CSCI 3022-002 Intro to Data Science
Hypotheses and P-Values

Example:

Company A produces circuit boards, but ~~10%~~ *baseline* of them are defective. Company B ~~claims that~~ they produce ~~fewer~~ defective circuit boards. What are the null and alternative hypotheses?

iterative

null H_0: proportion of boards from B is 10%

$$P = .1 \qquad (P = P_0) \qquad P_0 := \text{hypothesized value}$$

alternative H_a: proportion of boards from B is <u>less</u> than 10%

$$\text{testing } \ p < .1 \qquad (p < P_0)$$

Our data is a random sample of $n = 200$ boards from company B. What test procedure (or rule) could we devise to decide if the null hypothesis should be rejected?

Opening sol:

Example:
Company A produces circuit boards, but 10% of them are defective. Company B claims that they produce fewer defective circuit boards. What are the null and alternative hypotheses?

Our data is a random sample of $n = 200$ boards from company B. What test procedure (or rule) could we devise to decide if the null hypothesis should be rejected?

Opening sol:

Example:
Company A produces circuit boards, but 10% of them are defective. Company B claims that they produce fewer defective circuit boards. What are the null and alternative hypotheses?
Solution: Null: company B and A are equal. Alternative: company B produces fewer defective boards.
Our data is a random sample of $n = 200$ boards from company B. What test procedure (or rule) could we devise to decide if the null hypothesis should be rejected?
Solution: Reject if we observe a sample from the company B with a **very low** proportion of defects.

how low? Depends on: 1) error tolerance
2) sample size

Announcements and Reminders

- Short HW for next week. *Posted tonish t!*

- NB day Friday.

Logic of Hypothesis Testing

Notation and general process:

Logic of Hypothesis Testing

Notation and general process:

1. Assume the null hypothesis to be true, and state it: we propose that the parameter of interest θ satisfies $H_0 : \theta = \theta_0$.

θ: could be pop mean μ
" " " proportion p.

Could just be parameter: λ from exponential
Poisson

Logic of Hypothesis Testing

Notation and general process:

1. Assume the null hypothesis to be true, and state it: we propose that the parameter of interest θ satisfies $H_0 : \theta = \theta_0$.

2. State the alternative to be tested: H_a :
   $\theta > \theta_0$ **OR** $\theta < \theta_0$ **OR** $\theta \neq \theta_0$
   $\underbrace{\qquad}_{\text{greater}}$ $\underbrace{\qquad}_{\text{less}}$ $\underbrace{\qquad}_{\text{"far from"}}$

3. Draw a decision based on how improbable or probable the actual data looks if the null hypothesis is true. If the observed data is very unlikely, it might be because our hypothesis was wrong!

Why *assume* the null hypothesis?

Logic of Hypothesis Testing

Notation and general process:

1. Assume the null hypothesis to be true, and state it: we propose that the parameter of interest θ satisfies $H_0 : \theta = \theta_0$.

Why *assume* the null hypothesis?

1. Burden of proof

2. We know how to calculate probabilities when we *know* θ!

we know $\exp(2)$ $\rightarrow$ p

$\text{bin}\,((100, \cdot 6)$

Test Statistics: The Evidence

"given $H_0 =$ true"

Definition: *Test Statistic*

A test statistic is a a quantity derived based on <u>sample data</u> and calculated <u>under the null hypothesis</u>. It is used in a decision about whether to reject H_0.

→ from data!

We can think of a test statistic as our evidence. Next, we need to quantify whether we think our evidence is "rare" under the null hypothesis.

Ex: $H_0: p = .1$

data: 15 defective out of $200 = \frac{15}{200} < 10\%$

Prob of exactly $\frac{15}{200}$ if true value was $p \geq .1$? versus $\frac{19}{200}$

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

1. State hypothesis: H_0 : fair coin, or $p = .5$.   "assume coin is fair"
   H_a : unfair coin, or $p \neq .5$

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

1. State hypothesis: H_0 : fair coin, or $p = .5$.
   H_a : unfair coin, or $p \neq .5$

2. Get to flippin', collect some data

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

1. State hypothesis: H_0 : fair coin, or $p = .5$.
   H_a : unfair coin, or $p \neq .5$

2. Get to flippin', collect some data

3. Compute something from our data. Maybe a sample proportion of heads $\hat{p}$?

Sample = heads!

Test Statistics: The Evidence

Back to our Belgian Euro: how would you decide whether it was fair?

1. State hypothesis: H_0 : fair coin, or $p = .5$.
   H_a : unfair coin, or $p \neq .5$

2. Get to flippin', collect some data

3. Compute something from our data. Maybe a sample proportion of heads $\hat{p}$?

4. Decide whether $\hat{p}$ is **too far** from $p = .5$, and make a decision accordingly.

Test Statistics: The Evidence

Which test statistic is "best"?

