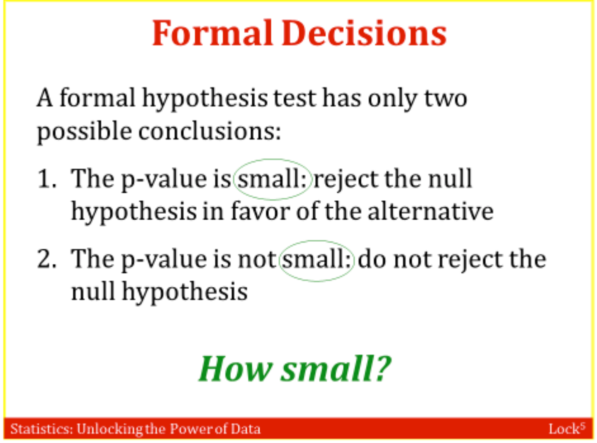
Quick Self-Quiz: Which P-value shows more evidence?

In each case, which p-value provides the strongest evidence against H_0 and for H_a?

a). p-value = 0.95 or p-value = 0.02

b). p-value = 0.008 or p-value = 0.02

The smaller the p-value, the stronger the evidence against H_0



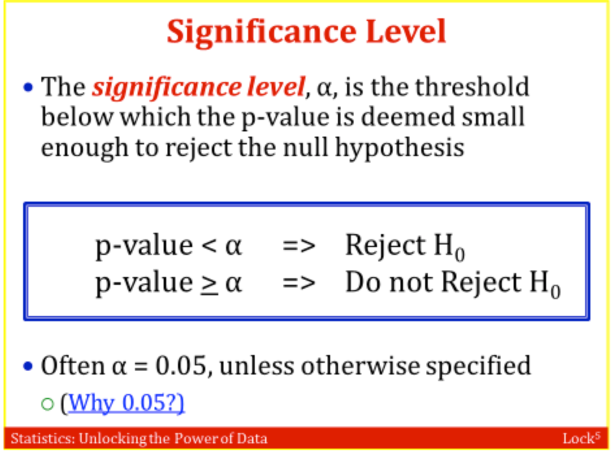
Formal Decisions

A formal hypothesis test has only two possible conclusions:

1. The p-value is small: reject the null hypothesis in favor of the alternative
2. The p-value is not small: do not reject the null hypothesis

How small?

Statistics: Unlocking the Power of Data Lock5



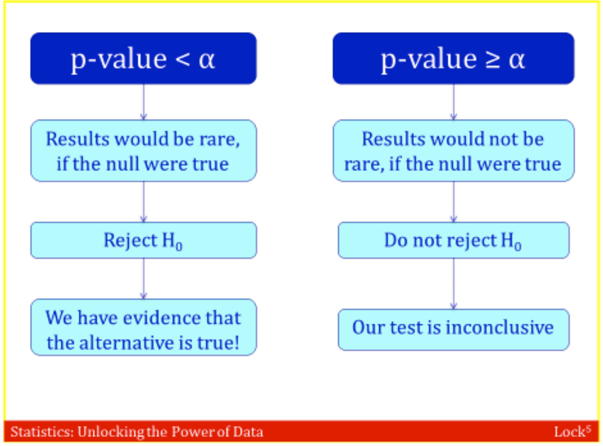
Significance Level

- The ***significance level***, α, is the threshold below which the p-value is deemed small enough to reject the null hypothesis

$$\boxed{\begin{array}{lll} \text{p-value} < \alpha & \Rightarrow & \text{Reject } H_0 \\ \text{p-value} \geq \alpha & \Rightarrow & \text{Do not Reject } H_0 \end{array}}$$

- Often α = 0.05, unless otherwise specified
  - (Why 0.05?)

Statistics: Unlocking the Power of Data Lock5



p-value < α

Results would be rare, if the null were true

Reject H_0

We have evidence that the alternative is true!

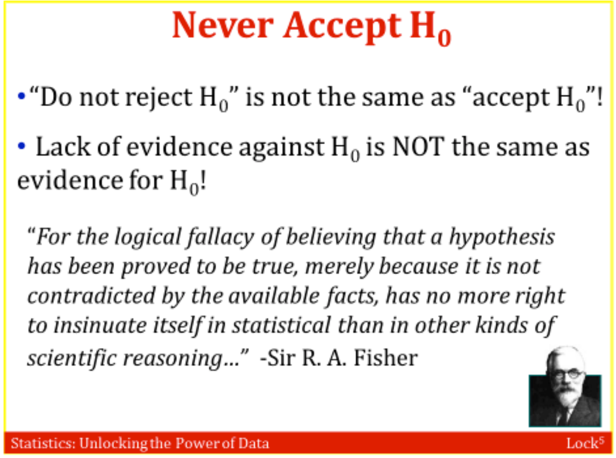
p-value ≥ α

Results would not be rare, if the null were true

Do not reject H_0

Our test is inconclusive

Statistics: Unlocking the Power of Data Lock[5]

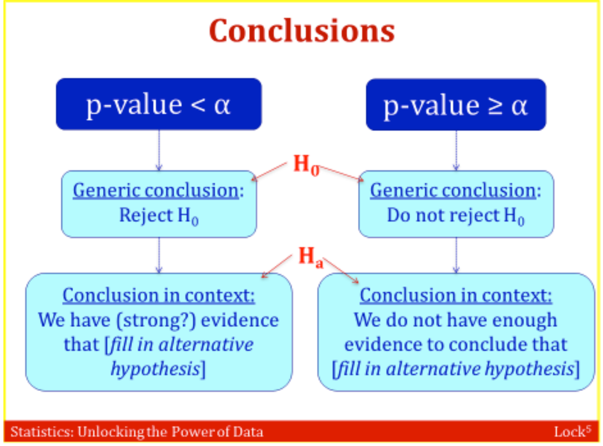


Never Accept H_0

- "Do not reject H_0" is not the same as "accept H_0"!
- Lack of evidence against H_0 is NOT the same as evidence for H_0!

