

# Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

# Lecture #19 – Hypothesis Testing II

Test for the mean of a normal population (variance known)

Relationship Between Confidence Intervals and Hypothesis Tests

P-values

## A Simple Example

Suppose  $X_1, \dots, X_{100} \sim \text{iid } N(\mu, \sigma^2 = 9)$  and we want to test

$$H_0: \mu = 2$$

$$H_1: \mu \neq 2$$

Step 1 – Specify Null Hypothesis  $H_0$  and alternative Hypothesis  $H_1$  ✓

Step 2 – Choose Test Statistic  $T_n$

If  $\bar{X}$  is far from 2 then  $\mu = 2$  is implausible. Why?



Suppose  $X_1, \dots, X_{100} \sim \text{iid } N(2, \sigma^2 = 9)$ . What is the sampling distribution of  $\bar{X}$ ?

- (a)  $N(0, 1)$
- (b)  $t(99)$
- (c)  $N(2, 0.3)$
- (d)  $N(2, 1)$
- (e)  $N(2, 0.09)$

If  $\bar{X}_n$  is far from 2, then  $\mu = 2$  is implausible

Since  $X_1, \dots, X_{100} \sim \text{iid } N(\mu, 9)$ , if  $\mu = 2$  then  $\bar{X} \sim N(2, 0.09)$

$$\begin{aligned} P(a \leq \bar{X} \leq b) &= P\left(\frac{a-2}{3/10} \leq \frac{\bar{X}-2}{3/10} \leq \frac{b-2}{3/10}\right) \\ &= P\left(\frac{a-2}{0.3} \leq Z \leq \frac{b-2}{0.3}\right) \end{aligned}$$

where  $Z \sim N(0, 1)$  so we see that if  $H_0: \mu = 2$  is true then

$$P(1.7 \leq \bar{X} \leq 2.3) = P(-1 \leq Z \leq 1) \approx 0.68$$

$$P(1.4 \leq \bar{X} \leq 2.6) = P(-2 \leq Z \leq 2) \approx 0.95$$

$$P(1.1 \leq \bar{X} \leq 2.9) = P(-3 \leq Z \leq 3) > 0.99$$

## Step 2 – Choose Test Statistic $T_n$

- ▶ Reject  $H_0: \mu = 2$  if the sample mean is far from 2.
- ▶  $\Rightarrow T_n$  should depend on the **distance** from  $\bar{X}$  to 2, i.e.  $|\bar{X} - 2|$ .
- ▶ We can make our subsequent calculations much easier if we choose a **scale for  $T_n$  that is convenient under  $H_0$** ...

$$\mu = 2 \Rightarrow \bar{X} - 2 \sim N(0, 0.09)$$

$$\frac{\bar{X} - 2}{0.3} \sim N(0, 1)$$

So we will set  $T_n = \left| \frac{\bar{X} - 2}{0.3} \right|$

A Simple Example:  $X_1, \dots, X_{100} \sim \text{iid } N(\mu, \sigma^2 = 9)$

Step 1 –  $H_0: \mu = 2, H_1: \mu \neq 2$  ✓

Step 2 –  $T_n = \left| \frac{\bar{X} - 2}{0.3} \right|$  ✓

Step 3 – If  $\mu = 2$  then  $\left( \frac{\bar{X} - 2}{0.3} \right) \sim N(0, 1)$  ✓

Step 4 – Choose Critical Value  $c$

- (i) Specify significance level  $\alpha$ .
- (ii) Choose  $c$  so that  $P(T_n > c) = \alpha$  under  $H_0: \mu = 2$ .

Choose  $c$  so that  $P(T_n > c) = \alpha$  under  $H_0$

$$T_n = \left| \frac{\bar{X} - 2}{0.3} \right| \text{ and } \mu = 2 \implies \frac{\bar{X} - 2}{0.3} \sim N(0, 1)$$

$$P\left(\left| \frac{\bar{X} - 2}{0.3} \right| > c\right) = \alpha$$

$$1 - P\left(\left| \frac{\bar{X} - 2}{0.3} \right| \leq c\right) = \alpha$$

$$P\left(\left| \frac{\bar{X} - 2}{0.3} \right| \leq c\right) = 1 - \alpha$$

$$P\left(-c \leq \frac{\bar{X} - 2}{0.3} \leq c\right) = 1 - \alpha$$

Hence:  $c = \text{qnorm}(1 - \alpha/2)$  which should look familiar!



## A Simple Example: $X_1, \dots, X_{100} \sim \text{iid } N(\mu, \sigma^2 = 9)$

Step 1 –  $H_0: \mu = 2, H_1: \mu \neq 2$  ✓

Step 2 –  $T_n = \left| \frac{\bar{X} - 2}{0.3} \right|$  ✓

Step 3 – If  $\mu = 2$  then  $\left( \frac{\bar{X} - 2}{0.3} \right) \sim N(0, 1)$  ✓

Step 4 –  $c = \text{qnorm}(1 - \alpha/2)$  ✓

Step 5 – Look at the data: if  $T_n > c$ , reject  $H_0$

- ▶ Suppose I choose  $\alpha = 0.05$ . Then  $c \approx 2$ .
- ▶ I observe a sample of 100 observations. Suppose  $\bar{x} = 1.34$

$$T_n = \left| \frac{\bar{x} - 2}{0.3} \right| = \left| \frac{1.34 - 2}{0.3} \right| = 2.2$$

- ▶ Since  $T_n > c$ , I reject  $H_0: \mu = 2$ .