There are an infinite number of possible tests that could be devised, so we have to limit this in some way or total statistical madness will ensue!

if we can, use sample means to estimate

In the previous example, we might use $\hat{p}$.   *pop. means*

why? C.L.T

Rejection Regions

How would we know when the test statistic is "sufficiently rare" under the null hypothesis such that we might regard the null as false?

When unfair of a sample

⤷ Fair Coin

We could define a **rejection region**: a range of values of the test statistic that leads a researcher to reject the null hypothesis.

⤷ how large
" small
" . far from baseline

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

▶ What would 10 heads mean?

$$P(10 \text{ heads} \mid \text{fair coin})$$

$$= P(X=10 \mid H_0) = \left(\tfrac{1}{2}\right)^{10}$$

pdf of a binomial $(n=10, p=0.5)$

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\binom{10}{10} p^{10} = \left(\tfrac{1}{2}\right)^{10}$$

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

- What would 10 tails mean? $P(10 \text{ tails} \mid \text{fair coin}) = \left(\frac{1}{2}\right)^{10} = (1-p)^{10}$

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

▶ What would 6 heads mean?

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

▶ Is there a difference between 60% heads if we flip 10 times and 60% heads if we flip 1000 times?

5 vs. 6.

C.L.T. : $\overline{X}$ → pop average $\overline{X}$ → np ≈ 500!

What is extreme: let's compute these!

The probability of a fixed number of H in a fixed number of trials are values of the binomial pmf.

- $P(10\,heads|fair\,coin) = \frac{1}{2^{10}}$.
- $P(10\,tails|fair\,coin) = \frac{1}{2^{10}}$.
- $P(10\,or\,more\,of\,the\,same\,type\,of\,flip|fair\,coin) = \frac{2}{2^{10}}$ *2-sided* $\frac{1}{2^{10}} + \frac{1}{2^{10}}$
- $P(9\,or\,more\,heads|fair\,coin) = \frac{1}{2^{10}} + \frac{10}{2^{10}}$ $\to P(10\,heads) + \to P(9\,heads) = \frac{4}{2^{10}}$
- $P(9\,or\,more\,tails|fair\,coin) = \frac{1}{2^{10}} + \frac{10}{2^{10}}$
- $P(9\,or\,more\,of\,the\,same\,type\,of\,flip|fair\,coin) = \frac{22}{2^{10}}$ $\approx \frac{22}{1000} \approx .02$

At some point, we decide that the probabilities are so unlikely (*given* the coin is fair), that if we *do* observe those outcomes, we conclude that the coin probably wasn't fair!

Bring back α!

"Which outcomes are too far from baseline to be likely to originate from baseline"

$\alpha = Prob(\text{those extreme outcomes})$

Definition: The **Significance level** α of a hypothesis test is the largest *probability* of a test statistic under the null hypothesis that would lead you to reject the null hypothesis.

Equivalently, it's the probability of the entire rejection region!

We thought of α last week during CIs as a term that widened or shrank as our tolerance for error grew, now it's very literally an *error rate*. Specifically, it's the probability of rejecting the null hypothesis when we were not supposed to do so.

Where we at?
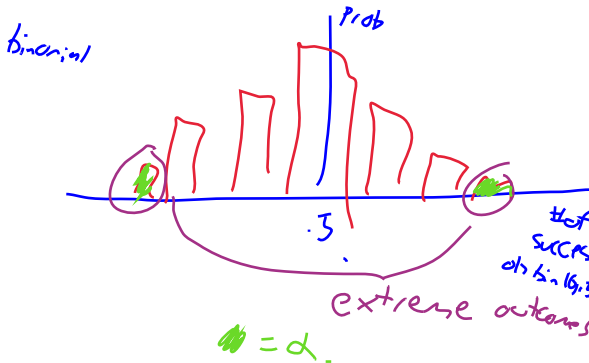
A summary of our process:

1. State hypothesis: H_0 : fair coin, or $p = .5$.
   H_a : unfair coin, or $p \neq .5$

2. Get to flippin', collect some data

3. Compute something from our data. Maybe a sample proportion of heads $\hat{p}$?

4. Decide whether $\hat{p}$ is **too far** from $p = .5$, and make a decision accordingly.

5. α is the value that describes the probability of rejecting a null hypotheses *given* that the hypothesis was true.

Rejection Regions

How would we know when the test statistic is "sufficiently rare" under the null hypothesis such that we might regard the null as false?

We could define a rejection region: a range of values that leads a researcher to reject the null hypothesis.

Rejection Regions

"we can be wrong a proportion of the time"

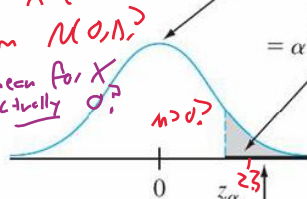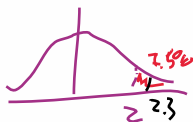*z* curve (probability distribution of test statistic Z when H_0 is true)

draw

$N(0,1)$.

we get

$X = 2.3$

did X come from $N(0,1)$?
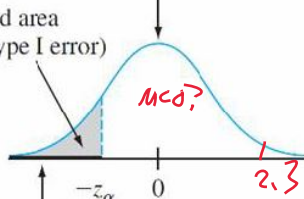
H_0: mean for X actually 0?