> *"For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in other kinds of scientific reasoning…"* -Sir R. A. Fisher

Statistics: Unlocking the Power of Data Lock[5]



Conclusions

p-value < α

Generic conclusion: Reject H_0

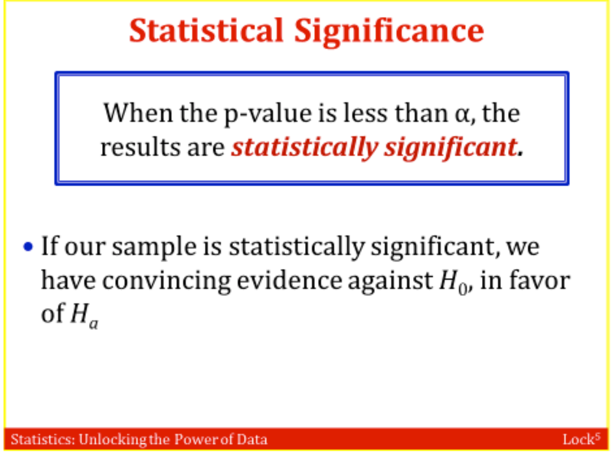
Conclusion in context: We have (strong?) evidence that [*fill in alternative hypothesis*]

p-value ≥ α

Generic conclusion: Do not reject H_0

Conclusion in context: We do not have enough evidence to conclude that [*fill in alternative hypothesis*]

Statistics: Unlocking the Power of Data Lock[5]



Statistical Significance

When the p-value is less than α, the results are ***statistically significant.***

- If our sample is statistically significant, we have convincing evidence against H_0, in favor of H_a

Statistics: Unlocking the Power of Data Lock[5]

Example: Red Wine and Weight Loss

Resveratrol, an ingredient in red wine and grapes, has been shown to promote weight loss in animals. In one study, a sample of lemurs had various measurements taken before and after receiving resveratrol supplements for 4 weeks. For each p-value given, indicate the formal generic conclusion as well as a conclusion in context. Use a 5% significance level.

(a) In the test to see if the mean resting metabolic rate is higher after treatment, the p-value is 0.013. *Reject H_0. There is evidence that metabolism is higher after receiving resveratrol.*

(b) In the test to see if the mean body mass is lower after treatment, the p-value is 0.007.
Reject H_0: There is strong evidence that body mass is lower after receiving resveratrol.

(c) In the test to see if locomotor activity changes after treatment, the p-value is 0.980.
Do not reject H_0. The data does not provide any evidence that resveratrol affects activity level.

(d) In the test to see if mean food intake changes after treatment, the p-value is 0.035.
Reject H_0: There is evidence that food intake is different after treatment (receiving resveratrol).

(e) Which of the results given in (a) - (d) above are significant at a 1% level?
Only the result in (b) on body mass. That p-value of 0.007 is very small and is significant at the 1% level

Quick Self-Quiz: Making Conclusions

1. In a hypothesis test of H_0: $\mu = 18$ vs H_a: $\mu > 18$, we obtain a p-value of 0.016. Using $\alpha = 0.05$, we conclude:

a). Reject H_0 b). Do not reject H_0 c). Reject H_a d). Do not reject H_a

Point out that options (c) and (d) are never viable options.

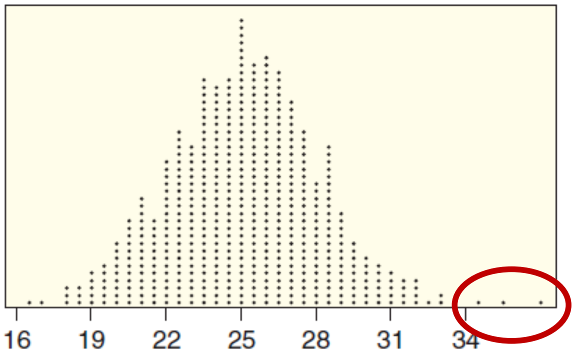
2. In a hypothesis test of H_0: $\mu = 18$ vs H_a: $\mu > 18$, we obtain a p-value of 0.016. Using α = 0.05, we conclude:

a). There is evidence that $\mu = 18$ b). There is evidence that $\mu > 18$ c). There is no evidence of anything

Point out that (a) is never a viable option.

