JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Sampling Variability and Confidence Intervals

John McGready
Johns Hopkins University

Lecture Topics

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

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section A

The Random Sampling Behavior of a Sample Mean Across Multiple Random Samples

Random Sample

- When a sample is randomly selected from a population, it is called a *random sample*
  - Technically speaking values in a random sample are representative of the distribution of the values in the population sample, regardless of size

- In a simple random sample, each individual in the population has an equal chance of being chosen for the sample

- Random sampling helps control systematic bias

- But even with random sampling, there is still *sampling variability* or error

Sampling Variability of a Sample Statistic

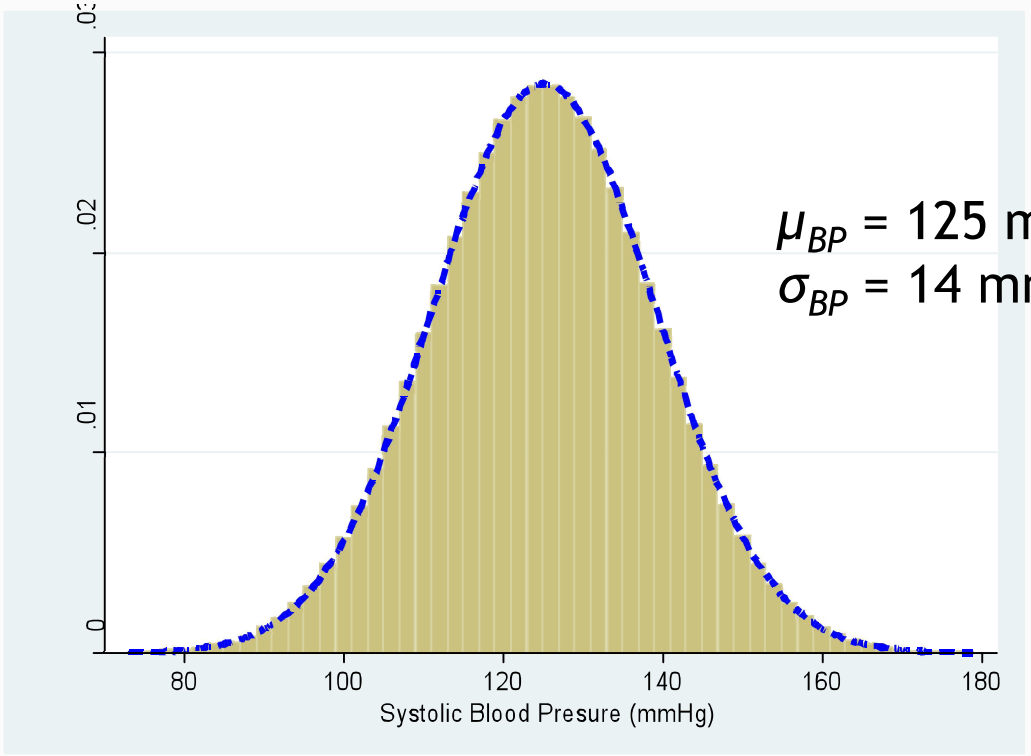
- If we repeatedly choose samples from the same population, a statistic will take different values in different samples

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

Example: Blood Pressure of Males

- Recall, we had worked with data on blood pressures using a random sample of 113 men taken from the population of all men

- Assume the population distribution is given by the following:

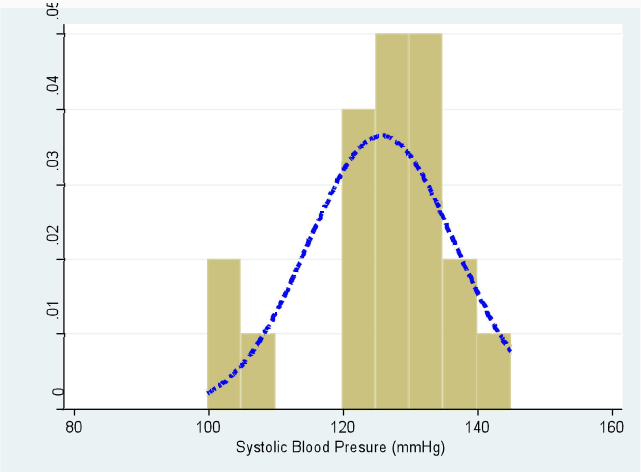


Example: Blood Pressure of Males

- Suppose we had all the time in the world
- We decide to do an experiment
- We are going to take 500 separate random samples from this population of men, each with 20 subjects
- For each of the 500 samples, we will plot a histogram of the sample BP values, and record the sample mean and sample standard deviation
- Ready, set, go . . .

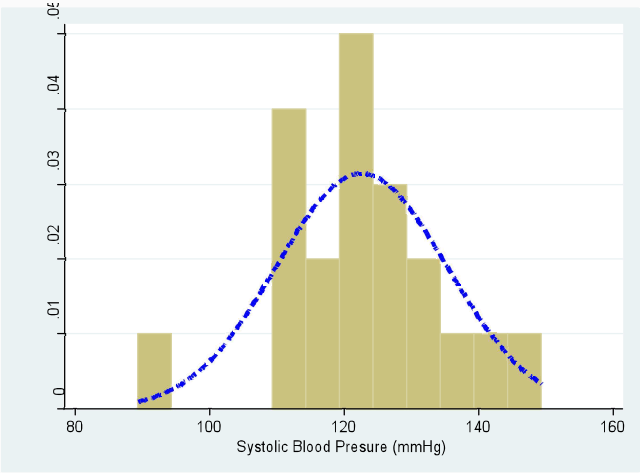
Random Samples

- Sample 1: n = 20



$\bar{x}_{BP}$ = 125.7 mmHg
s_{BP} = 10.9 mmHg

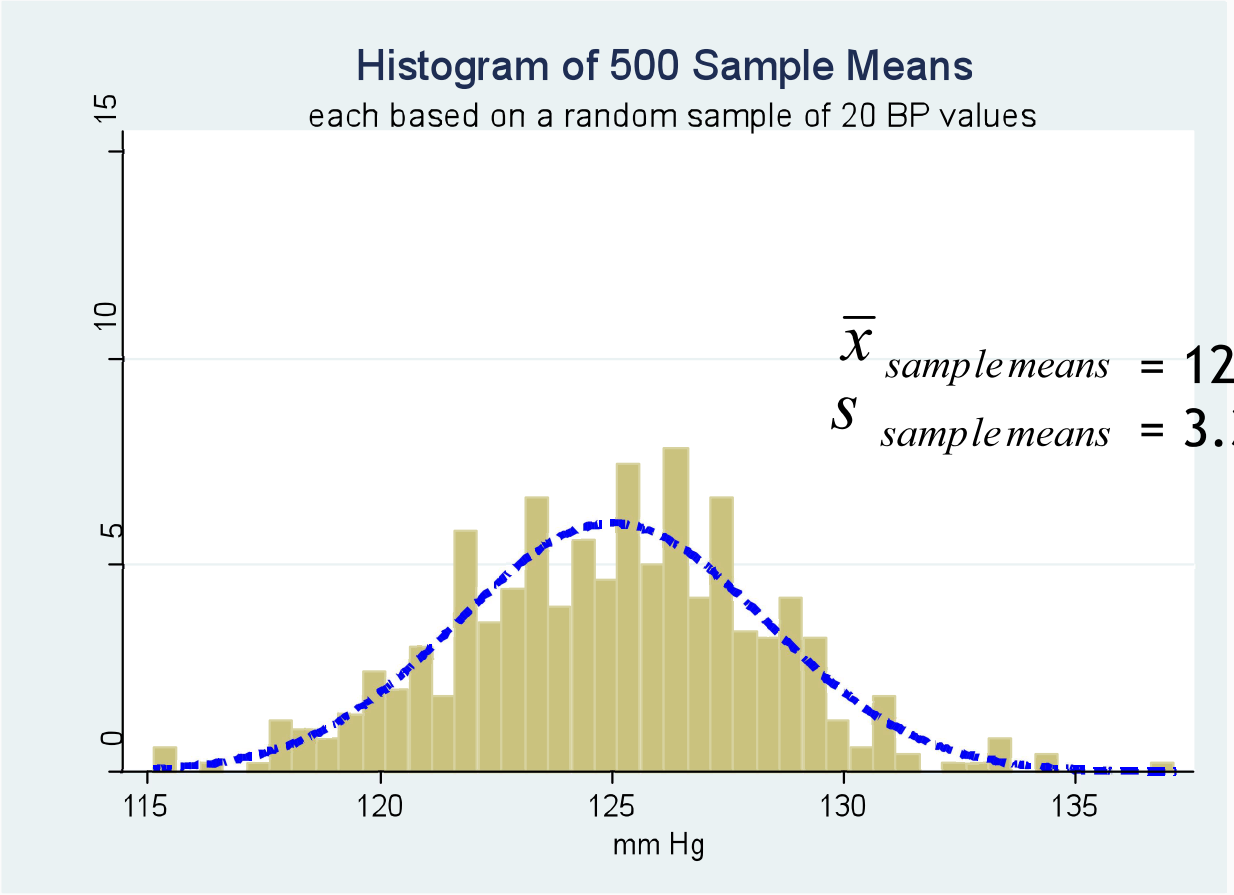
- Sample 2: n = 20



$\bar{x}_{BP}$ = 122.6 mmHg
s_{BP} = 12.7 mmHg

Example: Blood Pressure of Males

- So we did this 500 times: now let's look at a histogram of the 500 sample means

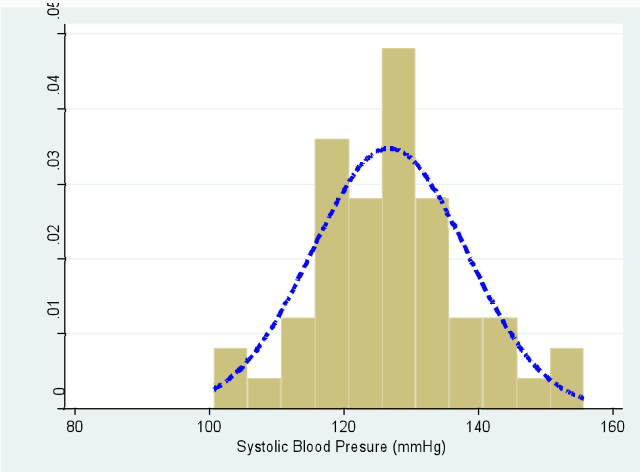


Example: Blood Pressure of Males

- We decide to do another experiment
- We are going to take 500 separate random samples from this population of me, each with 50 subjects
- For each of the 500 samples, we will plot a histogram of the sample BP values, and record the sample mean and sample standard deviation
- Ready, set, go . . .

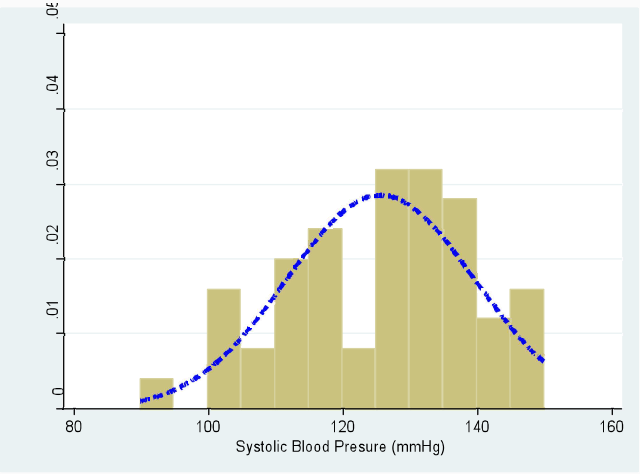
Random Samples

- Sample 1: n = 50



$\bar{x}_{BP}$ = 126.7 mmHg
s_{BP} = 11.5 mmHg

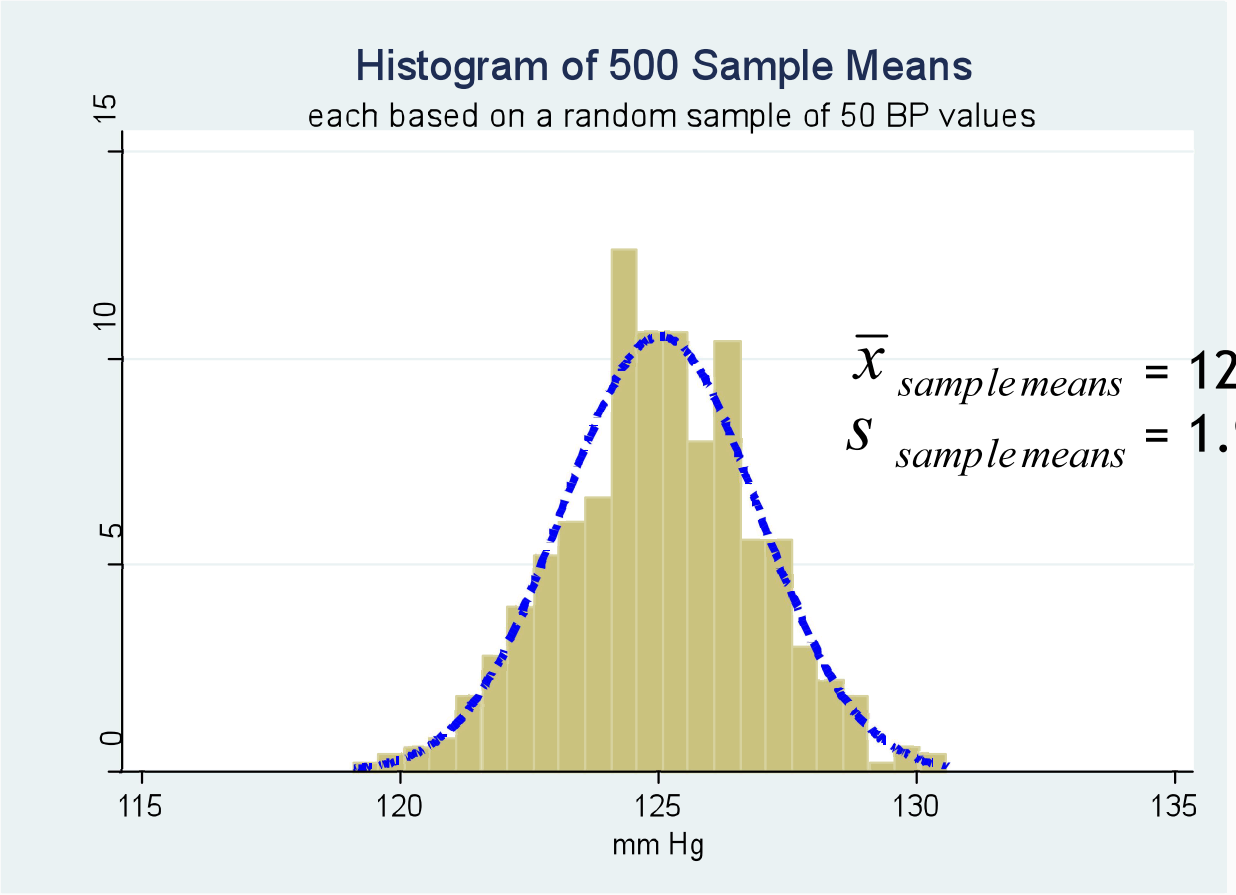
- Sample 2: n = 50



$\bar{x}_{BP}$ = 125.5 mmHg
s_{BP} = 14.0 mmHg

Example: Blood Pressure of Males

- So we did this 500 times: now let's look at a histogram of the 500 sample means

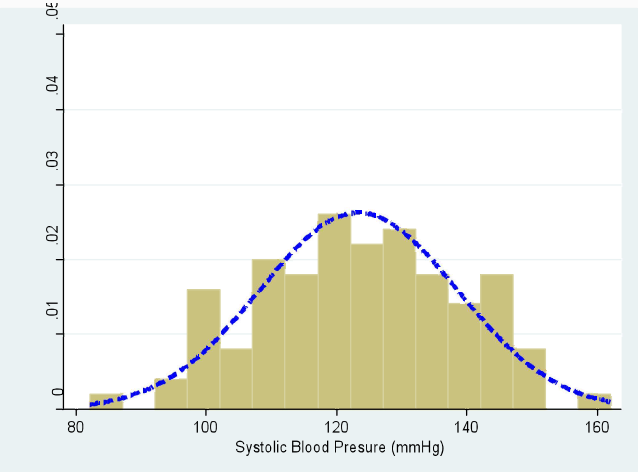


Example: Blood Pressure of Males

- We decide to do one more experiment
- We are going to take 500 separate random samples from this population of men, each with 100 subjects
- For each of the 500 samples, we will plot a histogram of the sample BP values, and record the sample mean, and sample standard deviation
- Ready, set, go . . .