Shaded area
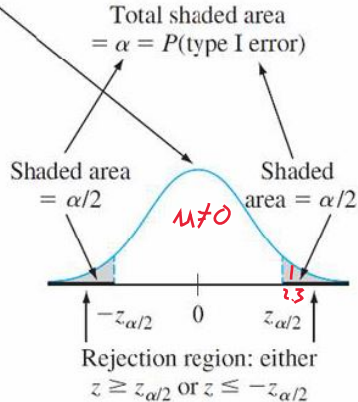$= \alpha = P(\text{type I error})$

$\mu > 0$?

Rejection region: $z \geq z_\alpha$

Rejection region: $z \leq -z_\alpha$

$\mu < 0$?

Total shaded area
$= \alpha = P(\text{type I error})$

Shaded area
$= \alpha/2$

$\mu \neq 0$

Shaded
area $= \alpha/2$

Rejection region: either
$z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$

Test for Population Proportion

Example To determine if the Belgian 1 Euro coin is fair, you flip it 100 times and record the number of heads. What is the test statistic? What are the null and alternative hypotheses?

Stat : $\hat{p}$ (we like proportions!)

H_0: $p = .5$ (fair!)
H_A: $p \neq .5$ (unfair!)

What is the distribution of the number of heads?

Assume: i.i.d, we also assume $P = P_0 = .5$.

$X \sim bin(100, .5)$

Dist of $\hat{p}$? Thats $\overline{X}$ for a sample of 100 Bernoulli's.

Test for Population Proportion

Example To determine if the Belgian 1 Euro coin is fair, you flip it 100 times and record the number of heads. What is the test statistic? What are the null and alternative hypotheses?

Test estimator: based on sample *proportion* of heads, $\hat{p}$

Test: $H_0 : p = .5$; fair coin

Against: $H_a : p \neq .5$; an unfair coin

What is the distribution of the number of heads?

$binomial(n, p)$.

We know everything about that thing! In particular, we know that

$$\hat{p} \overset{approx}{\sim} N\left(.5, \frac{.5(1 - .5)}{100}\right)$$

CLT.

Mean = fair

Var of binom:
$n p(1-p)$

Variance of binom/n
$\frac{p(1-p)}{n}$

Test for Population Proportion

Example To determine if the Belgian 1 Euro coin is fair, you flip it 100 times and record the number of heads. What is the test statistic? What are the null and alternative hypotheses?

What is the distribution of the number of heads?
$binomial(n, p)$.
We know everything about that thing! In particular, we know that

$$\hat{p} \overset{approx}{\sim} N\left(.5, \frac{.5(1-.5)}{100}\right)$$

$\hat{p} - .5 / \text{sd.dev.}$

$\hat{p}$ is a sample proportion.

We can use this to compute the probability given by $P(\hat{p} > \boxed{z_{\alpha/2}})$ critical value

Test for Population Proportion

Example (continued):

Null hypothesis: Assume baseline proportion: (p_0)

$$H_0: p = p_0$$
$$H_A: \text{not that}$$

Test statistic value:

Sample $\hat{p}$.

$$Z_{stat} = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Z is an x-value on the $N(0,1)$.

Alternative Hypothesis	Rejection Region for α level test:
H_A: $p > p_0$ (true vs. assumed)	reject if $Z_{stat} > Z_\alpha$ ← Z could be t
$p < p_0$	$Z_{stat} < -Z_\alpha$
$p \neq p_0$	reject if: $Z_{stat} > Z_{\alpha/2}$ OR $Z_{stat} < -Z_{\alpha/2}$ $\|Z_{stat}\| > Z_{\alpha/2}$

Test for Population Proportion

Example (continued):
Null hypothesis: $H_0 : p = p_0$

Test statistic value:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

<u>Alternative Hypothesis</u>   <u>Rejection Region</u> for α level test:

Test for Population Proportion

Example (continued):
Null hypothesis: $H_0 : p = p_0$

Test statistic value:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Alternative Hypothesis   Rejection Region for α level test:

$H_a : p > p_0$
$H_a : p < p_0$
$H_a : p \neq p_0$

Test for Population Proportion

Example (continued):
Null hypothesis: $H_0 : p = p_0$

to normalize

Test statistic value:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

point estimate − hyp. mean

standard error

Alternative Hypothesis | Rejection Region for α level test:

$H_a : p > p_0$ $z > z_\alpha$

$H_a : p < p_0$ $z < -z_\alpha$

$H_a : p \neq p_0$ $z < -z_{\alpha/2}$ **or** $z > z_{\alpha/2}$

Test for Population Proportion

Example To determine if the Belgian 1 Euro coin is fair, you flip it 250 times and it comes up heads 139 times. Do you reject the null at $\alpha = .1$ and conclude the coin is unfair?

