

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2006, The Johns Hopkins University and John McGready. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Confidence Intervals

John McGready
Johns Hopkins University

Lecture Topics

- ◆ Variability in the sampling distribution
- ◆ Standard error of the mean
- ◆ Standard error vs. standard deviation
- ◆ Confidence intervals for the population mean μ
- ◆ Standard error for a proportion
- ◆ Confidence intervals for a proportion



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section A

*Variability in the Sampling
Distribution;
Standard Error of a Sample
Statistic*

Random Sample

- ◆ When a sample is randomly selected from a population, it is called a *random sample*
- ◆ In a simple random sample, each individual in the population has an equal chance of being chosen for the sample

Random Sample

- ◆ Random sampling helps control systematic bias
- ◆ But even with random sampling, there is still *sampling variability* or error

Sampling Variability

- ◆ If we repeatedly choose samples from the same population, a statistic will take different values in different samples

Idea

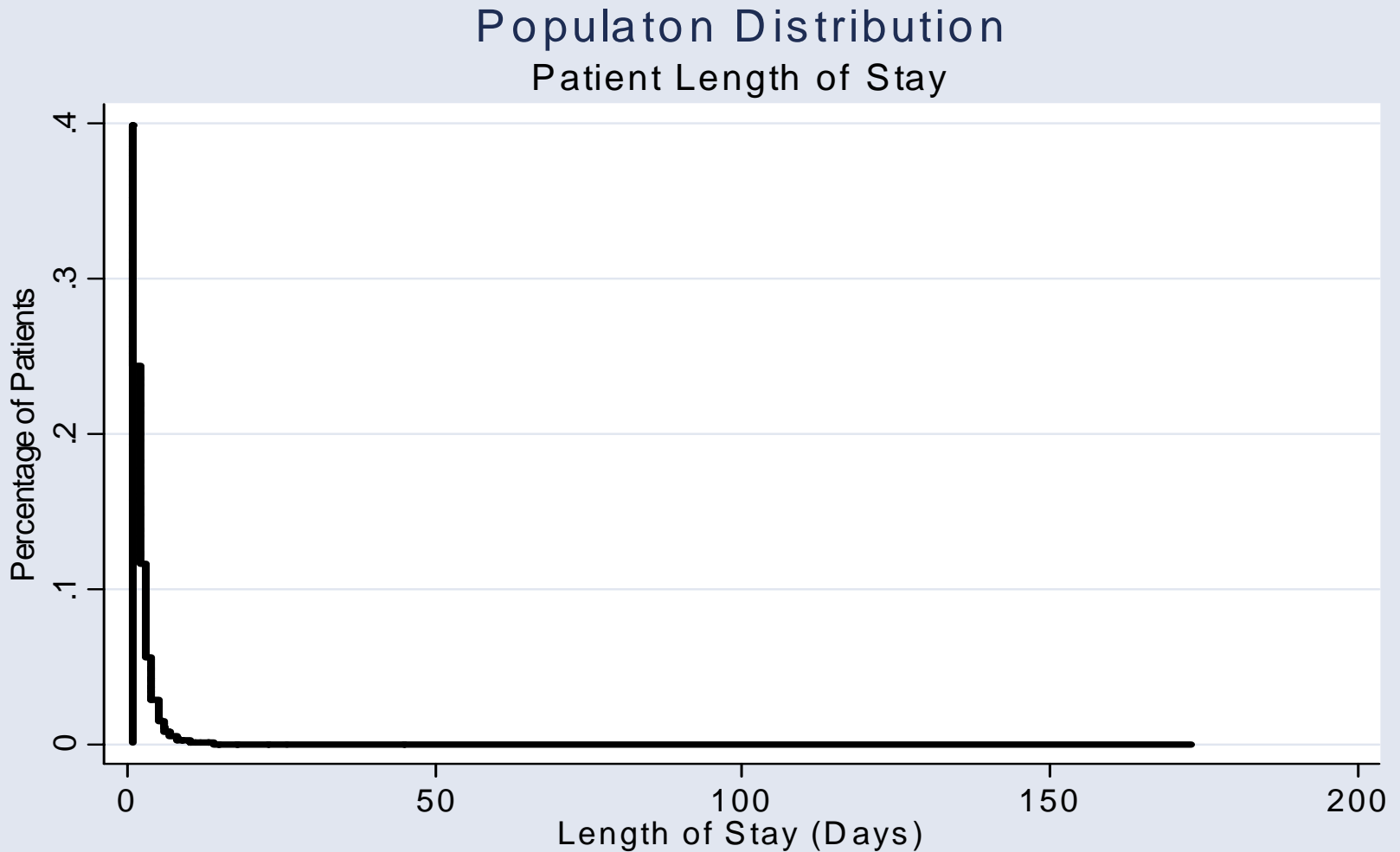
- ◆ If the statistic does not change much if you repeated the study (you get the similar answers each time), then it is fairly reliable (not a lot of variability)

Example:

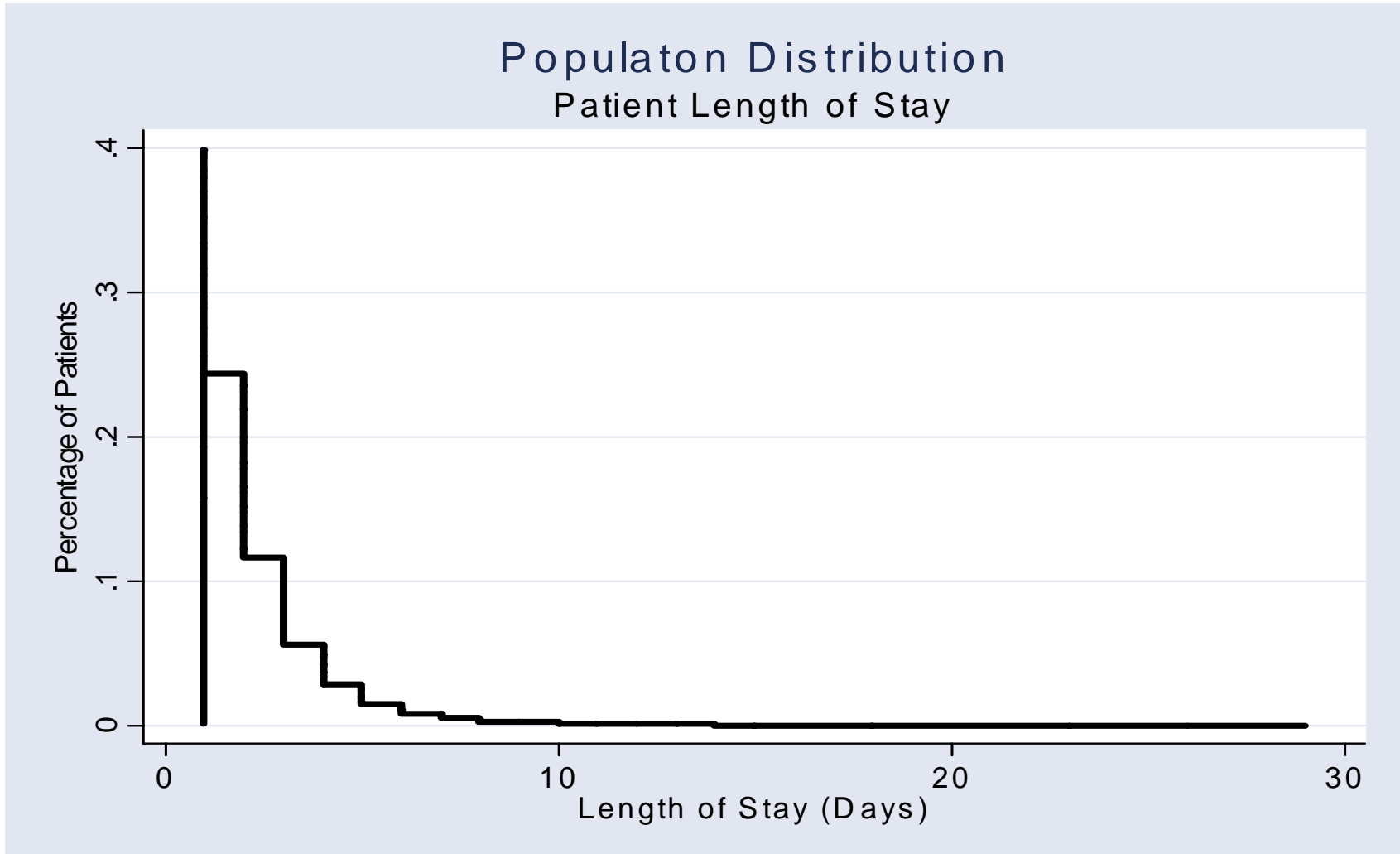
Hospital Length of Stay

- ◆ The distribution of the length of stay information for the population of patients discharged from a major teaching hospital in a one year period is a heavily right skewed distribution
 - Mean, 5.0 days, SD 6.9 days
 - Median, 3 days
 - Range 1 to 173 days

Population Distribution: Hospital Length of Stay



Population Distribution: Hospital Length of Stay



Hospital Length of Stay

- ◆ Suppose I have a random sample of 10 patients discharged from this hospital
- ◆ I wish to use the sample information to estimate average length of stay at the hospital
- ◆ The sample mean is 5.7 days
- ◆ How “good” an estimate is this of the population mean?

Hospital Length of Stay

- ◆ Suppose I take another random sample of 10 patients . . . and the sample mean length of stay for this sample is 3.9 days
- ◆ I do this a third time, and get a sample mean of 4.6 days

Hospital Length of Stay

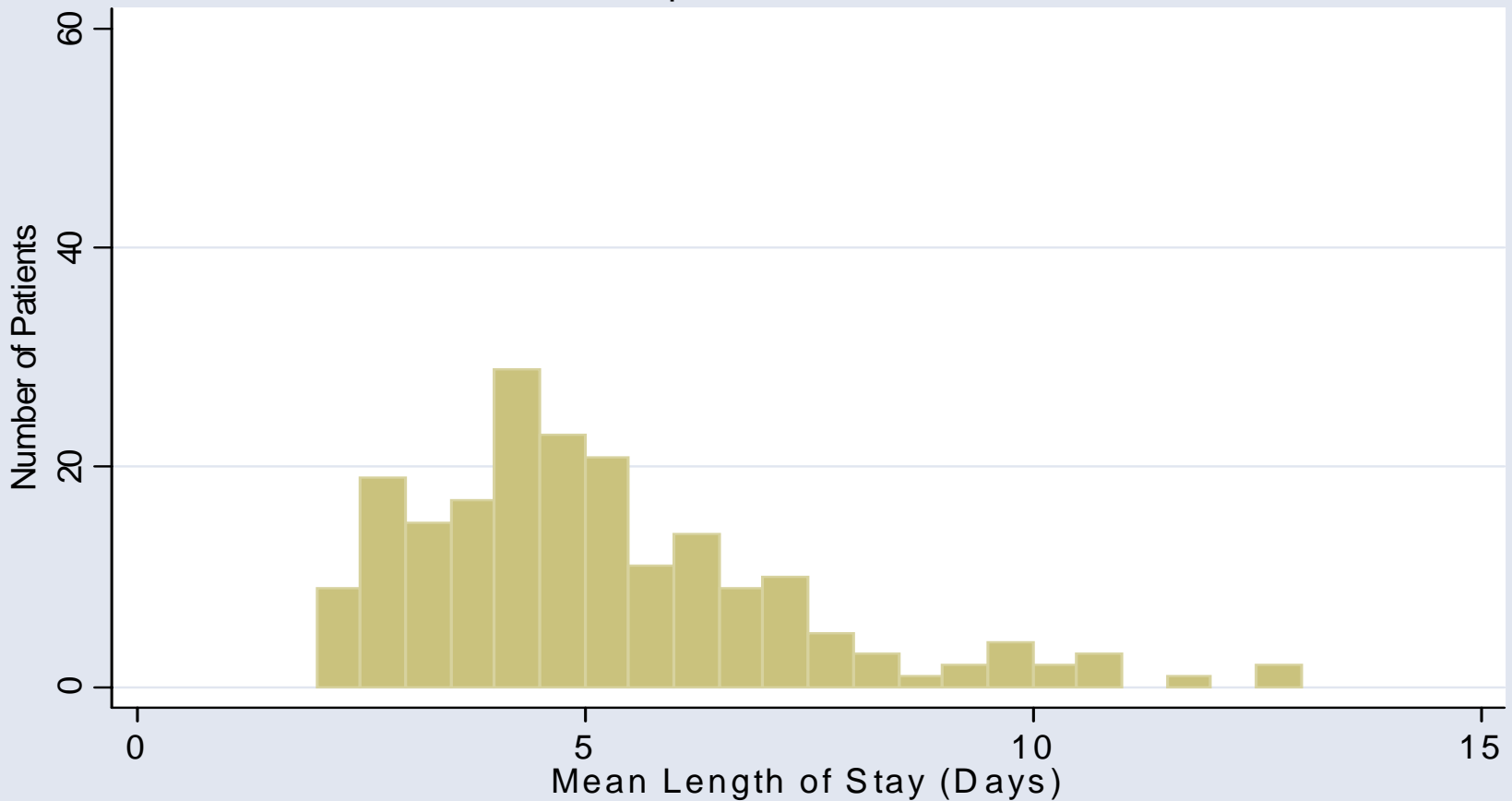
- ◆ Suppose I did this 200 times
- ◆ If I want to get a handle on the behavior of my sample mean estimate from sample to sample is to plot a histogram of my 200 sample mean values