Example 3: Sugar in Bottled Iced Tea

The nutrition label on a brand of iced tea says that the average amount of sugar per bottle is 25 grams. A chemical analysis of a sample of 30 bottles finds a mean of 33.8 grams of sugar per bottle. Test to see if this provides significant evidence that the true average is greater than 25. A randomization distribution for the test is shown, showing 1000 randomization statistics. [*Show all details: state hypotheses, give notation and value of the sample statistic, use the randomization distribution to estimate the p-value, give a formal conclusion at a 5% level, and give a conclusion in context.*]



H_0: μ = 25 where μ is mean grams of sugar for all bottles

H_a: μ > 25

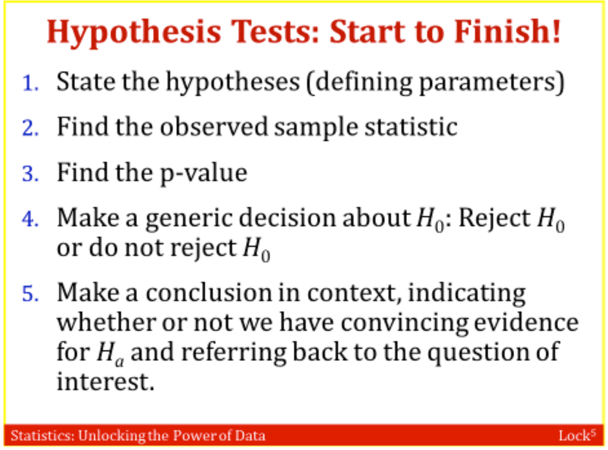
Statistic: $\bar{x}$ = 33.8

p-value is proportion of statistics to the right of 33.8, which appears to be 3/1000 = 0.003. We estimate p-value = 0.003.

(This is just an estimate, but students should know the p-value is small!)

Formal conclusion: Reject H_0

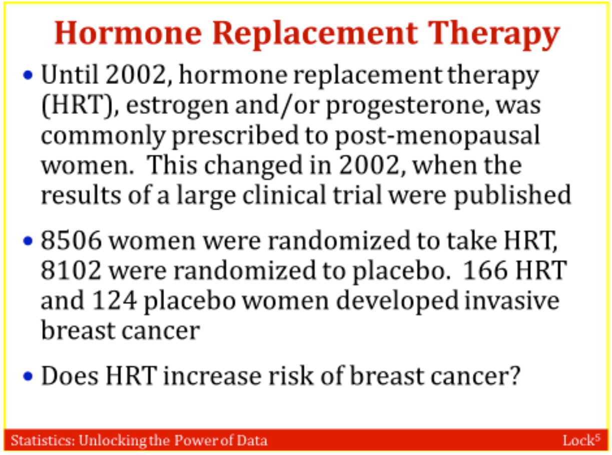
Conclusion in context: There is strong evidence that the mean number of grams of sugar in bottles of this iced tea is greater than 25



Hypothesis Tests: Start to Finish!

1. State the hypotheses (defining parameters)
2. Find the observed sample statistic
3. Find the p-value
4. Make a generic decision about H_0: Reject H_0 or do not reject H_0
5. Make a conclusion in context, indicating whether or not we have convincing evidence for H_a and referring back to the question of interest.

Statistics: Unlocking the Power of Data Lock[5]



Hormone Replacement Therapy

- Until 2002, hormone replacement therapy (HRT), estrogen and/or progesterone, was commonly prescribed to post-menopausal women. This changed in 2002, when the results of a large clinical trial were published
- 8506 women were randomized to take HRT, 8102 were randomized to placebo. 166 HRT and 124 placebo women developed invasive breast cancer
- Does HRT increase risk of breast cancer?

Statistics: Unlocking the Power of Data Lock[5]

Step 1: State Hypotheses

- Does HRT increase risk of breast cancer?

p_1 = proportion of women taking HRT who get invasive breast cancer
p_2 = proportion of women not taking HRT who get invasive breast cancer

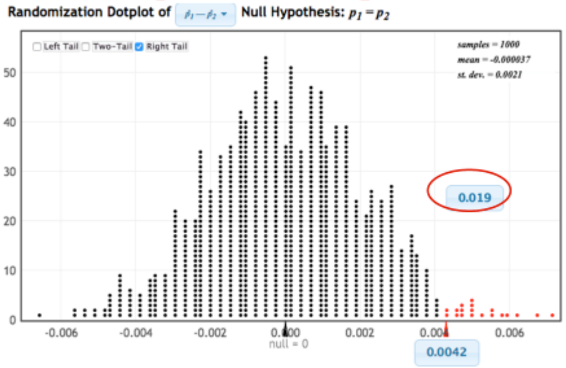
H_0: $p_1 = p_2$
H_a: $p_1 > p_2$

Step 2: Calculate Sample Statistic

- Does HRT increase risk of breast cancer?
  - ○ H_0: $p_1 = p_2$; H_a: $p_1 > p_2$
  - ○ HRT group: 166 of 8506 developed breast cancer
  - ○ Placebo group: 124 of 8102 developed breast cancer

$$\hat{p}_1 - \hat{p}_2 = \frac{166}{8506} - \frac{124}{8102} = 0.0042$$

Step 3: Find the p-value



Step 4: Make a Generic Conclusion

- Does HRT increase risk of breast cancer?
  - ○ H_0: $p_1 = p_2$; H_a: $p_1 > p_2$
  - ○ $\hat{p}_1 - \hat{p}_2 = 0.0042$
  - ○ p-value = 0.019

Using $\alpha = 0.05$:

Reject H_0

Step 5: Make a Conclusion in Context

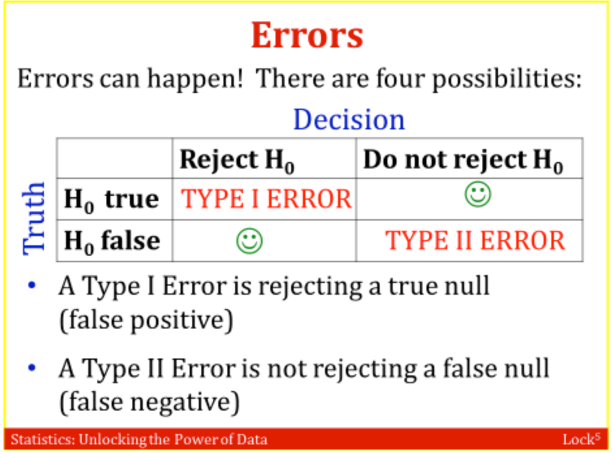
- Does HRT increase risk of breast cancer?
  - ○ H_0: $p_1 = p_2$; H_a: $p_1 > p_2$
  - ○ $\hat{p}_1 - \hat{p}_2 = 0.0042$
  - ○ p-value = 0.019
  - ○ Reject H_0

We have convincing evidence that taking hormone replacement therapy does increase risk of breast cancer.

(Because of this result, the trial was terminated early and HRT is no longer routinely recommended).

4.4: **A Closer Look at Testing**

- Statistics (unlike mathematics) is not an exact science! We will be wrong sometimes in a statistical test
- Conclusions based off p-values are not perfect
- Type I and Type II errors can happen



Errors

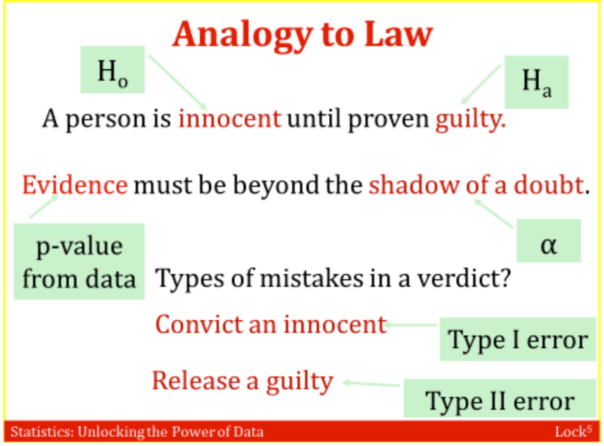
Errors can happen! There are four possibilities:

Decision

		Reject H_0	Do not reject H_0
Truth	H_0 true	TYPE I ERROR	☺
	H_0 false	☺	TYPE II ERROR

- A Type I Error is rejecting a true null (false positive)
- A Type II Error is not rejecting a false null (false negative)

Statistics: Unlocking the Power of Data Lock^5



Analogy to Law

H_0 H_a

A person is innocent until proven guilty.

Evidence must be beyond the shadow of a doubt.

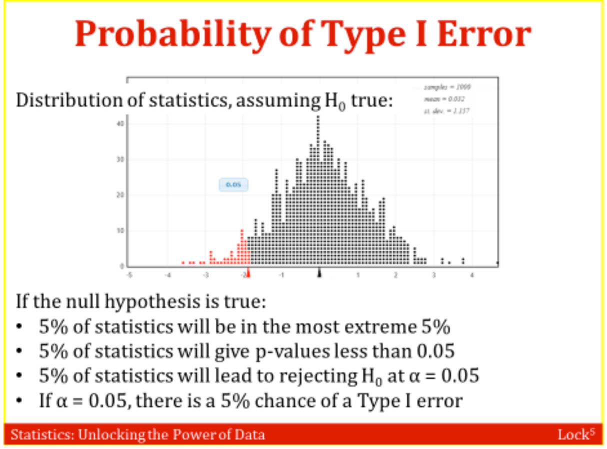
p-value from data α

Types of mistakes in a verdict?

Convict an innocent — Type I error

Release a guilty — Type II error

Statistics: Unlocking the Power of Data Lock^5



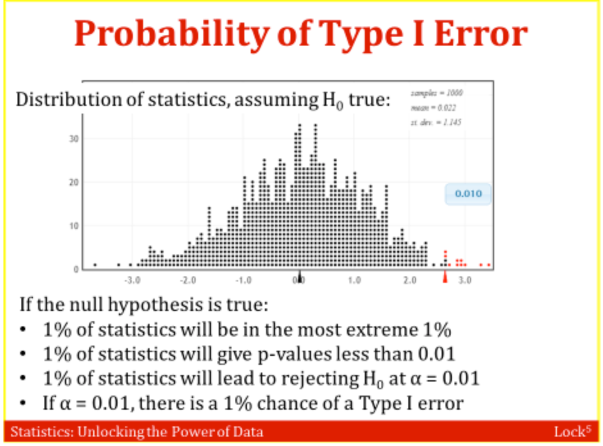
Probability of Type I Error

Distribution of statistics, assuming H_0 true:

If the null hypothesis is true:

- 5% of statistics will be in the most extreme 5%
- 5% of statistics will give p-values less than 0.05
- 5% of statistics will lead to rejecting H_0 at $\alpha = 0.05$
- If $\alpha = 0.05$, there is a 5% chance of a Type I error

Statistics: Unlocking the Power of Data Lock^5



Probability of Type I Error

Distribution of statistics, assuming H_0 true:

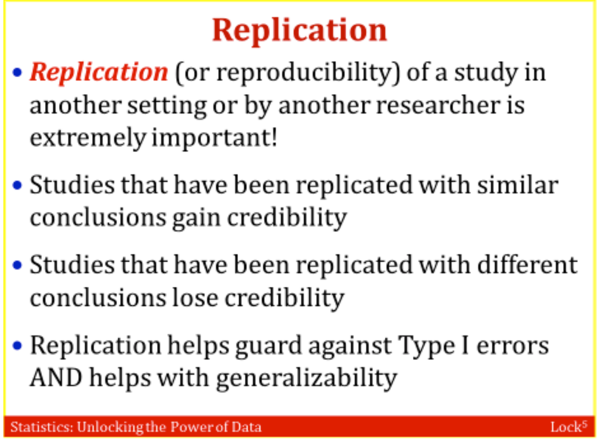
0.010
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0

If the null hypothesis is true:

- 1% of statistics will be in the most extreme 1%
- 1% of statistics will give p-values less than 0.01
- 1% of statistics will lead to rejecting H_0 at $\alpha = 0.01$
- If $\alpha = 0.01$, there is a 1% chance of a Type I error

Statistics: Unlocking the Power of Data Lock^5

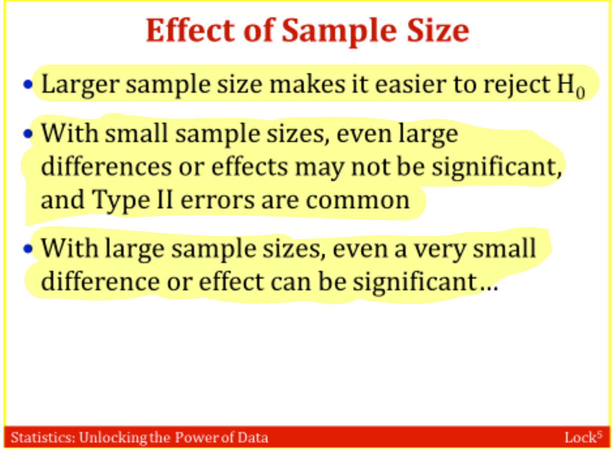
- **The probability of making a Type I error (rejecting a true null) is the significance level, , α**



Replication

- ***Replication*** (or reproducibility) of a study in another setting or by another researcher is extremely important!
- Studies that have been replicated with similar conclusions gain credibility
- Studies that have been replicated with different conclusions lose credibility
- Replication helps guard against Type I errors AND helps with generalizability

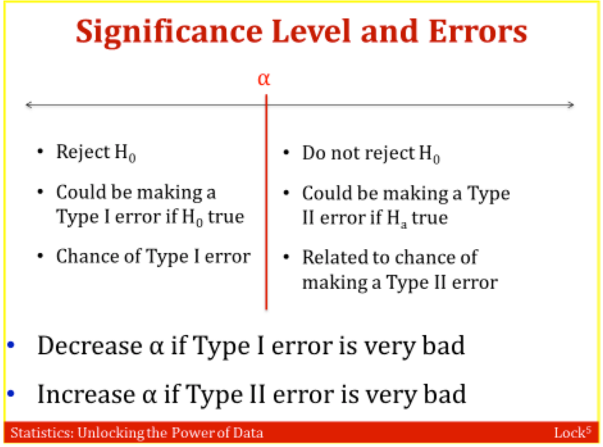
Statistics: Unlocking the Power of Data Lock^5



Effect of Sample Size

- Larger sample size makes it easier to reject H_0
- With small sample sizes, even large differences or effects may not be significant, and Type II errors are common
- With large sample sizes, even a very small difference or effect can be significant…

Statistics: Unlocking the Power of Data Lock^5



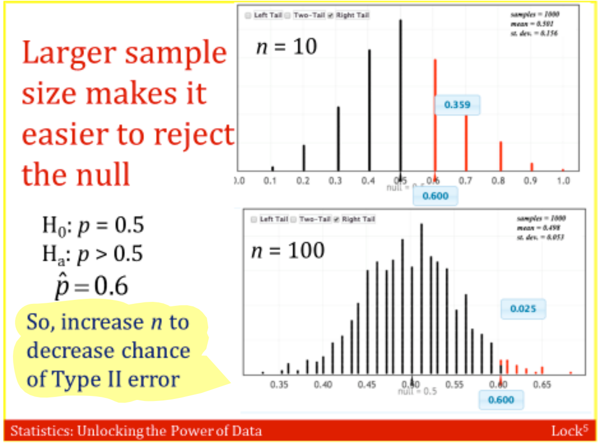
Significance Level and Errors

- Reject H_0
- Could be making a Type I error if H_0 true
- Chance of Type I error

- Do not reject H_0
- Could be making a Type II error if H_a true
- Related to chance of making a Type II error

- Decrease α if Type I error is very bad
- Increase α if Type II error is very bad

Statistics: Unlocking the Power of Data Lock5



Larger sample size makes it easier to reject the null

H_0: $p = 0.5$
H_a: $p > 0.5$
$\hat{p} = 0.6$

So, increase *n* to decrease chance of Type II error

Statistics: Unlocking the Power of Data Lock5

- If a Type I error (rejecting a true null) is much worse than a Type II error, In another words, if we don't want to reject H_0, we may choose a smaller α, like α = 0.01
- If a Type II error (not rejecting a false null) is much worse than a Type I error, In another words, if we want to reject H_0, we may choose a larger α, like α = 0.10(or use a larger sample size).