Test for Population Proportion

Example To determine if the Belgian 1 Euro coin is fair, you flip it 250 times and it comes up heads 139 times. Do you reject the null at $\alpha = .1$ and conclude the coin is unfair?

Test statistic:

$$Z = \frac{139/250 - .5}{\sqrt{\frac{.5(1-.5)}{250}}}$$

Test for Population Proportion

Example To determine if the Belgian 1 Euro coin is fair, you flip it 250 times and it comes up heads 139 times. Do you reject the null at $\alpha = .1$ and conclude the coin is unfair?

Test statistic:

$$Z = \frac{139/250 - .5}{\sqrt{\frac{.5(1-.5)}{250}}}$$

This is approximately 1.771. Compare to $z_{\alpha/2} = 1.645$.

Test for Population Proportion

Example To determine if the Belgian 1 Euro coin is fair, you flip it 250 times and it comes up heads 139 times. Do you reject the null at $\alpha = .1$ and conclude the coin is unfair?

Test statistic:

$$Z = \frac{139/250 - .5}{\sqrt{\frac{.5(1-.5)}{250}}}$$

This is approximately 1.771. Compare to $z_{\alpha/2} = 1.645$.

Reject, because our test statistic represents a more extreme observation under the null than our the corresponding value of $z_{\alpha/2}$.

Test for Population Mean (known variance)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

<u>Alternative Hypothesis</u>   <u>Rejection Region</u> for α level test:

Test for Population Mean (known variance)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

<u>Alternative Hypothesis</u>   <u>Rejection Region</u> for α level test:

Test for Population Mean (known variance)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Alternative Hypothesis   Rejection Region for α level test:

$H_a : \mu > \mu_0$

$H_a : \mu < \mu_0$

$H_a : \mu \neq \mu_0$

Test for Population Mean (known variance)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

<u>Alternative Hypothesis</u>   <u>Rejection Region</u> for α level test:

$H_a : \mu > \mu_0$ $z > z_\alpha$

$H_a : \mu < \mu_0$ $z < -z_\alpha$

$H_a : \mu \neq \mu_0$ $z < -z_{\alpha/2}$ **or** $z > z_{\alpha/2}$

Test for Population Mean (known variance)

Example: Suppose a company is considering putting a new type of coating on bearings that it produces. Let denote the true mean life for the new coating. The company would not want to make any (costly) changes unless evidence **strongly** suggested that exceeds 1000 hours.

State the null and alternative hypotheses:

Now, suppose the company tests $n = 25$ randomly selected bearings with the new coating, and finds that the sample mean for life of the new coating is 1090 hours. Suppose we know that the population standard deviation is 130 hours. Conduct a test for the mean.

Test for Population Mean (known variance)

Example: Suppose a company is considering putting a new type of coating on bearings that it produces. Let denote the true mean life for the new coating. The company would not want to make any (costly) changes unless evidence **strongly** suggested that exceeds 1000 hours.

State the null and alternative hypotheses:

Null hypothesis: $H_0 : \mu = \mu_0$

Now, suppose the company tests $n = 25$ randomly selected bearings with the new coating, and finds that the sample mean for life of the new coating is 1090 hours. Suppose we know that the population standard deviation is 130 hours. Conduct a test for the mean.

Test for Population Mean (known variance)

Example: Suppose a company is considering putting a new type of coating on bearings that it produces. Let denote the true mean life for the new coating. The company would not want to make any (costly) changes unless evidence **strongly** suggested that exceeds 1000 hours.

State the null and alternative hypotheses:

Null hypothesis: $H_0 : \mu = \mu_0$
Alt hypothesis: $H_a : \mu > \mu_0$

Now, suppose the company tests $n = 25$ randomly selected bearings with the new coating, and finds that the sample mean for life of the new coating is 1090 hours. Suppose we know that the population standard deviation is 130 hours. Conduct a test for the mean.

Test for Population Mean (known variance)

Example: Suppose a company is considering putting a new type of coating on bearings that it produces. Let denote the true mean life for the new coating. The company would not want to make any (costly) changes unless evidence **strongly** suggested that exceeds 1000 hours.

State the null and alternative hypotheses:

Null hypothesis: $H_0 : \mu = \mu_0$
Alt hypothesis: $H_a : \mu > \mu_0$

Now, suppose the company tests $n = 25$ randomly selected bearings with the new coating, and finds that the sample mean for life of the new coating is 1090 hours. Suppose we know that the population standard deviation is 130 hours. Conduct a test for the mean. Compare

$Z = \frac{1090 - 1000}{130/\sqrt{25}} = 3.45$ to $z_{.01} = 2.32$ and **Reject**

Errors in Hypothesis Testing

Definitions: *Type One and Type Two Errors*

A *type I error* is when the null hypothesis is rejected, but it is true. (False Positive).

A *type II error* is not rejecting H_0 when H_0 is false. (False Negative)

Errors in Hypothesis Testing

Definitions: *Type One and Type Two Errors*

A *type I error* is when the null hypothesis is rejected, but it is true. (False Positive).

We commit type I errors at a rate of α.