The Sampling Distribution

- ◆ The *sampling distribution of the sample mean* refers to what the distribution of the sample means would look like if we were to choose a large number of samples, each of the same size from the same population, and compute a mean for each sample

Sampling Distribution, $n = 10$

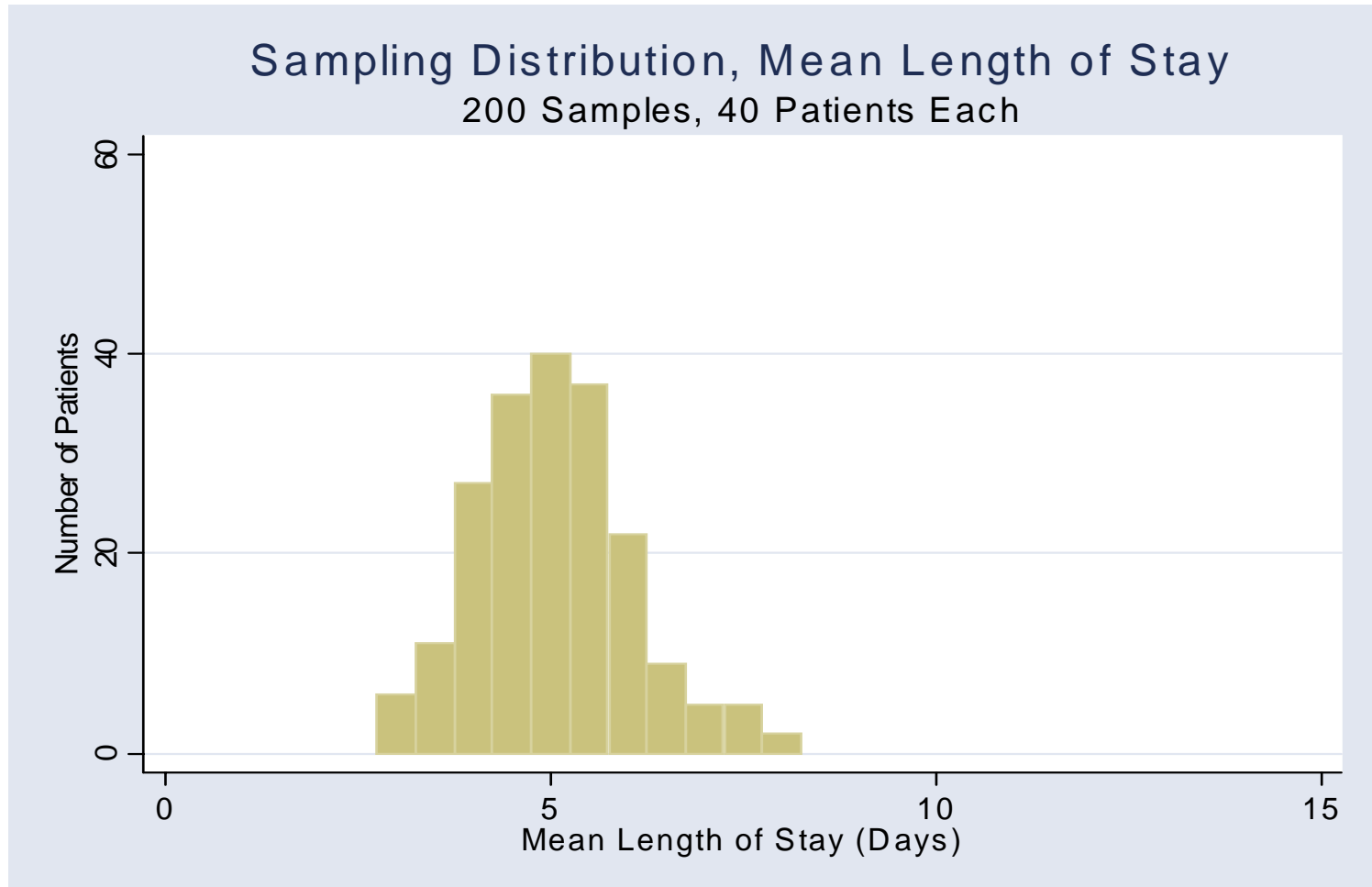
Sampling Distribution, Mean Length of Stay
200 Samples, 10 Patients Each



Sampling Distribution, $n = 40$

- ◆ Suppose I again took 200 random samples, but this time, each sample had 40 patients
- ◆ Again, I plot a histogram of the 200 sample mean values

Sampling Distribution, $n = 40$

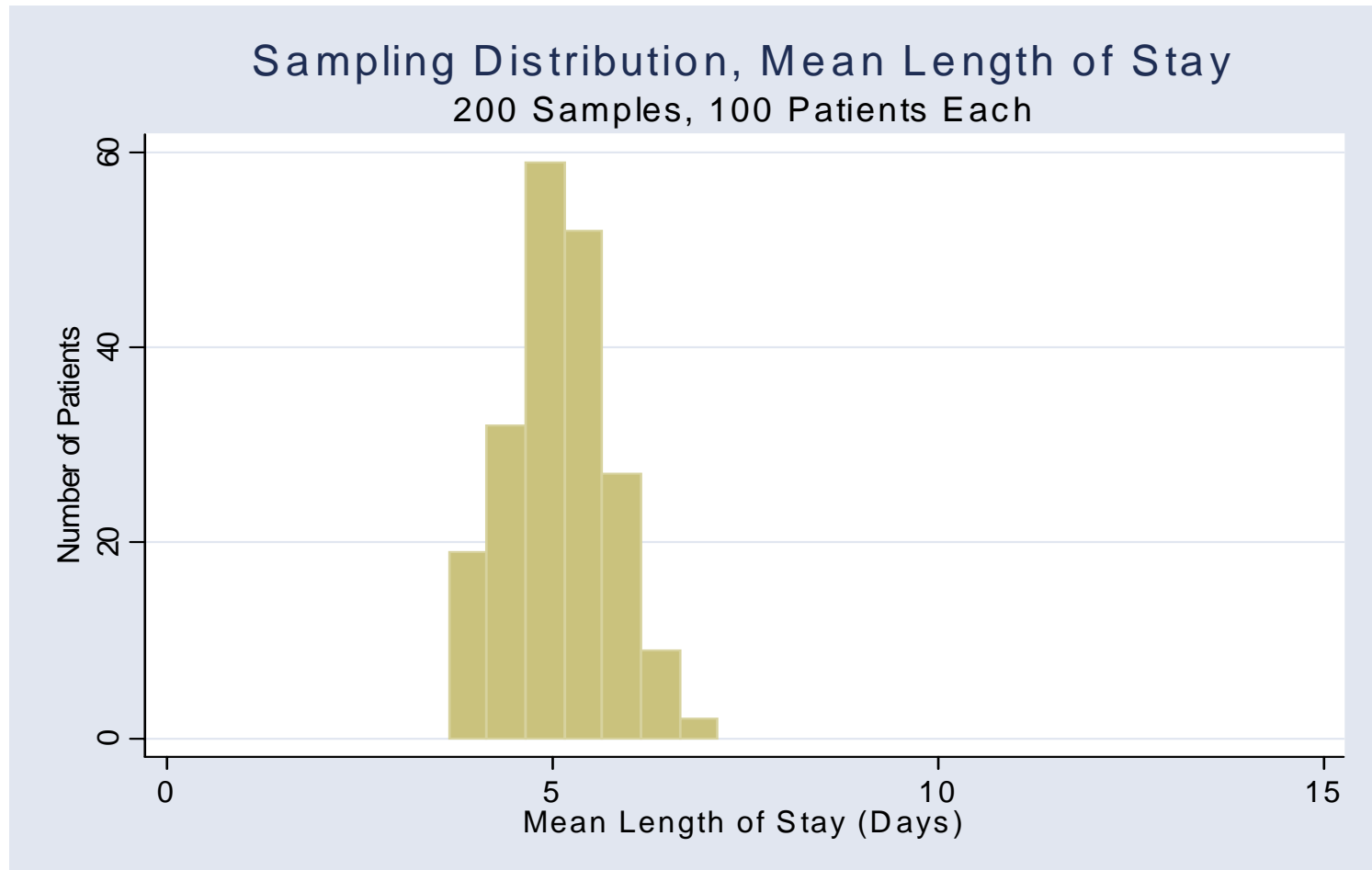


Mean length of stay from 200 samples, each of size $n = 40$

Sampling Distribution, $n = 40$

- ◆ Suppose I again took 200 random samples, but this time, each sample had 100 patients
- ◆ Again, I plot a histogram of the 200 sample mean values