# Reporting the Results of a Test

Our Example:  $X_1, \dots, X_{100} \sim \text{iid } N(\mu, 9)$

- ▶  $H_0: \mu = 2$  vs.  $H_1: \mu \neq 2$
- ▶  $T_n = |(\bar{X}_n - 2)/0.3|$
- ▶  $\alpha = 0.05 \implies c \approx 2$

Suppose  $\bar{x} = 1.34$

Then  $T_n = 2.2$ . Since this is greater than  $c$  for  $\alpha = 0.05$ , we **reject**  $H_0: \mu = 2$  at the 5% significance level.

Suppose instead that  $\bar{x} = 1.82$

Then  $T_n = 0.6$ . Since this is less than  $c$  for  $\alpha = 0.05$ , we **fail to reject**  $H_0: \mu = 2$  at the 5% significance level.

## General Version of Preceding Example

$X_1, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$  with  $\sigma^2$  known and we want to test:

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

where  $\mu_0$  is some specified value for the population mean.

- ▶  $|\bar{X}_n - \mu_0|$  tells how far sample mean is from  $\mu_0$ .
- ▶ Reject  $H_0: \mu = \mu_0$  if sample mean is far from  $\mu_0$ .
- ▶ Under  $H_0: \mu = \mu_0$ ,  $\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$ .
- ▶ Test statistic  $T_n = \left| \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \right|$
- ▶ Reject  $H_0: \mu = \mu_0$  if  $T_n > \text{qnorm}(1 - \alpha/2)$



Suppose  $X_1, \dots, X_{64} \sim \text{iid } N(\mu, \sigma^2 = 25)$  and we want to test  $H_0: \mu = 0$  vs.  $H_1: \mu \neq 0$  with  $\alpha = 0.32$ . If we observe  $\bar{x} = 0.5$  what is our decision?

- (a) Reject  $H_0$
- (b) Fail to Reject  $H_0$
- (c) Not enough information to determine.

$$T_n = \left| \frac{0.5 - 0}{5/8} \right| = 0.5 \times 8/5 = 0.8, \text{qnorm}(1 - 0.32/2) \approx 1$$

Fail to reject  $H_0$

## What is this test telling us to do?

Return to the example where  $H_0: \mu = 2$  vs.  $H_1: \mu \neq 2$  and  $X_1, \dots, X_{100} \sim \text{iid } N(\mu, 9)$  with  $\alpha = 0.05$ :

$$\text{Reject } H_0 \quad \text{if} \quad \left| \frac{\bar{X}_n - 2}{0.3} \right| > 2$$

$$\text{Reject } H_0 \quad \text{if} \quad |\bar{X}_n - 2| > 0.6$$

$$\text{Reject } H_0 \quad \text{if} \quad (\bar{X}_n < 1.4) \text{ or } (\bar{X}_n > 2.6)$$

Reject  $H_0: \mu = 2$  if  $\bar{X}_n$  is far from 2. How far? Depends on choice of  $\alpha$  along with sample size and population variance.

This looks suspiciously similar to a confidence interval...

$$X_1, \dots, X_n \sim \text{iid } N(\mu, \sigma^2) \text{ where } \sigma^2 \text{ is known}$$

$$T_n = \left| \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \right|, \quad c = \text{qnorm}(1 - \alpha/2), \quad \text{Reject } H_0: \mu = \mu_0 \text{ if } T_n > c$$

Another way of saying this is don't reject  $H_0$  if:

$$\begin{aligned} (T_n \leq c) &\iff \left( \left| \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \right| \leq c \right) \iff \left( -c \leq \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \leq c \right) \\ &\iff \left( \bar{X}_n - c \times \frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X}_n + c \times \frac{\sigma}{\sqrt{n}} \right) \end{aligned}$$

In other words, don't reject  $H_0: \mu = \mu_0$  at significance level  $\alpha$  if  $\mu_0$  lies inside the  $100 \times (1 - \alpha)\%$  confidence interval for  $\mu$ .

# CIs and Hypothesis Tests are Intimately Related

## Our Simple Example

$X_1, \dots, X_{100} \sim \text{iid } N(\mu, \sigma^2 = 9)$  and observe  $\bar{x} = 1.34$

Test  $H_0: \mu = 2$  vs.  $H_1: \mu \neq 2$  with  $\alpha = 0.05$

$T_n = 2.2$ ,  $c = \text{qnorm}(1 - 0.05/2) \approx 2$ . Since  $T_n > c$  we reject.

## 95% Confidence Interval for $\mu$

$1.34 \pm 2 \times 3/10$  i.e.  $1.34 \pm 0.6$  or equivalently  $(0.74, 1.94)$

## Another way to carry out the test...

Since 2 lies outside the 95% confidence interval for  $\mu$ , if our significance level is  $\alpha = 0.05$  we reject  $H_0: \mu = 2$ .

$X_1, \dots, X_{100} \sim \text{iid } N(\mu_X, 9)$  and  $Y_1, \dots, Y_{100} \sim \text{iid } N(\mu_Y, 9)$

Two researchers:  $H_0: \mu = 2$  vs.  $H_1: \mu \neq 2$  with  $\alpha = 0.05$

### Researcher 1

- ▶  $\bar{x} = 1.34$
- ▶  $T_n = 2.2 > 2$
- ▶ Reject  $H_0: \mu_X = 2$

### Researcher 2

- ▶  $\bar{y} = 11.3$
- ▶  $T_n = 31 > 2$
- ▶ Reject  $H_0: \mu_Y = 2$

Both researchers would report “reject  $H_0$  at the 5% level” but  
Researcher 2 found much stronger evidence against  $H_0$ ...