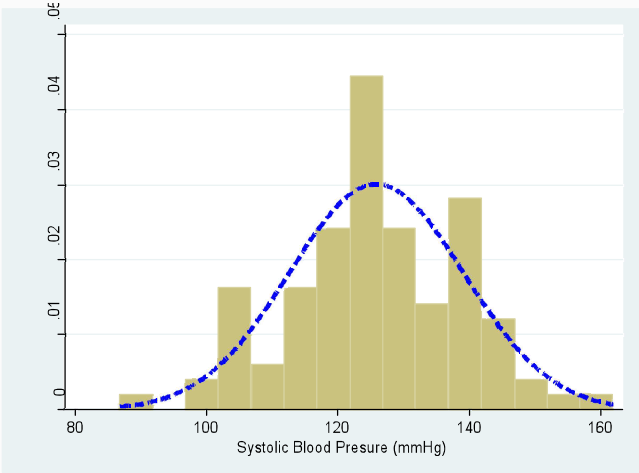
Random Samples

- Sample 1: *n* = 100



$\bar{x}_{BP}$ = 123.3 mmHg
s_{BP} = 15.2 mmHg

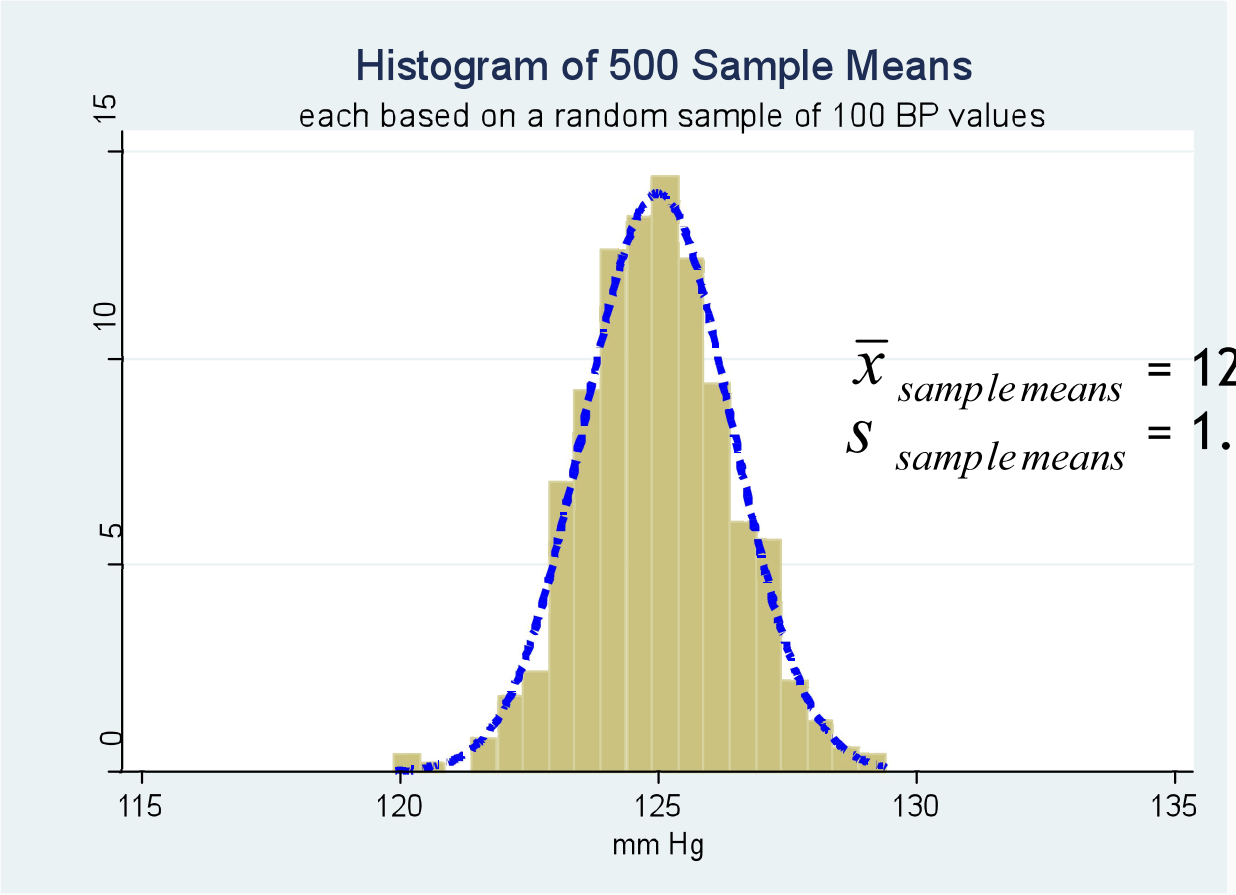
- Sample 2: *n* = 100



$\bar{x}_{BP}$ = 125.7 mmHg
s_{BP} = 13.2 mmHg

Example: Blood Pressure of Males

- So we did this 500 times: now let's look at a histogram of the 500 sample means



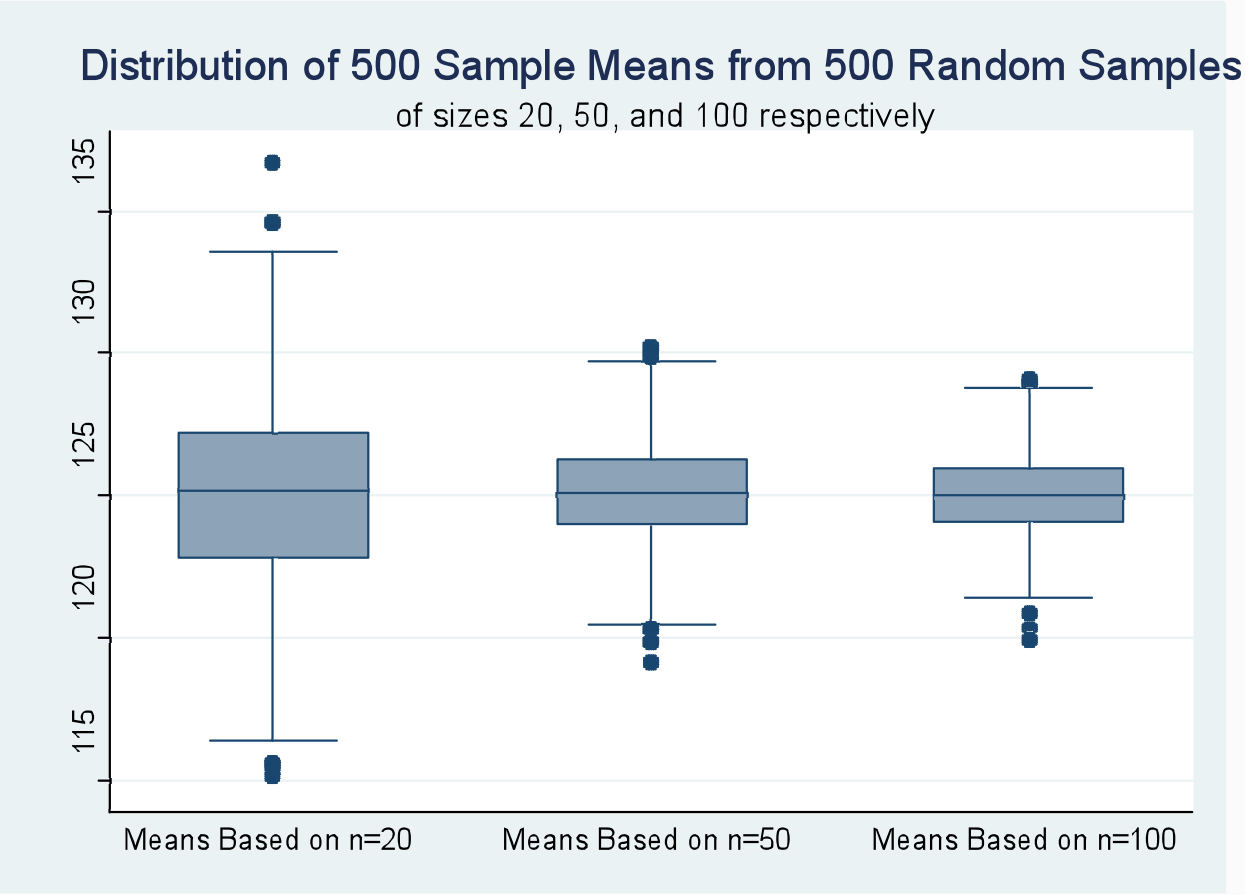
Example: Blood Pressure of Males

- Let's review the results
  - Population distribution of individual BP measurements for males: normal
  - True mean μ = 125 mmHg: σ = 14 mmHg
  - Results from 500 random samples:

Sample Sizes	Means of 500 Sample Means	SD of 500 Sample Means	Shape of Distribution of 500 sample means
n = 20	125 mmHg	3.3 mm Hg	Approx normal
n = 50	125 mmHg	1.9 mm Hg	Approx normal
n = 100	125 mmHg	1.4 mm Hg	Approx normal

Example: Blood Pressure of Males

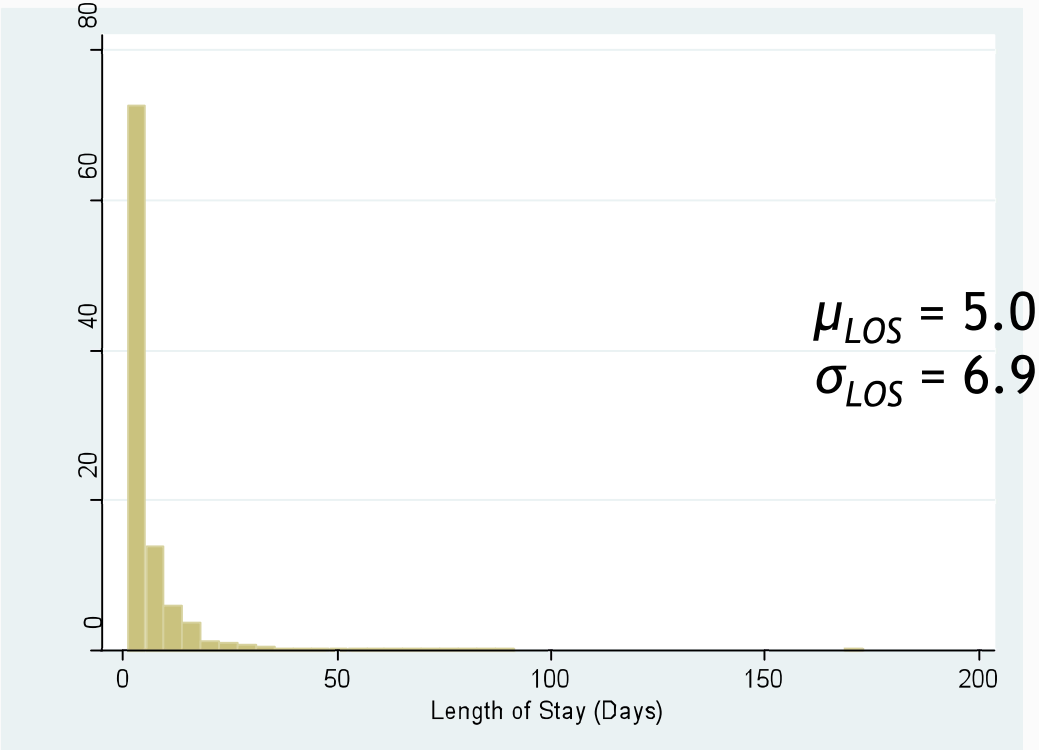
- Let's review the results



Example 2: Hospital Length of Stay

- Recall, we had worked with data on length of stay (LOS) using a random sample of 500 patients taken from sample of all patients discharged in year 2005

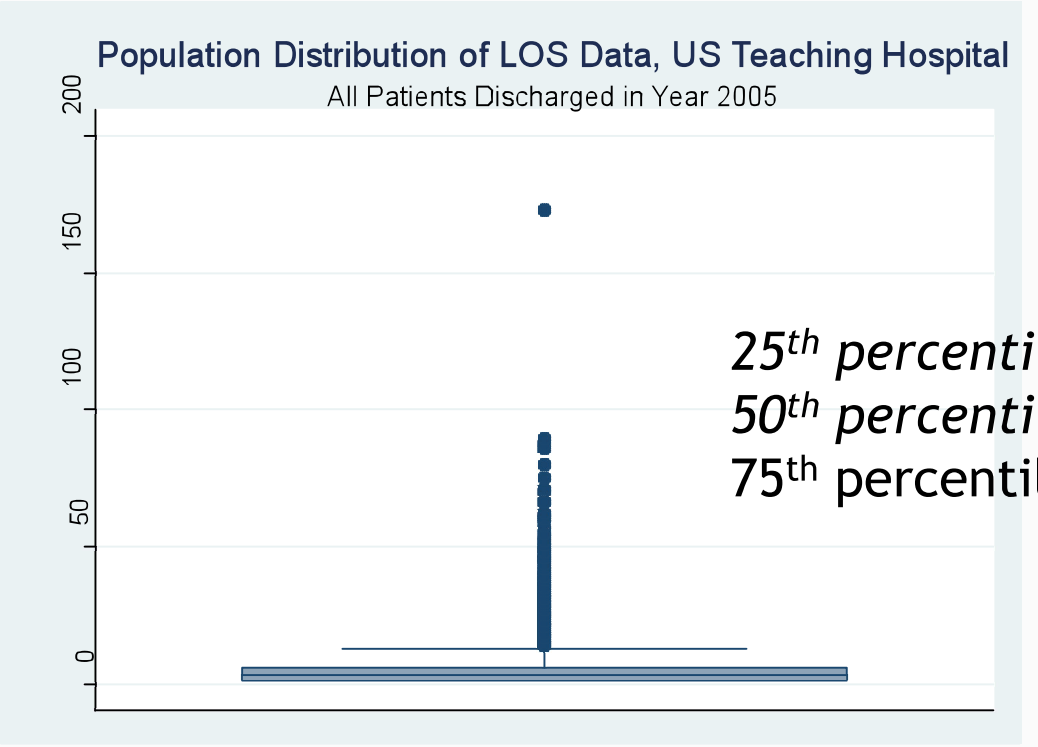
- Assume the population distribution is given by the following:



μ_{LOS} = 5.0 days
σ_{LOS} = 6.9 days

Example 2: Hospital Length of Stay

- Boxplot presentation



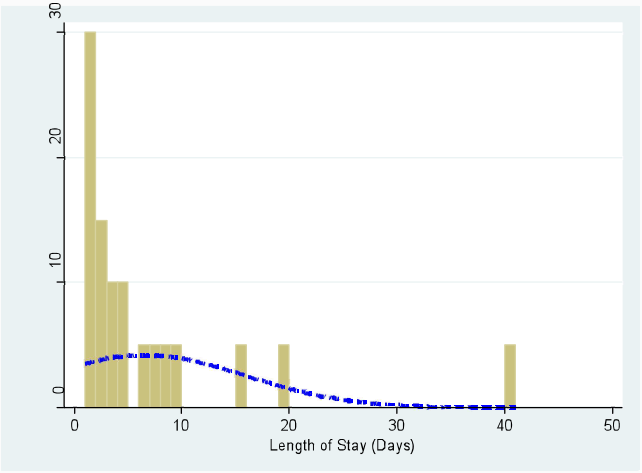
25th percentile: 1.0 days
50th percentile: 3.0 days
75th percentile: 6.0 days

Example 2: Hospital Length of Stay

- Suppose we had all the time in the world again
- We decide to do another set of experiments
- We are going to take 500 separate random samples from this population of patients, each with 20 subjects
- For each of the 500 samples, we will plot a histogram of the sample LOS values, and record the sample mean and sample standard deviation
- Ready, set, go . . .

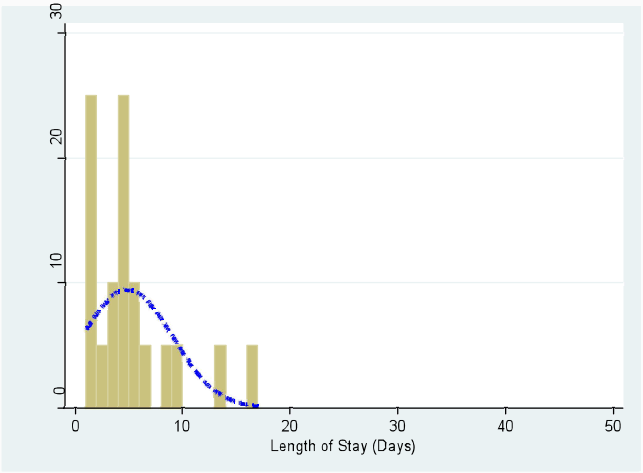
Random Samples

- Sample 1: n = 20



$\overline{x}_{LOS}$ = 6.6 days
s_{LOS} = 9.5 days

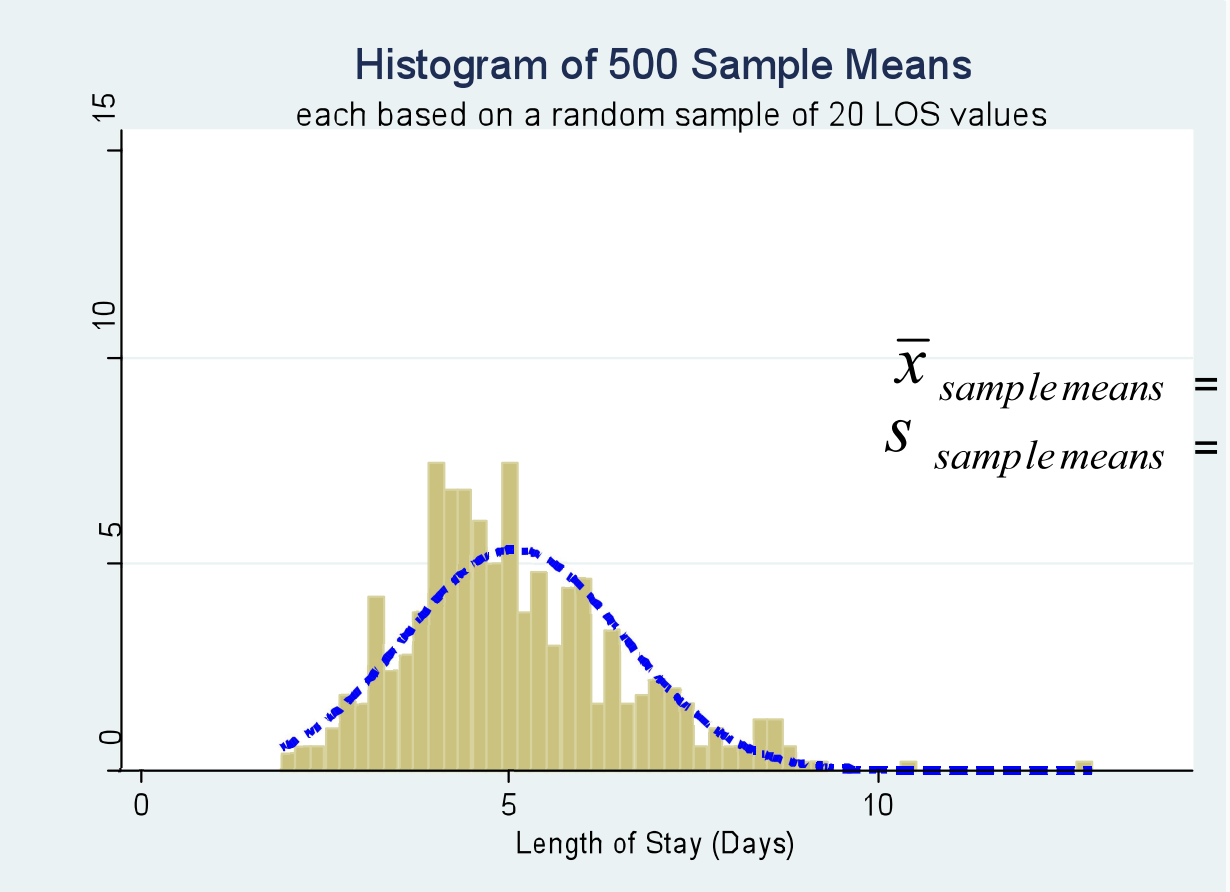
- Sample 2: n = 20



$\overline{x}_{LOS}$ = 4.8 days
s_{LOS} = 4.2 days

Example 2: Hospital Length of Stay

- So we did this 500 times: now let's look at a histogram of the 500 sample means

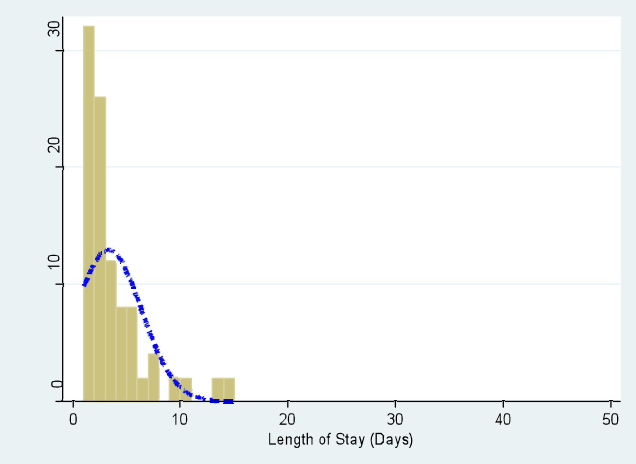


Example 2: Hospital Length of Stay

- Suppose we had all the time in the world again
- We decide to do one more experiment
- We are going to take 500 separate random samples from this population of me, each with 50 subjects
- For each of the 500 samples, we will plot a histogram of the sample LOS values, and record the sample mean and sample standard deviation
- Ready, set, go . . .