A *type II error* is not rejecting H_0 when H_0 is false. (False Negative)

We commit type II errors at a rate of β.

Errors in Hypothesis Testing

Typically, we specify the largest value of a type I error, α, that can be tolerated, and then find a rejection region with that α.

The resulting value of is often referred to as the **significance** level of the test.

Traditional levels of significance are .10, .05, and .01, though the level in any particular problem will depend on the seriousness of a type I error. The more serious the type I error, the smaller the significance level should be.

Errors in Hypothesis Testing

We can also obtain a smaller value of α - the probability of a type I error - by decreasing the size of the rejection region. However, this results in a larger value of β - the probability of a type II error - for all parameter values consistent with H_a.

No rejection region will simultaneously make α and β small at the same time. A region must be chosen to strike a **compromise** between these errors.

Errors in Hypothesis Testing

Suppose we are evaluating a medical procedure. It has the following outcomes:

1. Procedure works: subject is cured!

2. Procedure doesn't work: subject dies.

Do we want our test - which compared this procedure as H_a to a null hypothesis of "don't get the procedure" - to focus on decreasing α or β?

Errors in Hypothesis Testing

Suppose we are evaluating a medical procedure. It has the following outcomes:

1. Procedure works: subject is cured!

2. Procedure doesn't work: subject dies.

Do we want our test - which compared this procedure as H_a to a null hypothesis of "don't get the procedure" - to focus on decreasing α or β?

$\boxed{\alpha}$: probability of endorsing the test when it doesn't actually work. Result: subjects die.

$\boxed{\beta}$: probability of *not* endorsing the test when it really works! Result: subjects don't get treated.

Errors in Hypothesis Testing

Suppose we are evaluating a medical procedure. It has the following outcomes:

1. Procedure works: subject is cured!

2. Procedure doesn't work: subject dies.

Do we want our test - which compared this procedure as H_a to a null hypothesis of "don't get the procedure" - to focus on decreasing α or β?

$\boxed{\alpha}$: probability of endorsing the test when it doesn't actually work. Result: subjects die.

Where is your risk tolerance?

Example An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for more than 5 hours (300 minutes) on a single gallon of regular gasoline. (The leading brand lawnmower engine runs for 300 minutes on 1 gallon of gasoline.)

From his stock of engines, the inventor selects a simple random sample of 50 engines for testing. The engines run for an average of 305 minutes. The *sample* standard deviation s is found at 30 minutes.

Test the hypothesis that the mean run time is more than 300 minutes. Use a 0.01 level of significance.

Large Sample Testing

When the sample size is large, the z tests are easily modified to yield valid test procedures without requiring either a normal population distribution or known standard deviation.

Earlier, we used the key result to justify large-sample confidence intervals:
A large $n \, (> 30)$ implies that the standardized variable

has approximately a standard normal distribution.

Large Sample Testing

When the sample size is large, the z tests are easily modified to yield valid test procedures without requiring either a normal population distribution or known standard deviation.

Earlier, we used the key result to justify large-sample confidence intervals:
A large $n\,(>30)$ implies that the standardized variable

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has approximately a standard normal distribution.

Test for Population Mean (Large Sample)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

Alternative Hypothesis   Rejection Region for α level test:

Test for Population Mean (Large Sample)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Alternative Hypothesis   Rejection Region for α level test:

Test for Population Mean (Large Sample)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Alternative Hypothesis   Rejection Region for α level test:

$H_a : \mu > \mu_0$

$H_a : \mu < \mu_0$

$H_a : \mu \neq \mu_0$

Test for Population Mean (Large Sample)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Alternative Hypothesis	Rejection Region for α level test:
$H_a : \mu > \mu_0$	$z > z_\alpha$
$H_a : \mu < \mu_0$	$z < -z_\alpha$
$H_a : \mu \neq \mu_0$	$z < -z_{\alpha/2}$ **or** $z > z_{\alpha/2}$

Example, cont'd: An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for more than 5 hours (300 minutes) on a single gallon of regular gasoline. (The leading brand lawnmower engine runs for 300 minutes on 1 gallon of gasoline.)

$n = 50; \bar{X} = 305, s = 30.$ Is $\mu > 300$? at $\alpha = .01$?

Test the hypothesis that the mean run time is more than 300 minutes. Use a 0.01 level of significance.

Example, cont'd: An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for more than 5 hours (300 minutes) on a single gallon of regular gasoline. (The leading brand lawnmower engine runs for 300 minutes on 1 gallon of gasoline.)

$n = 50; \bar{X} = 305, s = 30.$ Is $\mu > 300$? at $\alpha = .01$?

Test the hypothesis that the mean run time is more than 300 minutes. Use a 0.01 level of significance.

Compare $Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{305 - 300}{30/\sqrt{50}} = 1.17$ to $z_{.01} = 2.32$ and **Fail to Reject**.