Example : BPA in Tomato Soup

A consumer protection agency is testing a sample of cans of tomato soup from a company. If they find evidence that the average level of the chemical bisphenol A (BPA) in tomato soup from this company is greater than 100 ppb (parts per billion), they will recall all the soup and sue the company.

a). State the null and alternative hypotheses.

This is a test for a single mean. The hypotheses are H_0: $\mu = 100$ vs H_a: $\mu > 100$

b). what does a Type I error mean in this situation?

A Type I error means the company's mean is within normal bounds of 100 (the null hypothesis is true) but the sample obtained happens to show(incorrectly) that the mean is too high and the agency ends up recalling all the soup and suing the company when it shouldn't have.

c). what does a Type II error mean in this situation?

A Type II error means the company's mean is too high (the null hypothesis is false) but the sample obtained doesn't give sufficient evidence to show that it is too high and the agency (incorrectly) decides not to recall the soup or sue the company.

d). which is more serious, a Type I error or a Type II error?

(There is no right answer to this one. It is a matter of opinion and one could argue either way.)

Both seem pretty serious so you really want to try to not make an error. (Good time to remind them of the benefits of a larger sample size!)

Example : Vitamin E and Heart Attacks?

Suppose 100 tests are conducted to determine whether taking vitamin E increases one's chances of having a heart attack. Suppose also that vitamin E has absolutely *no* effect on one's likelihood of having a heart attack. The tests will use a 5% significance level.

(a) How many of the tests are likely to show significance, just by random chance?

5% of the 100 tests, or 0.05(100) = 5 tests

(Remember that the significance level gives the probability of making a Type 1 error, so about 5% of the 100 tests will make a Type I error. In this case, that means showing significance when there really is nothing significant.)

(b) If only the significant tests are reported, what is the only information the public is likely to hear?

The public will hear the false information that vitamin E causes heart attacks! Emphasize that all significant tests should be replicated in further tests before we are confident in the results.

Quick Self-Quiz: Experimenting with Sample Size on StatKey

Suppose that we are testing a coin to see if it is fair, so our hypotheses are H_0: $p = 0.5$ vs H_a: $p \neq 0.5$. In each of (a) and (b) below, use the "Edit Data" option on StatKey to find the p-value for the sample results and give a conclusion in the test.

(a) We get 56 heads out of 100 tosses.

The p-value is about 0.28. An outcome of 56 heads in 100 tosses is relatively likely to happen by random chance, and we do not have evidence that the coin is not fair.

(b) We get 560 heads out of 1000 tosses.

The p-value is very small, close to zero. An outcome of 560 heads in 1000 tosses is very unlikely to happen just by random chance with a fair coin, so we have strong evidence that the coin is not fair.

(c) Compare the sample proportions in parts (a) and (b). Compare the p-values. Why are the p-values so different?

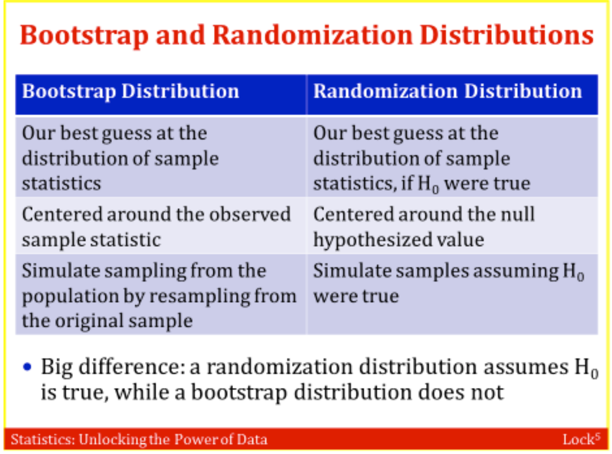
The sample proportions are the same, 0.56 in both (a) and (b).
The p-values are very different: 0.28 (not at all significant) to 0.000 (very significant!)
*The difference is due to the **sample size**. Sample size is very important in statistics, and a larger sample size can help us find significant results, such as a biased coin, if the coin really is biased.*

4.5: Making Connections

In Chapter 3 we examine methods to construct confidence intervals for population parameters. We sample (with replacement) from the original sample to create a *bootstrap distribution* of possible values for a sample statistic. Based on this distribution, we produce a range of plausible values for the parameter so that we have some degree of certainty that the interval will capture the actual parameter value for the population.

In this chapter, we develop methods to test claims about populations. After specifying null and alternative hypotheses, we assess the evidence in a sample by constructing a *randomization distribution* of possible sample statistics that we might see by random chance, if the null hypothesis were true. If the original sample statistic falls in an unlikely location of the randomization distribution, we have evidence to reject the null hypothesis in favor of the alternative.

In Chapter 3, we see that a confidence interval shows us the plausible values of the population parameter. In Chapter 4, we use a hypothesis test to determine whether a given parameter in a null hypothesis is plausible or not. Thus, we can use a confidence interval to make a decision in a hypothesis test, and we can use a hypothesis test to determine whether a given value will be inside a confidence interval!

