

Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

Lecture #15 – Confidence Intervals I

Quick Review: Sampling Distribution of \bar{X}_n

Confidence Interval for Mean of Normal Population (σ^2 Known)

Interpreting a Confidence Interval

Margin of Error and Width

Am I Taller Than The Average American Male?



Source: Centers for Disease Control (pg. 16)

My height is 73 inches. Based on a sample of US males aged 20 and over, the Centers for Disease Control (CDC) reported a mean height of about 69 inches in a recent report.

Clearly I'm taller than the average American male!

Do you agree or disagree?

- (a) Agree
- (b) Disagree
- (c) Not Sure

What affects the accuracy of the sample mean?

If $X_1, \dots, X_n \sim \text{iid}$ with population mean μ and variance σ^2 , then

$$\text{Var}(\bar{X}_n) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{\sigma^2}{n}$$

Sample Size

Smaller $n \implies$ easier to get an unrepresentative sample.

Population Variance

If everyone is very similar in height, we don't need a very large sample to estimate the population mean height accurately.

Am I Taller Than The Average American Male?

Source: Centers for Disease Control (pg. 16)

Table: Height in inches for Males aged 20 and over (approximate)

Sample Mean	69 inches
Sample Std. Dev.	6 inches
Sample Size	5647
My Height	73 inches

We'll return to this example later.

Today – Simplest Example of a Confidence Interval

- ▶ Suppose the population is $N(\mu, \sigma^2)$
- ▶ We know σ^2 but not μ
- ▶ Draw random sample $X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$
- ▶ Observe value of sample mean \bar{x}_n (e.g. 69 inches)
- ▶ What is a plausible range for μ ?
- ▶ How confident are we? Can we make this precise?

Next time we'll look at more realistic and interesting examples. . .



Suppose $X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$. What is the sampling distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$?

- (a) $N(\mu, \sigma^2)$
- (b) $N(0, 1)$
- (c) $N(0, \sigma)$
- (d) $N(\mu, 1)$
- (e) Not enough information to determine.

$$X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$$

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

Remember that we call the standard deviation of a sampling distribution the **standard error**, written SE , so

$$\frac{\bar{X}_n - \mu}{SE(\bar{X}_n)} \sim N(0, 1)$$

What happens if I rearrange?

$$P\left(-2 \leq \frac{\bar{X}_n - \mu}{SE(\bar{X}_n)} \leq 2\right) = 0.95$$

$$P(-2 \cdot SE \leq \bar{X}_n - \mu \leq 2 \cdot SE) = 0.95$$

$$P(-2 \cdot SE - \bar{X}_n \leq -\mu \leq 2 \cdot SE - \bar{X}_n) = 0.95$$

$$P(\bar{X}_n - 2 \cdot SE \leq \mu \leq \bar{X}_n + 2 \cdot SE) = 0.95$$

Confidence Intervals

Confidence Interval (CI)

Range (A, B) constructed from the **sample data** with specified probability of containing a **population parameter**:

$$P(A \leq \theta_0 \leq B) = 1 - \alpha$$

Confidence Level

The **specified probability**, typically denoted $1 - \alpha$, is called the confidence level. For example, if $\alpha = 0.05$ then the confidence level is 0.95 or 95%.

Confidence Interval for Mean of Normal Population

Population Variance Known

The interval $\bar{X}_n \pm 2\sigma/\sqrt{n}$ has approximately 95% probability of containing the population mean μ , provided that:

$$X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$$

But how are we supposed to interpret this?

Which quantities are random?



Suppose $X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$. Which quantities are random variables?

- (a) μ only
- (b) σ and μ
- (c) σ only
- (d) σ, μ and \bar{X}_n
- (e) \bar{X}_n only

\bar{X}_n only.

Confidence Interval is a Random Variable!

1. X_1, \dots, X_n are RVs $\Rightarrow \bar{X}_n$ is a RV (repeated sampling)
2. μ, σ and n are constants
3. Confidence Interval $\bar{X}_n \pm 2\sigma/\sqrt{n}$ is also a RV!

Meaning of Confidence Interval

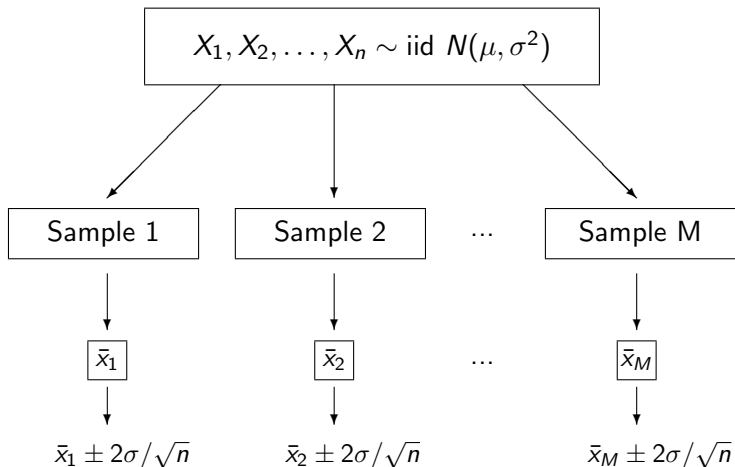
Meaning of Confidence Interval

If we sampled many times we'd get many different sample means, each leading to a **different** confidence interval. Approximately 95% of these intervals will contain μ .