Rejection Regions and CIs

Rejection regions have a lot in common with confidence intervals.
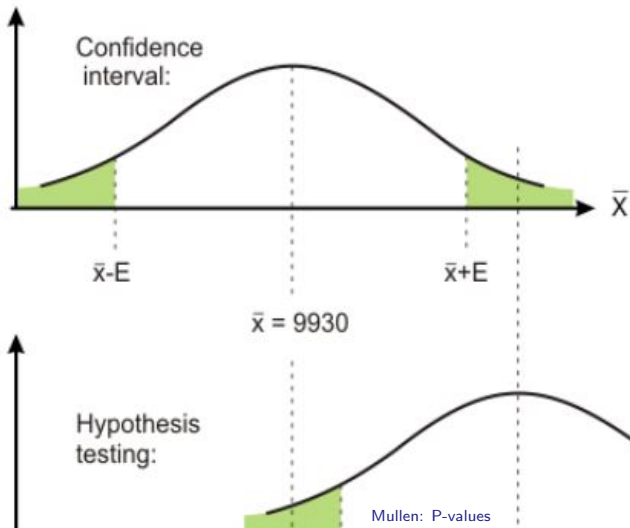
p-values

The p-value measures the "extremeness" of the test statistic.

Definition: *p-value:*

A *p-value* is the probability, under the null hypothesis, that we would get a test statistic at least as extreme as the one we calculated.

Idea: So, the smaller the p-value, the more evidence there is in the sample data against the null hypothesis (so the story goes. . .).

So what constitutes "sufficiently small" and "extreme enough" to make a decision about the null hypothesis?

p-values: a Hypothesis Testing decision

1. Select a significance level (as before, the desired type I error probability), which would define the rejection region.

p-values: a Hypothesis Testing decision

1. Select a significance level (as before, the desired type I error probability), which would define the rejection region.

2. Instead, calculate your test statistic (e.g. a z_{stat} or Z-score),

p-values: a Hypothesis Testing decision

1. Select a significance level (as before, the desired type I error probability), which would define the rejection region.

2. Instead, calculate your test statistic (e.g. a z_{stat} or Z-score),

3. ...and from it calculate the associated probability e.g. $p = P(Z > z_{stat})$ - or the less than, two sided variant - and use that.

p-values: a Hypothesis Testing decision

1. Select a significance level (as before, the desired type I error probability), which would define the rejection region.

2. Instead, calculate your test statistic (e.g. a z_{stat} or Z-score),

3. ...and from it calculate the associated probability e.g. $p = P(Z > z_{stat})$ - or the less than, two sided variant - and use that.

4. decision rule is:

$$\textbf{Reject if:} \quad p \leq \alpha$$

p-values: a Hypothesis Testing decision

1. Select a significance level (as before, the desired type I error probability), which would define the rejection region.

2. Instead, calculate your test statistic (e.g. a z_{stat} or Z-score),

4. decision rule is:

Reject if: $p \leq \alpha$

Thus if the p-value exceeds the chosen significance level, the null hypothesis cannot be rejected at that level.

Note, the p-value can be thought of as the smallest significance level at which H_0 can be rejected.

p-values

The p-value measures the "extremeness" of the test statistic.

Note:

1. This probability is calculated assuming that the null hypothesis is true.

2. Beware: The p-value is not the probability that H_0 is true, nor is it an error probability!

3. The p-value is between 0 and 1.

p-values

The calculation of the p-value depends on whether the test is upper-, lower-, or two-tailed.

Each of these is the probability of getting a value at least as extreme as what was obtained (assuming H_0 true).

p-values for Z tests

1. **Upper-tailed test**
   H_a contains the inequality $>$

   z curve

   P-value = area in upper tail
   $= 1 - \Phi(z)$

   0

   Calculated z
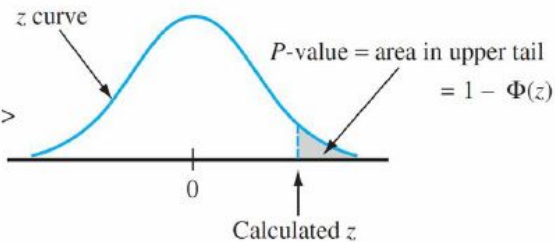
2. **Lower-tailed test**
   H_a contains the inequality $<$

   P-value = area in lower tail
   $= \Phi(z)$

   z curve

   0

   Calculated z

p-values for Z tests
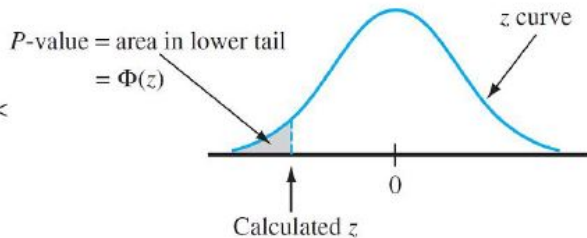
P-value = sum of area in two tails = $2[1 - \Phi(|z|)]$
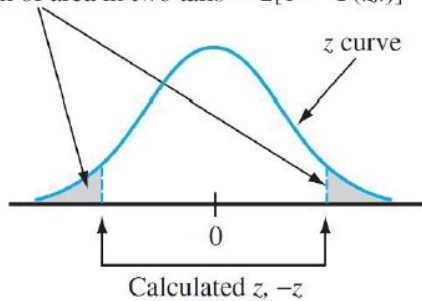
3. **Two-tailed test**

H_a contains the inequality $\neq$

z curve

0

Calculated $z, -z$

p-values for Z tests

Back to the lawnmower engine example: There, we had

$H_0 : \mu = 300$ vs $H_a : \mu > 300$

and $Z = 1.17$.