## What if we had chosen a different significance level $\alpha$ ?

$$T_n = 2.2, \quad c = \text{qnorm}(1 - \alpha/2), \quad \text{Reject } H_0: \mu = 2 \text{ if } T_n > c$$

$$\alpha = 0.32 \Rightarrow c = \text{qnorm}(1 - 0.32/2) \approx 0.99 \quad \text{Reject}$$

$$\alpha = 0.10 \Rightarrow c = \text{qnorm}(1 - 0.10/2) \approx 1.64 \quad \text{Reject}$$

$$\alpha = 0.05 \Rightarrow c = \text{qnorm}(1 - 0.05/2) \approx 1.96 \quad \text{Reject}$$

$$\alpha = 0.04 \Rightarrow c = \text{qnorm}(1 - 0.04/2) \approx 2.05 \quad \text{Reject}$$

$$\alpha = 0.03 \Rightarrow c = \text{qnorm}(1 - 0.03/2) \approx 2.17 \quad \text{Reject}$$

$$\alpha = 0.02 \Rightarrow c = \text{qnorm}(1 - 0.02/2) \approx 2.33 \quad \text{Fail to Reject}$$

$$\alpha = 0.01 \Rightarrow c = \text{qnorm}(1 - 0.01/2) \approx 2.58 \quad \text{Fail to Reject}$$

## Result of Test Depends on Choice of $\alpha$ !

$\alpha = 0.32 \Rightarrow$  Reject

$\alpha = 0.10 \Rightarrow$  Reject

$\alpha = 0.05 \Rightarrow$  Reject

$\alpha = 0.04 \Rightarrow$  Reject

$\alpha = 0.03 \Rightarrow$  Reject

$\alpha = 0.02 \Rightarrow$  Fail to Reject

$\alpha = 0.01 \Rightarrow$  Fail to Reject

- ▶ If you reject  $H_0$  at a given choice of  $\alpha$ , you would also have rejected at any **larger** choice of  $\alpha$ .
- ▶ If you fail to reject  $H_0$  at a given choice of  $\alpha$ , you would also have failed to reject at any **smaller** choice of  $\alpha$ .

### Question

If  $\alpha$  is large enough we will reject; if  $\alpha$  is small enough, we won't.

Where is the **dividing line** between reject and fail to reject?

## P-Value: Dividing Line Between Reject and Fail to Reject

$$T_n = 2.2, \quad c = \text{qnorm}(1 - \alpha/2), \quad \text{Reject } H_0: \mu = 2 \text{ if } T_n > c$$

### Question

Given that we observed a test statistic of 2.2, what choice of  $\alpha$  would put us **just at the cusp** of rejecting  $H_0$ ?

### Answer

Whichever  $\alpha$  makes  $c = 2.2$ ! At this  $\alpha$  we just **barely** fail to reject.

# Calculating the P-value

## Definition of a P-value

Significance level  $\alpha$  such that the critical value  $c$  **exactly equals** the observed value of the test statistic. Equivalently:  $\alpha$  that lies exactly on boundary between Reject and Fail to Reject.

## Our Example

The observed value of the test statistic is 2.2 and the critical value is  $\text{qnorm}(1 - \alpha/2)$ , so we need to solve:

$$2.2 = \text{qnorm}(1 - \alpha/2)$$

$$\text{pnorm}(2.2) = \text{pnorm}(\text{qnorm}(1 - \alpha/2))$$

$$\text{pnorm}(2.2) = 1 - \alpha/2$$

$$\alpha = 2 \times [1 - \text{pnorm}(2.2)] \approx 0.028$$

# How to use a p-value?

## Alternative to Steps 4–5

Rather than choosing  $\alpha$ , computing critical value  $c$  and reporting “Reject” or “Fail to Reject” at  $100 \times \alpha\%$  level, just report p-value.

## Example From Previous Slide

P-value for our test of  $H_0: \mu = 2$  against  $H_1: \mu \neq 2$  was  $\approx 0.028$

## Using P-value to Test $H_0$

Using the p-value we can test  $H_0$  for **any**  $\alpha$  without doing any new calculations! For p-value  $< \alpha$  reject; for p-value  $\geq \alpha$  fail to reject.

## Strength of Evidence Against $H_0$

P-value measures **strength of evidence against the null**. Smaller p-value = stronger evidence against  $H_0$ . **P-value does not measure size of effect.**

# Lecture #20 – Hypothesis Testing III

One-Sided Tests

Two-Sample Test For Difference of Means

Matched Pairs Test for Difference of Means

# One-sided Test: Different Decision Rule

Same Example as Last Time

$X_1, \dots, X_{100} \sim \text{iid } N(\mu, 9)$  and  $H_0: \mu = 2$ .

Three possible alternatives:

Two-sided

$$H_1: \mu \neq 2$$

One-sided ( $<$ )

$$H_1: \mu < 2$$

One-sided ( $>$ )

$$H_1: \mu > 2$$

Three corresponding decision rules:

- ▶ Two-sided: reject  $\mu = 2$  whenever  $|\bar{X}_n - 2|$  is too large.
- ▶ One-sided ( $<$ ): only reject  $\mu = 2$  if  $\bar{X}_n$  is far below 2.
- ▶ One-sided ( $>$ ): only reject  $\mu = 2$  if  $\bar{X}_n$  is far above 2.

# One-sided ( $>$ ) Example: $X_1, \dots, X_{100} \sim \text{iid } N(\mu, 9)$

## Null and Alternative

Test  $H_0: \mu = 2$  against  $H_0: \mu > 2$  with  $\alpha = 0.05$ .

## Test Statistic

Drop absolute value for one-sided test:  $T_n = \frac{\bar{X}_n - 2}{0.3}$

## Decision Rule

Reject  $H_0: \mu = 2$  if test statistic is **large and positive**:  $T_n > c$

## Critical Value

Choose  $c$  so that  $P(\text{type I error}) = P(T_n > c | \mu = 2) = 0.05$

Under  $H_0$ ,  $T_n \sim N(0, 1)$

If  $Z \sim N(0, 1)$  what value of  $c$  ensures  $P(Z > c) = 0.05$ ?