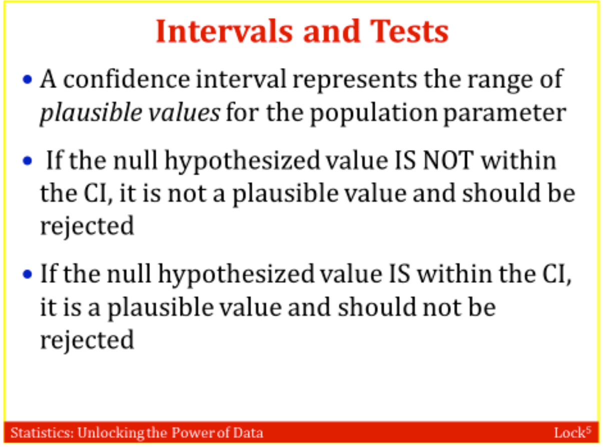


Bootstrap and Randomization Distributions

Bootstrap Distribution	Randomization Distribution
Our best guess at the distribution of sample statistics	Our best guess at the distribution of sample statistics, if H_0 were true
Centered around the observed sample statistic	Centered around the null hypothesized value
Simulate sampling from the population by resampling from the original sample	Simulate samples assuming H_0 were true

- Big difference: a randomization distribution assumes H_0 is true, while a bootstrap distribution does not

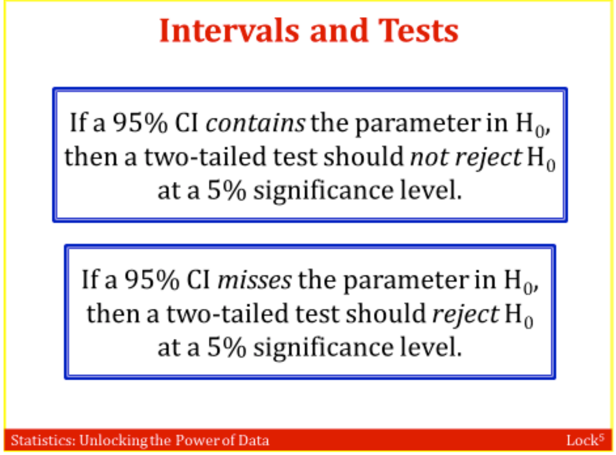
Statistics: Unlocking the Power of Data Lock[5]



Intervals and Tests

- A confidence interval represents the range of *plausible values* for the population parameter
- If the null hypothesized value IS NOT within the CI, it is not a plausible value and should be rejected
- If the null hypothesized value IS within the CI, it is a plausible value and should not be rejected

Statistics: Unlocking the Power of Data Lock[5]

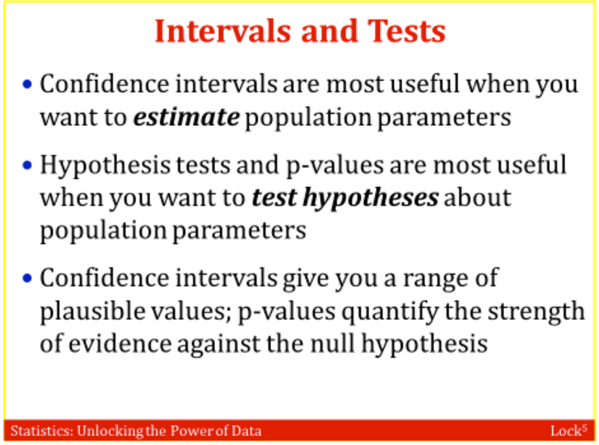


Intervals and Tests

If a 95% CI *contains* the parameter in H_0, then a two-tailed test should *not reject* H_0 at a 5% significance level.

If a 95% CI *misses* the parameter in H_0, then a two-tailed test should *reject* H_0 at a 5% significance level.

Statistics: Unlocking the Power of Data Lock[5]



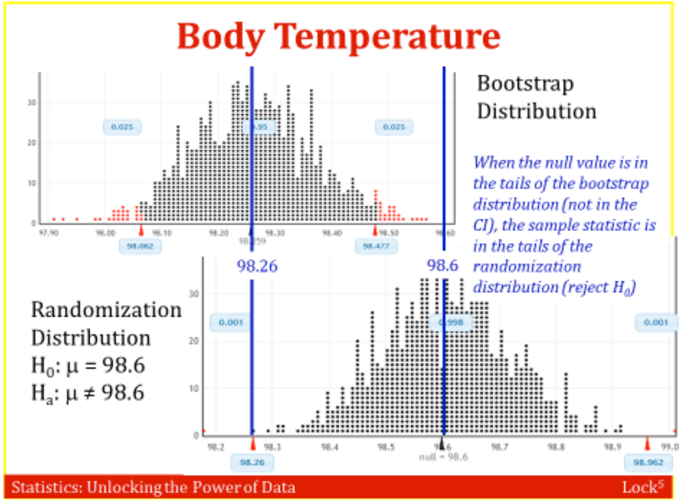
Intervals and Tests

- Confidence intervals are most useful when you want to ***estimate*** population parameters
- Hypothesis tests and p-values are most useful when you want to ***test hypotheses*** about population parameters
- Confidence intervals give you a range of plausible values; p-values quantify the strength of evidence against the null hypothesis

Statistics: Unlocking the Power of Data Lock[5]

Example 1: Normal Human Body Temperature

We find a 95% confidence interval for mean body temperature μ to be 98.05 to 98.47. What is the conclusion of a test of $H_0: \mu = 98.6$ vs $H_a: \mu \neq 98.6\$$? What significance level is used in making the conclusion?



Body Temperature

Bootstrap Distribution

When the null value is in the tails of the bootstrap distribution (not in the CI), the sample statistic is in the tails of the randomization distribution (reject H_0)

Randomization Distribution
$H_0: \mu = 98.6$
$H_a: \mu \neq 98.6$

Statistics: Unlocking the Power of Data

Lock5

The value 98.6 is not inside the confidence interval, so 98.6 is not a plausible value for μ and we reject H_0. There is evidence that mean body temperature is not 98.6°F. The significance level used is 5%, since the confidence level used was 95% for the interval.