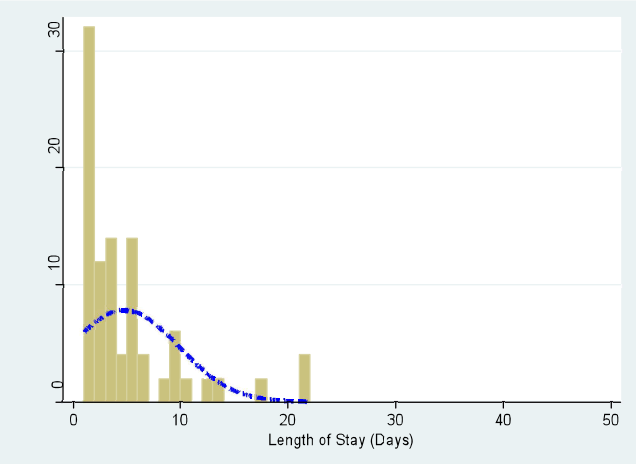
Random Samples

- Sample 1: n = 50



$\bar{x}_{LOS}$ = 3.3 days
s_{LOS} = 3.1 days

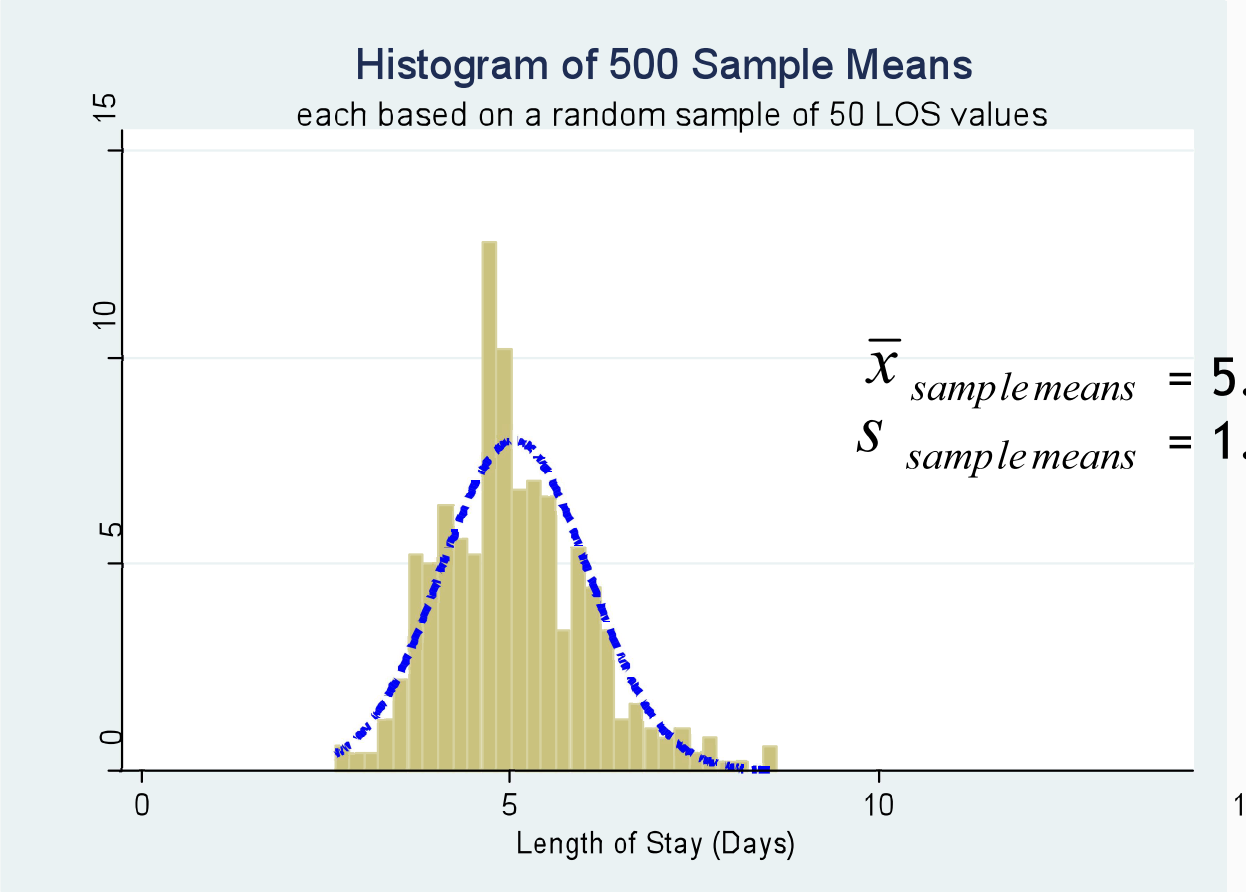
- Sample 2: n = 50



$\bar{x}_{LOS}$ = 4.7 days
s_{LOS} = 5.1 days

Distribution of Sample Means

- So we did this 500 times: now let's look at a histogram of the 500 sample means

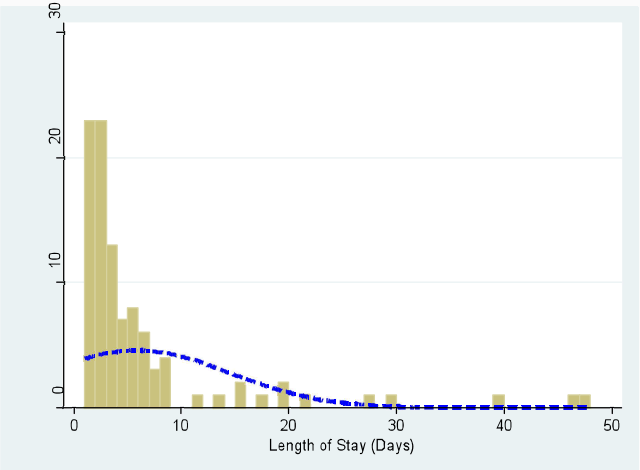


Example 2: Hospital Length of Stay

- Suppose we had all the time in the world again
- We decide to do one more experiment
- We are going to take 500 separate random samples from this population of me, each with 100 subjects
- For each of the 500 samples, we will plot a histogram of the sample BP values, and record the sample mean and sample standard deviation
- Ready, set, go . . .

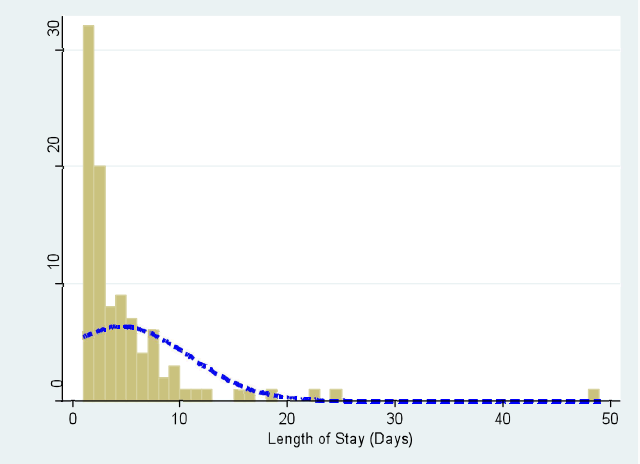
Random Samples

- Sample 1: *n* = 100



$\bar{x}_{LOS}$ = 5.8 days
s_{LOS} = 9.7 days

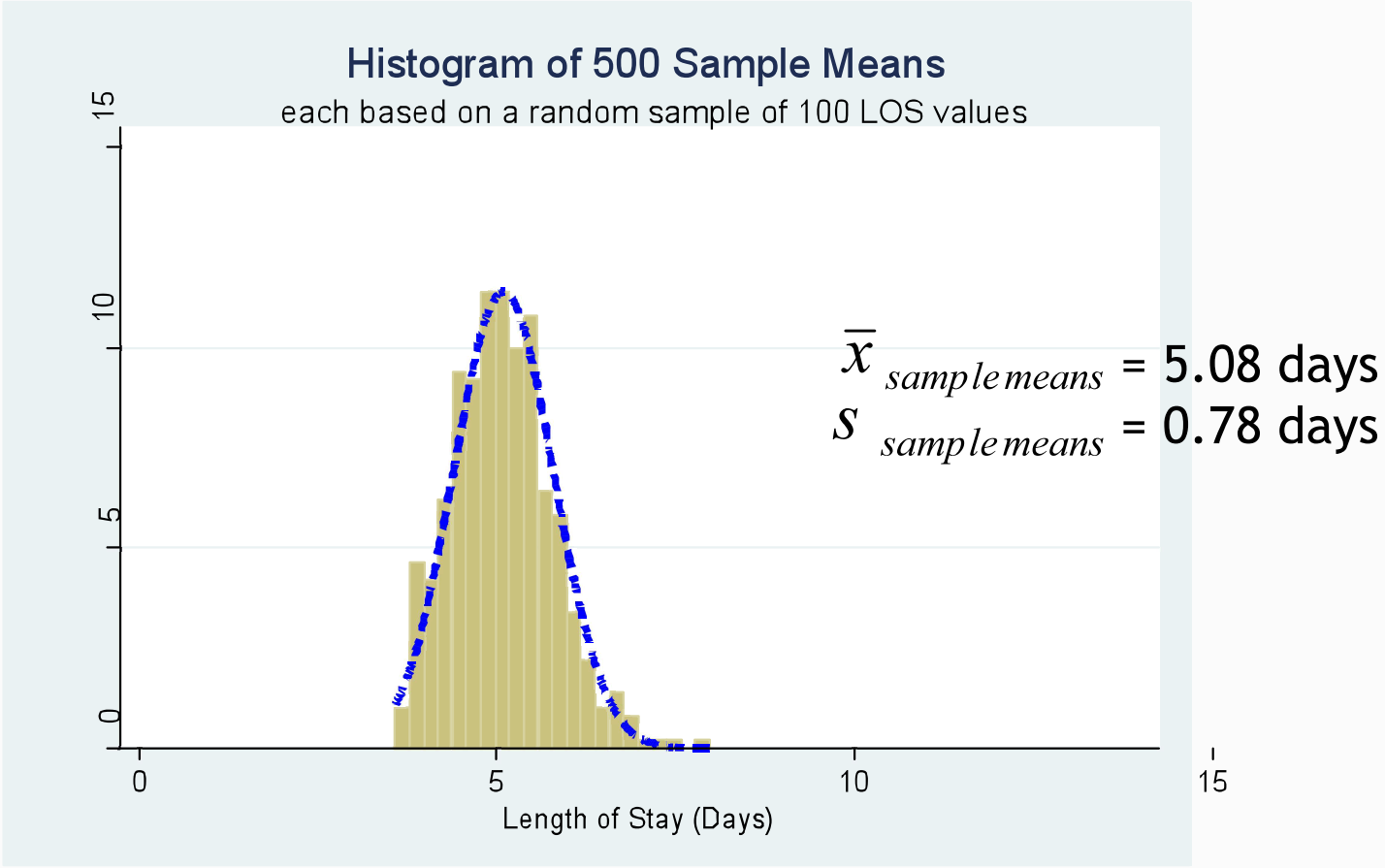
- Sample 2: *n* = 100



$\bar{x}_{LOS}$ = 4.5 days
s_{LOS} = 6.5 days

Distribution of Sample Means

- So we did this 500 times: now let's look at a histogram of the 500 sample means



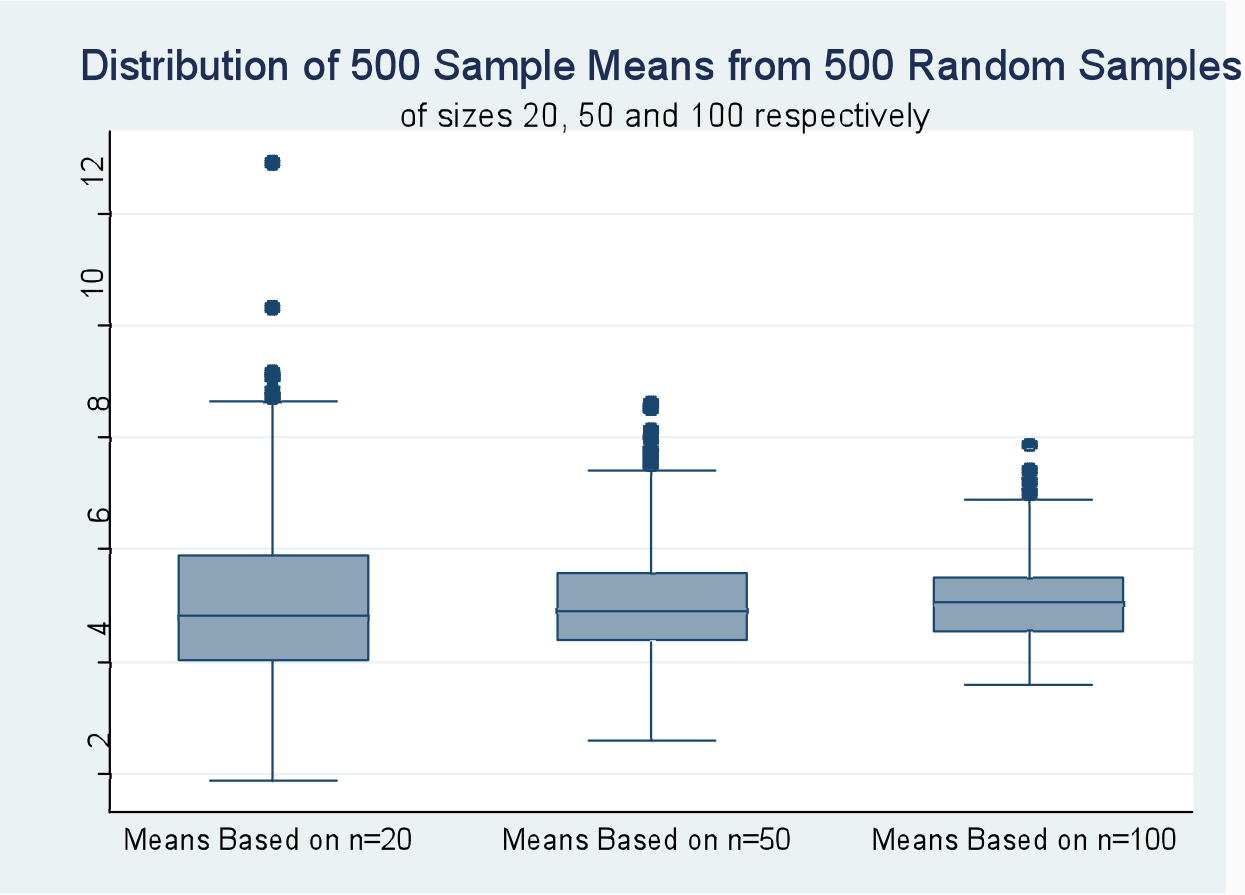
Example 2: Hospital Length of Stay

- Let's review the results
  - Population distribution of individual LOS values for population of patients: right skewed
  - True mean $\mu = 5.05$ days: $\sigma = 6.90$ days
  - Results from 500 random samples:

Sample Sizes	Means of 500 Sample Means	SD of 500 Sample Means	Shape of Distribution of 500 Sample Means
n = 20	5.05 days	1.49 days	Approx normal
n = 50	5.04 days	1.00 days	Approx normal
n = 100	5.08 days	0.70 days	Approx normal

Example 2: Hospital Length of Stay

- Let's review the results



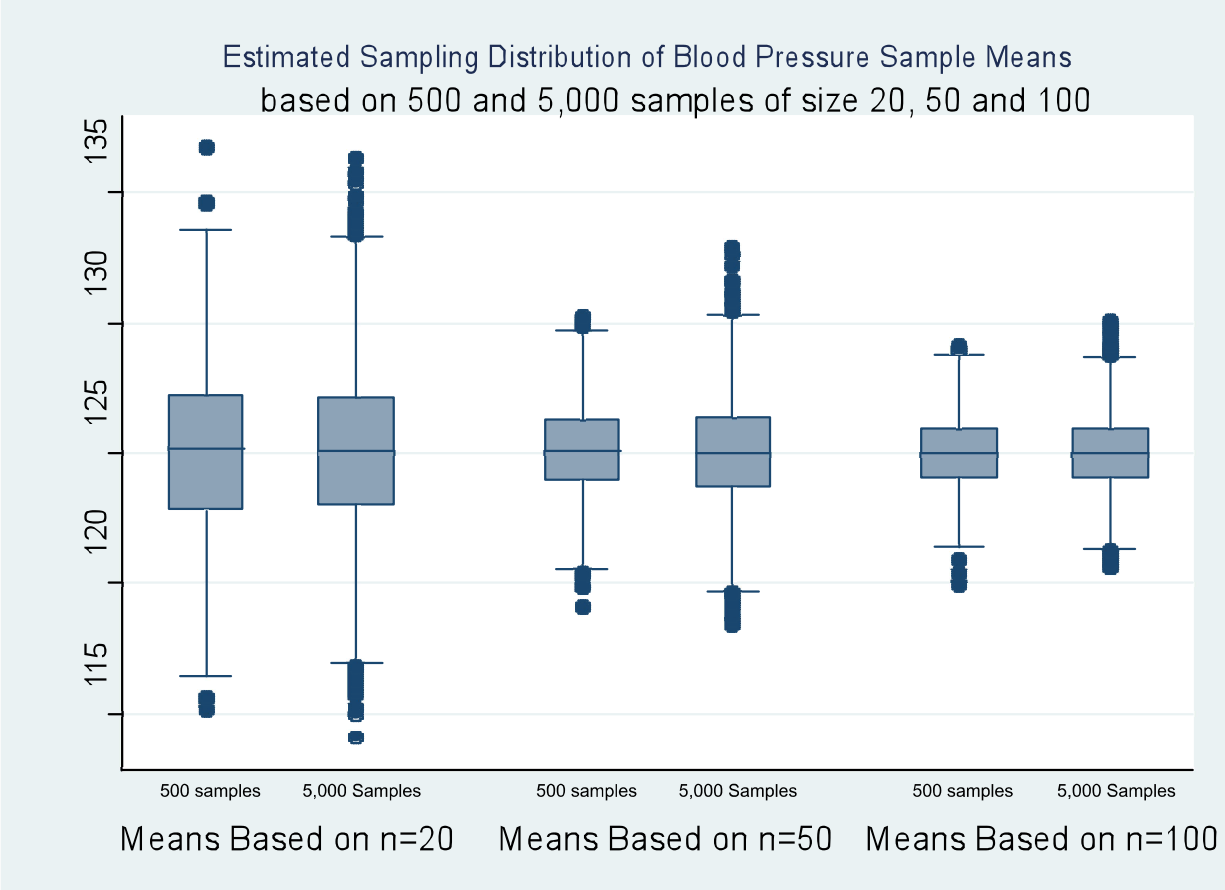
Summary

- What did we see across the two examples (BP of men, LOS for teaching hospital patients)?