Central Limit Theorem

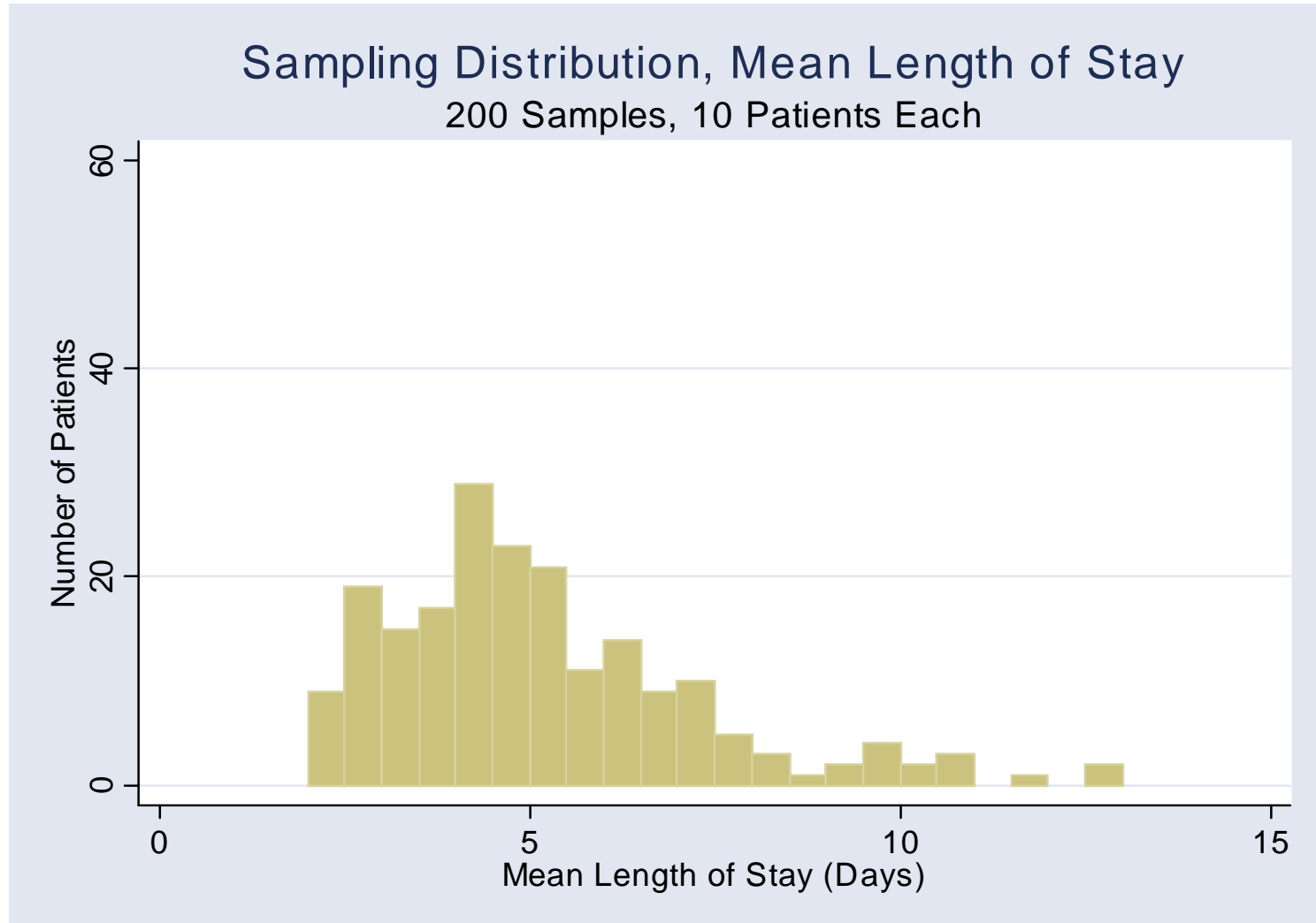


Mean length of stay from 200 samples, each of size $n = 100$

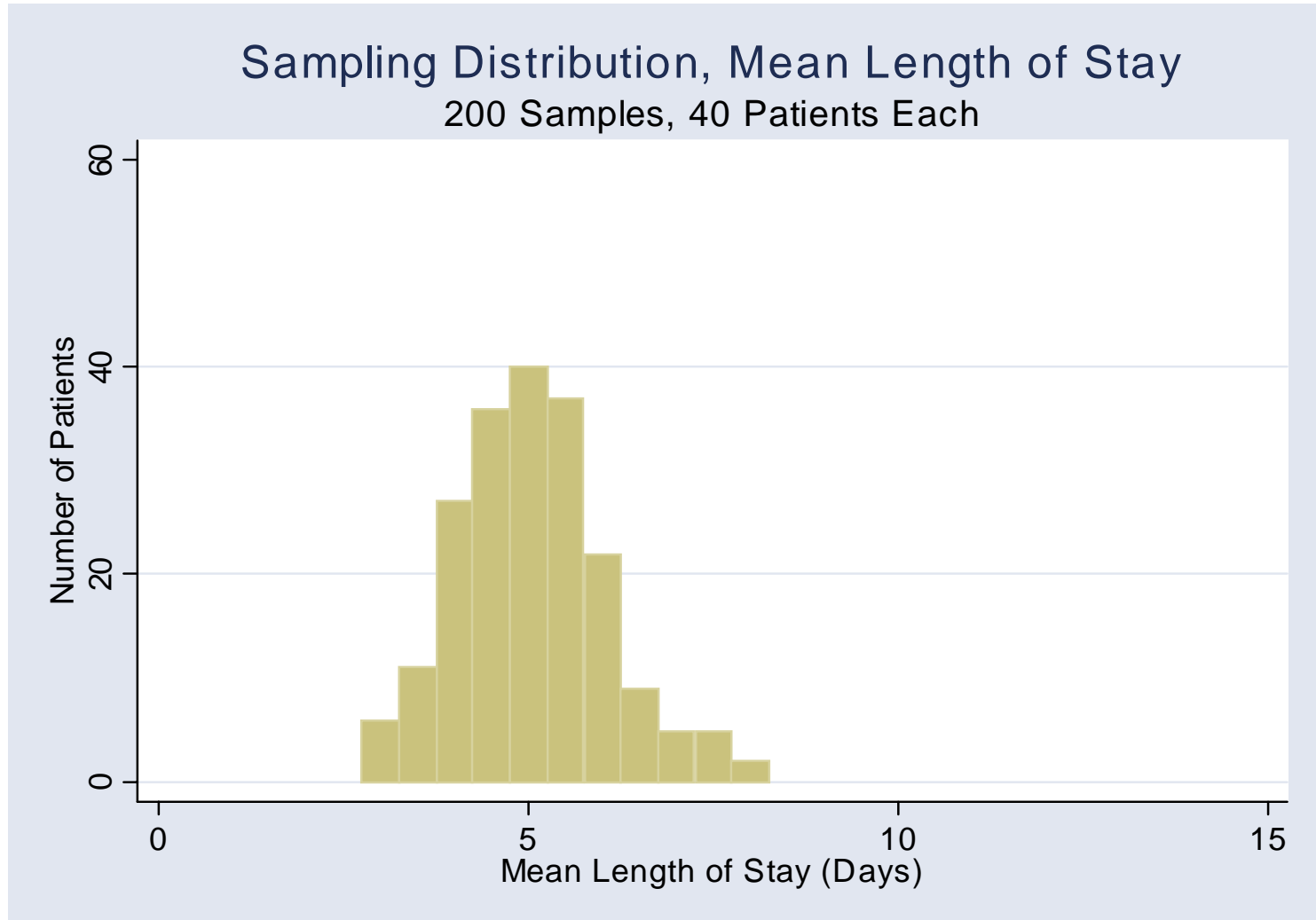
Comparing Sampling Distributions

- ◆ Did you notice any pattern regarding the sampling distributions and the size of the samples from which the means were computed?
 - Distribution gets “tighter” when means is based on larger samples
 - Distribution looks less like distribution of individual data, more like a “normal” curve

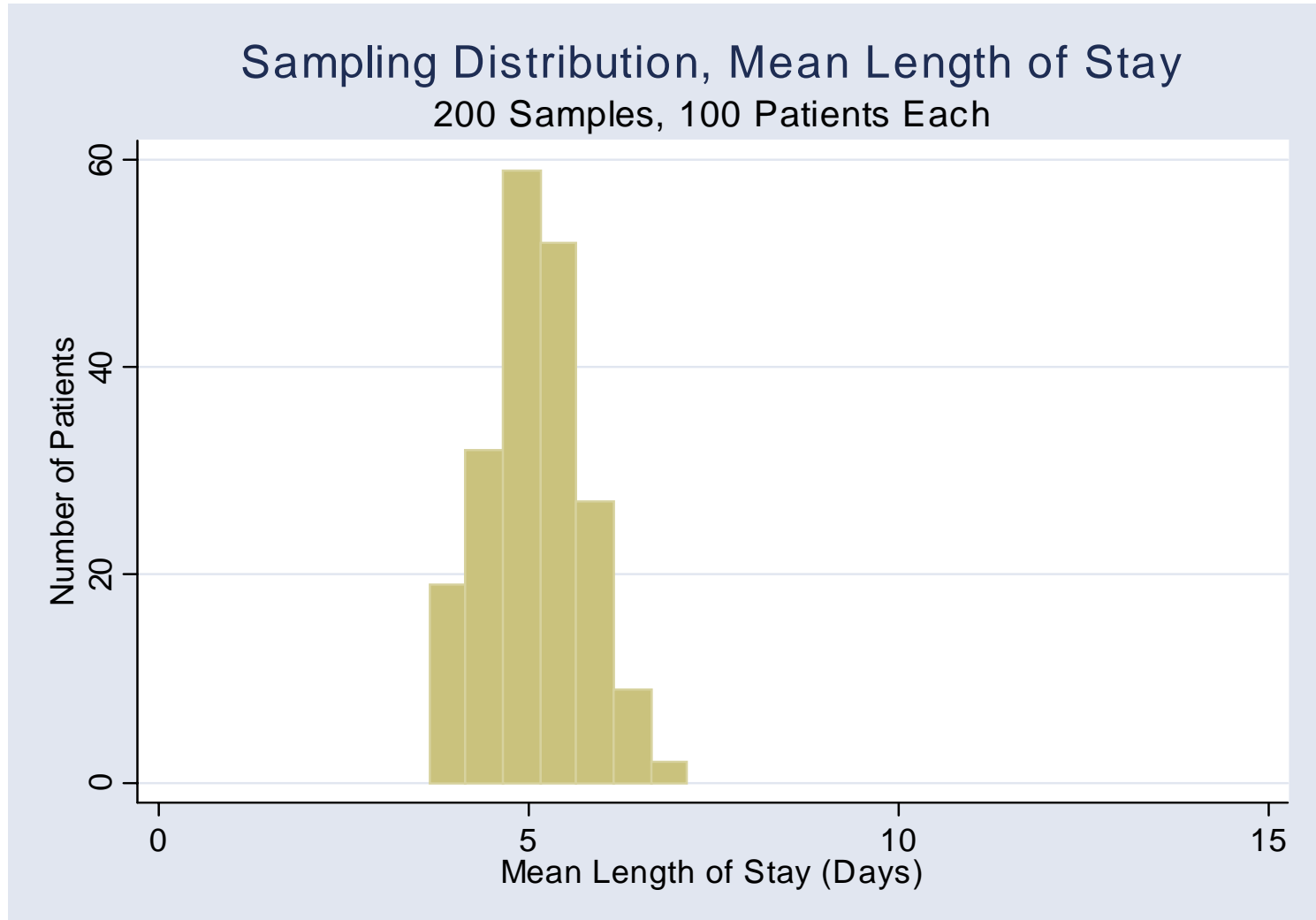
Sampling Distribution, $n = 10$



Sampling Distribution, $n = 40$



Sampling Distribution, $n = 100$



Amazing Result

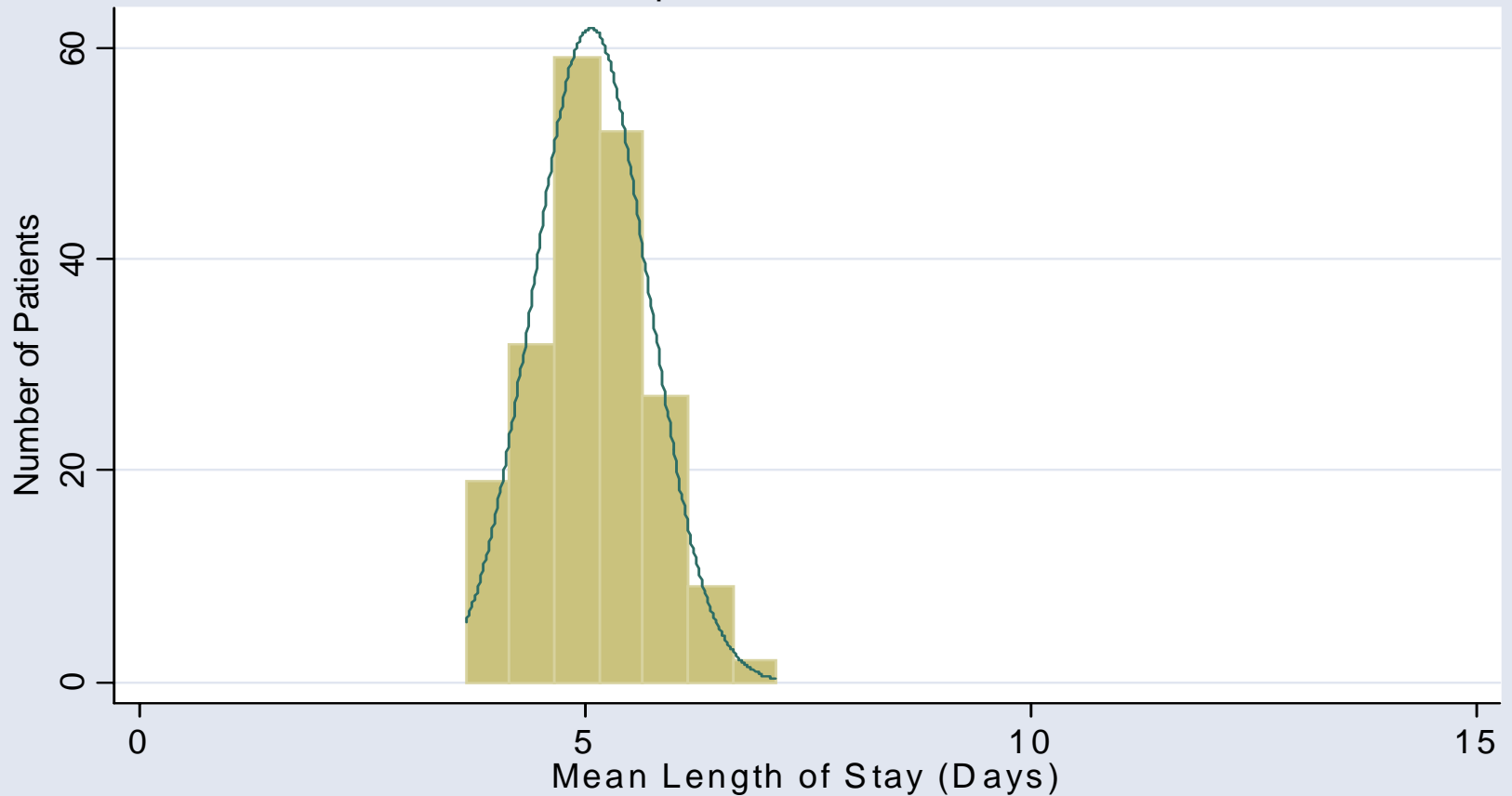
- ◆ Mathematical statisticians have figured out how to predict what the sampling distribution will look like without actually repeating the study numerous times and having to choose a sample each time