Rough Intuition

What values of μ are consistent with the data?

CI for Population Mean: Repeated Sampling



Repeat M times \rightarrow get M different intervals

Large $M \Rightarrow$ Approx. 95% of these Intervals Contain μ

Simulation Example: $X_1, \dots, X_5 \sim \text{iid } N(0, 1)$, $M = 20$

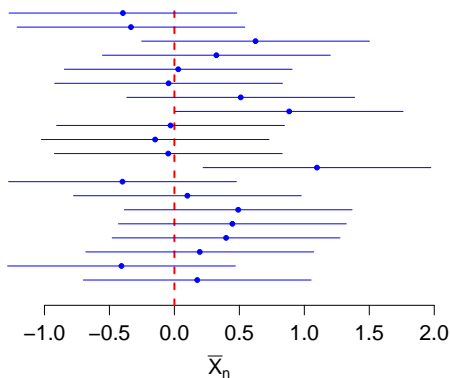


Figure: Twenty confidence intervals of the form $\bar{X}_n \pm 2\sigma/\sqrt{n}$ where $n = 5$, $\sigma^2 = 1$ and the true population mean is 0.

Meaning of Confidence Interval for θ_0

$$P(A \leq \theta_0 \leq B) = 1 - \alpha$$

Each time we sample we'll get a different confidence interval, corresponding to different realizations of the random variables A and B . If we sample many times, approximately $100 \times (1 - \alpha)\%$ of these intervals will contain the population parameter θ_0 .

Confidence Intervals: Some Terminology

Margin of Error

When a CI takes the form $\hat{\theta} \pm ME$, ME is the Margin of Error.

Lower and Upper Confidence Limits

The lower endpoint of a CI is the **lower confidence limit (LCL)**, while the upper endpoint is the **upper confidence limit (UCL)**.

Width of a Confidence Interval

The distance $|UCL - LCL|$ is called the **width** of a CI. This means exactly what it says.

What is the Margin of Error



In the preceding example of a 95% confidence interval for the mean of a normal population when the population variance is known, which of these is the margin of error?

- (a) σ/\sqrt{n}
- (b) \bar{X}_n
- (c) σ
- (d) $2\sigma/\sqrt{n}$
- (e) $1/\sqrt{n}$

$2\sigma/\sqrt{n}$, since the CI is $\bar{X}_n \pm 2\sigma/\sqrt{n}$

What is the Width?



In the preceding example of a 95% confidence interval for the mean of a normal population when the population variance is known, which of these is the width of the interval?

- (a) σ/\sqrt{n}
- (b) $2\sigma/\sqrt{n}$
- (c) $3\sigma/\sqrt{n}$
- (d) $4\sigma/\sqrt{n}$
- (e) $5\sigma/\sqrt{n}$

$4\sigma/\sqrt{n}$, since the CI is $\bar{X}_n \pm 2\sigma/\sqrt{n}$

Example: Calculate the Margin of Error



$X_1, \dots, X_{100} \sim \text{iid } N(\mu, 1)$ but we don't know μ .
Want to create a 95% confidence interval for μ .

What is the margin of error?

The confidence interval is $\bar{X}_n \pm 2\sigma/\sqrt{n}$ so

$$ME = 2\sigma/\sqrt{n} = 2 \cdot 1/\sqrt{100} = 2/10 = 0.2$$

Example: Calculate the Lower Confidence Limit



$X_1, \dots, X_{100} \sim N(\mu, 1)$ but we don't know μ .
Want to create a 95% confidence interval for μ .

We found that $ME = 0.2$. The sample mean $\bar{x} = 4.9$. What is the lower confidence limit?

$$LCL = \bar{x} - ME = 4.9 - 0.2 = 4.7$$

Example: Similarly for the Upper Confidence Limit...

$X_1, \dots, X_{100} \sim N(\mu, 1)$ but we don't know μ .
Want to create a 95% confidence interval for μ .

We found that $ME = 0.2$. The sample mean $\bar{x} = 4.9$. What is the upper confidence limit?

$$UCL = \bar{x} + ME = 4.9 + 0.2 = 5.1$$

Example: 95% CI for Normal Mean, Popn. Var. Known

$X_1, \dots, X_{100} \sim N(\mu, 1)$ but we don't know μ .

95% CI for $\mu = [4.7, 5.1]$

What values of μ are plausible?

The data actually came from a $N(5, 1)$ Distribution.

Want to be more certain? Use higher confidence level.

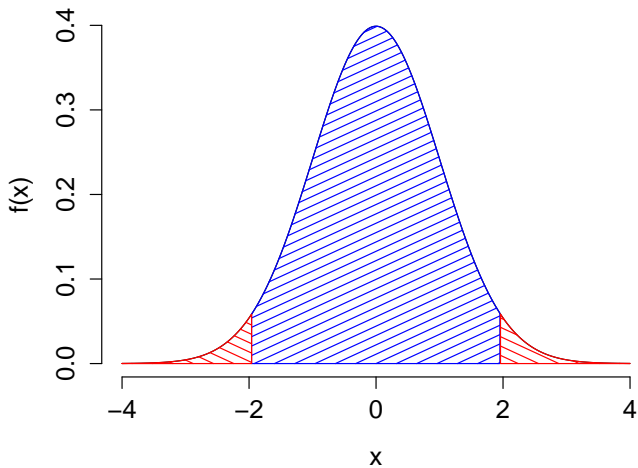
What value of c should we use to get a $100 \times (1 - \alpha)\%$ CI for μ ?