Example 2: Happy Family?

The Pew Research Center asked a random sample of US adults age 18 to 29 ``Does a child need both a father and a mother to grow up happily?” A 95% confidence interval is given below for p, the proportion of all US adults age 18 to 29 who say yes. Use the interval to state the conclusion to a hypothesis test of $H_0: p = 0.5$ vs $H_a: p \neq 0.5$.

(a) In 2010, the 95% confidence interval was 0.487 to 0.573.

Since 0.5 is in the confidence interval 0.487 to 0.573, and thus is a plausible value for p, we do not have evidence against the null hypothesis so we do not reject H_0. At a 5% level, we do not have evidence in 2010 that the proportion is different from 0.5.

(b) In 1997, the 95% confidence interval was 0.533 to 0.607.

Since 0.5 is not in the confidence interval 0.533 to 0.607, and thus is not a plausible value for p, we do have evidence against the null hypothesis, so we reject H_0. At a 5% level, we have evidence that the proportion in 1997 is different from 0.5.

Quick Self-Quiz: Intervals and Tests

Using the confidence interval given, indicate the conclusion of the test and indicate the significance level used.

(a) A 95% confidence interval for a mean μ is 12.5 to 17.1. Testing $H_0: \mu = 18$ vs $H_a: \mu \neq 18$.

18 is outside the interval so is not a plausible value for μ so we ***reject H_0*** *at a 5% level.*

(b) A 90% confidence interval for a proportion p is 0.62 to 0.80. Testing $H_0: p = 0.65$ vs $H_a: p \neq 0.65$.

*0.65 in inside the interval so is a plausible value for p so we **do not reject H_0** at a 10% level.*

(c) A 99% confidence interval for a difference in proportions is – 0.10 to 0.20. Testing H_0: $p_1 = p_2$ vs H_a: $p_1 \neq p_2$.

*A difference of 0 (no difference in proportions) is inside the interval, so is a plausible value for $p_1 – p_2$ so we **do not reject H_0** at a 1% level.*

Example 3: Normal Human Body Temperature, Revisited

Normal human body temperature is generally considered to be 98.6°F. We wish to test to see if there is evidence that mean body temperature is different from 98.6°F. We collect data from a random sample of 50 people and find $\bar{x}$ = 98.26.

a). State the null and alternative hypotheses.

H_0: $\mu = 98.6$ vs H_a: $\mu \neq 98.6$

b). Use *StatKey* to create a randomization distribution for this test, and then use it to find the p-value. Use the p-value to make a conclusion in the test.

*Notice that the distribution is centered at 98.6 as it should be. We see how extreme the sample statistic of 98.26 is in the tail of the randomization distribution **and** we remember to double it since this is a two-tail test. We see that the p-value is very small, so even doubling it, we still get a p-value very close to zero. There is very strong evidence that average human body temperature is not 98.6℉.*

Example 4: Evaluating Drugs to Fight Cocaine Addiction

In a randomized experiment on treating cocaine addiction, 48 cocaine addicts who were trying to quit were randomly assigned to take either desipramine (a new drug), or Lithium (an existing drug). The response variable is whether or not the person relapsed (which means the person was unable to break out of the cycle of addiction and returned to using cocaine.) We are testing to see if desipramine is better than lithium at treating cocaine addiction. The results are shown in the two-way table.

	Relapse	No relapse	Total
Desipramine	10	14	24
Lithium	18	6	24
Total	28	20	48

(a) Using p_D for the proportion of desipramine users who relapse and p_L for the proportion of lithium users who relapse, write the null and alternative hypotheses. [Example continued on reverse.]

H_0: $p_D = p_L$ vs H_a: $p_D < p_L$

(b) Compute the appropriate sample statistic.

We see that $\hat{p}_D = \frac{10}{24} = 0.417$ *and* $\hat{p}_L = \frac{18}{24} = 0.75$ *so we have*

$\hat{p}_D - \hat{p}_L = 0.417 – 0.75 = -0.333$

(c) We compute a randomization statistic by assuming **the null hypothesis is true.** What does that mean in this case?

It means that the two proportions are equal and the drug has no effect on the relapse rate. It doesn't matter what drug is taken.

(d) How might we compute a randomization sample for this data? What statistic would we compute as the randomization statistic?

Since drug doesn't matter, we combine all 48 patients together and see that 28 relapsed and 20 didn't. To see what happens by random chance, we randomly divide them into two groups and compute the difference in proportions of relapses between the two groups. The difference in proportions is the statistic.

(e) We can use StatKey to generate a randomization dotplot for the difference in proportions based on this sample and what we might see by random chance of the null hypothesis is true. Describe the resulting distribution. Where is it centered?

The resulting distribution will be bell-shaped and centered at the null hypothesis value, which is zero.

(f) How extreme is the sample statistic from part (b) in the randomization distribution? This tells us how unlikely the sample data is if the null hypothesis is true (which is the p-value!) Use the sample statistic calculated in (b) to find the p-value for this test. Use the p-value to make a conclusion.

This is a left-tail test, and we see on StatKey that the p-value is about 0.016. We reject the null hypothesis and conclude that despramine is significantly better at helping people kick the cocaine habit.