- A couple of trends:
  - Distributions of sample means tended to be approximately normal even when original, individual level data was not (LOS)
  - Variability in sample mean values decreased as size of sample of each mean based upon increased
  - Distributions of sample means centered at true, population mean

Clarification

- Variation in sample mean values tied to size of each sample selected in our exercise: NOT the number of samples



JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section B

The Theoretical Sampling Distribution of the Sample Mean and Its Estimate Based on a Single Sample

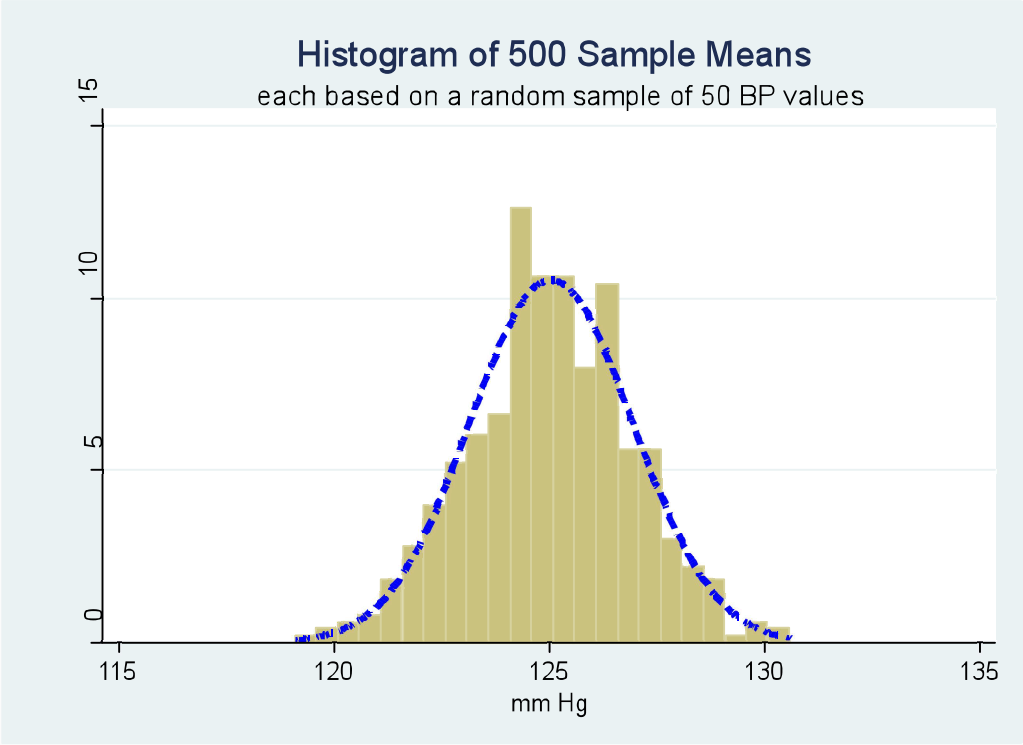
Sampling Distribution of the Sample Mean

- In the previous section we reviewed the results of simulations that resulted in estimates of what's formally called the sampling distribution of a sample mean

- The sampling distribution of a sample mean is a theoretical probability distribution; it describes the distribution of all sample means from all possible random samples of the same size taken from a population

Sampling Distribution of the Sample Mean

- For example: this histogram is an estimate of the sampling distribution of sample BP means based on random samples of $n = 50$ from the population of (BP measurements for) all men



Sampling Distribution of the Sample Mean

- In real research it is impossible to estimate the sampling distribution of a sample mean by actually taking multiple random samples from the same population, no research would ever happen if a study needed to be repeated multiple times to understand this sampling behavior

- Simulations are useful to illustrate a concept, but not to highlight a practical approach!

- Luckily, there is some mathematical machinery that generalizes some of the patterns we saw in the simulation results

The Central Limit Theorem (CLT)

- The Central Limit Theorem (CLT) is a powerful mathematical tool that gives several useful results
  - The sampling distribution of sample means based on all samples of same size *n* is approximately normal, regardless of the distribution of the original (individual level) data in the population/samples
  - The mean of all sample means in the sampling distribution is the true mean of the population from which the samples were taken, μ
  - Standard deviation in the sample means of size *n* is equal to $\frac{\sigma}{\sqrt{n}}$: this is often called the standard error of the sample mean and sometimes written as $SE(\bar{x})$

Example: Blood Pressure of Males

- Population distribution of individual BP measurements for males: normal

- True mean $\mu = 125$ mmHg: $\sigma = 14$ mmHg

Sample Sizes	Means of 500 Sample Means	Means of 5000 Sample Means	SD of 500 Sample Means	SD of 5000 Sample Means	SD of Sample Means (SE) by CLT
n = 20	124.98 mmHg	125.05 mmHg	3.31 mmHg	3.11 mmHg	3.13 mmHg
n = 50	125.03 mmHg	125.01 mmHg	1.89 mmHg	1.96 mmHg	1.98 mmHg
n = 100	124.99 mmHg	125.01 mmHg	1.43 mmHg	1.39 mmHg	1.40 mmHg

Example: Blood Pressure of Males

- Population distribution of individual BP measurements for males: normal

- True mean $\mu = 125$ mmHg: $\sigma = 14$ mmHg

Sample Sizes	Means of 500 Sample Means	Means of 5000 Sample Means	SD of 500 Sample Means	SD of 5000 Sample Means	SD of Sample Means (SE) by CLT
n = 20	124.98 mmHg	125.05 mmHg	3.31 mmHg	3.11 mmHg	3.13 mmHg
n = 50	125.03 mmHg	125.01 mmHg	1.89 mmHg	1.96 mmHg	1.98 mmHg
n = 100	124.99 mmHg	125.01 mmHg	1.43 mmHg	1.39 mmHg	1.40 mmHg

Example: Blood Pressure of Males

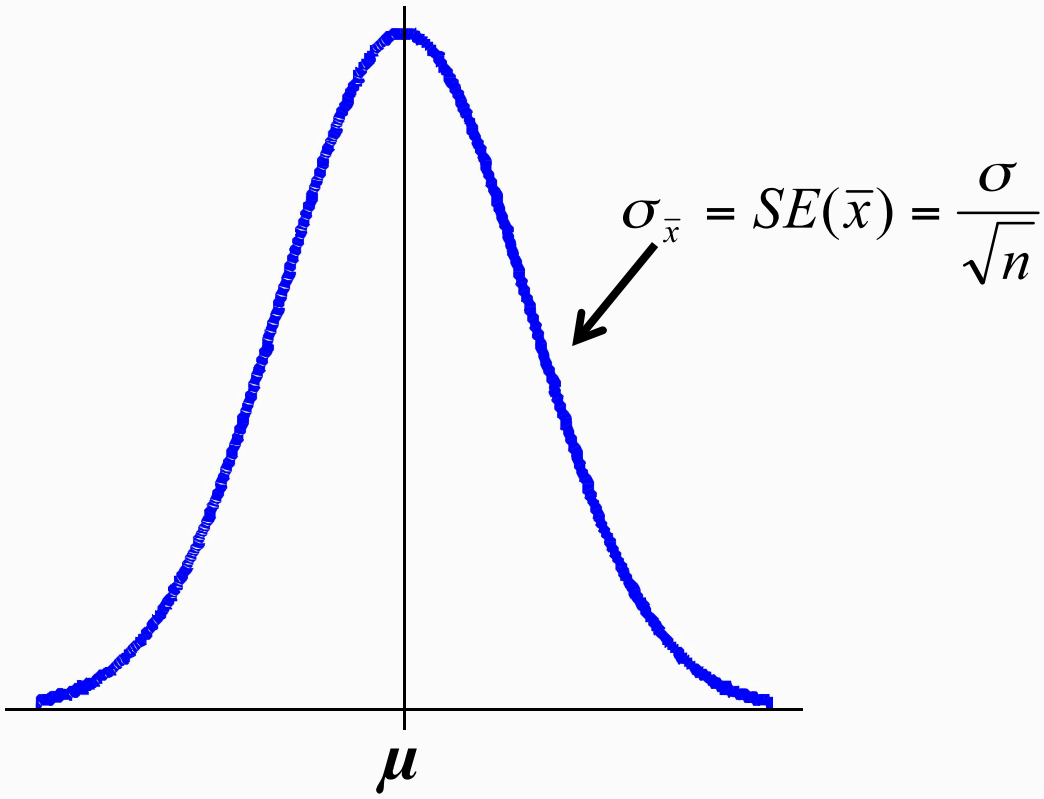
- Population distribution of individual BP measurements for males: Normal

- True mean $\mu = 125$ mmHg: $\sigma = 14$ mmHg

Sample Sizes	Means of 500 Sample Means	Means of 5000 Sample Means	SD of 500 Sample Means	SD of 5000 Sample Means	SD of Sample Means (SE) by CLT
n = 20	124.98 mmHg	125.05 mmHg	3.31 mmHg	3.11 mmHg	3.13 mmHg
n = 50	125.03 mmHg	125.01 mmHg	1.89 mmHg	1.96 mmHg	1.98 mmHg
n = 100	124.99 mmHg	125.01 mmHg	1.43 mmHg	1.39 mmHg	1.40 mmHg

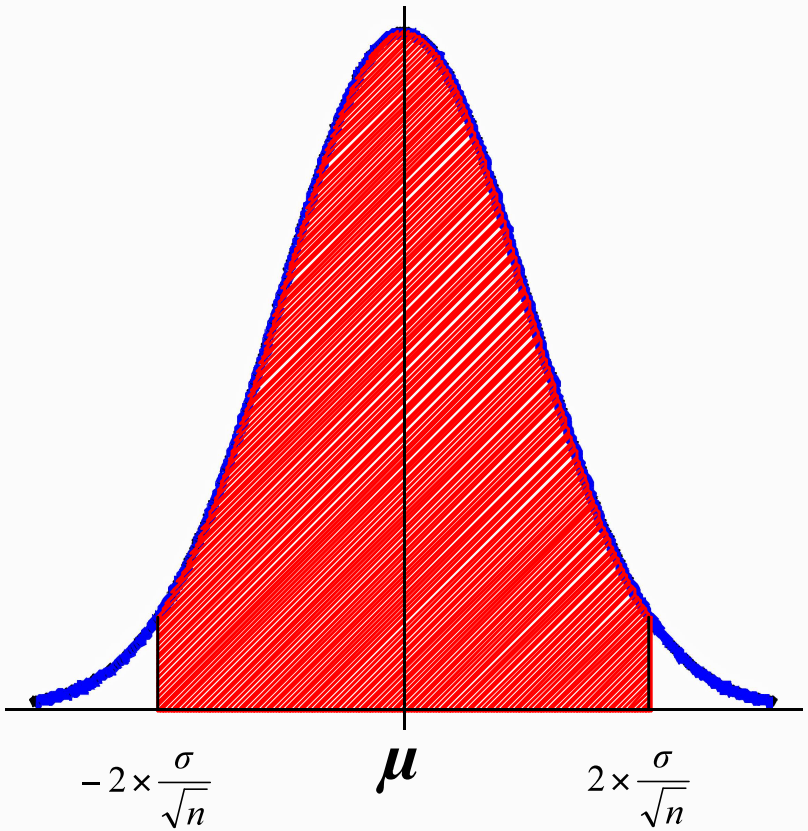
Recap: CLT

- So the CLT tells us the following:
  - When taking a random sample of continuous measures of size *n* from a population with true mean μ and true sd σ the theoretical sampling distribution of sample means from all possible random samples of size *n* is as follows:



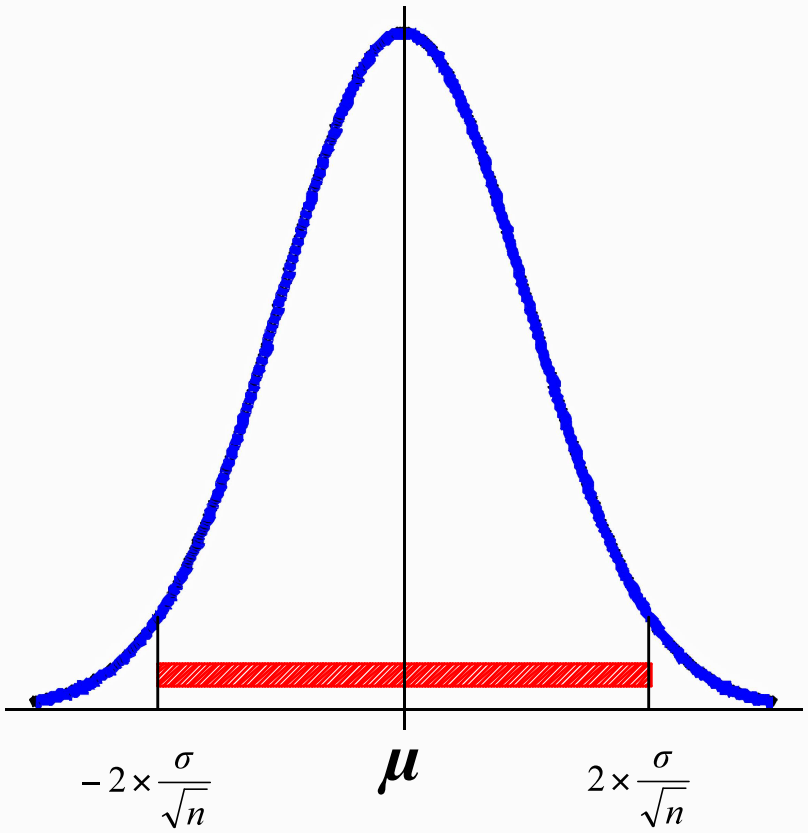
CLT: So What?

- So what good is this info?
  - Well using the properties of the normal curve, this shows that for most random samples we can take (95%), the sample mean will fall within 2 SEs of the true mean μ: $\overline{x}$



CLT: So What?

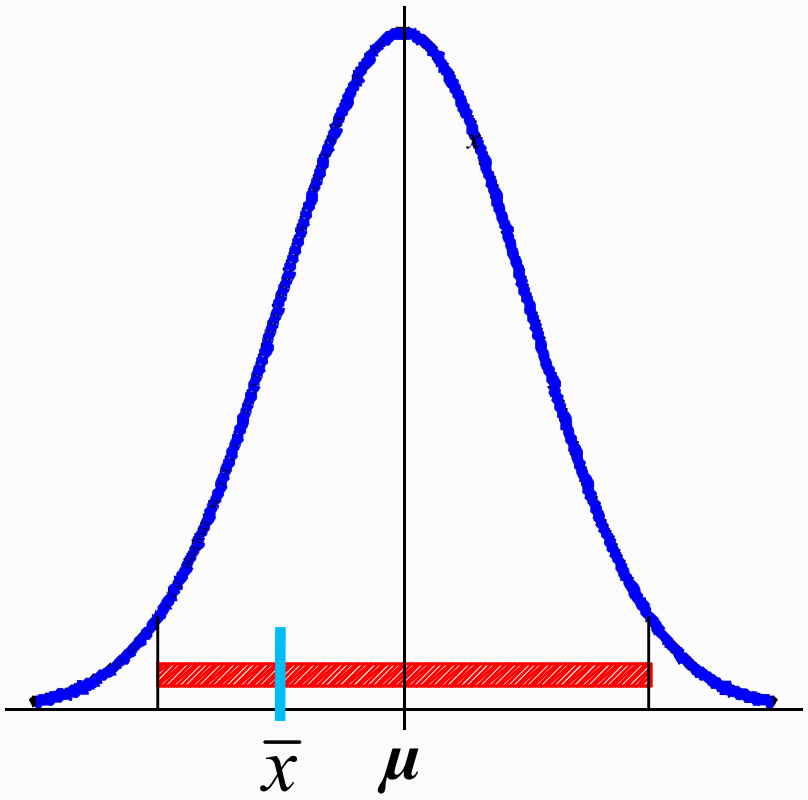
- So AGAIN what good is this info?
  - We are going to take a single sample of size *n* and get one $\overline{x}$
  - So we won't know μ, and if we did know μ why would we care about the distribution of estimates of μ from imperfect subsets of the population?



CLT: So What?

- We are going to take a single sample of size *n* and get one $\bar{x}$

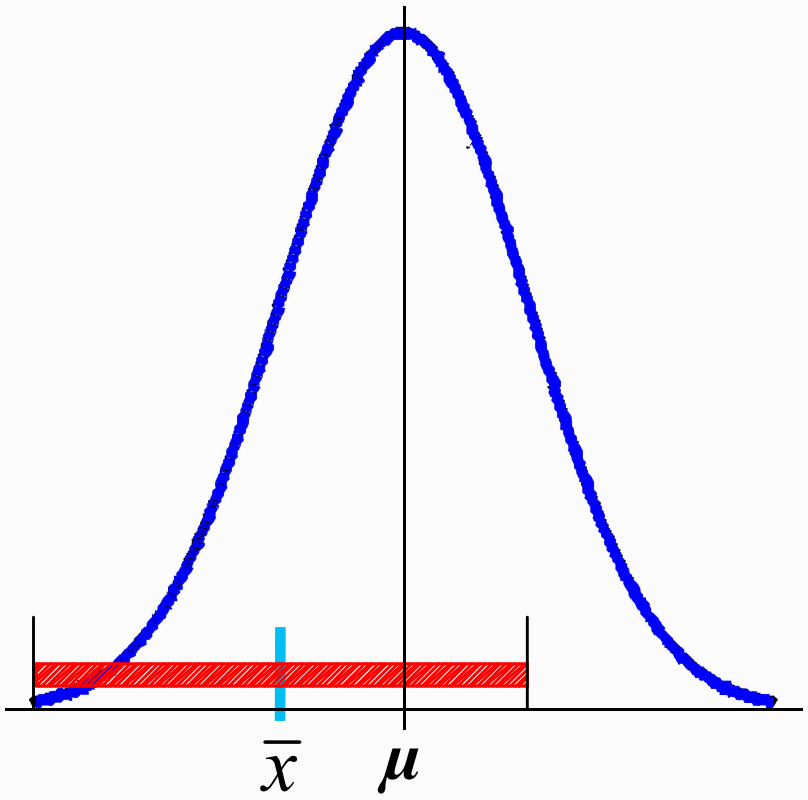
- But for most (95%) of the random samples we can get, our $\bar{x}$ will fall within +/- 2SEs of μ



CLT: So What?