What is the p-value for this result? What decision would you make?

p-values for Z tests

Back to the lawnmower engine example: There, we had

$H_0 : \mu = 300$ vs $H_a : \mu > 300$

and $Z = 1.17$.

What is the p-value for this result? What decision would you make?

Hopefully the same one! In particular,
$P(Z > 1.17) =$ `1-stats.norm.cdf(1.17)` $= .12 = p > \alpha$

Distribution of p-values

Figure below shows a histogram of the 10,000 P-values from a simulation experiment under a null $\mu = 20$ (with n = 4 and = 2).
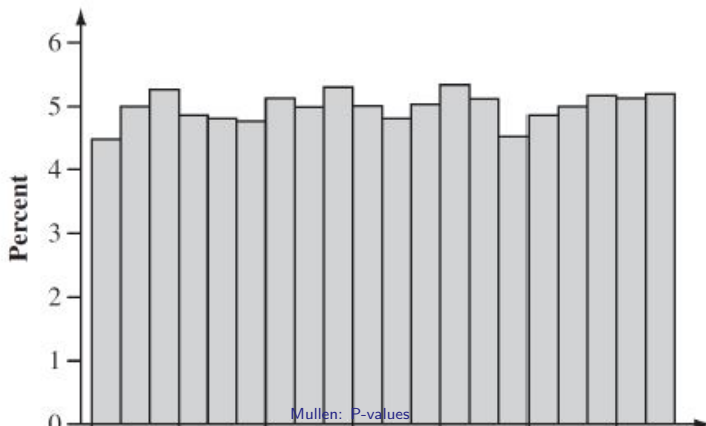
When H0 is true, the probability distribution of the P-value is a uniform distribution on the interval from 0 to 1.
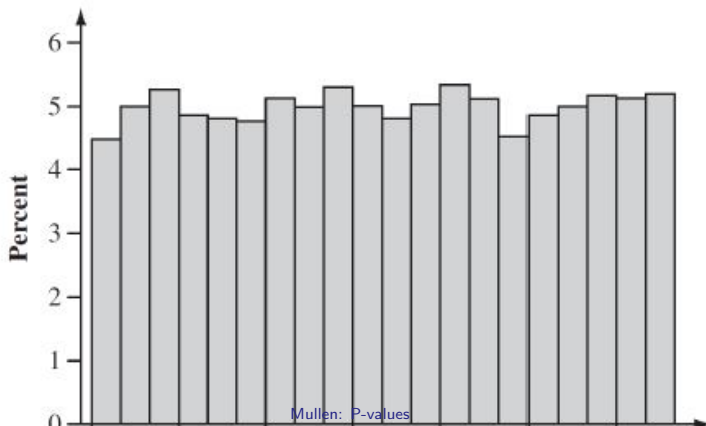
Distribution of p-values

Figure below shows a histogram of the 10,000 P-values from a simulation experiment under a null $\mu = 20$ (with n = 4 and = 2).

When H0 is true, the probability distribution of the P-value is a uniform distribution on the interval from 0 to 1.
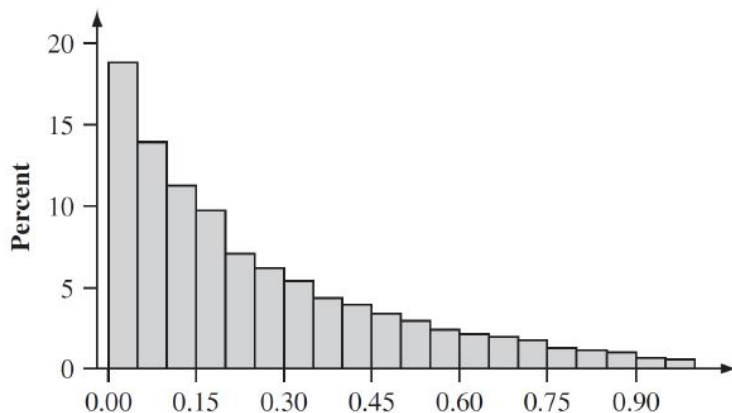
Distribution of p-values

About 4.5% of these p-values are in the interval from 0 to .05.

Thus when using a significance level of .05, the null hypothesis is rejected in roughly 4.5% of these 10,000 tests.