Amazing Result

- ◆ Often, the sampling distribution of a sample statistics will look “normally” distributed
 - This happens for sample means and sample proportions
 - This happens for sample mean differences and differences in sample proportions

Sampling Distribution

Sampling Distribution, Mean Length of Stay
200 Samples, 100 Patients Each



200 samples of size 100; a normal probability density is superimposed on the histogram

The Big Idea

- ◆ It's not practical to keep repeating a study to evaluate sampling variability and to determine the sampling distribution

The Big Idea

- ◆ Mathematical statisticians have figured out how to calculate it without doing multiple studies
- ◆ The sampling distribution of a statistic is often a *normal distribution*

The Big Idea

- ◆ This mathematical result comes from the *central limit theorem*
 - For the theorem to work, it requires the sample size (n) to be large
 - “Large sample size” means different things for different sample statistics
 - For sample means, the standard rule is $n > 60$ for the Central Limit Theorem to kick in

The Big Idea

- ◆ Statisticians have derived formulas to calculate the standard deviation of the sampling distribution
 - It's called the *standard error of the statistic*

Central Limit Theorem

- ◆ If the sample size is large, the distribution of sample means approximates a normal distribution

Beauty of Central Limit Theorem

- ◆ The central limit theorem (CLT) works even when the population is not normally distributed (or even continuous!)

Example

- ◆ Estimate the proportion of persons in a population who have health insurance; choose a sample of size $n = 100$
- ◆ The true proportion of individuals in this population is .80

Population Density



Example

- ◆ Sample 1

$$n = 100 \quad \hat{p} = \frac{83}{100} = .83$$

Example

- ◆ Is the sample proportion reliable?
 - If we took another sample of another 100 persons, would the answer bounce around a lot?

Example

Sample 1

$$\hat{p} = \frac{83}{100} = .83$$

Sample 2

$$\hat{p} = \frac{81}{100} = .81$$

Example

Sample 3

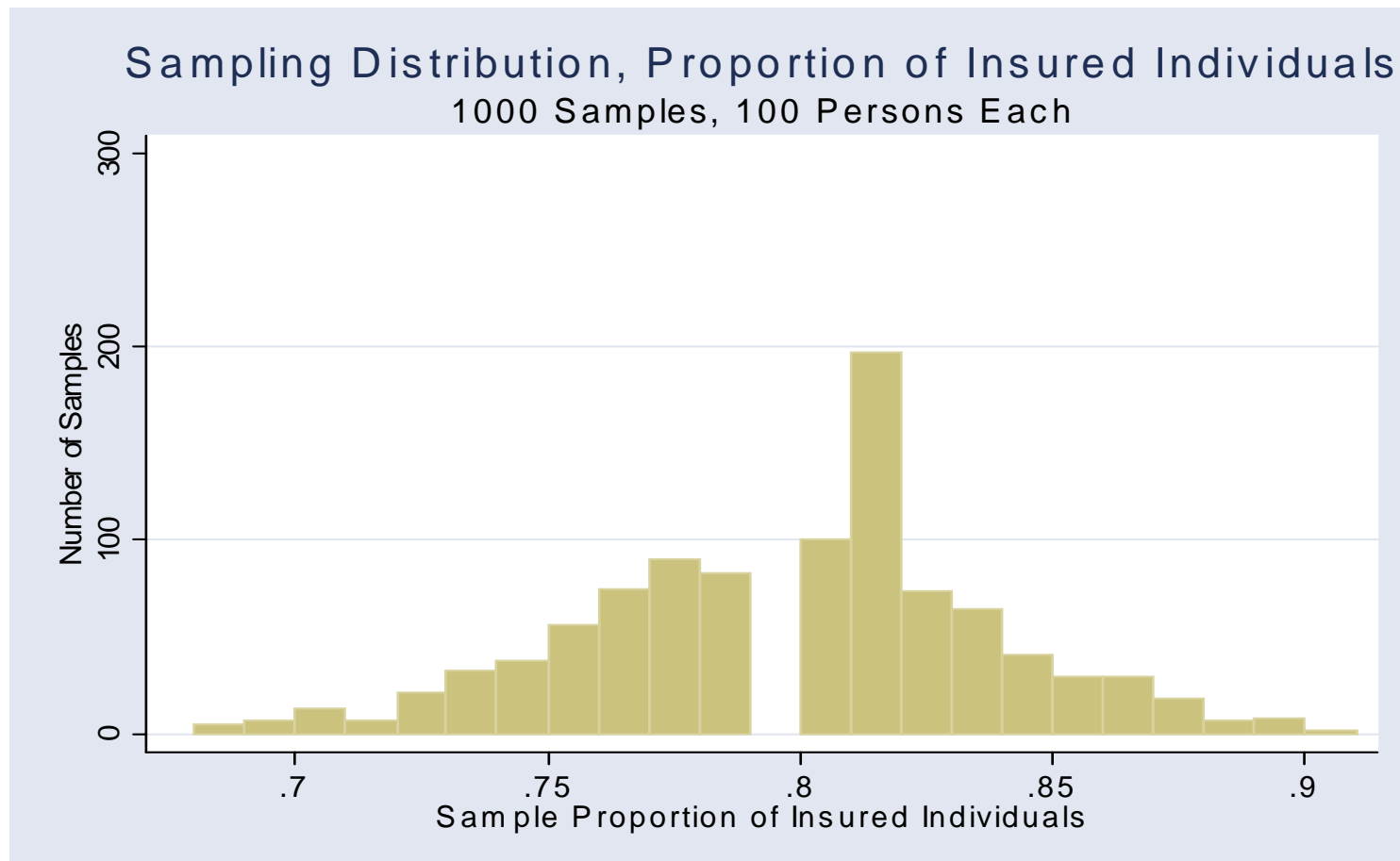
$$\hat{p} = .79$$

Sample 4

$$\hat{p} = .86$$

Sampling Distribution for p-hat

From 1,000 Samples of Size $n = 100$

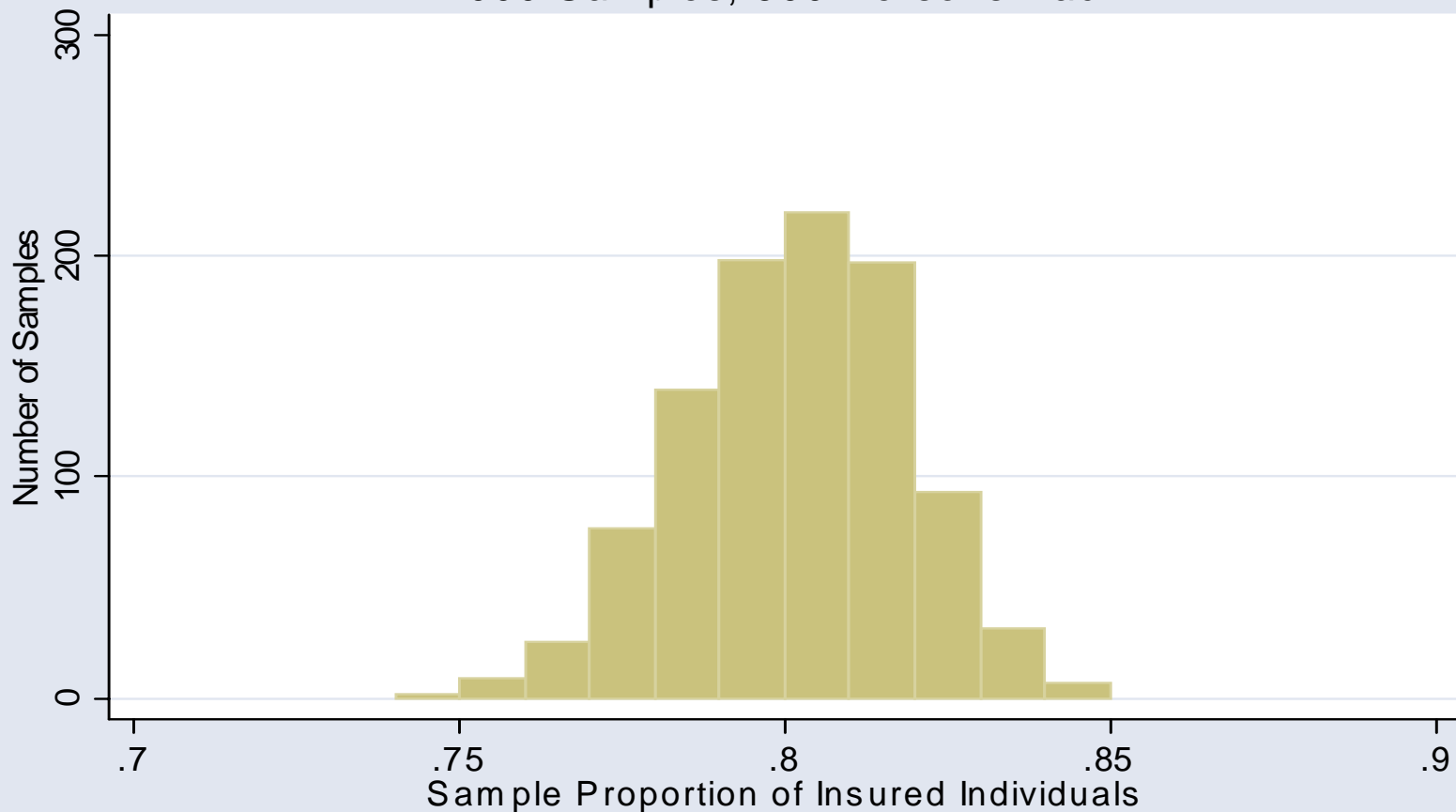


This is the sampling distribution of the sample proportion, based on 1,000 samples of size 100, when the population value is $p = .8$