# One-sided ( $<$ ) Example: $X_1, \dots, X_{100} \sim \text{iid } N(\mu, 9)$

## Null and Alternative

Test  $H_0: \mu = 2$  against  $H_1: \mu < 2$  with  $\alpha = 0.05$ .

## Test Statistic

Drop absolute value for one-sided test:  $T_n = \frac{\bar{X}_n - 2}{0.3}$

## Decision Rule

Reject  $H_0: \mu = 2$  if test statistic is **large and negative**:  $T_n < c$

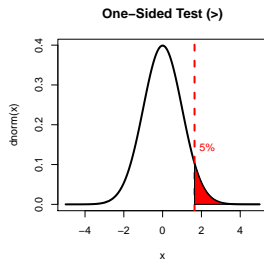
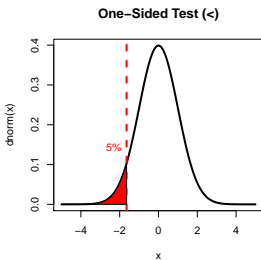
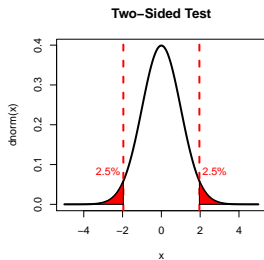
## Critical Value

Choose  $c$  so that  $P(\text{type I error}) = P(T_n < c | \mu = 2) = 0.05$

Under  $H_0$ ,  $T_n \sim N(0, 1)$

If  $Z \sim N(0, 1)$  what value of  $c$  ensures  $P(Z < c) = 0.05$ ?

## Critical Values – Two-sided vs. One-sided Tests: $\alpha = 0.05$



### Two-Sided

Splits  $\alpha = 0.05$  between two tails:  $c = \text{qnorm}(1 - 0.05/2) \approx 1.96$

### One-Sided

One tail:  $c = \text{qnorm}(0.05) \approx -1.64$  for (<);  $\text{qnorm}(0.95) \approx 1.64$  for (>)

Example:  $X_1, \dots, X_{100} \sim \text{iid } N(\mu, 9), \alpha = 0.05$

Suppose  $\bar{x} = 1.5 \implies (\bar{x} - 2)/0.3 \approx -1.67$

Two-sided

$$H_1: \mu \neq 2$$

Reject if  $|T_n| > 1.96$

$$T_n = 1.67$$

Fail to reject

One-sided ( $<$ )

$$H_1: \mu < 2$$

Reject if  $T_n < -1.64$

$$T_n = -1.67$$

Reject

One-sided ( $>$ )

$$H_1: \mu > 2$$

Reject if  $T_n > 1.64$

$$T_n = -1.67$$

Fail to reject

- ▶ If One-sided ( $<$ ) rejects, then one-sided ( $>$ ) doesn't and vice-versa.
- ▶ Two-sided and one-sided sometimes agree but sometimes disagree.
- ▶ One-sided test is "less stringent."

Testing  $H_0: \mu = \mu_0$  when  $X_1, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$

### Two-Sided

Reject  $H_0$  whenever  $\left| \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \right| > \text{qnorm}(1 - \alpha/2)$

### One-Sided ( $<$ )

Reject  $H_0$  whenever  $\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} < \text{qnorm}(\alpha)$

### One-Sided ( $>$ )

Reject  $H_0$  whenever  $\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} > \text{qnorm}(1 - \alpha)$

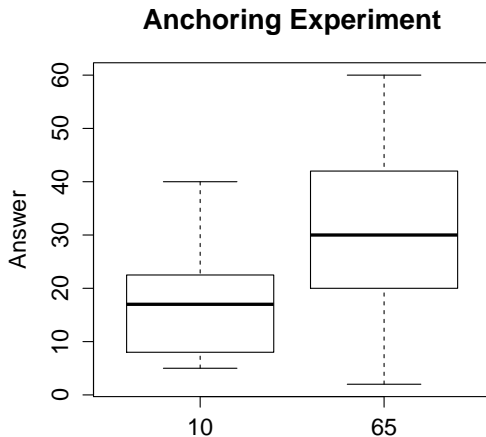
# One-sided P-value

- ▶ Only makes sense to calculate one-sided p-value when sign of test stat. agrees with alternative:
  - ▶ Preceding example:  $T_n = -1.67$
  - ▶ Calculate p-value for test vs.  $H_1: \mu < 2$  but **not**  $H_1: \mu > 2$
- ▶ Just as in two-sided test, p-value equals value of  $\alpha$  for which  $c$  exactly equals the observed test statistic:
  - ▶  $c = \text{qnorm}(\alpha)$  for ( $<$ )
  - ▶  $c = \text{qnorm}(1 - \alpha)$  for ( $>$ )
  - ▶ Example:  $-1.67 = \text{qnorm}(\alpha) \iff \alpha = 0.047$
- ▶ Use and report one-sided p-value in same way as two-sided p-value

# Comparing One-sided and Two-sided Tests

- ▶ Two-sided test is the default.
- ▶ Don't use one-sided unless you have a good reason!
- ▶ Relationship between CI and test **only holds for two-sided**.
- ▶ Why and when should we consider a one-sided test?
  - ▶ Suppose we know *a priori* that  $\mu < 2$  is crazy/uninteresting
  - ▶ Test of  $H_0: \mu = 2$  against  $H_1: \mu > 2$  with significance level  $\alpha$  has **lower type II error rate** than test against  $H_1: \mu \neq 2$ .
- ▶ If you use a one-sided test you **must choose ( $>$ ) or ( $<$ ) before looking at the data**. Otherwise the results are invalid.