If we continued to generate samples and carry out the test for each sample at significance level .05, in the long run, 5% of the p-values would be in the first class interval.
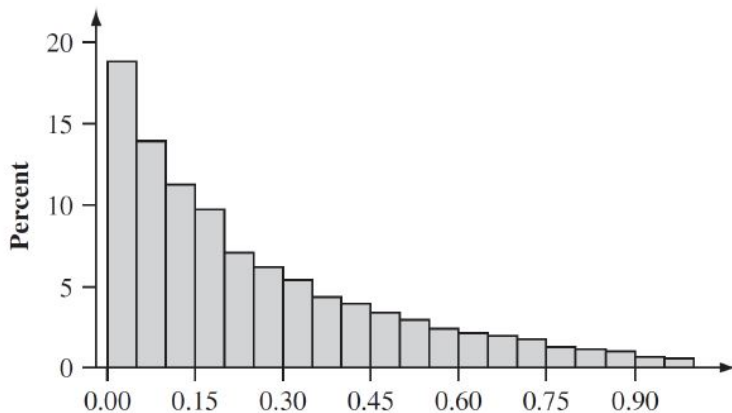
Distribution of p-values

A histogram of the p-values when we simulate under an alternative hypothesis. There is a much greater tendency for the p-value to be small (closer to 0) when $\mu = 21$ than when $\mu = 20$

Distribution of p-values

A histogram of the p-values when we simulate under an alternative hypothesis. There is a much greater tendency for the p-value to be small (closer to 0) when $\mu = 21$ than when $\mu = 20$

The p-histogram of p-values under the alternative hypothesis is TRUE. Rejecting the null
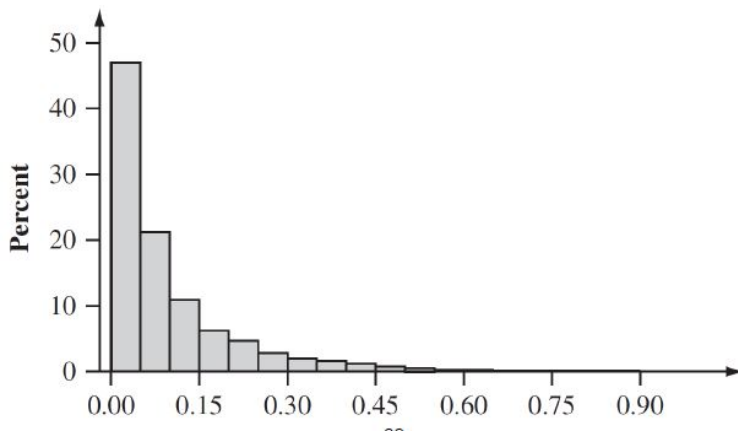
Distribution of p-values

Again, H_0 is rejected at significance level .05 whenever the p-value is at most .05 (in the first bin).

Unfortunately, this is the case for only about 19% of the p-values. So only about 19% of the 10,000 tests correctly reject the null hypothesis; for the other 81%, a type II error is committed.

The difficulty is that the sample size is quite small and 21 is not very different from the value asserted by the null hypothesis (the "effect size" is small).
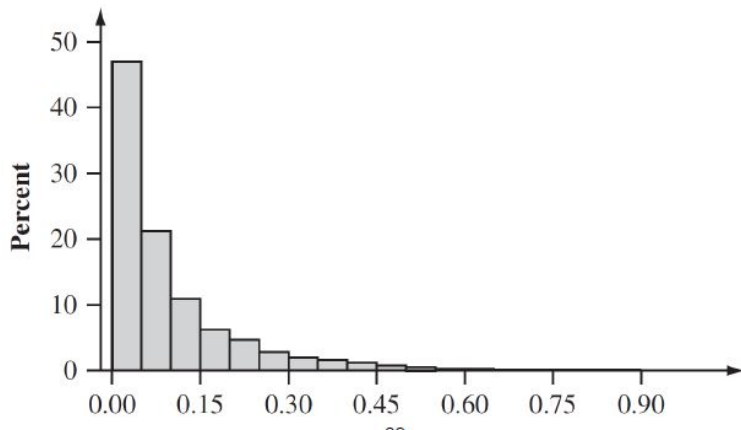
Distribution of p-values

The figure below illustrates what happens to the pvalue when H_0 is false because $\mu = 22$.

Distribution of p-values

The figure below illustrates what happens to the pvalue when H_0 is false because $\mu = 22$.

Again, these data come from H_a, but we reach the correct conclusion more often. The

Distribution of p-values

The histogram is even more concentrated toward values close to 0 than was the case when $\mu = 21$.

In general, as moves further to the right of the null value 20, the distribution of the p-value will become more and more concentrated on values close to 0.

Even here a bit fewer than 50% of the p-values are smaller than .05. So it is still slightly more likely than not that the null hypothesis is incorrectly not rejected. Only for values of much larger than 20 (e.g., at least 24 or 25) is it highly likely that the p-value will be smaller than .05 and thus give the correct conclusion.

Daily Recap

Today we learned

1. Comparing multiple large or normal samples for equivalence of the mean!

Moving forward:

- nb day Friday

Next time in lecture:

- Hypotheses!