Sampling Distribution for p-hat

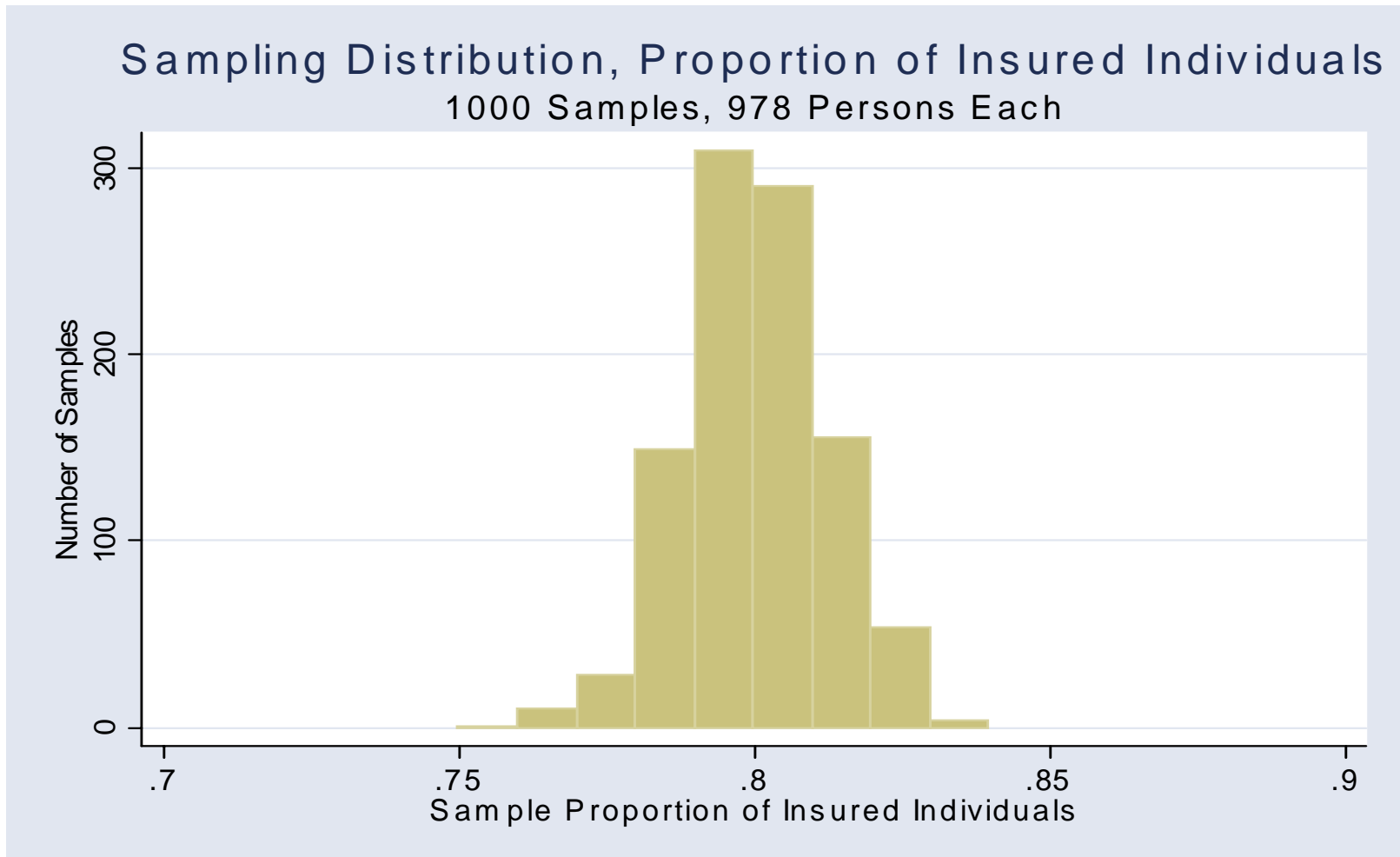
From 1,000 Samples of Size $n = 500$

Sampling Distribution, Proportion of Insured Individuals
1000 Samples, 500 Persons Each



Results of 1,000 Random Samples

Each of Size 978 from the Same Population



This is the sampling distribution of the sample proportion \hat{p} when the population value is $p = .8$

Normal Distribution

- ◆ **Why is the normal distribution so important in the study of statistics?**
- ◆ It's not because things in nature are always normally distributed (although sometimes they are)
- ◆ It's because of the central limit theorem: The sampling distribution of statistics—like a sample mean—often follows a normal distribution if the sample sizes are large

Sampling Distribution

- ◆ **Why is the sampling distribution important?**
- ◆ If a sampling distribution has a lot of variability (that is, a big standard error), then if you took another sample, it's likely you would get a very different result

Sampling Distribution

- ◆ About 95% of the time, the sample mean (or proportion) will be within two standard errors of the population mean (or proportion)
 - This tells us how “close” the sample statistic should be to the population parameter

Standard Errors

- ◆ *Standard errors (SE)* measure the precision of your sample statistic
- ◆ A small SE means it is more precise
- ◆ The SE is the standard deviation of the sampling distribution of the statistic

Calculating Standard Errors

- ◆ Mathematical statisticians have come up with formulas for the standard error; there are different formulas for:
 - Standard error of the mean (SEM)
 - Standard error of a proportion
- ◆ These formulas always involve the sample size n
 - As the sample size gets bigger, the standard error gets smaller

The standard deviation
is not
the standard error of a
statistic!

Standard Deviation vs. Standard Error

- ◆ *Standard deviation* measures the variability in the population
- ◆ *Standard error* measures the precision of a statistic—such as the sample mean or proportion—as an estimate of the population mean or population proportion



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section A

Practice Problems

Practice Problems

- ◆ Recall the income data on nine Internet-based MPHers (in thousands of \$):

37 102 34 12 111 56 72 17 33

- ◆ Recall, $\bar{X} = 52.67$ and $s = 35.6$

Practice Problems

1. How sure are we about our estimate of μ , the true mean income among online MPH students? Give an estimate of the standard error on our best estimate of μ , \bar{X}

Practice Problems

2. Suppose we took a random sample of 40 students, instead of nine. What is a sensible estimate for the standard deviation in this sample of 40?
3. What is a sensible estimate for the standard error of \bar{X} , the sample mean from the sample of 40 people?



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section A

Practice Problem Solutions

Solutions

1. How sure are we about our estimate of μ , the true mean income among online MPH students? Give an estimate of the standard error on our best estimate of μ , \bar{X}

Solutions

- ◆ Recall, in order to estimate the standard error of the sample mean (SEM), we just need the sample standard deviation, s , and the sample size n
 - In our sample, $s = 35.6$, and $n = 9$

Solutions

- ◆ So, the standard error of the mean—SEM, or $se(\bar{X})$ —estimate is . . .

$$SEM = \frac{s}{\sqrt{n}} = \frac{35.6}{\sqrt{9}} = \frac{35.6}{3} = 11.86$$

Solutions

2. Suppose we took a random sample of 40 students, instead of nine. What is a sensible estimate for the standard deviation in this sample of 40?

Solutions

- ◆ Recall, our sample standard deviation, s , is just an estimate of the population standard deviation
 - This should not change too much with a change in sample size
 - We have no other information about the sample of size 40, so our “guesstimate” of s is the value from the sample of size 9:
35.6

Solutions

3. What is a sensible estimate for the standard error of \bar{X} , the sample mean from the sample of 40 people?

Solutions

- ◆ Again, we have a “guesstimate” for s , and know the sample size:
 $n = 40$
 - The best estimate for the SEM would be:

$$\text{SEM} = \frac{s}{\sqrt{n}} = \frac{35.6}{\sqrt{40}} = \frac{35.6}{6.3} = 5.65$$

Solutions

- ◆ Remember s and SEM are not the same thing! They are estimating variability for two different distributions
- ◆ **S**—An estimate of the overall variability in the entire population
- ◆ **SEM**—An estimate of the variability of the value of the sample mean among samples of equal size



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section B

*Confidence Intervals
for the Population Mean μ*

Standard Error of the Mean

- ◆ The *standard error of the mean (SEM)* is a measure of the precision of the sample mean

$$SEM = \frac{s}{\sqrt{n}}$$

Example

- ◆ Measure systolic blood pressure on random sample of 100 students

Sample size $n = 100$

Sample mean $\bar{X} = 123.4$ mm Hg

Sample SD $s = 14.0$ mm Hg

$$\text{SEM} = \frac{14}{\sqrt{100}} = 1.4 \text{ mmHg}$$

Notes

- ◆ The smaller SEM is, the more precise \bar{X} is
- ◆ SEM depends on n and s
- ◆ SEM gets smaller if
 - s gets smaller
 - n gets bigger

Population Mean and Sample Mean

- ◆ How close to the population mean (μ) is the sample mean (\bar{X})?
- ◆ The standard error of the sample mean tells us!

Population Mean

- ◆ If we can calculate the sample mean \bar{x} and estimate its standard error, can that help us make a statement about the population mean?

Population Mean

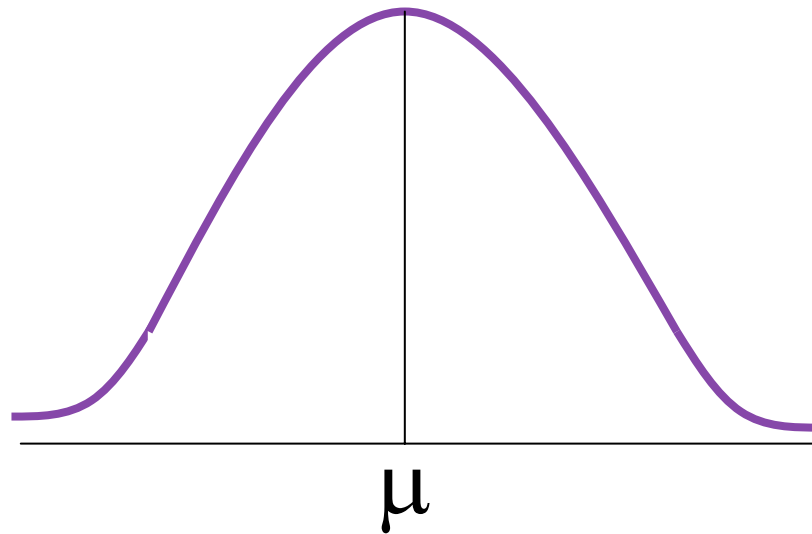
- ◆ The central limit theorem tells us that the sampling distribution for \bar{x} is approximately normal given enough data
- ◆ Additionally, the theorem tell us this sampling distribution should be centered about the true value of the population mean μ

Population Mean

- ◆ The standard error of \bar{x} gives us a measure of variability in the sampling distribution
 - We can then use properties of the normal distribution to make a statement about μ

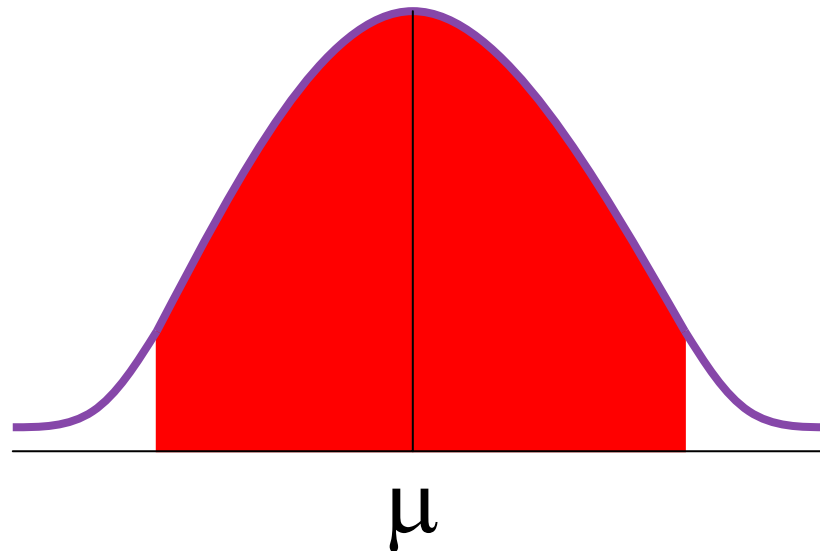
Sampling Distribution

- ◆ *Sampling distribution* is the distribution of all possible values of \bar{x} from samples of same size, n