# The Anchoring Experiment



# 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

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

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

## Test Statistic: Impose the Null

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

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

## 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?

$\alpha$	0.10	0.05	0.01
$\text{qnorm}(1 - \alpha)$	1.28	1.64	2.33
$\text{qnorm}(1 - \alpha/2)$	1.64	1.96	2.58

... Approximately 2



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

- (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  $H_1: \mu_{Lo} < \mu_{Hi}$

$$1 - \text{pnorm}(5) < 0.0000$$

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

$$2 * (1 - \text{pnorm}(5)) < 0.0000$$

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?
--------------------------



## 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, we'll use the CLT.

# One-Sample Hypothesis Test Using Differences

Let  $D_i = X_i - Y_i$  be (Midterm 2 Score - Midterm 1 Score) for student  $i$

## Null Hypothesis

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

## Two-Sided Alternative

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

## One-Sided Alternative

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

# Decision Rules

Let  $D_i = X_i - Y_i$  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  $\alpha = 0.1$ ?



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

$\alpha$	0.10	0.05	0.01
$\text{qnorm}(1 - \alpha)$	1.28	1.64	2.33
$\text{qnorm}(1 - \alpha/2)$	1.64	1.96	2.58

- (a) Reject
- (b) Fail to Reject
- (c) Not Sure

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



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

$\alpha$	0.10	0.05	0.01
$\text{qnorm}(1 - \alpha)$	1.28	1.64	2.33
$\text{qnorm}(1 - \alpha/2)$	1.64	1.96	2.58

- (a) Reject
- (b) Fail to Reject
- (c) Not Sure

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

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

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

$$1 - \text{pnorm}(1.36) = \text{pnorm}(-1.36) \approx 0.09$$

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

$$2 * (1 - \text{pnorm}(1.36)) = 2 * \text{pnorm}(-1.36) \approx 0.18$$

# Lecture #21 – Testing/CI Roundup

One-sample Test for Proportion

Test for Difference of Proportions

Statistical vs. Practical Significance

Data-Dredging

# Tests for Proportions

## Basic Idea

The population *can't be* normal (it's Bernoulli) so we use the CLT to get approximate sampling distributions (c.f. Lecture 18).

## There's a small twist!

Bernoulli has a *single* unknown parameter ( $p$ ) so  $SE(\hat{p})$  is *known* under  $H_0$ : don't have to estimate it. (C.f. Review Question #194)



# 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

$X_1, \dots, X_n \sim \text{iid Bernoulli}(p)$

## Sample Statistic

Sample Proportion:  $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$

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  $\hat{p}$ ?

(a) 1

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

(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!*

# One-Sample Test for a Population Proportion

## Sampling Model

$X_1, \dots, X_n \sim \text{iid Bernoulli}(p)$

## Null Hypothesis

$H_0: p = \text{Known Constant } p_0$

## Test Statistic

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

## One-Sample Example $H_0: p = 0.5$

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

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

### One-Sided p-value

$$1 - \text{pnorm}(2.2) \approx 0.014$$

### Two-Sided p-value

$$2 * (1 - \text{pnorm}(2.2)) \approx 0.028$$

# 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:  $X_1, \dots, X_n \sim \text{iid Bernoulli}(p)$  independent of

Democrats:  $Y_1, \dots, Y_m \sim \text{iid Bernoulli}(q)$

## Sample Statistics

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

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:

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

Under  $H_0: p = q$

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

# Pooled SE Estimator is More Efficient Under $H_0$

## Unpooled SE

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

## Pooled SE

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

## Why Pool?

- ▶ Under  $H_0$ ,  $p = q$ . Call their common value “ $\pi$ ”
- ▶ More accurate to estimate *1 parameter* ( $\pi$ ) with a *big* sample ( $n + m$ ) vs. *2 parameters* ( $p, q$ ) with smaller samples ( $n, m$ ).

# Two-Sample Test for Proportions

## Sampling Model

$X_1, \dots, X_n \sim \text{iid Bernoulli}(p)$  indep. of  $Y_1, \dots, Y_m \sim \text{iid Bernoulli}(q)$

## Sample Statistics

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

## Null Hypothesis

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

## Pooled Estimator of SE under $H_0$

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

## Test Statistic

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



## Two-Sample Example $H_0: p = q$

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

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

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

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

### One-Sided P-Value

$$1 - \text{pnorm}(1.78) \approx 0.04$$

### Two-Sided P-Value

$$2 * (1 - \text{pnorm}(1.78)) \approx 0.08$$

# Terminology: Statistical Significance

## Definition

If we reject  $H_0$  in a test with significance level  $\alpha$ , then we say that the results are “statistically significant at the  $\alpha\%$  level.

## Example: Anchoring Experiment

In a two-sided test, we found a difference between the “Hi” and “Lo” groups that was statistically significant at the 5% level.

## Example: Previous Slide

In a two-sided test, we found a difference between the share of Republicans and Democrats who knew that Romney is pro-life that was statistically significant at the 10% level.

# Statistical Significance $\neq$ Practical Importance

## Problem

People confuse “significance” in the statistical sense with the everyday English word meaning “important.”

## Statistically Significant Does Not Mean Important

- ▶ A difference can be unimportant but statistically significant.
- ▶ A difference can be important but statistically insignificant.

A p-value measures the *strength of evidence against  $H_0$* ; it does *not* measure 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  $H_1: p \neq 0.5$

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

Approximate 95% Confidence Interval

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

Actual  $p$  was 0.499

# Practically Important But Not Statistically Significant

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

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

# Data-Dredging and the Replication Crisis

## Reading Assignment

On Piazza: “The Economist - Trouble in the Lab.”

## Basic Idea

- ▶ Journals usually publish only “statistically significant” results.
- ▶ You test a large number of null hypotheses with  $\alpha = 0.05$ .
- ▶ Suppose all of these nulls are actually *TRUE*.
- ▶ You’ll reject 5% of the time: each rejection is a Type I error.
- ▶ Cheating in academia: carry out lots of ridiculous hypothesis tests and only report the “statistically significant” results.