- We are going to take a single sample of size *n* and get one $\bar{x}$

- So if we start at $\bar{x}$ and go 2SEs in either direction, the interval created will contain μ most (95 out of 100) of the time



Estimating a Confidence Interval

- Such an interval is a called a 95% confidence interval for the population mean μ
- Interval given by $\bar{x} \pm 2SE(\bar{x}) \rightarrow \bar{x} \pm 2 * \frac{\sigma}{n}$
- Problem: we don't know σ either
  - Can estimate with s, will detail this in next section
- What is interpretation of a confidence interval?

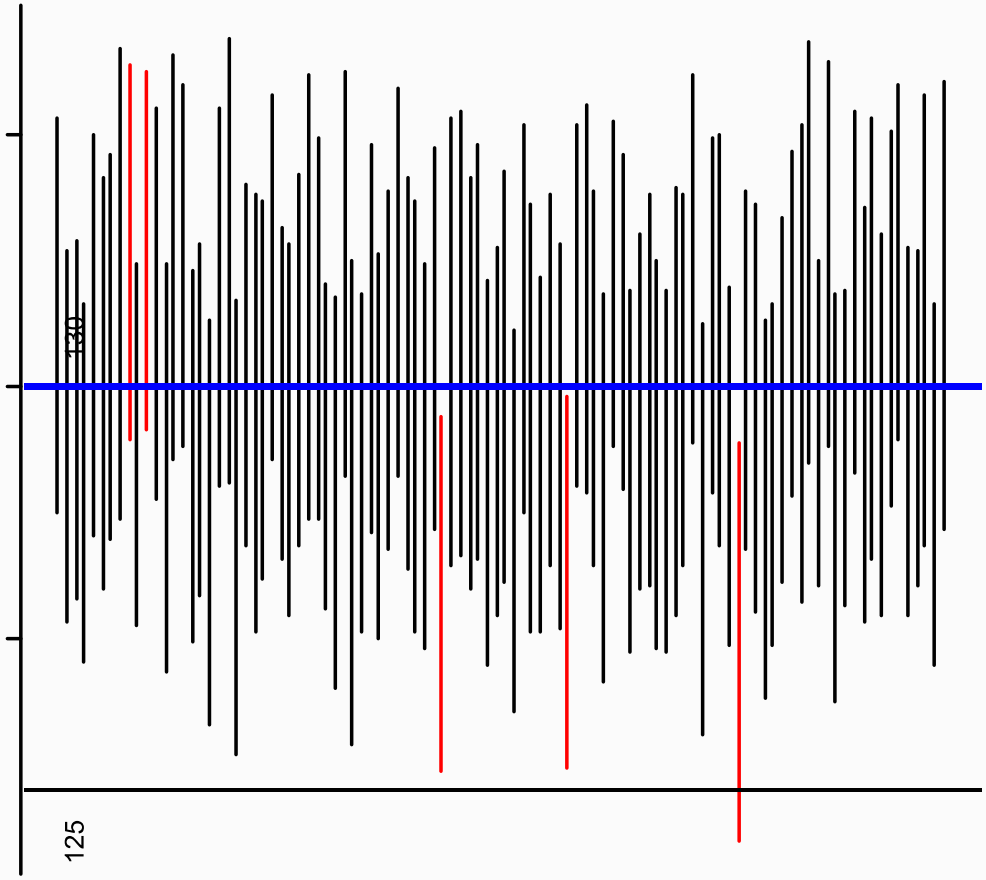
Interpretation of a 95% Confidence Interval (CI)

- Laypersons' range of "plausible" values for true mean
  - Researcher never can observe true mean μ
  - $\overline{x}$ is the best estimate based on a single sample
  - The 95% CI starts with this best estimate and additionally recognizes uncertainty in this quantity

- Technical
  - Were 100 random samples of size *n* taken from the same population, and 95% confidence intervals computed using each of these 100 samples, 95 of the 100 intervals would contain the values of true mean μ within the endpoints

Technical Interpretation

- One hundred 95% confidence intervals from 100 random samples of size n = 50: Blood Pressure for Males



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)

Notes on Confidence Intervals

- Are all CIs 95%?
  - No
  - It is the most commonly used
  - A 99% CI is wider
  - A 90% CI is narrower

- To change level of confidence adjust number of SE added and subtracted from $\overline{x}$
  - For a 99% CI, you need ± 2.6 SE
  - For a 95% CI, you need ± 2 SE
  - For a 90% CI, you need ± 1.65 SE

Semantic:Standard Deviation vs. Standard Error

- The term “standard deviation” refers to the variability in individual observations in a single sample (s) or population (σ)

- The standard error of the mean is also a measure of standard deviation, but not of individual values, rather variation in multiple sample means computed on multiple random samples of the same size, taken from the same population

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section C

Estimating Confidence Intervals for the Mean of a Population Based on a Single Sample of Size *n*: Some Examples

Estimating a 95% Confidence Interval

- In last section we defined a 95% confidence interval for the population mean μ
- Interval given by $\bar{x} \pm 2SE(\bar{x}) : \bar{x} \pm 2^* \frac{\sigma}{\sqrt{n}}$
- Problem: we don't know σ either
  - Can estimate with s, such that our estimated SE is
  - $S\hat{E}(\bar{x}) = \frac{s}{\sqrt{n}}$
- Estimated 95% CI for μ based on a single sample of size n
  - $\bar{x} \pm 2^* \frac{s}{\sqrt{n}}$

Example 1

- Suppose we had blood pressure measurements collected from a random sample of 100 Hopkins students collected in September 2008

- We wish to use the results of the sample to estimate a 95% CI for the mean blood pressure of all Hopkins students

- Results:
  - $\bar{x}$ = 123.4 mmHg; s = 13.7 mmHg
  - $\hat{SE}(\bar{x}) \frac{13}{\sqrt{100}} = 1.3 \text{ mmHg}$

- So a 95% CI for the true mean BP of all Hopkins Students:
  - 123.4 ± 2×1.3 → **123.4** ± 2.6
  - → (120.8 mmHg, 126.0 mmHg)

Example 2

- Data from the National Medical Expenditures Survey (1987):
  - U.S. Based Survey Administered by the Centers for Disease Control (CDC)

- Some results:

	Smoking History	No Smoking History
Mean 1987 Expenditures (U.S. $)	2,260	2,080
SD (U.S. $)	4,850	4,600
N	6,564	5,016

Example 2

- 95% CIs for 1987 medical expenditures by smoking history

- Smoking history: $2{,}260 \pm 2 \times \frac{4{,}850}{\sqrt{6{,}564}} \rightarrow 2{,}260 \pm 120 \rightarrow (\$2{,}140\,,\,\$2{,}380)$

- No smoking history: $2{,}080 \pm 2 \times \frac{4{,}600}{\sqrt{5{,}016}} \rightarrow 2{,}080 \pm 130 \rightarrow (\$1{,}950\,,\,\$2{,}210)$

Example 3

- Effect of lower targets for blood pressure and LDL cholesterol on atherosclerosis in diabetes: the SANDS Randomized Trial[1]
  - "**Objective**: To compare progression of subclinical atherosclerosis in adults with type 2 diabetes treated to reach aggressive targets of low-density lipoprotein cholesterol (LDL-C) of 70 mg/dL or lower and systolic blood pressure (SBP) of 115 mm Hg or lower vs standard targets of LDL-C of 100 mg/dL or lower and SBP of 130 mm Hg or lower."

Notes: [1] Howard, B., et al. *(2008)*. Effect of lower targets for blood pressure and LDL cholesterol on atherosclerosis in diabetes: The SANDS Randomized Trial. *Journal of the American Medical Association 299, no. 14.*

Example 3

- "**Design, setting, and participants**: a randomized, open-label, blinded-to-end point, three-year trial from April 2003-July 2007 at four clinical centers in Oklahoma, Arizona, and South Dakota. Participants were 499 American Indian men and women aged 40 years or older with type 2 diabetes and no prior CVD events."

- "**Interventions**: participants were randomized to aggressive (n = 252) vs. standard (n = 247) treatment groups with stepped treatment algorithms defined for both."

Example 3

- **Results mean**: target LDL-C and SBP levels for both groups were reached and maintained
  - Mean (95% confidence interval) levels for LDL-C in the last 12 months were 72 (69-75) and 104 (101-106) mg/dL and SBP levels were 117 (115-118) and 129 (128-130) mmHg in the aggressive vs. standard groups, respectively

Example 3

- Lots of 95% CIS!

Table 2. Differences in Mean Changes From Baseline to 36 Months, Aggressive vs Standard Groups[a]

	Mean (95% Confidence Interval)							
	Baseline		36 mo[b]		Change at 36 mo			*P* Value for Difference
	Aggressive	Standard	Aggressive	Standard	Aggressive	Standard	Difference	
Weight, kg	90 (88 to 93)	90 (88 to 92)	91 (89 to 94)	91 (88 to 93)	1.0 (−0.8 to 2.2)	1.0 (−0.3 to 2.3)	0.3 (−1.7 to 2.3)	.83
BMI[c]	34 (33 to 34)	33 (32 to 34)	34 (33 to 35)	34 (33 to 34.4)	0.3 (−0.3 to 0.9)	0.4 (−0.1 to 0.9)	0.1 (−0.6 to 0.9)	.77
Waist, cm	110 (108 to 112)	110 (108 to 112)	111 (109 to 113)	110 (108 to 112)	0.2 (−1.0 to 1.6)	0.6 (−0.7 to 2.0)	0.4 (−1.5 to 2.3)	.66
CRP mg/L[d]	2.7 (2.3 to 3.1)	2.8 (2.4 to 3.3)	2.2 (1.9 to 2.7)	3.3 (2.8 to 3.8)	−0.7 (11)[e]	0.9 (9)[e]	1.6 (−0.4 to 3.6)[e]	.12[e]
DBP, mm Hg	74 (73 to 76)	76 (75 to 78)	67 (66 to 68)	73 (72 to 74)	−7 (−8 to −6)	−3 (−4 to −1)	4.0 (2.5 to 5.5)[f]	<.001
SBP, mm Hg	128 (126 to 130)[g]	133 (131 to 135)[g]	117 (115 to 118)	129 (128 to 130)	−11 (−13 to −9)	−3 (−5 to −1)	8 (6 to 12)[f]	<.001
Glucose, mg/dL	159 (151 to 168)	156 (147 to 166)	169 (158 to 179)	169 (158 to 180)	11 (1 to 23)	14 (1 to 28)	4 (−14 to 22)	.68
HDL-C, mg/dL	46 (44 to 48)	46 (44 to 47)	48 (47 to 50)	48 (47 to 50)	3.0 (1.4 to 3.8)	3.0 (1.2 to 3.9)	0.1 (−1.9 to 1.8)	.94
LDL-C, mg/dL	104 (100 to 108)	104 (100 to 108)	72 (69 to 75)	104 (101 to 106)	−31 (−35 to −26)	1 (−3 to 6)	32 (26 to 38)[f]	<.001
Non-HDL-C, mg/dL	138 (134 to 142)	140 (136 to 144)	102 (98 to 106)	138 (135 to 141)	−35 (−40 to −30)	0.2 (−4.4 to 4.9)	35 (28 to −42)[f]	<.001
TC, mg/dL	184 (180 to 188)	185 (181 to 190)	150 (146 to 154)	187 (183 to 190)	−32 (−37 to −27)	3 (−2 to 8)	35 (27 to 42)[f]	<.001
TC/HDL-C, mg/dL	4.2 (4.1 to 4.4)	4.2 (4.1 to 4.4)	3.3 (3.1 to 3.4)	4.0 (3.9 to 4.2)	−1.0 (−1.1 to −0.8)	−0.1 (−0.3 to 0.0)	0.8 (0.6 to 1.0)[f]	<.001
Triglycerides, mg/dL[d]	158 (149 to 167)	168 (159 to 177)	137 (130 to 144)	160 (153 to 168)	−26 (78)[e]	−12 (84)[e]	14 (−3 to 29)[ef]	.06[e]
Hemoglobin A_{1c}	8.2 (7.9 to 8.4)	7.9 (7.6 to 8.1)	8.3 (8.0 to 8.6)	8.2 (7.8 to 8.5)	0.1 (−0.2 to 0.4)	0.3 (−0.1 to 0.6)	0.2 (−0.3 to 0.6)	.45

Abbreviations: BMI, body mass index; CRP, C-reactive protein; DBP, diastolic blood pressure; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; SBP, systolic blood pressure; TC, total cholesterol.

SI conversions: for CRP to nmol/L, multiply by 9.524; for glucose to mmol/L, multiply by 0.0555; for HDL-C, LDL-C, and TC to mmol/L multiply by 0.0259; for hemoglobin A_{1c} to proportion of total hemoglobin, multiply by 0.01; and for triglycerides to mmol/L, multiply by 0.0113.

[a] Twenty-three baseline variables are compared and presented in Tables 1 and 2.

[b] N for the 36-mo lipids variables was 458 and the mean values were based on the average of 24-, 30- and 36-month observations.

[c] BMI is calculated as weight in kilograms divided by height in meters squared.

[d] Geometric mean (95% confidence interval).

[e] *P* value is based on arithmetic mean.

[f] Significant mean difference at 36 mo: DBP, LDL-C, non-HDL-C, SBP, TC, TC/LDL-C, and triglycerides ($P < .001$).

[g] Significant differences at baseline for SBP, $P = .003$.

Example 3

- Lots of 95% CIS!

Table 2. Differences in Mean Changes From Baseline to 36 Months, Aggressive vs Standard Groups[a]

	Mean (95% Confidence Interval)							
	Baseline		36 mo[b]		Change at 36 mo			*P* Value for Difference
	Aggressive	Standard	Aggressive	Standard	Aggressive	Standard	Difference	
Weight, kg	90 (88 to 93)	90 (88 to 92)	91 (89 to 94)	91 (88 to 93)	1.0 (−0.8 to 2.2)	1.0 (−0.3 to 2.3)	0.3 (−1.7 to 2.3)	.83
BMI[c]	34 (33 to 34)	33 (32 to 34)	34 (33 to 35)	34 (33 to 34.4)	0.3 (−0.3 to 0.9)	0.4 (−0.1 to 0.9)	0.1 (−0.6 to 0.9)	.77
Waist, cm	110 (108 to 112)	110 (108 to 112)	111 (109 to 113)	110 (108 to 112)	0.2 (−1.0 to 1.6)	0.6 (−0.7 to 2.0)	0.4 (−1.5 to 2.3)	.66
CRP mg/L[d]	2.7 (2.3 to 3.1)	2.8 (2.4 to 3.3)	2.2 (1.9 to 2.7)	3.3 (2.8 to 3.8)	−0.7 (11)[e]	0.9 (9)[e]	1.6 (−0.4 to 3.6)[e]	.12[e]
DBP, mm Hg	74 (73 to 76)	76 (75 to 78)	67 (66 to 68)	73 (72 to 74)	−7 (−8 to −6)	−3 (−4 to −1)	4.0 (2.5 to 5.5)[f]	<.001
SBP, mm Hg	128 (126 to 130)[g]	133 (131 to 135)[g]	117 (115 to 118)	129 (128 to 130)	−11 (−13 to −9)	−3 (−5 to −1)	8 (6 to 12)[f]	<.001
Glucose, mg/dL	159 (151 to 168)	156 (147 to 166)	169 (158 to 179)	169 (158 to 180)	11 (1 to 23)	14 (1 to 28)	4 (−14 to 22)	.68
HDL-C, mg/dL	46 (44 to 48)	46 (44 to 47)	48 (47 to 50)	48 (47 to 50)	3.0 (1.4 to 3.8)	3.0 (1.2 to 3.9)	0.1 (−1.9 to 1.8)	.94
LDL-C, mg/dL	104 (100 to 108)	104 (100 to 108)	72 (69 to 75)	104 (101 to 106)	−31 (−35 to −26)	1 (−3 to 6)	32 (26 to 38)[f]	<.001
Non-HDL-C, mg/dL	138 (134 to 142)	140 (136 to 144)	102 (98 to 106)	138 (135 to 141)	−35 (−40 to −30)	0.2 (−4.4 to 4.9)	35 (28 to −42)[f]	<.001
TC, mg/dL	184 (180 to 188)	185 (181 to 190)	150 (146 to 154)	187 (183 to 190)	−32 (−37 to −27)	3 (−2 to 8)	35 (27 to 42)[f]	<.001
TC/HDL-C, mg/dL	4.2 (4.1 to 4.4)	4.2 (4.1 to 4.4)	3.3 (3.1 to 3.4)	4.0 (3.9 to 4.2)	−1.0 (−1.1 to −0.8)	−0.1 (−0.3 to 0.0)	0.8 (0.6 to 1.0)[f]	<.001
Triglycerides, mg/dL[d]	158 (149 to 167)	168 (159 to 177)	137 (130 to 144)	160 (153 to 168)	−26 (78)[e]	−12 (84)[e]	14 (−3 to 29)[ef]	.06[e]
Hemoglobin A_{1c}	8.2 (7.9 to 8.4)	7.9 (7.6 to 8.1)	8.3 (8.0 to 8.6)	8.2 (7.8 to 8.5)	0.1 (−0.2 to 0.4)	0.3 (−0.1 to 0.6)	0.2 (−0.3 to 0.6)	.45

Abbreviations: BMI, body mass index; CRP, C-reactive protein; DBP, diastolic blood pressure; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; SBP, systolic blood pressure; TC, total cholesterol.

SI conversions: for CRP to nmol/L, multiply by 9.524; for glucose to mmol/L, multiply by 0.0555; for HDL-C, LDL-C, and TC to mmol/L multiply by 0.0259; for hemoglobin A_{1c} to proportion of total hemoglobin, multiply by 0.01; and for triglycerides to mmol/L, multiply by 0.0113.

[a] Twenty-three baseline variables are compared and presented in Tables 1 and 2.

