Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

Lecture #13 – Sampling Distributions and Estimation I

Candy Weighing Experiment

Random Sampling Redux

Unbiasedness of Sample Mean

Standard Error of the Mean

Some More Intuition for Sampling Distributions

Estimator versus Estimate

Weighing a Random Sample

Bag Contains 100 Candies

Estimate total weight of candies by weighing a random sample of size 5 and multiplying the result by 20.

Your Chance to Win

The bag of candies and a digital scale will make their way around the room during the lecture. Each student gets a chance to draw 5 candies and weigh them.

Student with closest estimate wins the bag of candy!

Weighing a Random Sample

Procedure

When the bag and scale reach you, do the following:

1. Fold the top of the bag over and shake to randomize.
2. Randomly draw 5 candies without replacement.
3. Weigh your sample and record the result in grams along with your name on the sign-up sheet.
4. Replace your sample and shake again to re-randomize.
5. Pass bag and scale to next person.

Sampling and Estimation

Questions to Answer

1. How accurately do sample statistics estimate population parameters?
2. How can we quantify the uncertainty in our estimates?
3. What's so good about random sampling?

Random Sample

Verbal Definition from Lecture #1

Each member of population is chosen strictly by chance, so that: (1) selection of one individual doesn't influence selection of any other, (2) each individual is just as likely to be chosen, (3) every possible sample of size n has the same chance of selection.

Mathematical Definition

$X_1, X_2, \ldots, X_n \sim$ iid $f(x)$ if continuous

$X_1, X_2, \ldots, X_n \sim$ iid $p(x)$ if discrete

Random Sample Means *Sample With Replacement*

- Sampling *without replacement* creates dependence between samples (Extension Problem #11).
- But if the population is large relative to the sample, this dependence is negligible: candy experiment isn't bogus!

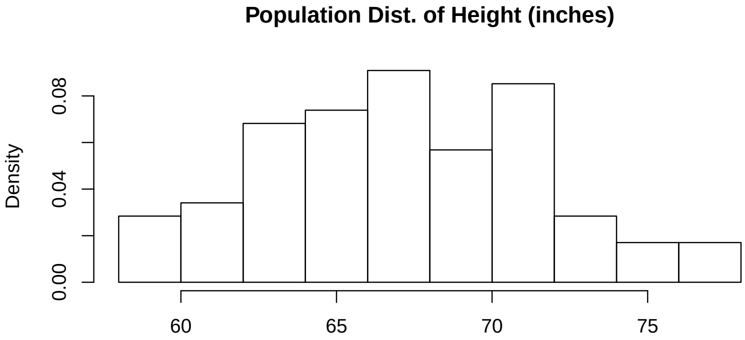
Example: Sampling from Econ 103 Class List

- Pretend the students in this class are a population of interest.
- What is the population mean height?
- In reality I know this since I know all of your heights!
- Suppose I didn't: I could take a random sample of n students and use the sample mean to estimate the population mean.
- I know all of your heights, so I can simulate this in R.

Use this idea to explore the properties of random sampling. . .

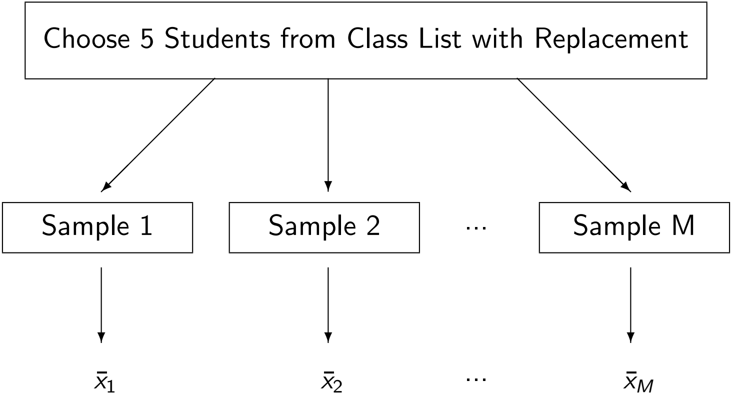
Example: Sampling from the Econ 103 Class List

```
survey <- read.csv('http://ditraglia.com/econ103/old_survey.csv')
height <- na.omit(survey$height)
hist(height, freq = FALSE, xlab = '',
     main = 'Population Dist. of Height (inches)')
```



```
# What is the population mean?
mean(height)

## [1] 67.54545

# Draw a random sample of n = 5 and compute the sample mean
set.seed(3827)
random_sample <- sample(height, 5, replace = FALSE)
random_sample

## [1] 65 75 69 71 60

mean(random_sample)

## [1] 68
```

Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$



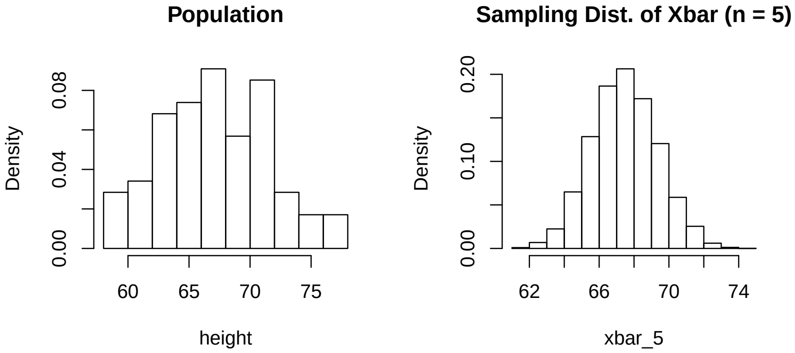
Repeat M times → get M different sample means