$$P\left(-c \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq c\right) = 1 - \alpha$$

$$P\left(\bar{X}_n - c\sigma/\sqrt{n} \leq \mu \leq \bar{X}_n + c\sigma/\sqrt{n}\right) = 1 - \alpha$$

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

$$\bar{X}_n \pm \text{qnorm}(1 - \alpha/2) \times \sigma/\sqrt{n}$$



What Affects the Margin of Error?

$$\bar{X}_n \pm \text{qnorm}(1 - \alpha/2) \times \sigma / \sqrt{n}$$

Sample Size n

ME decreases with n : bigger sample \implies tighter interval

Population Std. Dev. σ

ME increases with σ : more variable population \implies wider interval

Confidence Level $1 - \alpha$

ME increases with $1 - \alpha$: higher conf. level \implies wider interval

Conf. Level	90%	95%	99%
α	0.1	0.05	0.01
$\text{qnorm}(1 - \alpha/2)$	1.64	1.96	2.56

Lecture #16 – Confidence Intervals II

Comparing intervals with different confidence levels

What if the population is normal but σ is unknown?

What if the population isn't normal? – The Central Limit Theorem

CI for a Proportion Using the Central Limit Theorem

How Much Narrower is a 68% CI?



Suppose we're constructing a confidence interval for the mean of a normal population with σ^2 known:

$$\bar{X}_n \pm \text{qnorm}(1 - \alpha/2) \times \sigma / \sqrt{n}$$

Approximately what is the *ratio* of the width of a 95% interval divided by that of a 68% interval?

$$\text{qnorm}(1 - 0.05/2) \approx 2$$

$$\text{qnorm}(1 - 0.32/2) \approx 1$$

$$\Rightarrow \frac{2 \times \text{qnorm}(1 - 0.05/2) \times \sigma / \sqrt{n}}{2 \times \text{qnorm}(1 - 0.32/2) \times \sigma / \sqrt{n}} \approx 2$$

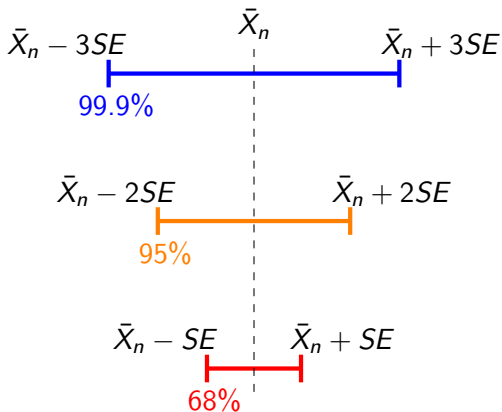


Figure: Each CI gives a range of “plausible” values for the population mean μ , centered at the sample mean \bar{X}_n . Values near the middle are “more plausible” in the sense that a small reduction in confidence level gives a much shorter interval centered in the same place. This is because the sample mean is unlikely to take on values far from the population mean in repeated sampling.

Assume that: $X_1, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$

σ Known

$$P \left[-\text{qnorm}(1 - \alpha/2) \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq \text{qnorm}(1 - \alpha/2) \right] = 1 - \alpha$$

\implies Confidence Interval: $\bar{X}_n \pm \text{qnorm}(1 - \alpha/2) \times \sigma/\sqrt{n}$

σ Unknown

Idea: estimate σ with S . Unfortunately:

$$\frac{\bar{X}_n - \mu}{S/\sqrt{n}} \quad \text{IS NOT A NORMAL RV!}$$

50000 Simulation replications: $X_1, \dots, X_5 \sim \text{iid } N(\mu, \sigma^2)$

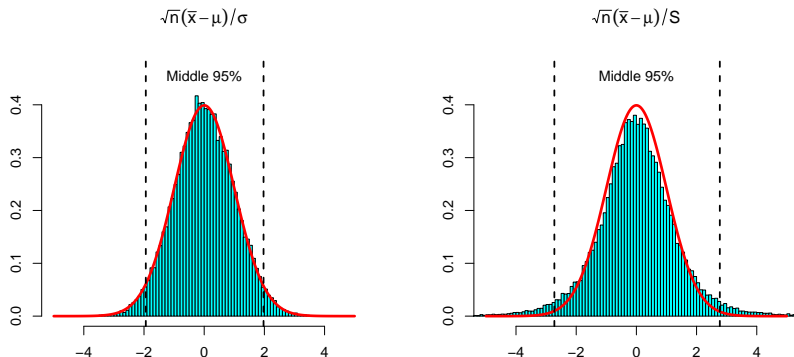


Figure: In each plot the red curve is the pdf of the standard normal RV.
At left: the sampling distribution of $\sqrt{5}(\bar{X}_5 - \mu)/\sigma$ is standard normal.
At right: the sampling distribution of $\sqrt{5}(\bar{X}_5 - \mu)/S$ clearly isn't!