[b] N for the 36-mo lipids variables was 458 and the mean values were based on the average of 24-, 30- and 36-month observations.

[c] BMI is calculated as weight in kilograms divided by height in meters squared.

[d] Geometric mean (95% confidence interval).

[e] *P* value is based on arithmetic mean.

[f] Significant mean difference at 36 mo: DBP, LDL-C, non-HDL-C, SBP, TC, TC/LDL-C, and triglycerides ($P < .001$).

[g] Significant differences at baseline for SBP, $P = .003$.

Using Stata to Create 95% CI for a Mean

- The "cii" command
  - Syntax "cii *n* $\bar{x}$ *s*"
  - For example 1:
    $\bar{x}$ = 123.4 mm Hg; *s* = 13.7 mmHg; *n* = 100

```
. cii 100 123.4 1 3

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |        100       123.4         1.3       120.8205    125.9795
```

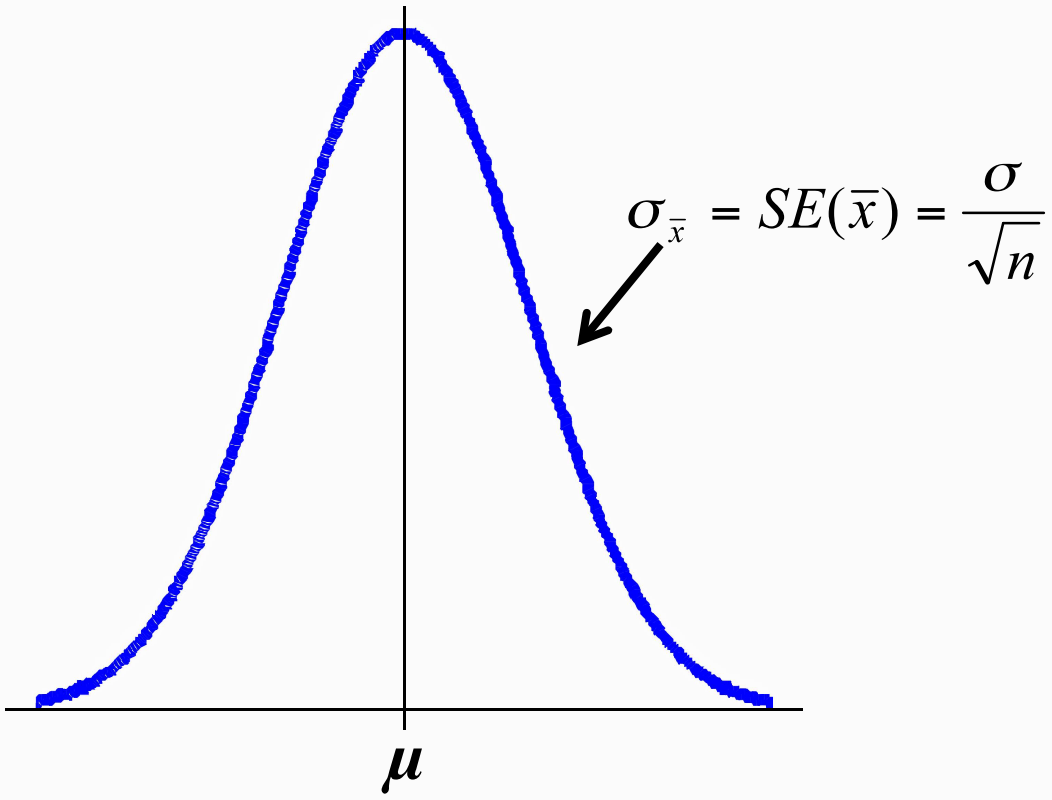
JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section D

True Confessions Biostat Style: What We Mean by Approximately Normal and What Happens to the Sampling Distribution of the Sample Mean with Small *n*

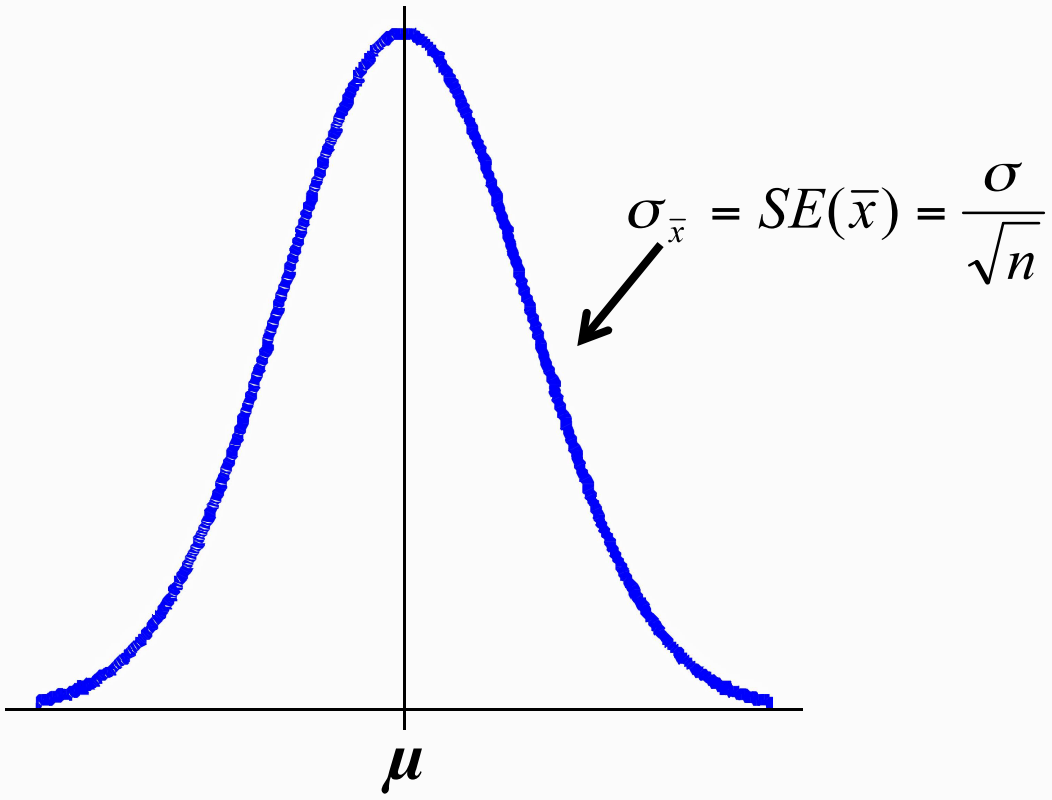
Recap: CLT

- So the CLT tells us the following: when taking a random sample of continuous measures of size *n* from a population with true mean μ and true sd σ the theoretical sampling distribution of sample means from all possible random samples of size *n* is:



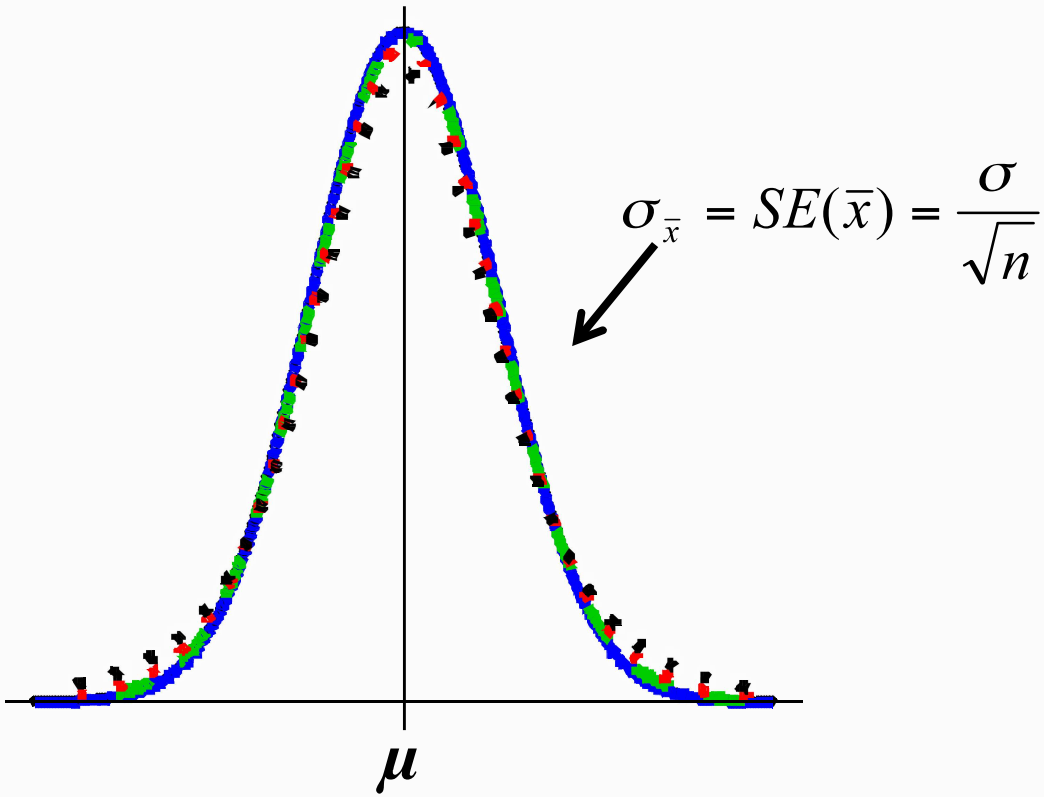
Recap: CLT

- Technically this is true for "large *n*": for this course, we'll say *n* > 60; but when *n* is smaller, sampling distribution is not quite normal, but follows a *t-distribution*



t-distributions

- The t-distribution is the “fatter, flatter cousin” of the normal: t-distribution is uniquely defined by degrees of freedom



Why the t?

- Basic idea: remember, the true SE($\bar{x}$) is given by the formula

$$\sigma_{\bar{x}} = SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

- But of course we don't know σ, and replace with s to estimate

$$\hat{SE}(\bar{x}) = \frac{s}{\sqrt{n}}$$

- In small samples, there is a lot of sampling variability in s as well: so this estimate is less precise

- To account for this additional uncertainty, we have to go slightly more than $\pm 2 \times \hat{SE}(\bar{x})$ to get 95% coverage under the sampling distribution

Underlying Assumptions

- How much bigger the *2* 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 SEs 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*
- One option: you can look up the correct number in a "t-table" or "t-distribution" with *n – 1* degrees of freedom

$$\bar{x} \pm t_{.95,n-1} \times S\hat{E}(\bar{x}) \Rightarrow$$

$$\bar{x} \pm t_{.95,n-1} \times \frac{s}{\sqrt{n}}$$

Notes on the t-Correction

- The particular t-table gives the number of SEs needed to cut off 95% under the sampling distribution

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. 360	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

- You can easily find a t-table for other cutoffs (90%, 99%) in any stats text or by searching the internet
- Also, using the *cii* command takes care of this little detail
- The point is not to spend a lot of time looking up t-values: more important is a basic understanding of why slightly more needs to be added to the sample mean in smaller samples to get a valid 95% CI
- The interpretation of the 95% CI (or any other level) is the same as discussed before

Example

- Small study on response to treatment among 12 patients with hyperlipidemia (high LDL cholesterol) given a treatment

- Change in cholesterol post–pre treatment computed for each of the 12 patients

- Results: $\bar{x}_{change} = -1.4\ mmol/L$

  $s_{change} = 0.55\ mmol/L$

Example

- 95% confidence interval for true mean change

$$\bar{x} \pm t_{.95,11} \times S\hat{E}(\bar{x}) \Rightarrow$$

$$\bar{x} \pm 2.2 \times S\hat{E}(\bar{x}) \Rightarrow$$

$$-1.4 \pm 2.2 \times \frac{0.55}{\sqrt{12}} \Rightarrow$$

$$(-1.75\ , mmol/L, -1.05\ mmol/L)$$

Using Stata to Create Other CIs for a Mean

- The “cii” command,

```
. cii 12  -1.4 .55

    Variable |        Obs        Mean    Std   . Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |         12        -1.4    .1587713        -1.749453   -1.050547
```

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section E

The Sample Proportion as a Summary Measure for Binary Outcomes and the CLT

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
- 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" responses divided by total sample size
  - This is the key summary measure for binary data, analogous to a mean for continuous data
  - There is a formula for the standard deviation of a proportion, but the quantity lacks the "physical interpretability" that it has for continuous data

Example 1

- Proportion of dialysis patients with national insurance in 12 countries (only six shown..)[1]

EXHIBIT 1
Descriptive Measures Of The Prevalent Cross-Sectional Patient Sample, Dialysis Patients In Twelve Countries, 2002–2004

	A/NZ (n = 561)	BEL (n = 468)	CAN (n = 503	FRA (n = 481)	GER (n = 524)	ITA (n = 540)
Mean age (years)	59.9 (14.7)	66.2 (13.4)	62.1 (14.7)	64.1 (14.5)	61.7 (14.1)	64 (13.7)
Minority[a]	21.5%	6.3%	18.7%	7.1%	0.4%	0.4%
Income ($US)						
<$20,000	85.0%	73.4%	71.8%	67.0%	59.7%	78.3%
$20,000–$39,000	9.1	17.5	20.8	21.8	27.1	17.4
≥$40,000	5.9	9.1	7.4	11.2	13.1	4.2
Insurance type						
National only	69.8%	74.1%	79.6%	45.5%	95.4%	99.6%
Private only	5.4	0.4	0.2	0.2	2.9	0.0
Mean number of comorbid conditions[b]	3.7 (2)	3.9 (2.1)	4.1 (2.1)	3.1 (1.9)	3.4 (2.1)	2.7 (1.9)
Mean number of prescribed medications	8.7 (3.6)	9.9 (4.1)	12.6 (4.8)	7.7 (3.5)	9.7 (3.5)	6.4 (3.6)

- Example: Canada: $\hat{p} = \frac{400}{503} = 0.796$

Notes: [1] Hirth, R., et al. (2008). Out-of-pocket spending and medication adherence among dialysis patients in twelve countries, *Health Affairs*, 27 (1).

Example 2

- Maternal/infant transmission of HIV[1]

- HIV-infection status was known for 363 births (180 in the zidovudine [AZT] group and 183 in the placebo group); thirteen infants in the zidovudine group and 40 in the placebo group were HIV-infected

$$\hat{p}_{AZT} = \frac{13}{180} = 0.07 = 7\%$$

$$\hat{p}_{PLAC} = \frac{40}{183} = 0.22 = 22\%$$

Notes: [1]Spector, S., et al. (1994). A controlled trial of intravenous immune globulin for the prevention of serious bacterial infections in children receiving zidovudine for advanced human immunodeficiency virus infection, *New England Journal of Medicine* 331 (18).

Proportions (p)

- What is the sampling behavior of a sample proportion?

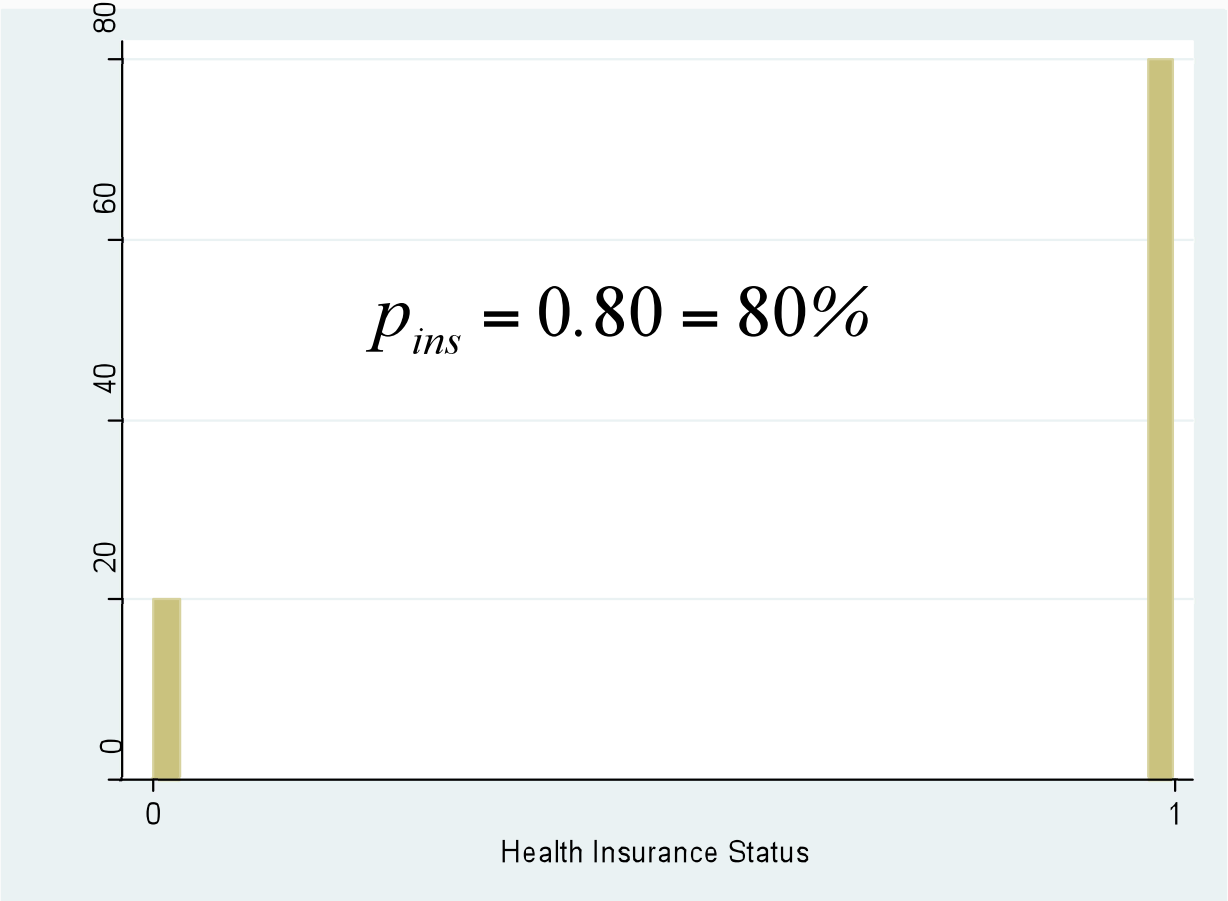
- In other words, how do sample proportions, estimated from random samples of the same size from the same population, behave?