Sampling Dist: relative frequencies of the $\bar{x}_i$ when $M = \infty$

```r
set.seed(2985)
# Function: take a random sample of size n, compute sample mean
draw_xbar <- function(n) {
  random_sample <- sample(height, size = n, replace = FALSE)
  mean(random_sample)
}
# Calculate the mean of 10000 random samples with n = 5
M <- 10000
xbar_5 <- replicate(M, draw_xbar(5))
# Compare simulated sample means to population mean: 67.5454 in.
head(xbar_5)
```

```
## [1] 64.8 66.4 68.2 68.6 65.4 71.2
```

```
# Compare popn. dist. of height to histogram of the simulated x-bars
par(mfrow = c(1,2))
hist(height, freq = FALSE, main = 'Population')
hist(xbar_5, freq = FALSE, main = 'Sampling Dist. of Xbar (n = 5)')
```

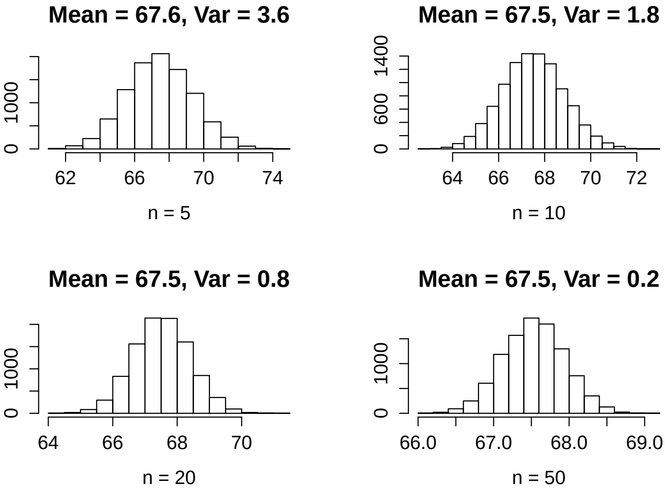


```
par(mfrow = c(1,1))
```

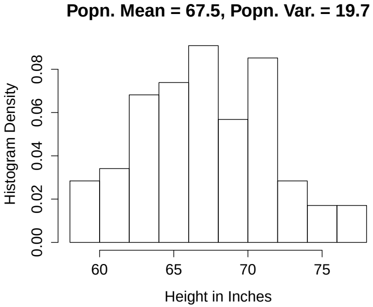
```
# Population mean height
mean(height)

## [1] 67.54545

# Mean of sampling dist. of x-bar (n = 5)
mean(xbar_5)

## [1] 67.56044

# Population variance
var(height)

## [1] 19.74504

# Variance of sampling dist of x-bar (n = 5)
var(xbar_5)

## [1] 3.573168
```

Histograms of sampling distribution of sample mean $\bar{X}_n$

Random Sampling With Replacement, 10000 Reps. Each



Population Distribution vs. Sampling Distribution of $\bar{X}_n$



	Sampling Dist. of $\bar{X}_n$	
n	Mean	Variance
5	67.6	3.6
10	67.5	1.8
20	67.5	0.8
50	67.5	0.2

Things to Notice:

1. Sampling dist. "correct on average"
2. Sampling variability decreases with n
3. Sampling dist. bell-shaped even though population isn't!

Mean of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{n\mu}{n} = \mu$$

Hence, sample mean is "correct on average." The formal term for this is *unbiased*.

Variance of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$\begin{aligned} Var[\bar{X}_n] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

The sampling variance of $\bar{X}_n$ *decreases linearly with sample size.*

Standard Error

Std. Dev. of a sampling distribution is called a standard error.

Standard Error of the Sample Mean

$$SE(\bar{X}_n) = \sqrt{Var\left(\bar{X}_n\right)} = \sqrt{\sigma^2/n} = \sigma/\sqrt{n}$$

Step 1: Population as RV rather than List of Objects

Old Way	New Way
In the 2016 election, 65,853,625 out of 137,100,229 voters voted for Hillary Clinton	Bernoulli($p = 0.48$) RV

Old Way	New Way
List of heights for 97 million US adult males with mean 69 in and std. dev. 6 in	$N(\mu = 69, \sigma^2 = 36)$ RV

Second example assumes distribution of height is bell-shaped.

Step 2: iid RVs Represent Random Sampling from Popn.

Hillary Voters Example

Poll random sample of 1000 people who voted in 2016:

$$X_1, \ldots, X_{1000} \sim \text{ iid Bernoulli}(p = 0.48)$$

Height Example

Measure the heights of random sample of 50 US males:

$$Y_1, \ldots, Y_{50} \sim \text{ iid } N(\mu = 69, \sigma^2 = 36)$$

Key Question

What do the properties of the population imply about the properties of the sample?

The rest of the probabilities. . .

Suppose that exactly half of US voters plan to vote for Hillary Clinton and we poll a random sample of 4 voters.

$$\begin{aligned}
P\,(\text{Exactly 0 Hillary Voters in the Sample}) &= 0.0625\\
P\,(\text{Exactly 1 Hillary Voters in the Sample}) &= 0.25\\
P\,(\text{Exactly 2 Hillary Voters in the Sample}) &= 0.375\\
P\,(\text{Exactly 3 Hillary Voters in the Sample}) &= 0.25\\
P\,(\text{Exactly 4 Hillary Voters in the Sample}) &= 0.0625
\end{aligned}$$

You should be able to work these out yourself. If not, review the lecture slides on the Binomial RV.