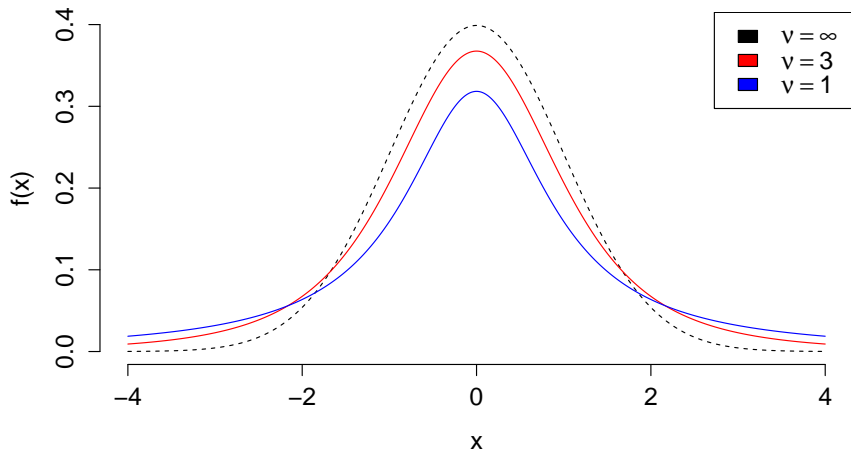
Student-t Random Variable

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

$$\frac{\bar{X}_n - \mu}{S/\sqrt{n}} \sim t(n-1)$$

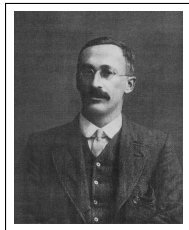
- ▶ Parameter: $\nu = n - 1$ “degrees of freedom”
- ▶ Support = $(-\infty, \infty)$
- ▶ Symmetric around zero, but mean and variance may not exist!
- ▶ Degrees of freedom ν control “thickness of tails”
- ▶ As $\nu \rightarrow \infty$, $t \rightarrow$ Standard Normal.

Student-t PDFs



Who was “Student?”

“Guinnessometrics: The Economic Foundation of Student's t”



“Student” is the pseudonym used in 19 of 21 published articles by William Sealy Gosset, who was a chemist, brewer, inventor, and self-trained statistician, agronomer, and designer of experiments ... [Gosset] worked his entire adult life ... as an experimental brewer for one employer: Arthur Guinness, Son & Company, Ltd., Dublin, St. James's Gate. Gosset was a master brewer and rose in fact to the top of the top of the brewing industry: Head Brewer of Guinness.

CI for Mean of Normal Distribution, Popn. Var. Unknown

Same argument as we used when the variance was known, except with $t(n - 1)$ rather than standard normal distribution:

$$P\left(-c \leq \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \leq c\right) = 1 - \alpha$$

$$P\left(\bar{X}_n - c\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X}_n + c\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

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

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

Comparison of CIs for Mean of Normal Distribution

$100 \times (1 - \alpha)\%$ Confidence Level

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

Known Population Std. Dev. (σ)

$$\bar{X}_n \pm \text{qnorm}(1 - \alpha/2) \frac{\sigma}{\sqrt{n}}$$

Unknown Population Std. Dev. (σ)

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

Comparison of Normal and t CIs

Table: Values of $qt(1 - \alpha/2, df = n - 1)$ for various choices of n and α .

n	1	5	10	30	100	∞
$\alpha = 0.10$	6.31	2.02	1.81	1.70	1.66	1.64
$\alpha = 0.05$	12.71	2.57	2.23	2.04	1.98	1.96
$\alpha = 0.01$	63.66	4.03	3.17	2.75	2.63	2.58

As $n \rightarrow \infty$, $t(n - 1) \rightarrow N(0, 1)$

In a sense, using the t -distribution involves making a “small-sample correction.” In other words, it is only when n is fairly small that this makes a practical difference for our confidence intervals.

Am I Taller Than The Average American Male?

Source: Centers for Disease Control (pg. 16)

Assuming the population is normal,

$$\bar{X}_n \pm qt(1 - \alpha/2, df = n - 1) \widehat{SE}(\bar{X}_n)$$

What is the approximate value of
 $qt(1-0.05/2, df = 5646)$?

For large n , $t(n - 1) \approx N(0, 1)$, so the answer is approximately 2

What is the ME for the 95% CI?

$$ME \approx 0.16 \implies 69 \pm 0.16$$

Sample Mean	69 inches
Sample Std. Dev.	6 inches
Sample Size	5647
My Height	73 inches

$$\begin{aligned}\widehat{SE}(\bar{X}_n) &= s/\sqrt{n} \\ &= 6/\sqrt{5647} \\ &\approx 0.08\end{aligned}$$

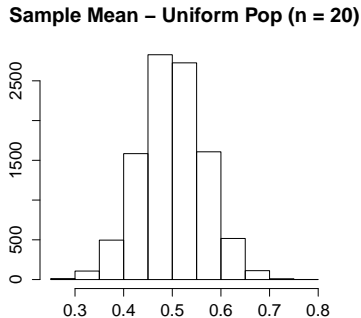
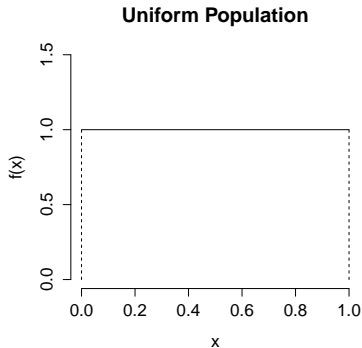
The Central Limit Theorem