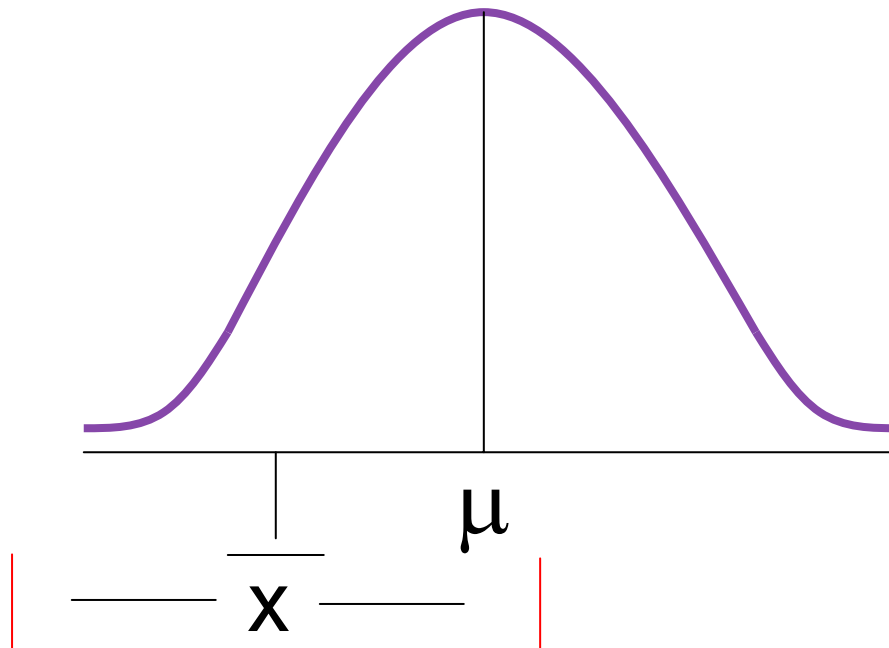
Sampling Distribution

- ◆ 95% of possible values for \bar{x} will fall within approximately two standard errors of μ



Sampling Distribution

- ◆ The “reverse” is also true—95% of the time μ will fall within two standard errors of a given \bar{x}



Sampling Distribution

- ◆ 95% of the time, the population mean will lie within about two standard errors of the sample mean
 - $\bar{x} \pm 2SEM$
- ◆ Why is this true?
 - Because of the central limit theorem

Interpretation

- ◆ We are 95% confident that the sample mean is within two standard errors of the population mean

Confidence Interval

- ◆ A 95% confidence interval for population mean μ is

$$\bar{x} \pm 2 SEM$$

- ◆ The confidence interval gives the range of plausible values for μ

Example

- ◆ Blood pressure
 $n = 100$, $\bar{X} = 125$ mm Hg, $s = 14$
- ◆ 95% CI for μ (mean blood pressure in the population) is . . .
 - $125 \pm 2 \times 1.4$
 - 125 ± 2.8

Ways to Write a Confidence Interval

- ◆ *122.2 to 127.8*
- ◆ *(122.2, 127.8)*
- ◆ *(122.2–127.8)*
- ◆ *We are highly confident that the population mean falls in the range 122.2 to 127.8*
- ◆ *The 95% error bound on \bar{x} is 2.8*

Using Stata to Create 95% CI for A Mean

- ◆ The "cii" command
 - Syntax "`cii n \bar{x} s` "

```
. cii 100 125 14
```

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]	
	100	125	1.4	122.2221	127.7779

Notes on Confidence Intervals

◆ Interpretation

- Plausible values for the population mean μ with high confidence

◆ Are all CIs 95%?

- No
- It is the most commonly used
- A 99% CI is wider
- A 90% CI is narrower

Notes on Confidence Intervals

- ◆ To be “more confident” you need a bigger interval
 - For a 99% CI, you need ± 2.6 SEM
 - For a 95% CI, you need ± 2 SEM
 - For a 90% CI, you need ± 1.65 SEM

Notes on Confidence Intervals

- ◆ **The length of CI decreases when . . .**
 - n increases
 - s decreases
 - Level of confidence decreases—for example, 90%, 80% vs 95%

Notes on Confidence Intervals

- ◆ **Random sampling error**
 - Confidence interval only accounts for random sampling error—not other systematic sources of error or bias

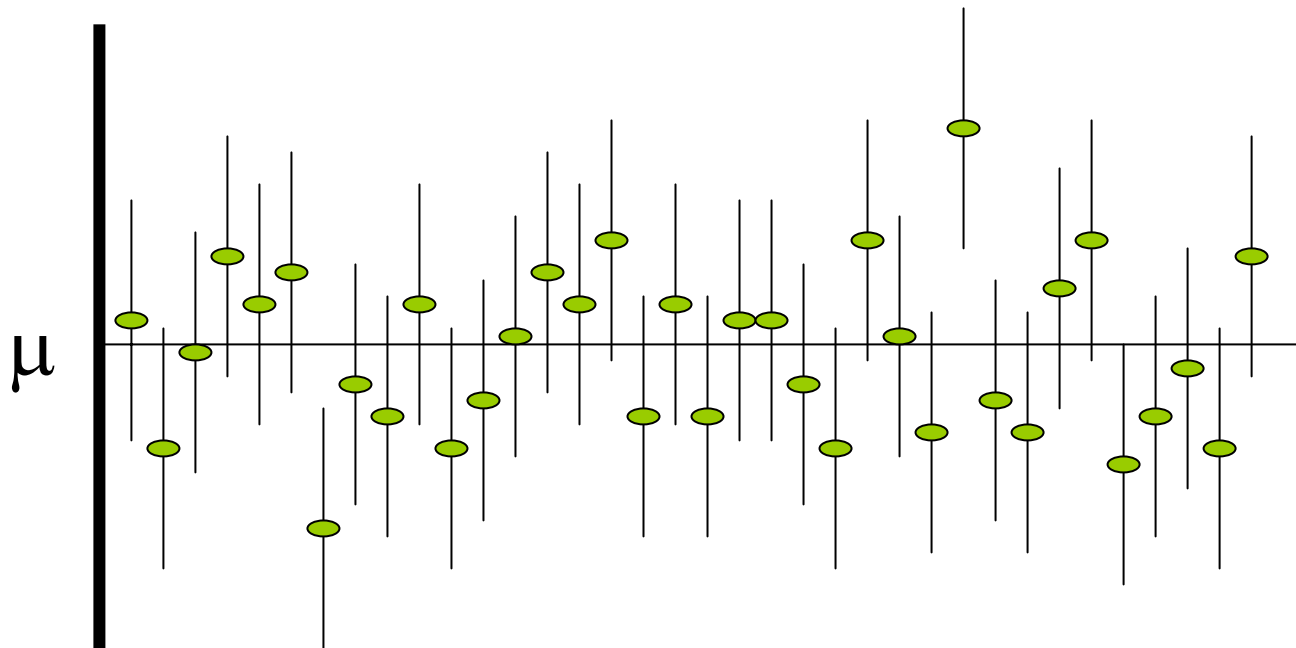
Examples of Systematic Bias

- ◆ BP measurement is always +5 too high (broken instrument)
- ◆ Only those with high BP agree to participate (non-response bias)

Confidence Interval Interpretation

- ◆ Technical interpretation
 - The CI works (includes μ) 95% of the time
 - If we were to take 100 random samples each of the same size, approximately 95 of the CIs would include the true value of μ

Confidence Interval



Each bar represents a 95% CI created from a random sample of size n

Underlying Assumptions

- ◆ In order to be able to use the formula

$$\bar{x} \pm 2SEM$$

- ◆ The data must meet a few conditions that satisfy the underlying assumptions necessary to use this result

Underlying Assumptions

- ◆ Random sample of population—important!
- ◆ Observations in sample independent
- ◆ Sample size n is at least 60 to use ± 2 SEM
 - Central limit theorem requires large n !

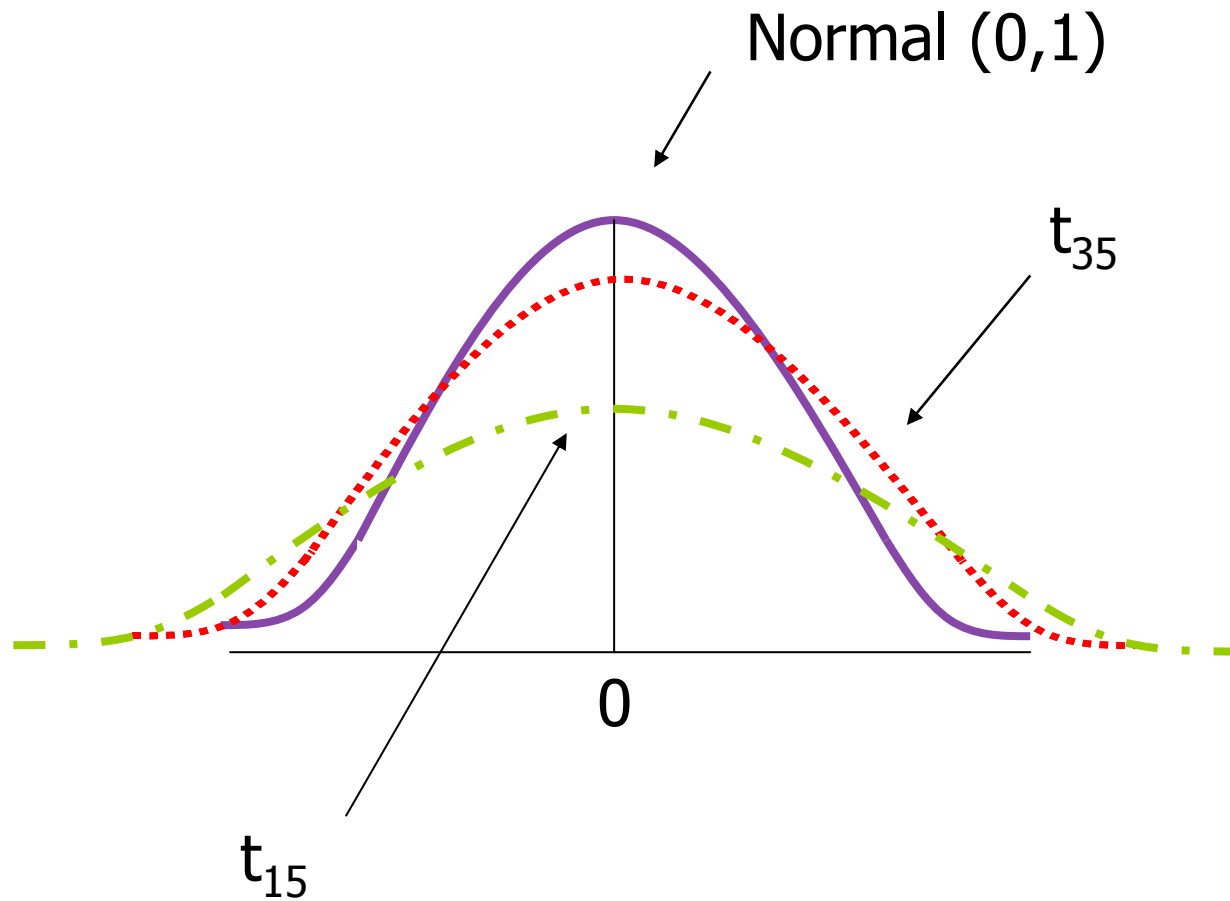
Underlying Assumptions

- ◆ If sample size is smaller than 60
 - The sampling distribution is not quite normally distributed
 - The sampling distribution instead approximates a “t-distribution”

The t-distribution

- ◆ The t-distribution looks like a standard normal curve that has been “stepped on”—it’s a little flatter and fatter
- ◆ A t-distribution is solely determined by its degrees of freedom—the lower the degrees of freedom, the flatter and fatter it is

The t-distribution



Underlying Assumptions

- ◆ If sample size is smaller than 60
 - There needs to be a small correction—called the **t-correction**
 - The number 2 in the formula $\bar{x} \pm 2SEM$ needs to be slightly bigger

Underlying Assumptions

- ◆ How much bigger the z needs to be depends on the sample size
- ◆ You can look up the correct number in a “t-table” or “t-distribution” with $n-1$ degrees of freedom

The t-distribution

- ◆ So if we have a smaller sample size, we will have to go out more than 2 SEMs to achieve 95% confidence
- ◆ How many standard errors we need to go depends on the degrees of freedom—this is linked to sample size
- ◆ The appropriate degrees of freedom are $n - 1$