# Green Jelly Beans Cause Acne!

xkcd #882

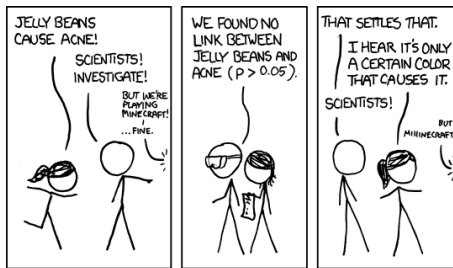


Figure: Reading this comic strip is part of your homework!

And now a simulation example of Data Dredging using R...

```
# Function to calculate the p-value for a two-sided  
#test for difference of means  
get_p_value <- function(x, y) {  
  xbar <- mean(x)  
  ybar <- mean(y)  
  n <- length(x)  
  m <- length(y)  
  s_x <- sd(x)  
  s_y <- sd(y)  
  SE <- sqrt(s_x^2 / n + s_y^2 / m)  
  test_stat <- abs(xbar - ybar) / SE  
  return(2 * (1 - pnorm(test_stat)))  
}
```



```
# Test get_p_value using the anchoring experiment  
# example from our previous lecture  
data_url <- 'http://ditraglia.com/econ103/old_survey.csv'  
survey <- read.csv(data_url)  
anchoring <- survey[, c('rand.num', 'africa.percent')]  
rand_num <- na.omit(anchoring$rand.num)  
africa_percent <- na.omit(anchoring$africa.percent)  
  
x <- subset(africa_percent, rand_num == 65)  
y <- subset(africa_percent, rand_num == 10)  
get_p_value(x, y)  
  
## [1] 6.682931e-07
```

```
# Use *real* student test scores as the outcome
data_url <- 'http://ditraglia.com/econ103/midterms.csv'
midterms <- read.csv(data_url)
scores <- na.omit(midterms$Midterm1)
n_students <- length(scores)

# Generate fake "grouping variables" (0/1) indep. of scores
set.seed(987654321)

n_fake <- 500
# Empty matrix to store grouping variables:
fake <- matrix(NA, nrow = n_students, ncol = n_fake)

for(i in 1:n_fake) {
  fake[, i] <- rbinom(n_students, size = 1, prob = 0.5)
}
```

```
# Use each grouping variable to split students into x and y  
# and calculate p-value for test of difference of means  
  
p_values <- rep(NA, n_fake) # empty vector to store results  
  
for(i in 1:n_fake) {  
  group_indicator <- fake[,i]  
  x <- subset(scores, group_indicator == 1)  
  y <- subset(scores, group_indicator == 0)  
  p_values[i] <- get_p_value(x, y)  
}  
  
# How many of the tests were statistically significant?  
sum(p_values < 0.05)  
  
## [1] 20
```

```
# Grouping variable with the lowest p-value
group_indicator <- fake[, which.min(p_values)]
x <- subset(scores, group_indicator == 1)
y <- subset(scores, group_indicator == 0)

# These results look convincing, but are spurious!
mean(x) - mean(y)

## [1] -7.974127

sqrt(var(x) / length(x) + var(y) / length(y))

## [1] 2.240852
```

# Lecture #22 – Regression II

The Population Regression Model

Inference for Regression

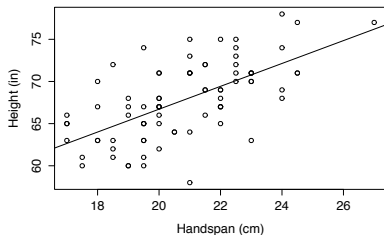
Inference for Regression: Predicting Height

Residual Standard Deviation and  $R^2$

Multiple Regression

## Beyond Regression as a Data Summary

Based on a sample of Econ 103 students, we made the following graph of handspan against height, and fitted a linear regression:



The estimated slope was about 1.4 inches/cm and the estimated intercept was about 40 inches.

What if anything does this tell us about the relationship between height and handspan *in the population*?

# The Population Regression Model

## Question

If we want to predict  $Y$  using  $X$  in the *population* using a line, how should we choose the slope and intercept?

## Optimization Problem

Choose  $\beta_0, \beta_1$  to minimize  $E[(Y - \beta_0 - \beta_1 X)^2]$

## Solution

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}, \quad \beta_0 = E[Y] - \beta_1 E[X]$$

... you will derive this as an extension problem.

# The Regression Error Term: $\varepsilon$

## Definition

$$\varepsilon \equiv Y - \beta_0 - \beta_1 X \quad (\text{Hence: } Y = \beta_0 + \beta_1 X + \varepsilon)$$

## Interpretation

$\varepsilon$  is the part of  $Y$  that isn't predicted by  $X$

## Properties

- ▶  $E[\varepsilon] = 0$
- ▶  $Cov(X, \varepsilon) = 0$
- ▶  $Var(\varepsilon) = Var(Y) - Cov(X, Y)^2 / Var(X)$

... using the expressions for  $\beta_0$  and  $\beta_1$  from the previous slide.



# The Population Regression Coefficients: $\beta_0, \beta_1$

## Recall

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}, \quad \beta_0 = E[Y] - \beta_1 E[X]$$

## Interpretation

- ▶  $\beta_0, \beta_1$  are population parameters: *unknown constants*
- ▶ If  $X = 0$ , we predict  $Y = \beta_0$ .
- ▶ If two people differ by one unit in  $X$ , we predict that they will differ by  $\beta_1$  units in  $Y$ .

The only problem is, we don't know  $\beta_0, \beta_1 \dots$

# Estimating $\beta_0, \beta_1$

## Random Sample

Observe  $(Y_1, X_1), \dots, (Y_n, X_n) \sim \text{iid}$  with  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ .

## Estimators of $\beta_0, \beta_1$

$$\hat{\beta}_1 = \frac{S_{XY}}{S_X^2} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}, \quad \hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$$

Under random sampling, the estimators  $(\hat{\beta}_0, \hat{\beta}_1)$  have sampling distributions. . .

# Sampling Uncertainty: Pretend the Class is our Population

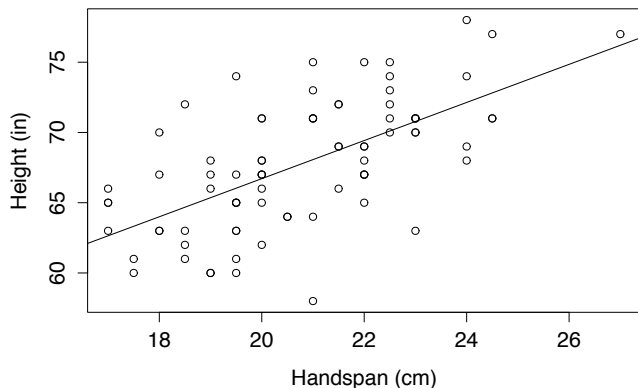
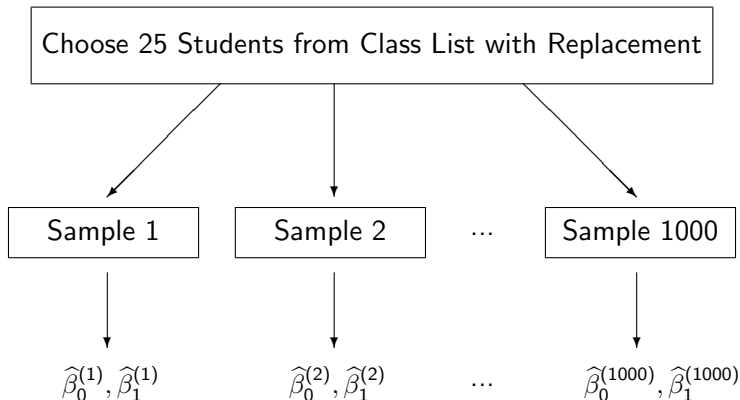


Figure: Estimated Slope = 1.4, Estimated Intercept = 40

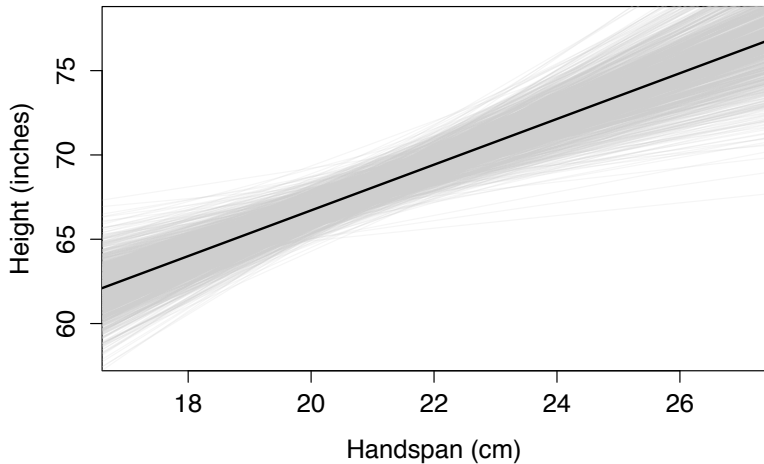
# Sampling Distribution of Regression Coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$



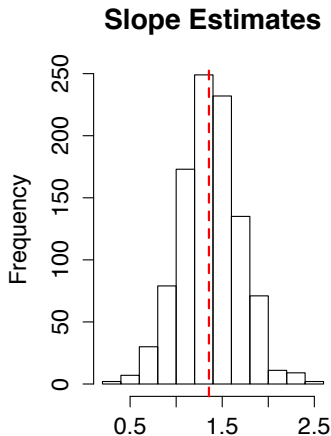
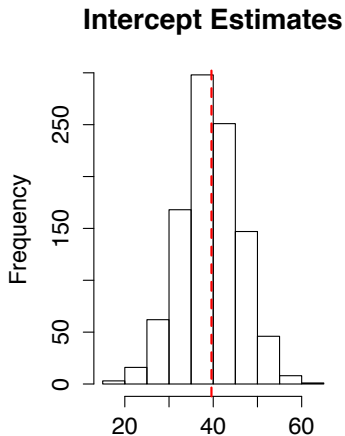
Repeat 1000 times → get 1000 different pairs of estimates

**Sampling Distribution: long-run relative frequencies**

## 1000 Replications, $n = 25$



Population: Intercept = 40, Slope = 1.4



Based on 1000 Replications,  $n = 25$

# Inference for Linear Regression

## Central Limit Theorem

$$\frac{\hat{\beta} - \beta}{\widehat{SE}(\hat{\beta})} \approx N(0, 1)$$

How to calculate  $\widehat{SE}$ ?

R will do this for us, but we won't cover the details in Econ 103.

You'll have to wait for Econ 104!