Suppose that X_1, \dots, X_n are a random sample from a some population that is **not necessarily normal** and has an unknown mean μ . Then, provided that n is *sufficiently large*,

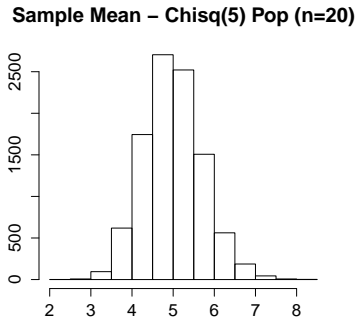
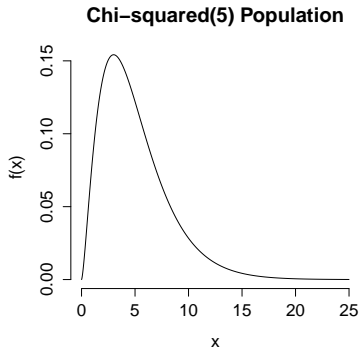
$$\frac{\bar{X}_n - \mu}{S/\sqrt{n}} \approx N(0, 1)$$

We will use this fact to create *approximate* CIs for population mean even if we know *nothing* about the population.

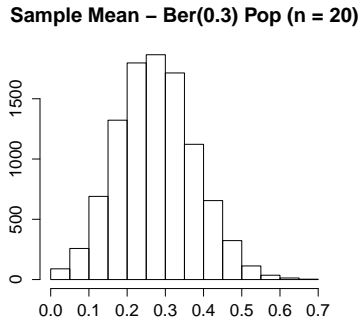
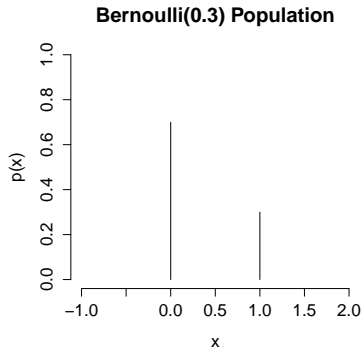
Example: Uniform(0,1) Population, $n = 20$



Example: $\chi^2(5)$ Population, $n = 20$



Example: Bernoulli(0.3) Population, $n = 20$



Who is the Chief Justice of the US Supreme Court?



- (a) Harry Reid
- (b) John Roberts
- (c) William Rehnquist
- (d) Stephen Breyer

Are US Voters Really That Ignorant?

Pew: "What Voters Know About Campaign 2012"

The Data

Of 771 registered voters polled, only 39% correctly identified John Roberts as the current chief justice of the US Supreme Court.

Research Question

Is the majority of voters unaware that John Roberts is the current chief justice, or is this just sampling variation?

Assume Random Sampling...

Confidence Interval for a Proportion

What is the appropriate probability model for the sample?

$X_1, \dots, X_n \sim \text{iid Bernoulli}(p)$, 1 = Know Roberts is Chief Justice

What is the parameter of interest?

p = Proportion of voters *in the population* who know Roberts is Chief Justice.

What is our estimator?

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

Sample Proportion *is* the Sample Mean!

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

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n$$

$$E[\hat{p}] = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{np}{n} = p$$

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

$$SE(\hat{p}) = \sqrt{\text{Var}(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}$$

$$\widehat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Central Limit Theorem Applied to Sample Proportion

Central Limit Theorem: Intuition

Sample means are approximately normally distributed provided the sample size is large even if the population is non-normal.

CLT For Sample Mean

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

CLT for Sample Proportion

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx N(0, 1)$$

In this example, the population is Bernoulli(p) rather than normal. The sample mean is \hat{p} and the population mean is p .

Approximate 95% CI for Population Proportion

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx N(0, 1)$$

$$P\left(-2 \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq 2\right) \approx 0.95$$

$$P\left(\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 0.95$$

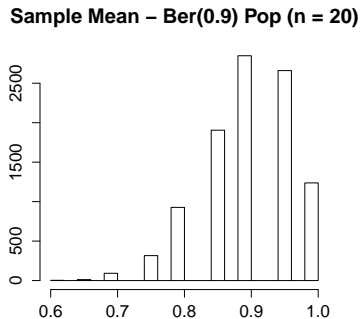
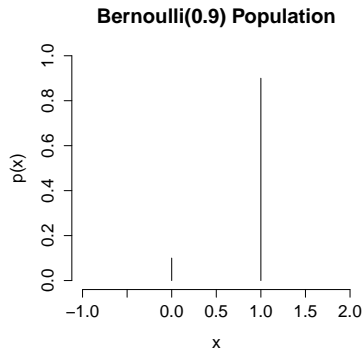
$100 \times (1 - \alpha)$ CI for Population Proportion (p)

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

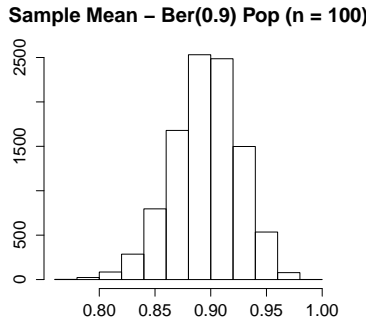
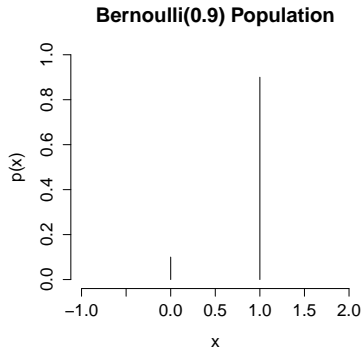
$$\hat{p} \pm \text{qnorm}(1 - \alpha/2) \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Approximation based on the CLT. Works well provided n is large and p isn't too close to zero or one.