Population Size is Irrelevant Under Random Sampling

Crucial Point

None of the preceding calculations involved the population size: I didn't even tell you what it was! We'll never talk about population size again in this course.

Why?

Draw with replacement $\implies$ only the sample size and the *proportion* of Hillary supporters in the population matter.

(Sample) Statistic

Any function of the data *alone*, e.g. sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Used to estimate a population parameter: e.g. $\bar{x}$ estimates of μ.

Step 3: Random Sampling ⇒ *Sample Statistics* are RVs

This is *the crucial point of the course*: if we draw a random sample, the dataset we get is random. Since a statistic is a function of the data, it is a random variable!

Sampling Distribution

Under random sampling, a statistic is a RV so it has a PDF if continuous or PMF if discrete: this is its sampling distribution.

Sampling Dist. of Sample Mean in Polling Example

$$\begin{aligned} p(0) &= 0.0625 \\ p(0.25) &= 0.25 \\ p(0.5) &= 0.375 \\ p(0.75) &= 0.25 \\ p(1) &= 0.0625 \end{aligned}$$

Contradiction? No, but we need better terminology. . .

- Under random sampling, a statistic is a RV
- Given dataset is *fixed* so statistic is a *constant number*
- Distinguish between: Estimator vs. Estimate

Estimator

Description of a general procedure.

Estimate

Particular result obtained from applying the procedure.

$\bar{X}_n$ is an Estimator = Procedure = Random Variable

1. Take a random sample: $X_1, \ldots, X_n$
2. Average what you get: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$\bar{x}$ is an Estimate = Result of Procedure = Constant

- Result of taking a random sample was the dataset: $x_1, \ldots, x_n$
- Result of averaging the observed data was $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sampling Distribution of $\bar{X}_n$

Thought experiment: suppose I were to repeat the procedure of taking the mean of a random sample over and over forever. What relative frequencies would I get for the sample means?

Lecture #14 – Sampling Distributions and Estimation II

Bias of an Estimator

Why divide by $n-1$ in sample variance?

Biased Sampling and the Candy-Weighing Experiment

Efficiency: Choosing between Unbiased Estimators

Mean-Squared Error: Choosing Between Biased Estimators

Consistency and the Law of Large Numbers

Unbiased means "Right on Average"

Bias of an Estimator

Let $\widehat{\theta}_n$ be a sample estimator of a population parameter θ_0. The *bias* of $\widehat{\theta}_n$ is $E[\widehat{\theta}_n] - \theta_0$.

Unbiased Estimator

A sample estimator $\widehat{\theta}_n$ of a population parameter θ_0 is called *unbiased* if $E[\widehat{\theta}_n] = \theta_0$

Why $(n-1)$ for sample variance?

We will show that having $n-1$ in the denominator ensures:

$$E[S^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \sigma^2$$

under random sampling.

Why $(n-1)$ for sample variance?

Step #1 – Extension Problem #3(b) gives:

$$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2=\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-n(\bar{X}-\mu)^2$$

Why $(n-1)$ for sample variance?

Step # 2 – Take Expectations of Step # 1:

$$
\begin{aligned}
E\left[\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right] &= E\left[\left\{\sum_{i=1}^{n}(X_i-\mu)^2\right\}-n(\bar{X}-\mu)^2\right] \\
&= E\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-E\left[n(\bar{X}-\mu)^2\right] \\
&= \sum_{i=1}^{n}E\left[(X_i-\mu)^2\right]-n\,E\left[(\bar{X}-\mu)^2\right]
\end{aligned}
$$

Where we have used the linearity of expectation.

Why $(n-1)$ for sample variance?

Step # 3 – Use assumption of random sampling:

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$\begin{aligned}
E\left[\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2\right] &= \sum_{i=1}^{n} E\left[(X_i - \mu)^2\right] - n\,E\left[(\bar{X} - \mu)^2\right] \\
&= \sum_{i=1}^{n} Var(X_i) - n\,E\left[(\bar{X} - E[\bar{X}])^2\right] \\
&= \sum_{i=1}^{n} Var(X_i) - n\,Var(\bar{X}) = n\sigma^2 - \sigma^2 \\
&= (n-1)\sigma^2
\end{aligned}$$

Since $E[\bar{X}] = \mu$ and $Var(\bar{X}) = \sigma^2/n$ under random sampling.

Why $(n-1)$ for sample variance?

Finally – Divide Step # 3 by $(n-1)$:

$$E[S^2] \;\; = \;\; E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right] = \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$

Hence, having $(n-1)$ in the denominator ensures that the sample variance is "correct on average," that is *unbiased*.

A Different Estimator of the Population Variance

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

$$E[\widehat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \frac{(n-1)\sigma^2}{n}$$

Bias of $\widehat{\sigma}^2$

$$E[\widehat{\sigma}^2] - \sigma^2 = \frac{(n-1)\sigma^2}{n} - \sigma^2 = \frac{(n-1)\sigma^2}{n} - \frac{n\sigma^2}{n} = -\sigma^2/n$$

How Large is the Average Family?

How many brothers and sisters are in your family, including yourself?

What's Going On Here?

Twenty years ago the average number of children per family was about 2.0. But our average was much higher!

Biased Sample!