Adjustment for Small Sample Sizes

$$\bar{X} \pm t^* (\text{SEM})$$

$$\bar{X} \pm t^* (s/\sqrt{n})$$

Adjustment for Small Sample Sizes

Value of $T_{.95}$ Used for 95% Confidence Interval for Mean

df	T	df	T
1	12.706	12	2.179
2	4.303	13	2.160
3	3.182	14	2.145
4	2.776	15	2.131
5	2.571	20	2.086
6	2.447	25	2.060
7	2.365	30	2.042
8	2.036	40	2.021
9	2.262	60	2.000
10	2.228	120	1.980
11	2.201	∞	1.960

Notes on the t-Correction

- ◆ The value of t that you need depends on the level of confidence you want as well as the sample size

Notes on the t-Correction

- ◆ With really small sample sizes ($n < 15$, or so), you also need to pay attention to the underlying distribution of the data in your sample
 - Needs to be “well behaved” for us to use $\bar{X} \pm t^*SEM$ for creating confidence intervals

Example: Blood Pressure

- ◆ $n = 5, \bar{X} = 99 \text{ mm Hg}, s = 16$
- ◆ 95% CI is $\bar{X} \pm 2.78 \text{ SEM}$
 - 2.78 from t-distribution with 4 degrees of freedom

Example: Blood Pressure

- ◆ $n = 5, \bar{X} = 99 \text{ mm Hg}, s = 16$
- ◆ 95% CI is $\bar{X} \pm 2.78 \text{ SEM}$



$$99 \pm 2.78 \times \frac{16}{\sqrt{5}}$$

Example: Blood Pressure

- ◆ $n = 5, \bar{X} = 99 \text{ mm Hg}, s = 16$
- ◆ 95% CI is $\bar{X} \pm 2.78 \text{ SEM}$

$$99 \pm 2.78 \times \frac{16}{\sqrt{5}}$$



$$99 \pm 2.78 * (7.16)$$

Example: Blood Pressure

- ◆ $n = 5, \bar{X} = 99 \text{ mm Hg}, s = 16$
- ◆ 95% CI is $\bar{X} \pm 2.78 \text{ SEM}$

$$99 \pm 2.78 \times \frac{16}{\sqrt{5}}$$



$$99 \pm 19.9$$

Example: Blood Pressure

- ◆ $n = 5, \bar{X} = 99 \text{ mm Hg}, s = 16$
- ◆ 95% CI is $\bar{X} \pm 2.78 \text{ SEM}$

$$99 \pm 2.78 \times \frac{16}{\sqrt{5}}$$



$$(79.1, 118.9)$$

Example: Blood Pressure

- ◆ The 95% CI for mean blood pressure is . . .
 - (79.1, 118.9)
 - 79.1–118.9
- ◆ Rounding off is okay, too
 - (79, 119)

Using Stata to Create 95% CI for a Mean

- ◆ Same "cii" command as before, same syntax

```
. cii 5 99 16
```

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]	
	5	99	7.155418	79.13338	118.8666



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Part B

Practice Problems

Practice Problems

1. In the last set of exercises, you calculated the SEM for the income information on nine Internet-based MPHers. Use this information to construct a 95% CI for the mean income among all Internet-based MPH students (assume income data “well behaved”, i.e., approximately normal at population level).

Practice Problems

- ◆ Suppose a pilot study is conducted using data from 15 smokers. The study measures the blood cholesterol level on each of the 12 smokers:

$$\bar{X} = 205 \text{ mg/100 ml, and } s = 43 \text{ mg/100ml}$$

Practice Problems

2. Suppose you want launch a more formal study of cholesterol levels in smokers based on the results of the pilot study. In your grant application, you promise an error bound of ± 5 mg/ 100ml. Approximately how many smokers will you need to recruit?



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Part B

Practice Problem Solutions

Solutions

1. In the last set of exercises, you calculated the SEM for the income information on nine Internet-based MPHers. Use this information to construct a 95% CI for the mean income among all Internet-based MPH students (assume income data “well behaved”, i.e., approximately normal at population level).

Solutions

- ◆ Here, we have a n of 9, so we'll need to appeal to the t-table to do our CI
 - We need to seek out the appropriate t-value with $n-1=8$ degrees of freedom

Solutions

- Using our formula for a 95% CI we would get:

$$\bar{X} \pm T^*(SEM)$$


- Where $\bar{X} = 52.7$

- SEM =


$$\frac{s}{\sqrt{n}} = \frac{35.6}{\sqrt{9}} = 11.9$$

- And $T = 2.3$

Solutions

- ◆ By our formula, a 95% CI for μ
 - $52.7 \pm 2.3 \cdot (11.9)$
- ◆ By our formula, a 95% CI for μ
 - $52.7 \pm 2.3 \cdot (11.9)$
 - 52.7 ± 27.4 

Solutions

- ◆ By our formula, a 95% CI for μ
 - $52.7 \pm 2.3 \cdot (11.9)$
- ◆ By our formula, a 95% CI for μ
 - $52.7 \pm 2.3 \cdot (11.9)$
 - 52.7 ± 27.4
 - $(25.3, 80.1)$ 
 - (in units of thousands of dollars)

Solutions

2. Suppose a pilot study is conducted using data from 15 smokers. The study measures the blood cholesterol level on each of the 12 smokers:
 $\bar{X} = 205$ mg/100 ml, and $s = 43$ mg/100ml

Solutions

2. Suppose you want to launch a more formal study of cholesterol levels in smokers based on the results of the pilot study. In your grant application, you promise an error bound of ± 5 mg/ 100ml. Approximately how many smokers will you need to recruit?

Solutions

- ◆ Recall, the “error bound” is the $\pm T^*(SEM)$ portion of the CI
 - We want this bound to equal 5 mg/100ml
 - From the pilot study, we can estimate s with 43 mg/100ml. Recall, $SEM =$

$$\frac{s}{\sqrt{n}}$$

Solutions

- ◆ For our situation, $SEM = \frac{43}{\sqrt{n}}$
 - In order for our error bound to be ± 5 , we need to choose n such that:

$$1.96 * \frac{43}{\sqrt{n}} = 5$$

- (We can use 1.96 for the 95% error bound making the assumption that n will be larger than 60)

Solutions

- A little algebra yields:

$$\sqrt{n} = \frac{1.96 * 43}{5}$$

$$\sqrt{n} = \frac{1.96 * 43}{5} = 16.86$$

$$n = (16.86)^2 = 284.12$$

Solutions

- You would need about 285 people to deliver on your promise!



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section C

*Standard Error for a Proportion;
Confidence Intervals for a Proportion*

Proportions (P)

- ◆ Proportion of individuals with health insurance
- ◆ Proportion of patients who became infected
- ◆ Proportion of patients who are cured
- ◆ Proportion of individuals who are hypertensive

Proportions (P)

- ◆ Proportion of individuals positive on a blood test
- ◆ Proportion of adverse drug reactions
- ◆ Proportion of premature infants who survive

Proportions (P)

- ◆ For each individual in the study, we record a binary outcome (Yes/No; Success/Failure) rather than a continuous measurement

Proportions (P)

- ◆ Compute a sample proportion, \hat{p} (pronounced “p-hat”), by taking observed number of “yes’s” divided by total sample size

Proportions (P)

- ◆ Example: 978 persons polled to see if each currently has health insurance—793 of the 978 surveyed have insurance

$$\hat{p} = \frac{793}{978} = 81\%$$

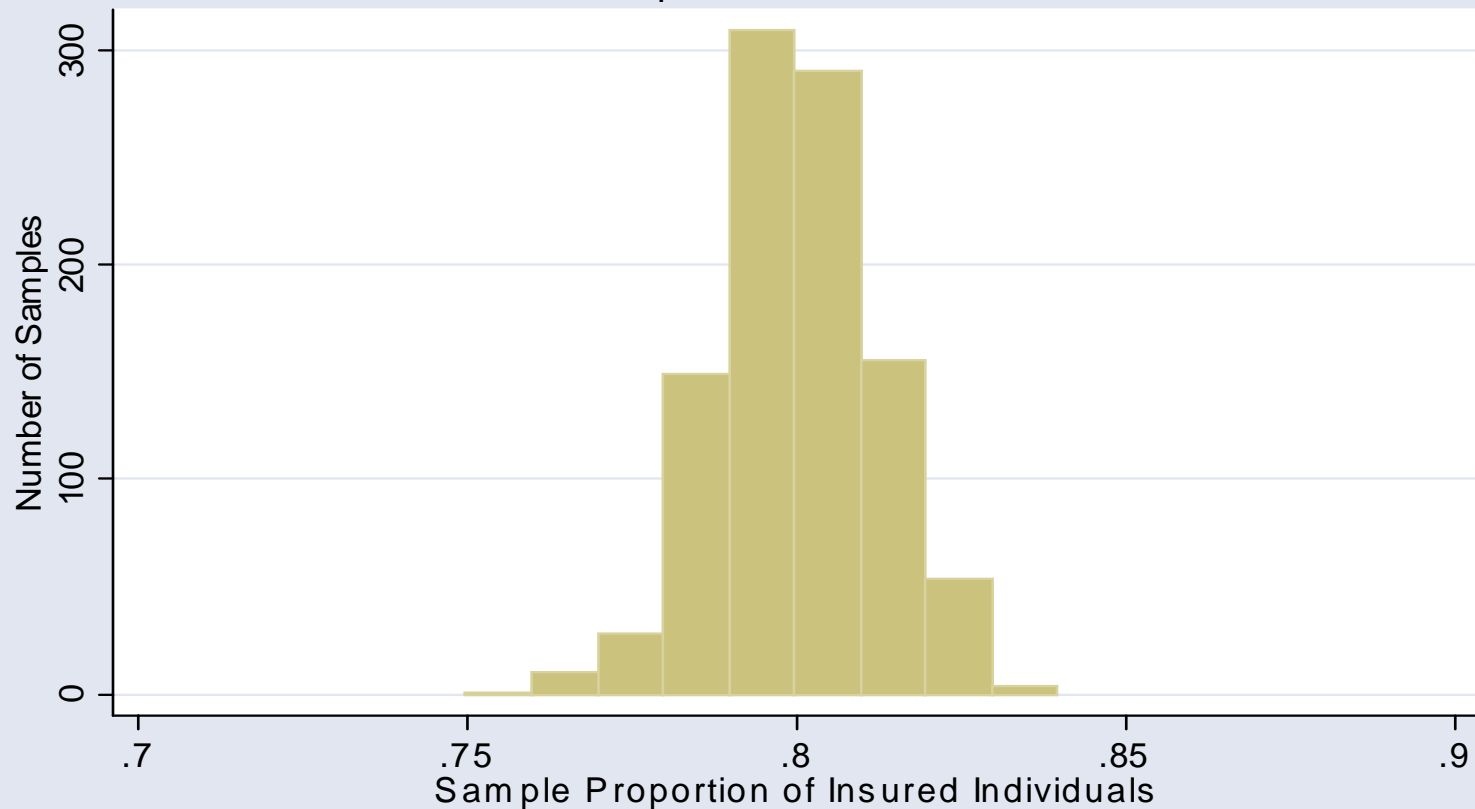
- ◆ Where \hat{p} is the proportion of persons with insurance

Proportions (P)

- ◆ How accurate of an estimate is the sample proportion of the population proportion?
- ◆ What is the standard error of a proportion?

The Sampling Distribution of a Proportion

Sampling Distribution, Proportion of Insured Individuals
1000 Samples, 978 Persons Each



The Sampling Distribution of a Proportion

- ◆ The standard error of a sample proportion is estimated by:

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

Example

- ◆ $n = 200$ patients
- ◆ $X = 90$ adverse drug reaction
- ◆ The estimated proportion who experience an adverse drug reaction is . . .

$$\hat{p} = \frac{90}{200} = .45$$

Notes

- ◆ There is uncertainty about this rate because it involved only $n = 200$ patients
- ◆ If we had studied another sample of 200 patients, would we have gotten a much different answer?