Example: Bernoulli(0.9) Population, $n = 20$



Example: Bernoulli(0.9) Population, $n = 100$



Approximate 95% CI for Population Proportion



39% of 771 Voters Polled Correctly Identified Chief Justice Roberts

$$\begin{aligned}\widehat{SE}(\hat{p}) &= \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = \sqrt{\frac{(0.39)(0.61)}{771}} \\ &\approx 0.018\end{aligned}$$

What is the ME for an approximate 95% confidence interval?

$$ME \approx 2 \times \widehat{SE}(\bar{X}_n) \approx 0.04$$

What can we conclude?

Approximate 95% CI: (0.35, 0.43)

Lecture #17 – Confidence Intervals III

Sampling Dist. of $(\bar{X} - \bar{Y})$ – Normal Populations, Variances Known

CI for Difference of Population Means Using CLT

CI for Difference of Population Proportions Using CLT

Matched Pairs versus Independent Samples

Sampling Dist. of $(\bar{X}_n - \bar{Y}_m)$ – Normal Popns. Vars. Known

Suppose $X_1, \dots, X_n \sim \text{iid } N(\mu_x, \sigma_x^2)$ indep. of $Y_1, \dots, Y_m \sim \text{iid } N(\mu_y, \sigma_y^2)$

$$SE(\bar{X}_n - \bar{Y}_m) = \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$$

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_x - \mu_y)}{SE(\bar{X}_n - \bar{Y}_m)} \sim N(0, 1)$$

You should be able to prove this using what we've learned about RVs.

CI for $(\mu_X - \mu_Y)$ – Indep. Normal Popns. σ_X^2, σ_Y^2 Known

$$(\bar{X}_n - \bar{Y}_m) \pm \text{qnorm}(1 - \alpha/2) SE(\bar{X}_n - \bar{Y}_m)$$

$$SE(\bar{X}_n - \bar{Y}_m) = \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$$

CI for Difference of Population Means Using CLT

Setup: Independent Random Samples

$X_1, \dots, X_n \sim \text{iid}$ with unknown mean μ_X & unknown variance σ_X^2
 $Y_1, \dots, Y_m \sim \text{iid}$ with unknown mean μ_Y & unknown variance σ_Y^2
where each sample is independent of the other

We Do Not Assume the Populations are Normal!

Difference of Sample Means $\bar{X}_n - \bar{Y}_m$ and the CLT

What We Have

Approx. sampling dist. for *individual* sample means from CLT:

$$\bar{X}_n \approx N(\mu_X, S_X^2/n), \quad \bar{Y}_m \approx N(\mu_Y, S_Y^2/m)$$

What We Want

Sampling Distribution of the *difference* $\bar{X}_n - \bar{Y}_m$

Use Independence of the Two Samples

$$\bar{X}_n - \bar{Y}_m \approx N\left(\mu_X - \mu_Y, \frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)$$

CI for Difference of Pop. Means (Independent Samples)

$X_1, \dots, X_n \sim \text{iid}$ with mean μ_X and variance σ_X^2

$Y_1, \dots, Y_m \sim \text{iid}$ with mean μ_Y and variance σ_Y^2

where each sample is independent of the other

$$(\bar{X}_n - \bar{Y}_m) \pm \text{qnorm}(1 - \alpha/2) \widehat{SE}(\bar{X}_n - \bar{Y}_m)$$

$$\widehat{SE}(\bar{X}_n - \bar{Y}_m) = \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$$

Approximation based on the CLT. Works well provided n, m large.

The Anchoring Experiment

At the beginning of the semester you were each shown a “random number.” In fact the numbers weren’t random: there was a “Hi” group that was shown 65 and a “Lo” group that was shown 10. You were randomly assigned to one of these two groups and shown your “random” number. You were then asked what proportion of UN member states are located in Africa.

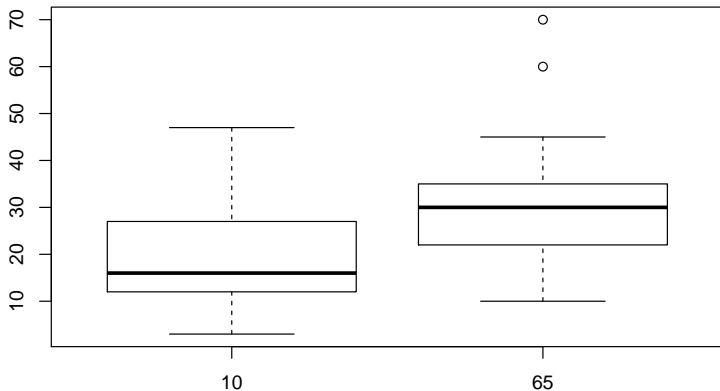
Load Data for Anchoring Experiment

```
data.url <- "http://www.ditraglia.com/econ103/survey_2016_spring.csv"
survey <- read.csv(data.url)
anchoring <- survey[,c("rand.num", "africa.percent")]
head(anchoring)
```

##	rand.num	africa.percent
## 1	65	22
## 2	10	22
## 3	NA	NA
## 4	65	15
## 5	65	20
## 6	10	9

Boxplot of Anchoring Experiment

```
boxplot(africa.percent ~ rand.num, data = anchoring)
```



Anchoring Experiment

From what population is our sample drawn?