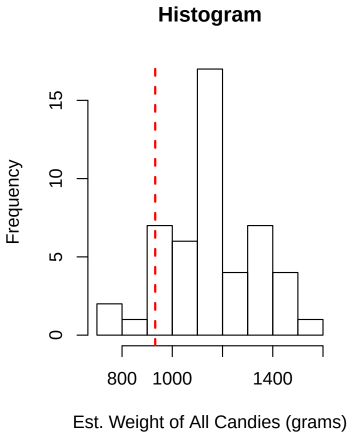
- Zero children ⇒ didn't send any to college
- Sampling by *children* so large families oversampled

Candy Weighing: 49 Estimates, Each With $n = 5$

$$\widehat{\theta} = 20 \times (X_1 + \ldots + X_5)$$

Summary of Sampling Dist.	
Overestimates	45
Exactly Correct	0
Underestimates	4
$E[\hat{\theta}]$	1164 grams
$\text{SD}(\widehat{\theta})$	189 grams

Actual Mass: θ_0 =932 grams



What was in the bag?

100 Candies Total:

- 20 Fun Size Snickers Bars (large)
- 30 Reese's Miniatures (medium)
- 50 Tootsie Roll "Midgees" (small)

So What Happened?

Not a random sample! The Snickers bars were *oversampled*.

Could we have avoided this? How?

Let $X_1, X_2, \ldots X_n \sim$ *iid* mean μ, variance σ^2. True or False:

X_1 *is an unbiased estimator of* μ

(a) True

(b) False

TRUE!

How to choose between two unbiased estimators?

Suppose $X_1, X_2, \ldots X_n \sim iid$ with mean μ and variance σ^2

From Last Lecture:

$$E[\bar{X}_n] = \mu, \quad Var(\bar{X}_n) = \sigma^2/n$$

Compared To:

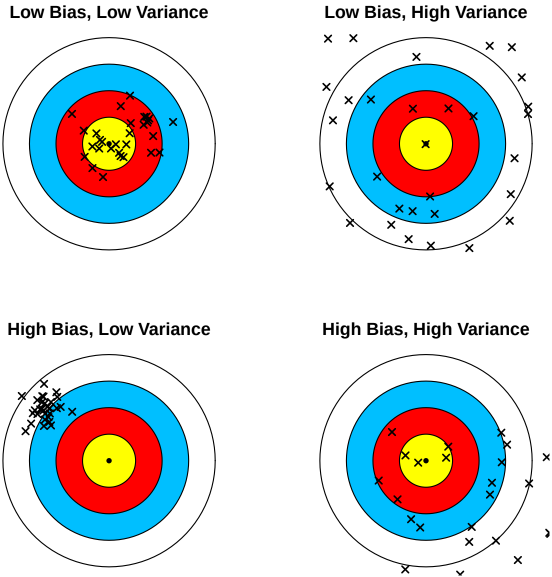
$$E[X_1] = \mu, \quad Var(X_1) = \sigma^2$$

Both $\bar{X}_n$ and X_1 are unbiased estimators of μ, but $\bar{X}_n$ has a lower variance!

Efficiency - Compare Unbiased Estimators by Variance

Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be unbiased estimators of θ_0. We say that $\widehat{\theta}_1$ is *more efficient* than $\widehat{\theta}_2$ if $Var(\widehat{\theta}_1) < Var(\widehat{\theta}_2)$.

Bias and Variance are Both Bad Things



Low Bias, Low Variance

Low Bias, High Variance

High Bias, Low Variance

High Bias, High Variance

Mean-Squared Error: Trading Bias Against Variance

- Unbiased estimator with a huge bias is bad.
- Highly biased estimator with a low variance is bad.
- Often there is a “tradeoff” between bias and variance:
  - Low bias estimators often have high variance.
  - Low variance estimators often have high bias.

Mean-Squared Error (MSE):

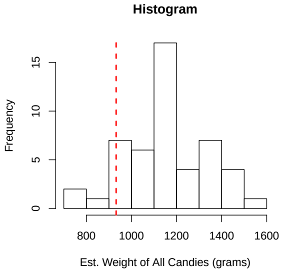
Compare estimators accounting for both bias and variance:

$$MSE(\widehat{\theta}) = \text{Bias}(\widehat{\theta})^2 + Var(\widehat{\theta})$$

Root Mean-Squared Error (RMSE): $\sqrt{\text{MSE}}$

Calculate MSE for Candy Experiment

$E[\hat{\theta}]$	1164 grams
θ_0	932 grams
$SD(\widehat{\theta})$	189 grams



$$\begin{aligned}
\text{Bias} &= 1164 \text{ grams} - 932 \text{ grams} \\
&= 232 \text{ grams} \\
\text{MSE} &= \text{Bias}^2 + \text{Variance} \\
&= (232^2 + 189^2) \text{ grams}^2 \\
&= 8.9545 \times 10^4 \text{ grams}^2 \\
\text{RMSE} &= \sqrt{\text{MSE}} = 299 \text{ grams}
\end{aligned}$$

Finite Sample versus Asymptotic Properties of Estimators

Finite Sample Properties

For *fixed sample size n* what are the properties of the sampling distribution of $\widehat{\theta}_n$? (E.g. bias and variance.)

Asymptotic Properties

What happens to the sampling distribution of $\widehat{\theta}_n$ *as the sample size n gets larger and larger?*

1. Law of Large Numbers (today)
2. Central Limit Theorem (Lecture 16)

Consistency

Definition

We say that an estimator $\widehat{\theta}_n$ is *consistent for* a parameter θ_0 if $\lim_{n\to\infty} \text{MSE}(\widehat{\theta}_n) = 0$, in other words, if both the bias and variance of $\widehat{\theta}_n$ disappear as the sample size grows.

Intuitively, this means $\widehat{\theta}_n$ becomes "less random" as the sample size increases, eventually converging to a constant: θ_0.

Law of Large Numbers

Let $X_1, X_2, \ldots X_n \sim iid$ mean μ, variance σ^2. Then the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is consistent for the population mean μ.