Proportions (p)

- Suppose we have a population in which 80% of persons have some form of health insurance and 20% have no health insurance

Example: Health Insurance Coverage

- Assume the population distribution is given by the following:

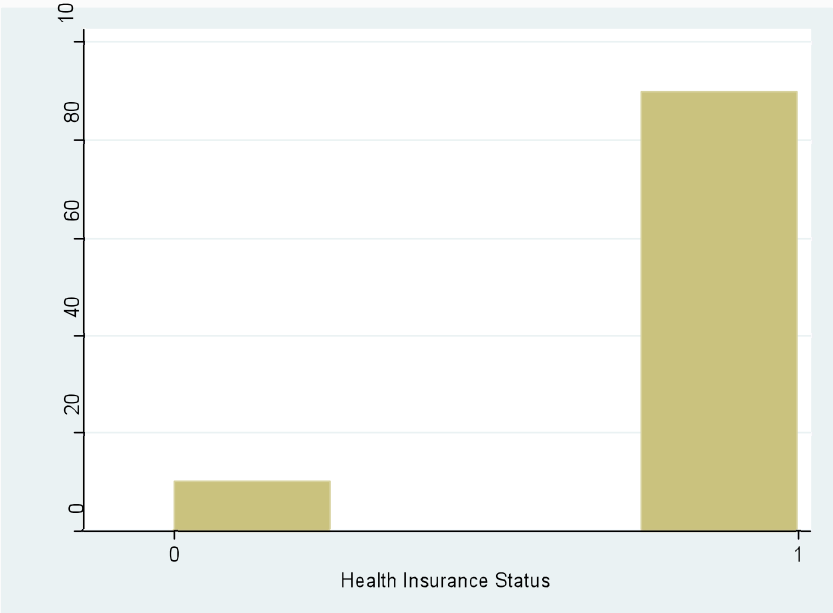


Example: Health Insurance Coverage

- Suppose we had all the time in the world (leftover from last time)
- We decide to do another set of experiments
- We are going to take 500 separate random samples from this population, each with 20 subjects
- For each of the 500 samples, we will plot a histogram of the sample proportion of insured individuals and record the sample proportion
- Ready, set, go . . .

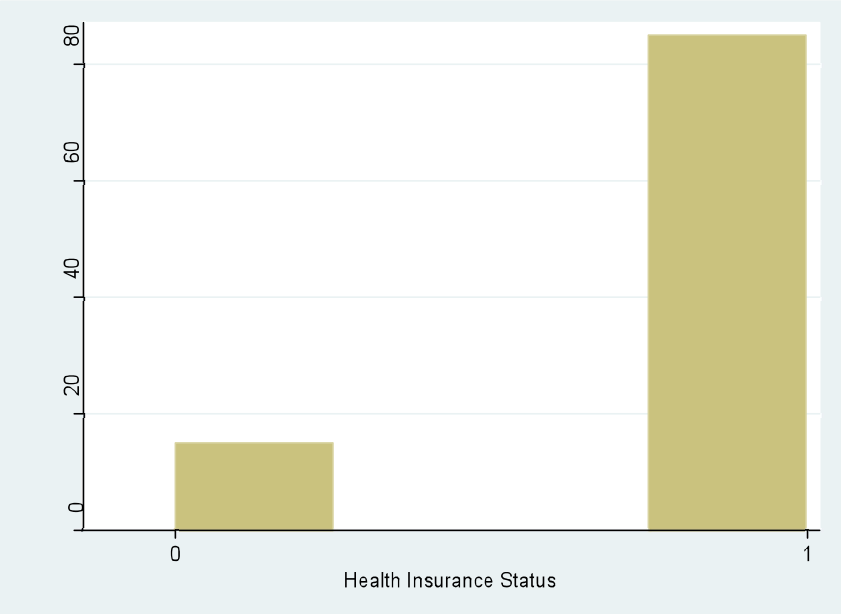
Random Samples

- Sample 1: $n = 20$



$$\hat{p}_{ins} = 0.90 = 90\%$$

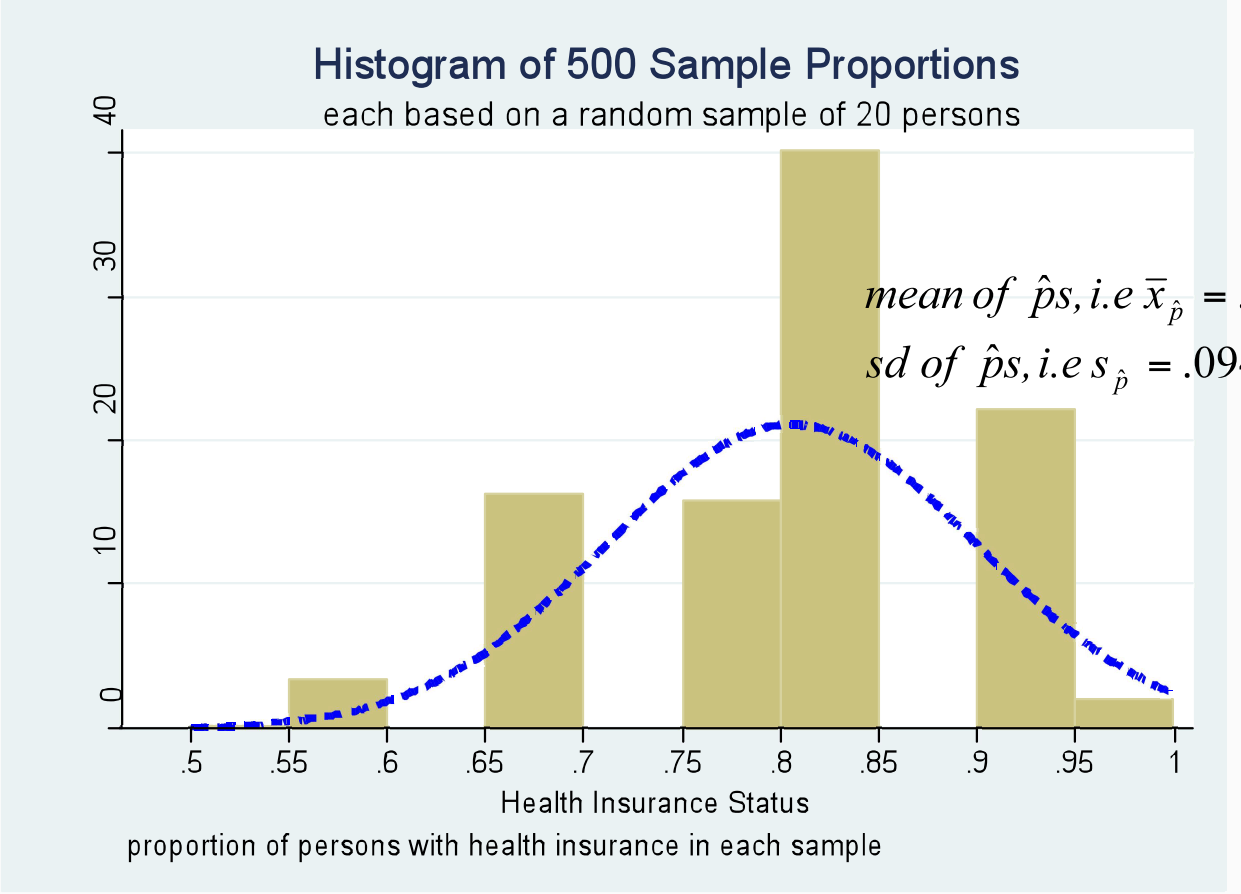
- Sample 2: $n = 20$



$$\hat{p}_{ins} = 0.85 = 85\%$$

Estimated Sampling Distribution

- So we did this 500 times: now let's look at a histogram of the 500 proportions

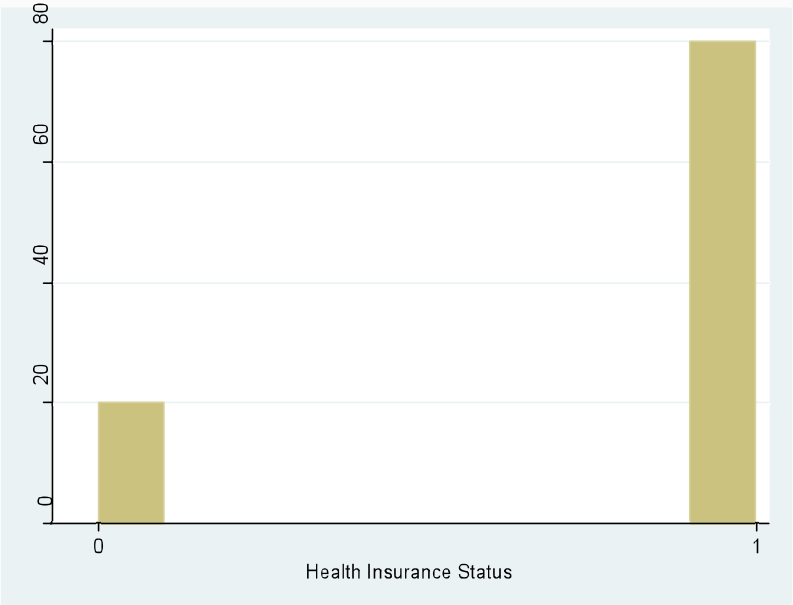


Example: Health Insurance Coverage

- We decide to do one more experiment
- We are going to take 500 separate random samples from this population, each with 100 subjects
- For each of the 500 samples, we will plot a histogram of the sample proportioned of insured individuals and record the sample proportion
- Ready, set, go . . .

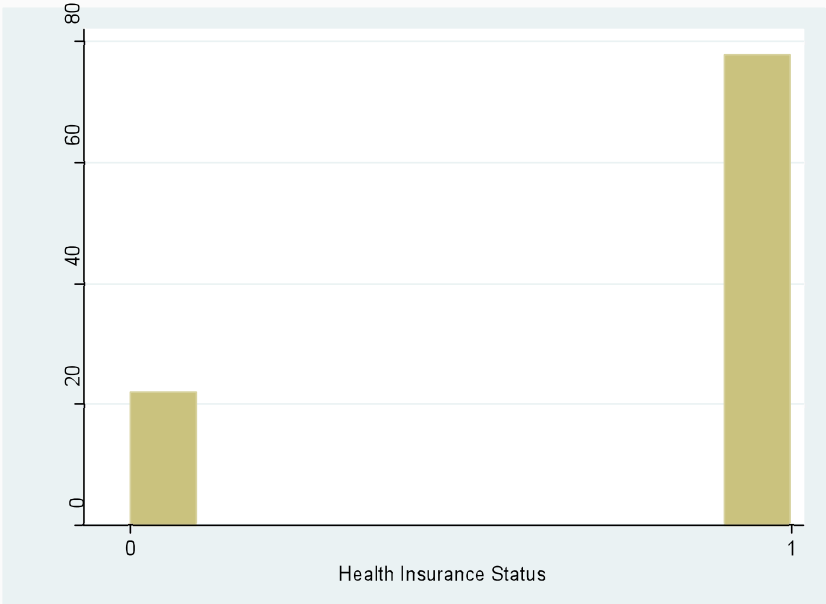
Random Samples

- Sample 1: $n = 100$



$$\hat{p}_{ins} = 0.80 = 80\%$$

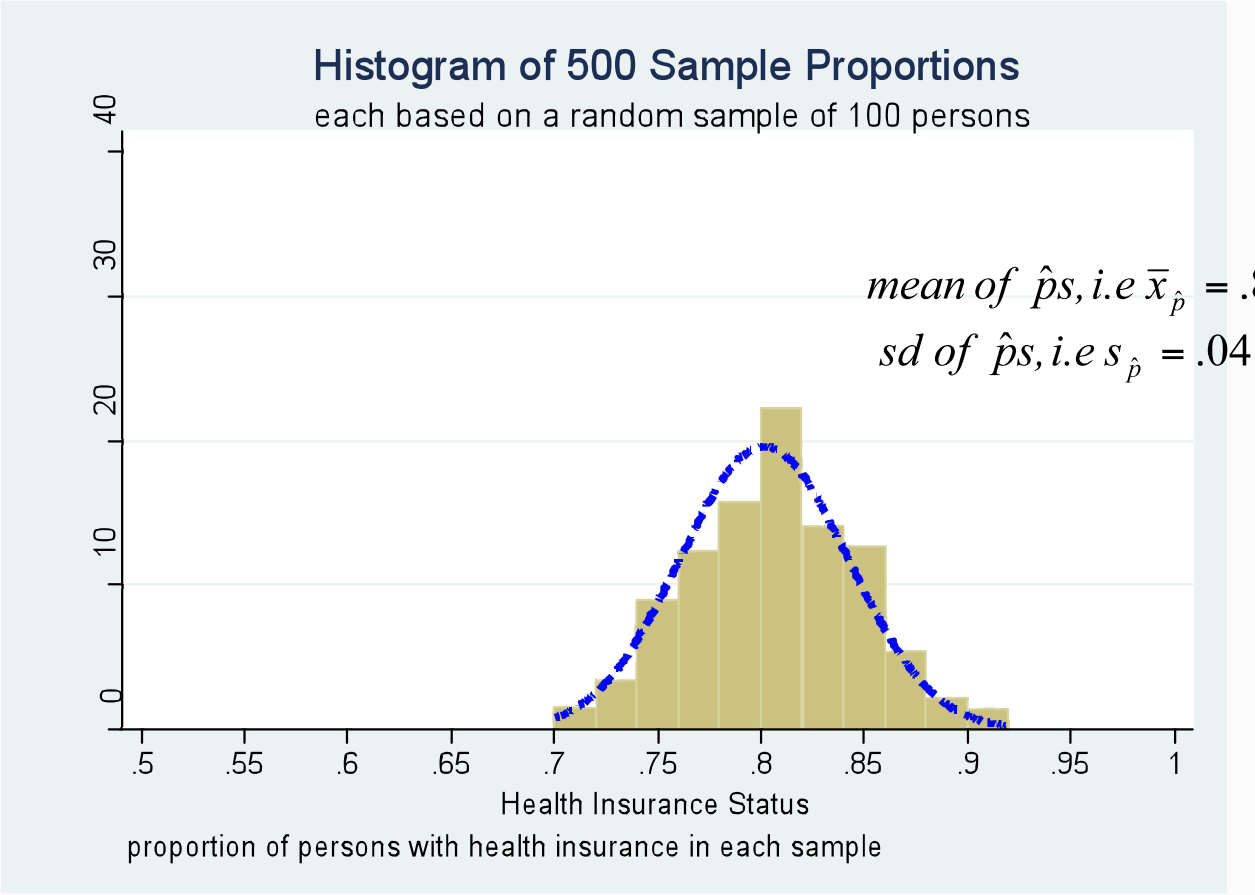
- Sample 2: $n = 100$



$$\hat{p}_{ins} = 0.78 = 78\%$$

Example: Blood Pressure of Males

- So we did this 500 times: now let's look at a histogram of the 500 proportions

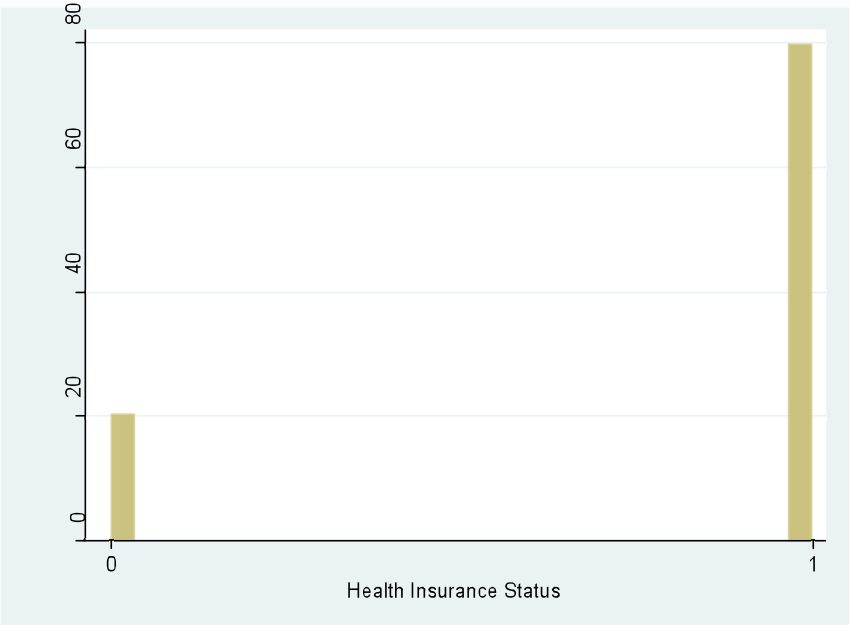


Example: Health Insurance Coverage

- We decide to do one more experiment
- We are going to take 500 separate random samples from this population, each with 1,000 subjects
- For each of the 500 samples, we will plot a histogram of the sample proportioned of insured individuals, and record the sample proportion
- Ready, set, go . . .