US College Students? Penn Students? Penn Econ Majors?

Do We Have a Random Sample?

Definitely not a random sample of US College Students. Possibly a random sample of Penn Econ Majors since Econ 103 is required.

Observational or Experimental Data?

Randomized Experiment drew from a bag of “random” numbers

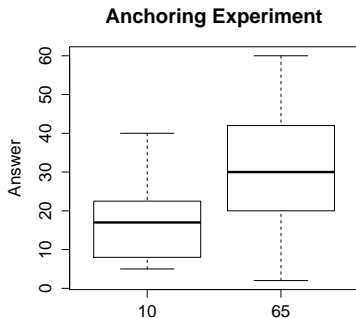
Are the two samples independent?

Yes: I told you not to show your number to any other students or consult with them in any way.

What is the Research Question?

Does “anchoring” cause of bias in decision-making?

Past Semester's Anchoring Experiment



“Lo” Group – Shown 10

$$m_{Lo} = 43$$

$$\bar{y}_{Lo} = 17.1$$

$$s_{Lo}^2 = 86$$

“Hi” Group – Shown 65

$$n_{Hi} = 46$$

$$\bar{x}_{Hi} = 30.7$$

$$s_{Hi}^2 = 253$$

ME for approx. 95% for Difference of Means

“Lo” Group

$$\begin{aligned}\bar{y}_{Lo} &= 17.1 \\ m_{Lo} &= 43 \\ s_{Lo}^2 &= 86 \\ \widehat{SE}(\bar{y}_{Lo})^2 &= \frac{s_{Lo}^2}{m_{Lo}} = 2\end{aligned}$$

“Hi” Group

$$\begin{aligned}\bar{x}_{Hi} &= 30.7 \\ n_{Hi} &= 46 \\ s_{Hi}^2 &= 253 \\ \widehat{SE}(\bar{x}_{Hi})^2 &= \frac{s_{Hi}^2}{n_{Hi}} = 5.5\end{aligned}$$

$$\bar{X}_{Hi} - \bar{Y}_{Lo} = 30.7 - 17.1 = 13.6$$

$$\widehat{SE}(\bar{X}_{Hi} - \bar{Y}_{Lo}) = \sqrt{\widehat{SE}(\bar{X}_{Hi})^2 + \widehat{SE}(\bar{Y}_{Lo})^2} = \sqrt{7.5} \approx 2.7 \Rightarrow ME \approx 5.4$$

Approximate 95% CI (8.2, 19)

What can we conclude?

Confidence Interval for a Difference of Proportions via CLT

What is the appropriate probability model for the sample?

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

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

What is the parameter of interest?

The difference of population proportions $p - q$

What is our estimator?

The difference of sample proportions: $\hat{p} - \hat{q}$ where:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i \qquad \hat{q} = \frac{1}{m} \sum_{i=1}^m Y_i$$

Difference of Sample Proportions $\hat{p} - \hat{q}$ and the CLT

What We Have

Approx. sampling dist. for *individual* sample proportions from CLT:

$$\hat{p} \approx N\left(p, \frac{\hat{p}(1 - \hat{p})}{n}\right), \quad \hat{q} \approx N\left(q, \frac{\hat{q}(1 - \hat{q})}{m}\right)$$

What We Want

Sampling Distribution of the *difference* $\hat{p} - \hat{q}$

Use Independence of the Two Samples

$$\hat{p} - \hat{q} \approx N\left(p - q, \frac{\hat{p}(1 - \hat{p})}{n} + \frac{\hat{q}(1 - \hat{q})}{m}\right)$$

Approximate CI for Difference of Popn. Proportions ($p - q$)

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

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

where each sample is independent of the other

$$(\hat{p} - \hat{q}) \pm \text{qnorm}(1 - \alpha/2) \widehat{SE}(\hat{p} - \hat{q})$$

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

Approximation based on the CLT. Works well provided n, m large and p, q aren't too close to zero or one.

Are Republicans Better Informed Than Democrats?

Pew: "What Voters Know About Campaign 2012"

Of the 239 Republicans surveyed, 47% correctly identified John Roberts as the current chief justice. Only 31% of the 238 Democrats surveyed correctly identified him. Is this difference meaningful or just sampling variation?

Again, assume random sampling.

ME for approx. 95% for Difference of Proportions

47% of 239 Republicans vs. 31% of 238 Democrats identified Roberts

Republicans

$$\hat{p} = 0.47$$

$$n = 239$$

$$\widehat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \approx 0.032$$

Democrats

$$\hat{q} = 0.31$$

$$m = 238$$

$$\widehat{SE}(\hat{q}) = \sqrt{\frac{\hat{q}(1 - \hat{q})}{m}} \approx 0.030$$

Difference: (Republicans - Democrats)

$$\hat{p} - \hat{q} = 0.47 - 0.31 = 0.16$$

$$\widehat{SE}(\hat{p} - \hat{q}) = \sqrt{\widehat{SE}(\hat{p})^2 + \widehat{SE}(\hat{q})^2} \approx 0.044 \implies ME \approx 0.09$$

Approximate 95% CI (0.07, 0.25)

What can we conclude?