How do we know this?

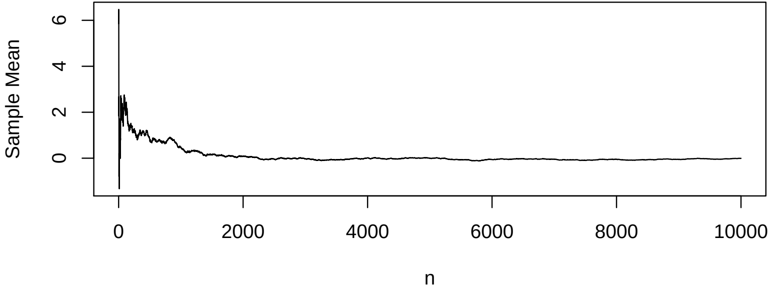
From our last lecture:

$$E[\bar{X}_n] = \mu, \quad Var(\bar{X}_n) = \sigma^2/n$$

and hence:

$$\begin{aligned}
\text{MSE}(\bar{X}_n) &= \text{Bias}(\bar{X}_n)^2 + Var(\bar{X}_n) \\
&= \left(E[\bar{X}_n] - \mu\right)^2 + Var(\bar{X}_n) \\
&= 0 + \sigma^2/n \to 0
\end{aligned}$$

```r
set.seed(12345)
n <- 10000
x <- rnorm(n, mean = 0, sd = 10)
xbar_n <- cumsum(x) / (1:n)
plot(xbar_n, type = 'l', xlab = 'n', ylab = 'Sample Mean')
```



Lecture #15 – Confidence Intervals I

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

Interpreting a Confidence Interval

Margin of Error and Width

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 $X_1, X_2, \ldots, X_n \sim$ 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 $X_1, X_2, \ldots, X_n \sim$ 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.

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

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

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?

$$
\begin{aligned}
P\left(-2 \leq \frac{\bar{X}_n - \mu}{SE(\bar{X}_n)} \leq 2\right) &= 0.95 \\
P\left(-2 \cdot SE \leq \bar{X}_n - \mu \leq 2 \cdot SE\right) &= 0.95 \\
P\left(-2 \cdot SE - \bar{X}_n \leq -\mu \leq 2 \cdot SE - \bar{X}_n\right) &= 0.95 \\
P\left(\bar{X}_n - 2 \cdot SE \leq \mu \leq \bar{X}_n + 2 \cdot SE\right) &= 0.95
\end{aligned}
$$

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

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

But how are we supposed to interpret this?

Confidence Interval is a Random Variable!

1. $X_1, \ldots, X_n$ 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

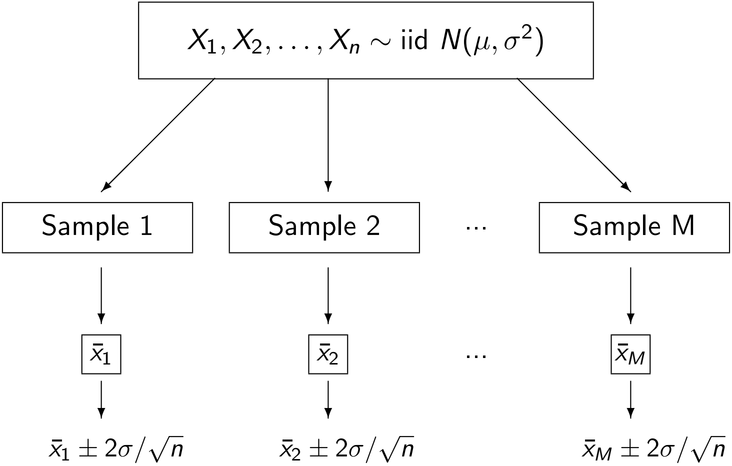
Formal Meaning

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



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

Sample 1

Sample 2

$\cdots$

Sample M

$\boxed{\bar{x}_1}$

$\boxed{\bar{x}_2}$

$\cdots$

$\boxed{\bar{x}_M}$

$\bar{x}_1 \pm 2\sigma/\sqrt{n}$

$\bar{x}_2 \pm 2\sigma/\sqrt{n}$

$\bar{x}_M \pm 2\sigma/\sqrt{n}$

Repeat M times → get M different intervals

Large M ⇒ Approx. 95% of these Intervals Contain μ

Simulation Example: $X_1, \ldots, X_5 \sim$ iid $N(0,1)$, $M = 20$

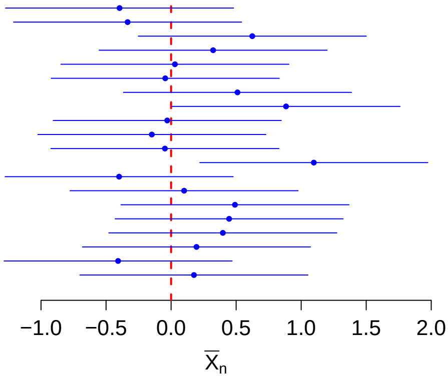


Figure: Twenty confidence intervals of the form $\bar{X}_n \pm 2\sigma/\sqrt{n}$ where $n = 5$, $\sigma^2 = 1$ and the true population mean is 0.

Meaning of Confidence Interval for θ_0

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

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 $\widehat{\theta} \pm ME$, 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 $|\text{UCL} - \text{LCL}|$ 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) $\sigma/\sqrt{n}$

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

(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

$X_1, \ldots, X_{100} \sim$ 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

$$\boxed{\begin{array}{l} X_1, \ldots, X_{100} \sim N(\mu, 1) \text{ but we don't know } \mu. \\ \text{Want to create a 95\% confidence interval for } \mu. \end{array}}$$

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

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

Example: Similarly for the Upper Confidence Limit...