## How to get R to display standard errors?

```
data_url <- 'http://ditraglia.com/econ103/old_survey.csv'
survey <- read.csv(data_url)
survey <- na.omit(survey)
reg1 <- lm(height ~ handspan, data = survey)
reg1 # Gives estimates but not SE

##
## Call:
## lm(formula = height ~ handspan, data = survey)
##
## Coefficients:
## (Intercept)      handspan
##      39.596         1.356
```



## summary gives too much information...

```
summary(reg1)

##
## Call:
## lm(formula = height ~ handspan, data = survey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0680  -2.4238   0.2204   2.7073   7.9657
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.5962     3.9596  10.000 1.26e-15 ***
## handspan       1.3558     0.1898   7.143 4.20e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.556 on 78 degrees of freedom
## Multiple R-squared:  0.3955, Adjusted R-squared:  0.3877
## F-statistic: 51.03 on 1 and 78 DF,  p-value: 4.201e-10
```

display, from my website, cuts the less important info...

```
source('http://ditraglia.com/econ103/display.R')
display(reg1)

## lm(formula = height ~ handspan, data = survey)
##               coef.est coef.se
## (Intercept) 39.60      3.96
## handspan      1.36      0.19
## ---
## n = 80, k = 2
## residual sd = 3.56, R-Squared = 0.40
```

We will learn what everything in this table means...

# Regression with only an intercept: sample mean

See Review Problem #46

```
reg2 <- lm(height ~ 1, data = survey)
display(reg2)

## lm(formula = height ~ 1, data = survey)
##               coef.est coef.se
## (Intercept) 67.74      0.51
## ---
## n = 80, k = 1
## residual sd = 4.54, R-Squared = 0.00
```

## Dummy Variable (aka Binary Variable)

A predictor variable that takes on only two values: 0 or 1. Used to represent two categories, e.g. Male/Female.

## Regression with intercept & dummy variable: Male/Female

```
reg3 <- lm(height ~ sex, data = survey)
display(reg3)

## lm(formula = height ~ sex, data = survey)
##               coef.est coef.se
## (Intercept)  64.46      0.56
## sexMale      6.10      0.76
## ---
## n = 80, k = 2
## residual sd = 3.38, R-Squared = 0.45
```

# Height & Handspan Regression

```
display(reg1)

## lm(formula = height ~ handspan, data = survey)
##           coef.est coef.se
## (Intercept) 39.60      3.96
## handspan      1.36      0.19
## ---
## n = 80, k = 2
## residual sd = 3.56, R-Squared = 0.40
```

What are n, k, residual sd and R-Squared?

# Fitted Values and Residuals

## Fitted Value $\hat{y}_i$

Predicted  $y$ -value for person  $i$  given her  $x$ -variables using estimated regression coefficients:  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik}$

## Residual $\hat{\epsilon}_i$

Person  $i$ 's *vertical deviation* from regression line:  $\hat{\epsilon}_i = y_i - \hat{y}_i$ .

The residuals are *stand-ins* for the unobserved errors  $\epsilon_i$ .

## Residual Standard Deviation: $\hat{\sigma}$

- ▶ Idea: use residuals  $\hat{\epsilon}_i$  to estimate  $\sigma$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{n - k}}$$

- ▶ Measures avg. distance of  $y_i$  from regression line.
  - ▶ E.g. if  $Y$  is points scored on a test and  $\hat{\sigma} = 16$ , the regression predicts to an accuracy of about 16 points.
- ▶ Same units as  $Y$  (Exam practice: verify this)
- ▶ Denominator  $(n - k) = (\# \text{ Datapoints} - \# \text{ of } X \text{ variables})$



## Proportion of Variance Explained: $R^2$

$$R^2 \approx 1 - \frac{\hat{\sigma}^2}{s_y^2}$$

- ▶  $R^2$  = proportion of  $\text{Var}(Y)$  “explained” by the regression.
  - ▶ Higher value  $\implies$  greater proportion explained
- ▶ Unitless, between 0 and 1
- ▶ Harder to interpret than  $\hat{\sigma}$ , but...
- ▶ Simple Linear Regression:  $R^2 = (r_{xy})^2$  and this where its name comes from!

```

display(reg1)

## lm(formula = height ~ handspan, data = survey)
##               coef.est coef.se
## (Intercept) 39.60      3.96
## handspan    1.36      0.19
## ---
## n = 80, k = 2
## residual sd = 3.56, R-Squared = 0.40

cor(survey$height, survey$handspan)^2

## [1] 0.3954669

sqrt(sum(reg1$residuals^2) / (80 - 2))

## [1] 3.555941

```

# Which Gives Better Predictions: Sex or Handspan?

```
## lm(formula = height ~ handspan, data = survey)
##               coef.est coef.se
## (Intercept)  39.60      3.96
## handspan      1.36      0.19
## ---
## n = 80, k = 2
## residual sd = 3.56, R-Squared = 0.40
```

```
## lm(formula = height ~ sex, data = survey)
##               coef.est coef.se
## (Intercept)  64.46      0.56
## sexMale       6.10      0.76
## ---
## n = 80, k = 2
## residual sd = 3.38, R-Squared = 0.45
```

# Simple vs. Multiple Regression

## Terminology

$Y$  is the “outcome” and  $X$  is the “predictor.”

## Simple Regression

One predictor variable:  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

## Multiple Regression

More than one predictor variable:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \epsilon_i$$

# Multiple Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \epsilon_i$$

## Ceteris Paribus Interpretation

If two individuals differ by one unit in  $X_j$  but have the *same values for all other predictors*, we predict they will differ by  $\beta_j$  units in  $Y$ .

## Estimating $\beta_0, \beta_1, \dots, \beta_k$

The formulas require matrix algebra: R will do it for us.

## Inference for Multiple Regression

$\frac{\hat{\beta}_j - \beta_j}{\widehat{SE}(\hat{\beta}_j)} \approx N(0, 1)$  if  $n$  is large. R will calculate the SE for us.

## Multiple Regression Example: Sex and Handspan

```
reg4 <- lm(height ~ sex + handspan, data = survey)
display(reg4)

## lm(formula = height ~ sex + handspan, data = survey)
##               coef.est coef.se
## (Intercept)  49.95      4.16
## sexMale      4.18      0.89
## handspan     0.75      0.21
## ---
## n = 80, k = 3
## residual sd = 3.16, R-Squared = 0.53
```