Notes

- ◆ The sample proportion is *.45 or 45%*
- ◆ But it is not the true rate of adverse drug reactions in the population

95% Confidence Interval for a Proportion

$$\hat{p} \pm 2SE(\hat{p})$$

$$\hat{p} \pm 2\sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

\hat{p} is the sample proportion
 n is the sample size

Example

- ◆ $n = 200$ patients
 $X = 90$ adverse drug reactions
 $= \hat{p} \ 90/200 = .45$

$$.45 \pm 2 \sqrt{\frac{.45 \times .55}{200}}$$

Example

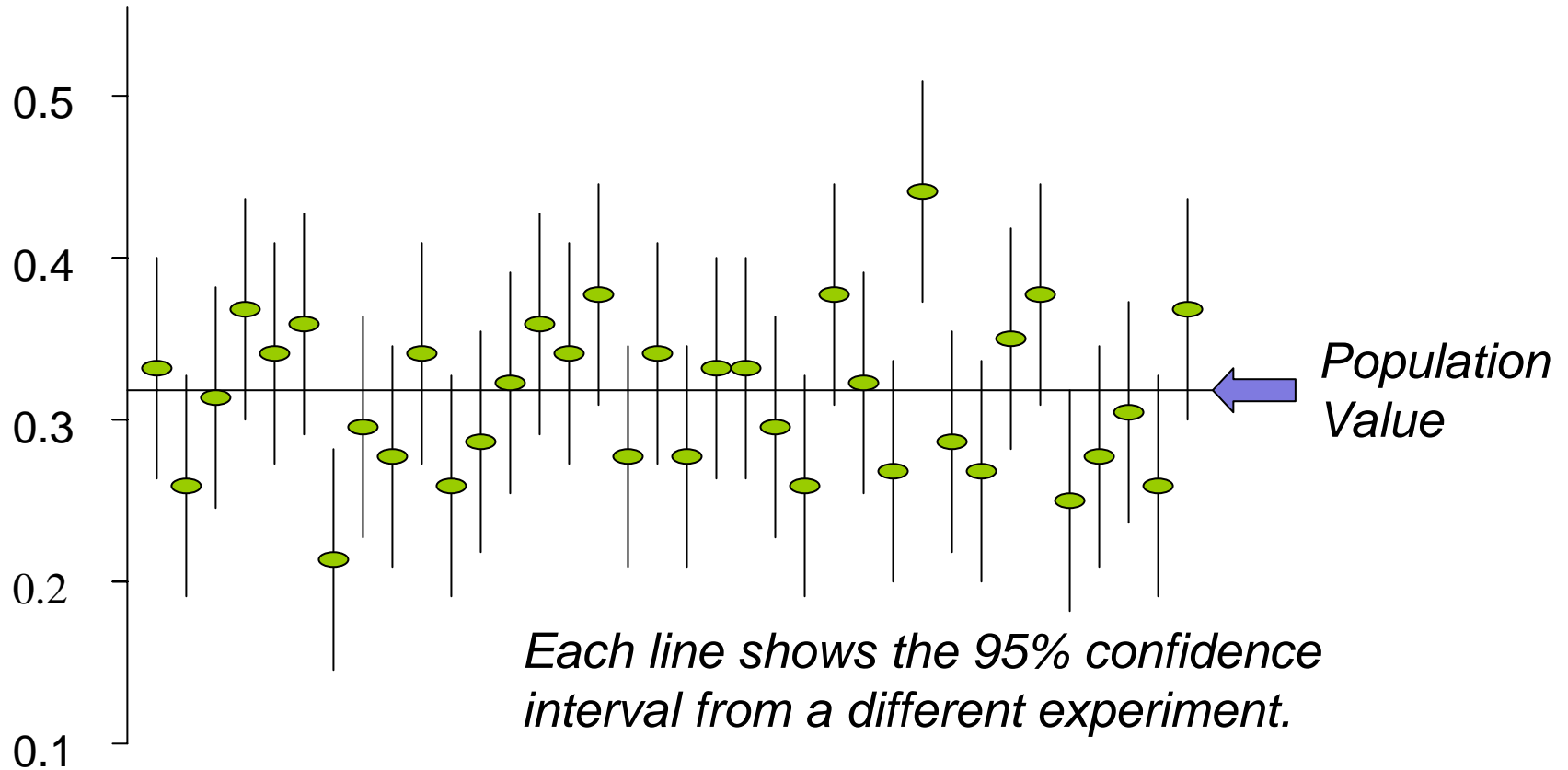
- ◆ $.45 \pm 2 \times .035$
 $.45 \pm .07$

The 95% confidence interval is . . .
(.38 – .52)

CI and Proportion

- ◆ **How do we interpret a 95% confidence interval for a proportion?**
 - Plausible range of values for population proportion
 - Highly confident that population proportion is in the interval
 - The method works 95% of the time

CI and Proportion



In this example, the true proportion of “yes” (p) is 0.32; 33 of 35 CI calculated from 35 samples contain p

Notes on 95% Confidence Interval for a Proportion

- ◆ The confidence interval does not address your definition of drug reaction and whether that's a good or bad definition; it accounts only for sampling variation
- ◆ Can also have CI with different levels of confidence

Notes on 95% Confidence Interval for a Proportion

- ◆ Sometimes $\pm 2 SE(\hat{p})$ is called
 - *95% error bound*
 - *Margin of error*

Notes on 95% Confidence Interval for a Proportion

- ◆ The formula for a 95% CI is only *approximate*; it works very well if you have enough data in your sample
- ◆ The “rule”:
 - If $n \times \hat{p} \times (1 - \hat{p}) \geq 5$ then the approximation is good

Notes on 95% Confidence Interval for a Proportion

- ◆ The “rule” applied to drug failures data
 - $n \times \hat{p} \times (1 - \hat{p}) =$
 - $200 * (.45) * (.55) =$
 - ≈ 50

Notes on 95% Confidence Interval for a Proportion

- ◆ You do *not* use the t-correction for small sample sizes like we did for sample means
 - We use exact binomial calculations

Exact Confidence Intervals for a Proportion Using a Computer

- ◆ Stata command (done at command line):

```
ci i N X
```

- Where n is the sample size, and X is the number of “yes outcomes”
- This will give a 95% CI

Exact Confidence Intervals for the Drug Failures Example

```
. cii 200 90
```


```
                                -- Binomial Exact  --  
Variable |   Obs   Mean  Std. Err.   [95% Conf. Interval]  
-----+-----  
         |  200    .45   .0351781   .3797562   .5217545
```

Recall our example: 200 patients (N), and 90 drug failures (X)

Exact Confidence Intervals for the Drug Failures Example

```
. cii 200 90
```

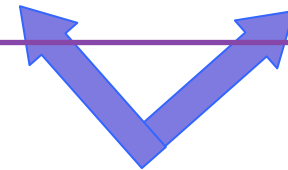
Variable	Obs	Mean	Std. Err.	-- Binomial Exact -- [95% Conf. Interval]	
	200	.45	.0351781	.3797562	.5217545


$$\sqrt{\frac{(.45)(1-.45)}{200}}$$

Exact Confidence Intervals for the Drug Failures Example

```
. cii 200 90
```

Variable	Obs	Mean	Std. Err.	-- Binomial Exact -- [95% Conf. Interval]	
	200	.45	.0351781	.3797562	.5217545



*Exact 95% CI for
proportion of patients
with drug failure*

Exact Confidence Intervals of Different Width (Not 95%)

```
. cii 200 90, level(99)
```

```
-- Binomial Exact --
```

```
Variable | Obs Mean Std. Err. [99% Conf. Interval]
```

```
-----+-----
```

```
| 200 .45 .0351781 .3590332 .5434235
```

Example

- ◆ In a study of patients hospitalized after myocardial infarction and treated with streptokinase, two of fifteen patients died within twelve months
- ◆ The one-year mortality rate was 13% (95% CI 1.7 – 40.5)

“Behind the Scenes” Stata Calculation

```
. cii 15 2
```

```
                -- Binomial Exact --  
Variable |   Obs   Mean  Std. Err.   [95% Conf. Interval]  
-----+-----  
         |   15   .133333  .0877707   .0165771   .4045898
```

Sample Size and the Margin of Error

- ◆ The 95% error bound (margin of error) is . . .

$$\hat{p} \pm 2\sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

Sample Size and the Margin of Error

- ◆ In the myocardial infarction example, what do you think the margin of error would turn out to be if we did a larger study, such *as* $n = 50$?

Sample Size and the Margin of Error

- ◆ Before the study, we don't know P
 - “Guesstimate”: For example, use the earlier study result ($\hat{p} = .13$)

$$2 \times \sqrt{\frac{.13 \times .87}{50}} = .09$$

Sample Size and the Margin of Error

N	Margin of error
50	$\pm .09$
100	$\pm .07$
200	$\pm .05$
300	$\pm .04$
500	$\pm .03$

Sample Size and the Margin of Error

- ◆ We would need a sample size of about 500 to estimate the death rate following MI to within $\pm 3\%$
 - That is, the 95% error bound for the death rate (or margin or error) is $\pm .03$

Example

- ◆ Study of survival of premature infants
 - All premature babies born at Johns Hopkins during a three-year period (Allen, et al., *NEJM*, 1993)
 - $N = 39$ infants born at 25 weeks gestation
 - 31 survived six months

Example

$$\hat{p} = \frac{31}{39} = 0.79$$

95% CI .63-.91

Necessity of Confidence Intervals

Intervals

- ◆ Are confidence intervals needed even though all infants were studied?
- ◆ Are the 39 infants a sample?
- ◆ Seems like it's the whole population—but do you really want to talk just about these infants, or those at similar urban hospitals, for example?
- ◆ Do you view this as a sample from a random, underlying process?

Sampling Error Is Not the Only Kind of Error

- ◆ Remember, these methods only account for sampling error!



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section C

Practice Problems

Practice Problems

1. Use the Stata output to compute an approximate 95% CI for population proportion of patients with drug failure. How does your calculation compare to the exact 95% CI?

Practice Problems

```
. cii 200 90
```

```
-- Binomial Exact --
```

```
Variable | Obs      Mean   Std. Err.   [95% Conf. Interval]
```

```
-----+-----
```

```
| 200      .45   .0351781   .3797562   .5217545
```

*Recall our example: 200 patients (N)
and 90 drug failures (X)*

Practice Problems

2. In a study of patients hospitalized after myocardial infarction and treated with streptokinase, two of fifteen patients died within twelve months. The one-year mortality rate was 13% (95% exact CI 1.7 – 40.5). Calculate an approximate 95% CI for the one-year mortality rate and compare to the exact 95% CI.

Practice Problems

3. Devise a one sentence “recipe” for calculating an approximate 95% CI for a parameter, whether it be a proportion or a mean (assume a large sample).



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section C

Practice Problem Solutions

Solutions

1. Use the Stata output to compute an approximate 95% CI for population proportion of patients with drug failure. How does your calculation compare to the exact 95% CI?

Solutions

```
. cii 200 90
```

```
                -- Binomial Exact  --  
Variable |   Obs   Mean  Std. Err.   [95% Conf. Interval]  
-----+-----  
         |  200    .45   .0351781   .3797562   .5217545
```

*Recall our example: 200 patients (N)
and 90 drug failures (X)*

Solutions

- The formula $\hat{p} \pm 1.96 * SE(\hat{p})$ yields:
 $.45 \pm 1.96 * (.035)$
 $.45 \pm (.0686)$
 $(.3814, .5186)$
- The approximate 95% CI is similar to the exact CI in this situation! Why?

Solutions

2. In a study of patients hospitalized after myocardial infarction and treated with streptokinase, two of fifteen patients died within twelve months. The one-year mortality rate was 13% (95% exact CI 1.7–40.5). Calculate an approximate 95% CI for the one-year mortality rate and compare to the exact 95% CI.

Solutions

- The formula $\pm 1.96 * SE(P)$ yields:

$$.13 \pm 1.96 * (.0878)$$

$$.31 \pm (.1721)$$

$$(-.0421, .3021)$$

- The approximate 95% CI is different than the exact CI in this situation! Why?

Solutions

3. Devise a one sentence “recipe” for calculating an approximate 95% CI for a parameter, whether it be a proportion or a mean (assume a large sample)
 - *(Our estimate) $\pm 2*(SE\ of\ our\ estimate)$*