$X_1, \ldots, X_{100} \sim N(\mu, 1)$ 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?

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

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

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

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.

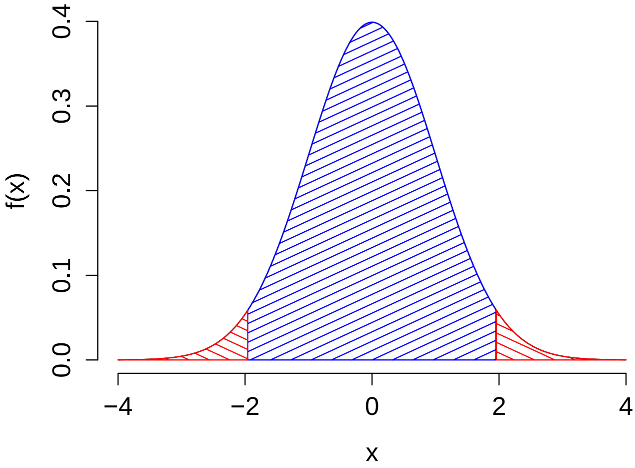
What value of c should we use to get a $100\times(1-\alpha)\%$ CI for μ?

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

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

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

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


0.0
0.1
0.2
0.3
0.4
f(x)
−4
−2
0
2
4
x

What Affects the Margin of Error?

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

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
$\texttt{qnorm}(1-\alpha/2)$	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

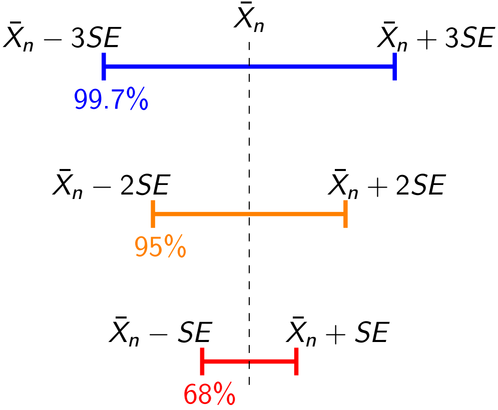


Figure: Each CI gives a range of "plausible" values for the population mean μ, centered at the sample mean $\bar{X}_n$. Values near the middle are "more plausible" in the sense that a small reduction in confidence level gives a much shorter interval centered in the same place. This is because the sample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

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

σ Known

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

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

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: $X_1, \ldots, X_5 \sim$ iid N(μ, σ^2)

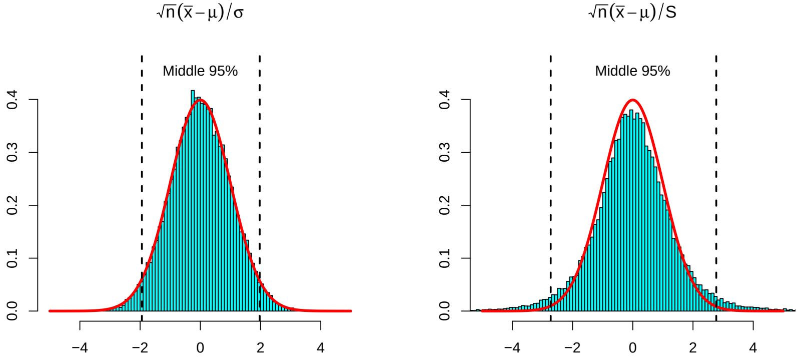


Figure: In each plot the red curve is the pdf of the standard normal RV.
At left: the sampling distribution of $\sqrt{5}(\bar{X}_5 - \mu)/\sigma$ is standard normal.
At right: the sampling distribution of $\sqrt{5}(\bar{X}_5 - \mu)/S$ clearly isn't!

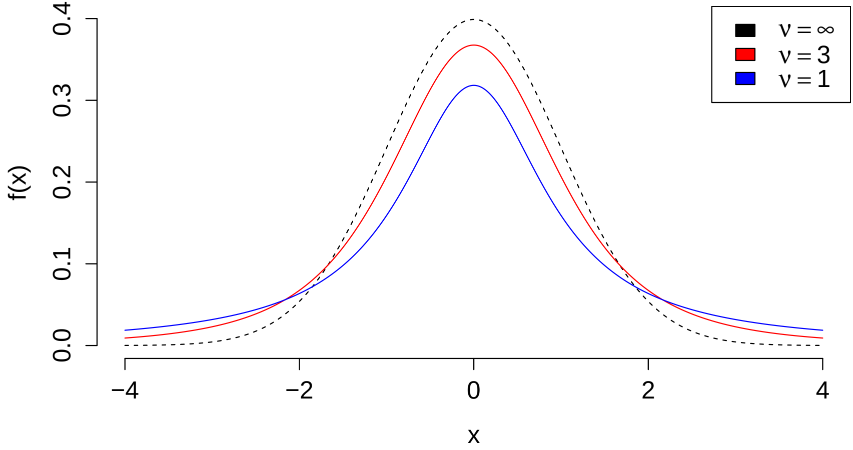
Student-t Random Variable

If $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$, then

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

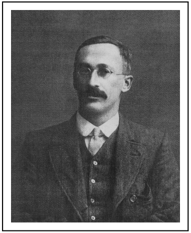
- Parameter: $\nu = n - 1$ “degrees of freedom”
- Support $= (-\infty, \infty)$
- Symmetric around zero, but mean and variance may not exist!
- Degrees of freedom ν control “thickness of tails”
- As $\nu \to \infty$, $t \to$ Standard Normal.