Which is the Harder Exam?

Here are the scores from two midterms:

Student	Exam 1	Exam 2	Difference
1	57.1	60.7	3.6
2	77.1	77.9	0.7
3	83.6	93.6	10.0
\vdots	\vdots	\vdots	\vdots
69	75.0	74.3	-0.7
70	96.4	86.4	-10.0
71	78.6	82.9	4.3
Sample Mean:	79.6	81.4	1.8

Is it true that students score, on average, better on Exam 2 or is this just sampling variation?

Are the two samples independent?



Suppose we treat the scores on the first midterm as one sample and the scores on the second as another. Are these samples independent?

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

Matched Pairs Data – Dependent Samples

The samples are dependent: each includes **the same students**:

Student	Exam 1	Exam 2	Difference
1	57.1	60.7	3.6
⋮	⋮	⋮	⋮
71	78.6	82.9	4.3
Sample Mean:	79.6	81.4	1.8
Sample Corr.	0.54		

This is really a **one-sample** problem if we consider the **difference** between each student's score on Exam 2 and Exam 1. This setup is referred to as **matched pairs data**.

Solving this as a One-Sample Problem

Let $D_i = X_i - Y_i$ be the difference of student i 's exam scores.

I calculated the following in R:

$$\bar{D}_n = \frac{1}{n} \sum_{i=1}^n D_i \approx 1.8$$

$$S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2 \approx 124$$

$$\widehat{SE}(\bar{D}_n) = (S_D/\sqrt{n}) \approx \sqrt{124/71} \approx 1.3$$

Approximate 95% CI Based on the CLT:

$$1.8 \pm 2.6 = (-0.8, 4.4)$$

What is our conclusion?

How are the Independent Samples and Matched Pairs Problems Related?

Difference of Means = Mean of Differences?



Let $D_i = X_i - Y_i$ be the difference of student i 's exam scores.

True or False:

$$\bar{D}_n = \bar{X}_n - \bar{Y}_n$$

- (a) True
- (b) False
- (c) Not Sure

Difference of Means Equals Mean of Differences

$$\bar{D}_n = \frac{1}{n} \sum_{i=1}^n D_i = \frac{1}{n} \sum_{i=1}^n (X_i - Y_i) = \bar{X}_n - \bar{Y}_n$$

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

$$\bar{D}_n = 1.8$$

$$\bar{X}_n - \bar{Y}_n = 81.4 - 79.6 = 1.8 \checkmark$$

...But Correlation Affects the Variance

$$\begin{aligned}S_D^2 &= \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D}_n)^2 = \frac{1}{n-1} \sum_{i=1}^n [(X_i - Y_i) - (\bar{X}_n - \bar{Y}_n)]^2 \\&= \frac{1}{n-1} \sum_{i=1}^n [(X_i - \bar{X}_n) - (Y_i - \bar{Y}_n)]^2 \\&= \frac{1}{n-1} \sum_{i=1}^n [(X_i - \bar{X}_n)^2 + (Y_i - \bar{Y}_n)^2 - 2(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)] \\&= S_X^2 + S_Y^2 - 2S_{XY} \\&= S_X^2 + S_Y^2 - 2S_X S_Y r_{XY}\end{aligned}$$

$$r_{XY} > 0 \implies S_D^2 < S_X^2 + S_Y^2$$

$$r_{XY} = 0 \implies S_D^2 = S_X^2 + S_Y^2$$

$$r_{XY} < 0 \implies S_D^2 > S_X^2 + S_Y^2$$

Dependent Samples – Calculating the ME

Student	Exam 1	Exam 2	Difference
1	57.1	60.7	3.6
⋮	⋮	⋮	⋮
71	78.6	82.9	4.3
Sample Var.	117	151	?
Sample Corr.	0.54		

$$117 + 151 - 2 \times 0.54 \times \sqrt{117 \times 151} \approx 124 \checkmark$$

This agrees with our calculations based on the differences.

The “Wrong CI” (Assuming Independence)

Student	Exam 1	Exam 2	Difference
Sample Size	71	71	71
Sample Mean	79.6	81.4	1.8
Sample Var.	117	151	124
Sample Corr.	0.54		

Wrong Interval – Assumes Independence

$$1.8 \pm 2 \times \sqrt{117/71 + 151/71} \Rightarrow (-2.1, 5.7)$$

Correct Interval – Matched Pairs

$$1.8 \pm 2 \times \sqrt{124/71} \Rightarrow (-0.8, 4.4)$$

Top CI is too wide: since exam scores are positively correlated the variance of the differences is less than the sum of the variances.

CI for a Difference of Means – Two Cases

Independent Samples

Two independent samples: X_1, \dots, X_n and Y_1, \dots, Y_m .

Matched Pairs

Matched pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ where X_i is **not independent** of Y_i but each pair (X_i, Y_i) is independent of the other pairs.

Crucial Points

- ▶ Learn to recognize matched pairs and independent samples setups since the CIs are different!
- ▶ Two equivalent ways to construct matched pairs CI:
 1. Method 1: use sample mean and std. dev. of $D_i = X_i - Y_i$
 2. Method 2: use \bar{X}_n , \bar{Y}_n , along with S_X , S_Y and r_{XY}