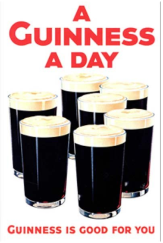
Student-t PDFs



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. Jamess 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:

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

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

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

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

Comparison of CIs for Mean of Normal Distribution

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

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

Known Population Std. Dev. (σ)

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

Unknown Population Std. Dev. (σ)

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

Comparison of Normal and t CIs

Table: Values of $\texttt{qt}(1-\alpha/2, \texttt{df} = n-1)$ 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 $n \to \infty$, $t(n-1) \to N(0,1)$

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)

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

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

Assuming the population is normal,

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

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?

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

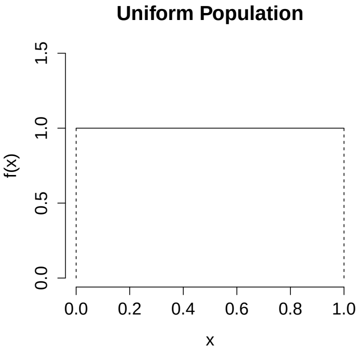
The Central Limit Theorem

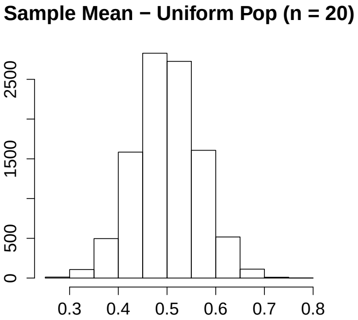
Suppose that $X_1, \ldots, X_n$ are a random sample from a some population that is not necessarily normal and has an unknown mean μ. Then, provided that n is *sufficiently large*,

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

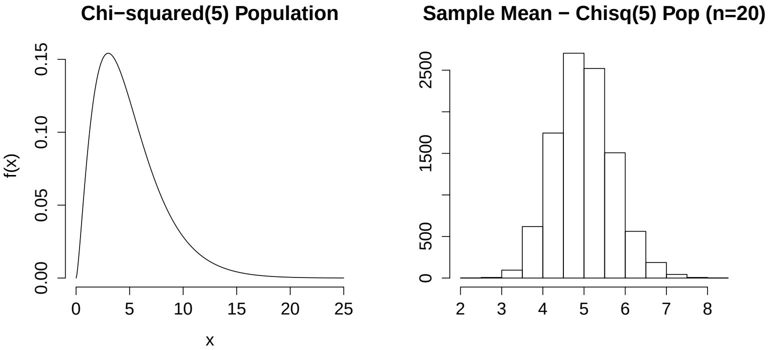
We will use this fact to create *approximate* CIs for population mean even if we know *nothing* about the population.

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





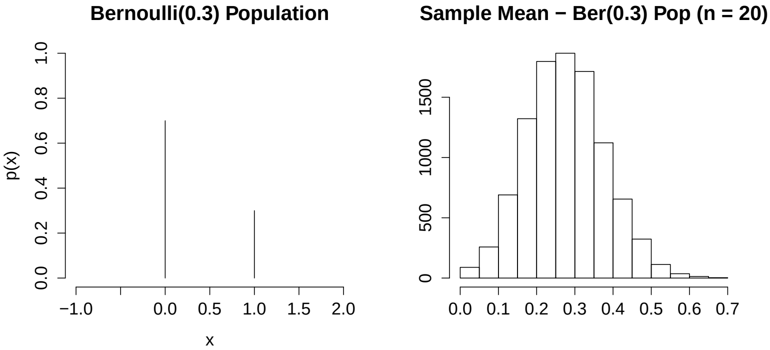
Example: $\chi^2(5)$ Population, $n = 20$



Chi−squared(5) Population

Sample Mean − Chisq(5) Pop (n=20)

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



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?

$X_1, \ldots, X_n \sim$ 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: $\widehat{p} = (\sum_{i=1}^{n} X_i)/n$

Sample Proportion *is* the Sample Mean!

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

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

$$\begin{aligned}
E[\widehat{p}] &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{np}{n} = p \\
Var(\widehat{p}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \\
SE(\widehat{p}) &= \sqrt{Var(\widehat{p})} = \sqrt{\frac{p(1-p)}{n}} \\
\widehat{SE}(\widehat{p}) &= \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}
\end{aligned}$$

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

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

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

Approximate 95% CI for Population Proportion

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

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

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

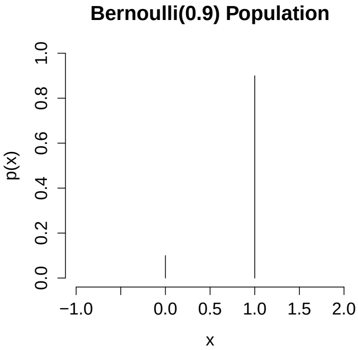
$100 \times (1 - \alpha)$ CI for Population Proportion (p)

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

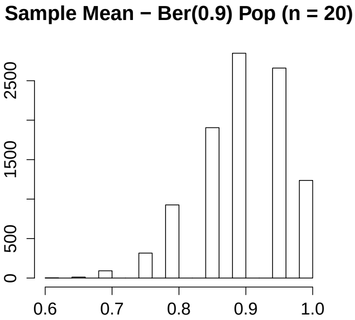
$$\widehat{p} \pm \texttt{qnorm}(1 - \alpha/2)\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

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, $n = 20$

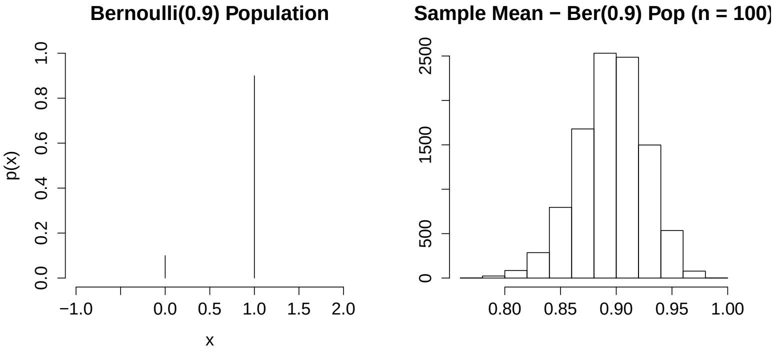


Bernoulli(0.9) Population



Sample Mean − Ber(0.9) Pop (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

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

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?

$$\boxed{\text{Approximate 95\% CI: } (0.35, 0.43)}$$

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 $X_1, \ldots, X_n \sim$ iid $N(\mu_x, \sigma_x^2)$ indep. of $Y_1, \ldots, Y_m \sim$ 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

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

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

CI for Difference of Population Means Using CLT

Setup: Independent Random Samples

$X_1, \ldots, X_n \sim$ iid with unknown mean μ_X & unknown variance σ_X^2

$Y_1, \ldots, Y_m \sim$ 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:

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

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)

$X_1, \ldots, X_n \sim$ iid with mean μ_X and variance σ_X^2

$Y_1, \ldots, Y_m \sim$ iid with mean μ_Y and variance σ_Y^2

where each sample is independent of the other

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

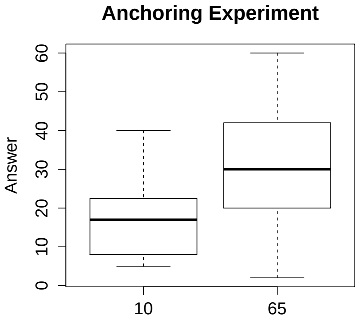
$$\widehat{SE}(\bar{X}_n - \bar{Y}_m) = \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$$

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.

Past Semester's Anchoring Experiment



"Lo" Group – Shown 10

$m_{Lo} = 43$

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

$s^2_{Lo} = 86$

"Hi" Group – Shown 65

$n_{Hi} = 46$

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

$s^2_{Hi} = 253$

ME for approx. 95% for Difference of Means

“Lo” Group

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

“Hi” Group

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

$$\begin{aligned} \bar{X}_{Hi} - \bar{Y}_{Lo} &= 30.7 - 17.1 = 13.6 \\ \widehat{SE}(\bar{X}_{Hi} - \bar{Y}_{Lo}) &= \sqrt{\widehat{SE}(\bar{X}_{Hi})^2 + \widehat{SE}(\bar{Y}_{Lo})^2} = \sqrt{7.5} \approx 2.7 \Rightarrow ME \approx 5.4 \end{aligned}$$

$\boxed{\text{Approximate 95\% CI} \quad (8.2, 19)}$ What can we conclude?

Confidence Interval for a Difference of Proportions via CLT

What is the appropriate probability model for the sample?

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

$Y_1, \ldots, Y_m \sim$ 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: $\widehat{p} - \widehat{q}$ where:

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

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

What We Have

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

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

What We Want

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

Use Independence of the Two Samples

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

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

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

$Y_1, \ldots, Y_m \sim$ iid Bernoulli(q)

where each sample is independent of the other

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

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

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

$$\begin{aligned} \widehat{p} &= 0.47 \\ n &= 239 \\ \widehat{SE}(\widehat{p}) &= \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \approx 0.032 \end{aligned}$$

Democrats

$$\begin{aligned} \widehat{q} &= 0.31 \\ m &= 238 \\ \widehat{SE}(\widehat{q}) &= \sqrt{\frac{\widehat{q}(1-\widehat{q})}{m}} \approx 0.030 \end{aligned}$$

Difference: (Republicans - Democrats)

$$\begin{aligned} \widehat{p}-\widehat{q} &= 0.47-0.31=0.16 \\ \widehat{SE}(\widehat{p}-\widehat{q}) &= \sqrt{\widehat{SE}(\widehat{p})^2+\widehat{SE}(\widehat{q})^2} \approx 0.044 \implies ME \approx 0.09 \end{aligned}$$

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
⋮	⋮	⋮	⋮
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

$$\boxed{\text{Let } D_i = X_i - Y_i \text{ be the difference of student } i\text{'s exam scores.}}$$

I calculated the following in R:

$$\begin{aligned}
\bar{D}_n &= \frac{1}{n}\sum_{i=1}^{n} D_i \approx 1.8 \\
S_D^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2 \approx 124 \\
\widehat{SE}(\bar{D}_n) &= (S_D/\sqrt{n}) \approx \sqrt{124/71} \approx 1.3
\end{aligned}$$

Approximate 95% CI Based on the CLT:

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

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

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

$$\begin{aligned} \bar{D}_n &= 1.8 \\ \bar{X}_n - \bar{Y}_n &= 81.4 - 79.6 = 1.8 \checkmark \end{aligned}$$

...But Correlation Affects the Variance

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

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

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

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

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	$\boxed{?}$
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

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

Correct Interval – Matched Pairs

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

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

CIs for a Difference of Means – Two Cases

Independent Samples

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

Matched Pairs

Matched pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ 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}