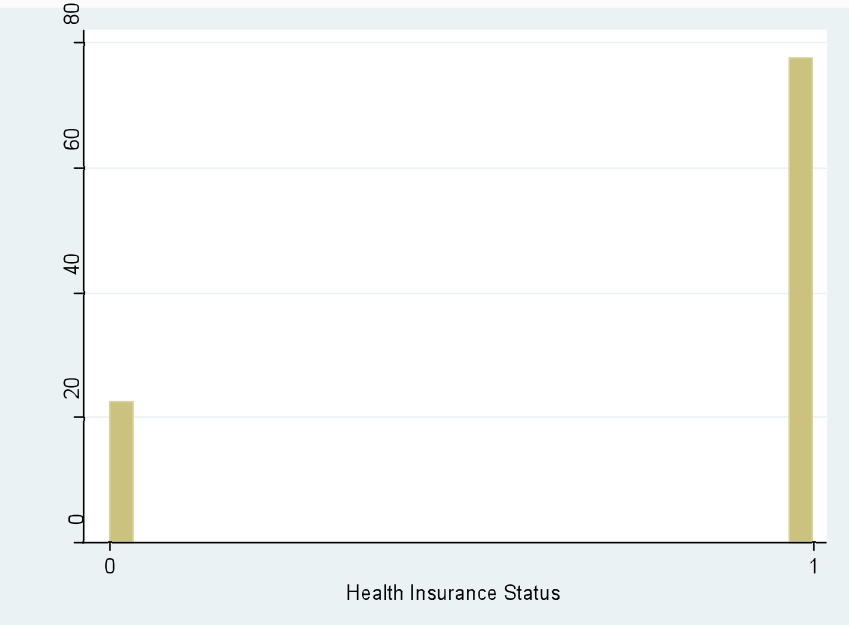
Random Samples

- Sample 1: n = 1,000



$$\hat{p}_{ins} = 0.798 = 79.8\%$$

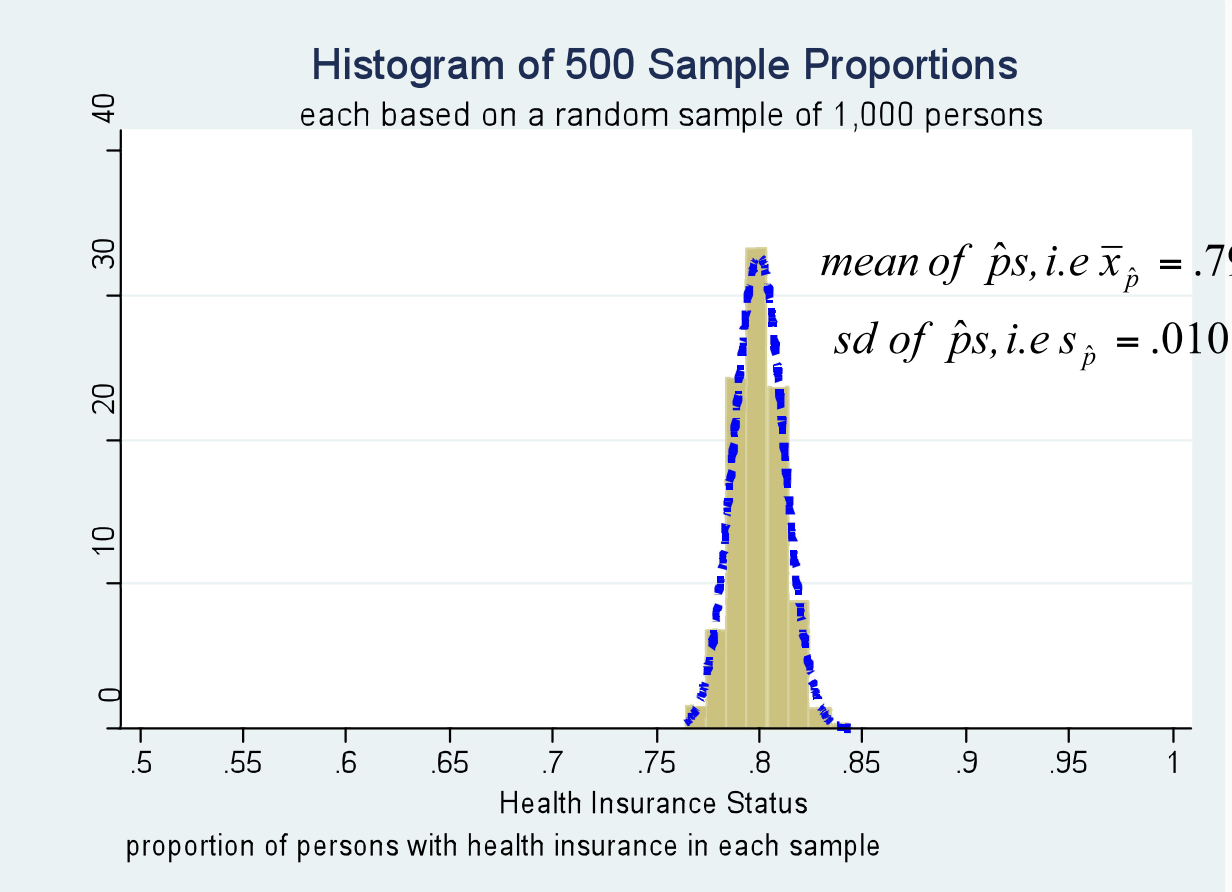
- Sample 2: n = 100



$$\hat{p}_{ins} = 0.777 = 77.7\%$$

Example: Blood Pressure of Males

- So we did this 500 times: now let's look at a histogram of the 500 proportions



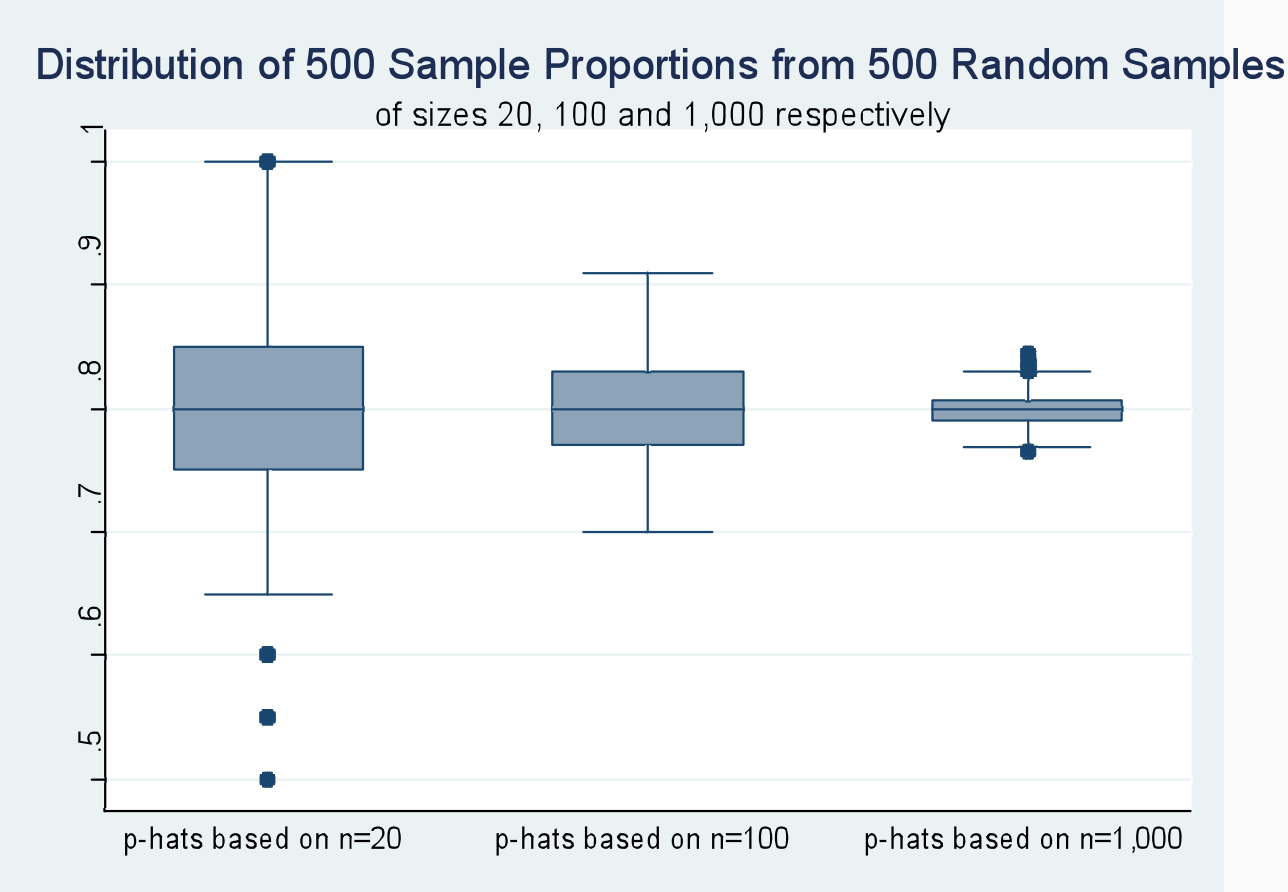
Example 2: Hospital Length of Stay

- Let's review the results
- True proportion of insured: p = 0.80
- Results from 500 random samples:

Sample Sizes	Means of 500 Sample Proportions	SD of 500 Sample Proportions	Shape of Distribution of 500 Sample Proportions
n = 20	0.805	0.094	Approaching normal?
n = 100	0.801	0.041	Approximately normal
n = 1,000	0.799	0.012	Approximately normal

Example 2: Hospital Length of Stay

- Let's review the results



JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section F

The Theoretical Sampling Distribution of the Sample Proportion and Its Estimate Based on a Single Sample

Sampling Distribution of the Sample Mean

- In the previous section we reviewed the results of simulations that resulted in estimates of what was formally called the sampling distribution of a sample proportion

- The sampling distribution of a sample proportion is a theoretical probability distribution
  - It describes the distribution of all sample proportions from all possible random samples of the same size taken from a population

Sampling Distribution of the Sample Mean

- In real research it is impossible to estimate the sampling distribution of a sample mean by actually taking multiple random samples from the same population (no research would ever happen if a study needed to be repeated multiple times) to understand this sampling behavior

- Simulations are useful to illustrate a concept, but not to highlight a practical approach!

- Luckily, there is some mathematical machinery that generalizes some of the patterns we saw in the simulation results

The Central Limit Theorem (CLT)

- The Central Limit Theorem (CLT) is a powerful mathematical tool that gives several useful results
  - The sampling distribution of sample proportions based on all samples of same size *n* is approximately normal
  - The mean of all sample proportions in the sampling distribution is the true mean of the population from which the samples were taken, *p*
  - The standard deviation in the sample proportions of size *n* is equal to $\sqrt{\frac{p \times (1-p)}{n}}$
  - This is often called the standard error of the sample proportion and sometimes written as $SE(\hat{p})$

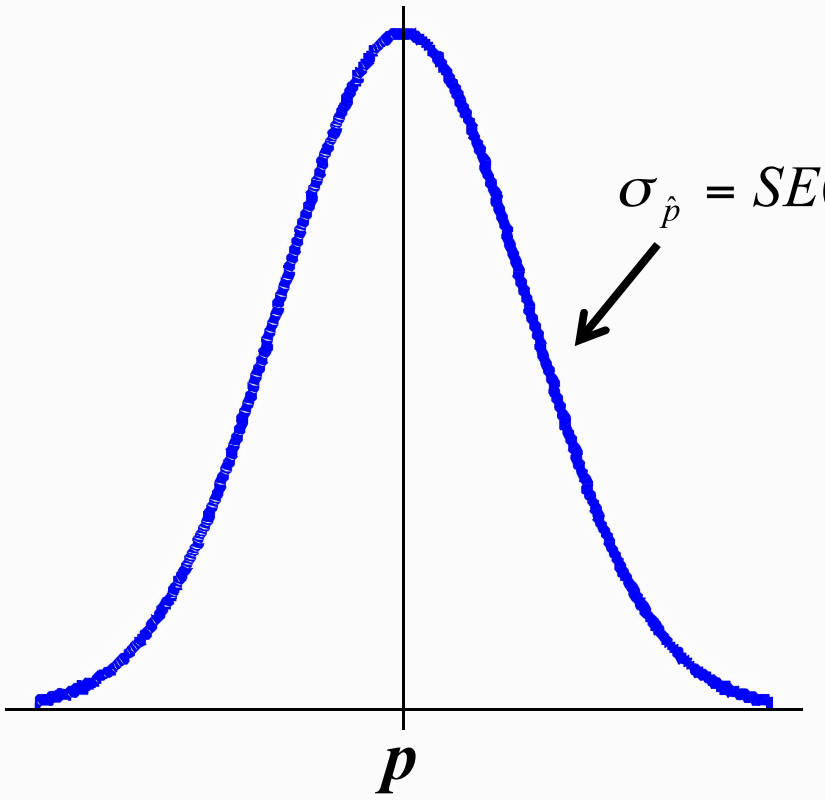
Example: Blood Pressure of Males

- Population distribution of individual insurance status
  - True proportion $p = 0.8$

Sample Sizes	Means of 500 Sample Proportions	Means of 5000 Sample Proportions	SD of 500 Sample Proportion	SD of 5000 Sample Proportions	SD of Sample Proportions (SE) by CLT
n = 20	0.805	0.799	0.094	0.090	0.089
n = 100	0.801	0.799	0.041	0.040	0.040
n = 1,000	0.799	0.80	0.012	0.012	0.012

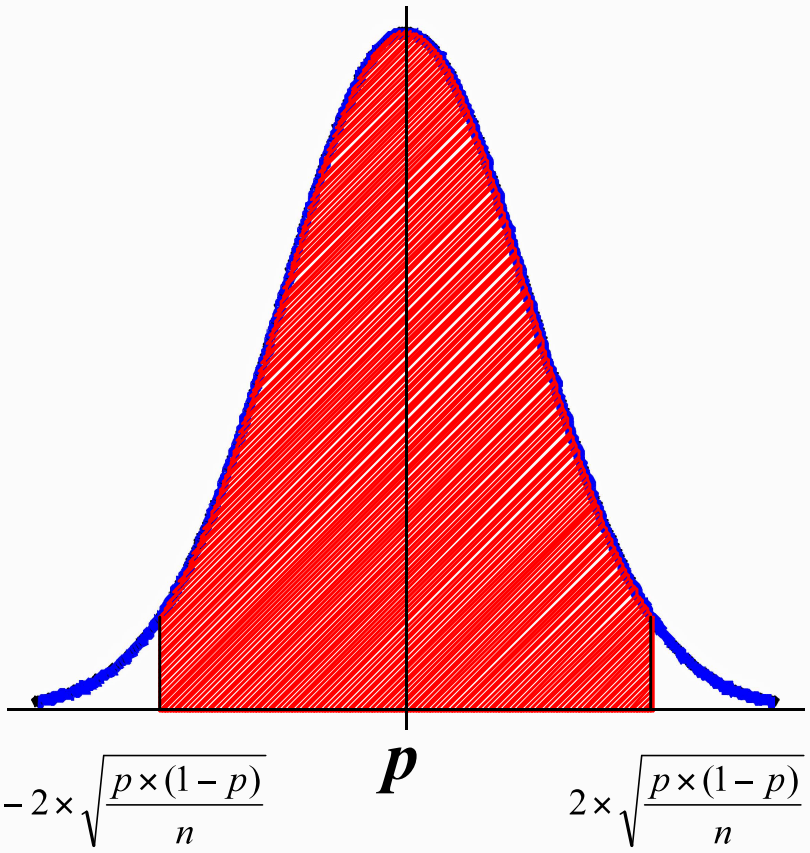
Recap: CLT

- So the CLT tells us the following:
  - When taking a random sample of binary measures of size *n* from a population with true proportion *p* the theoretical sampling distribution of sample proportions from all possible random samples of size *n* is:



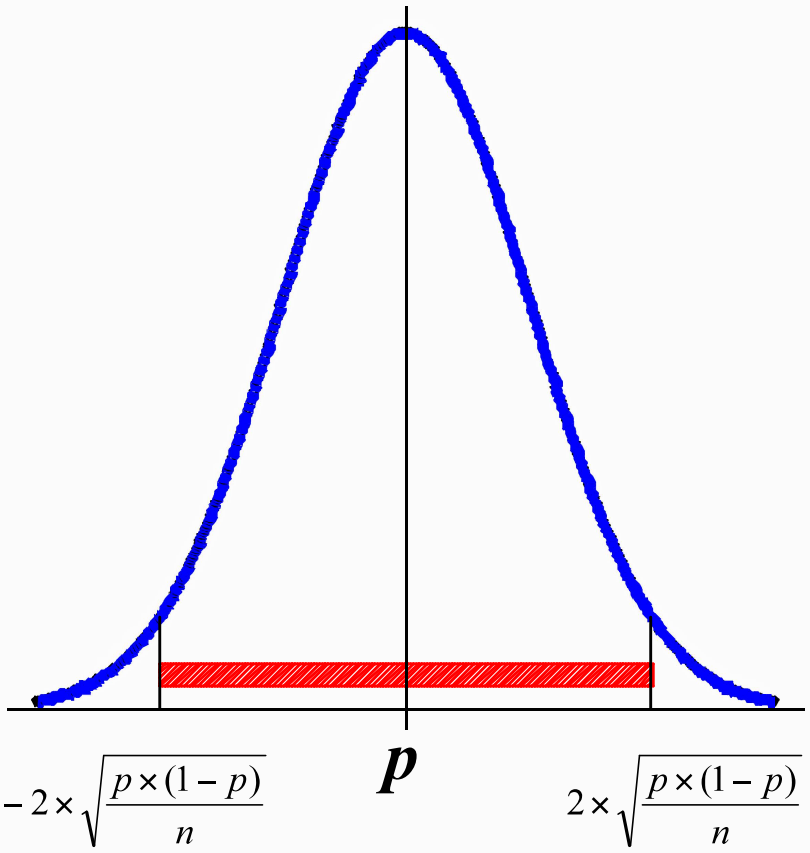
CLT: So What?

- So what good is this info?
  - Well using the properties of the normal curve, this shows that for most random samples we can take (95%), the sample proportion $\hat{p}$ will fall within 2 SEs of the true proportion *p*:



CLT: So What?

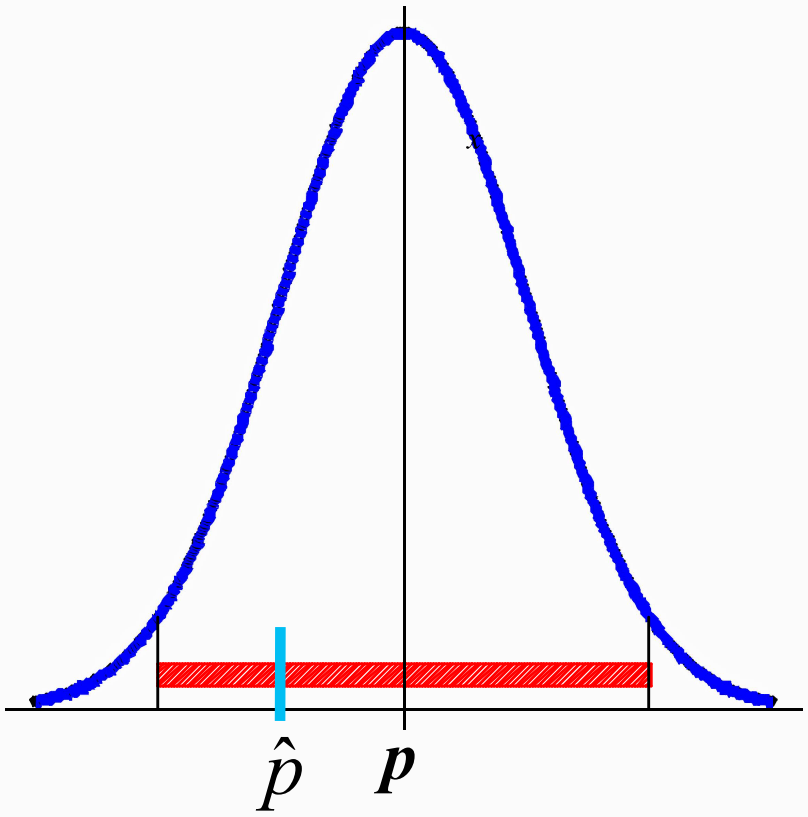
- So AGAIN what good is this info?
  - We are going to take a single sample of size *n* and get one $\hat{p}$
  - So we won't know *p* and if we did know *p* why would we care about the distribution of estimates of *p* from imperfect subsets of the population?



CLT: So What?

- We are going to take a single sample of size *n* and get one $\hat{p}$

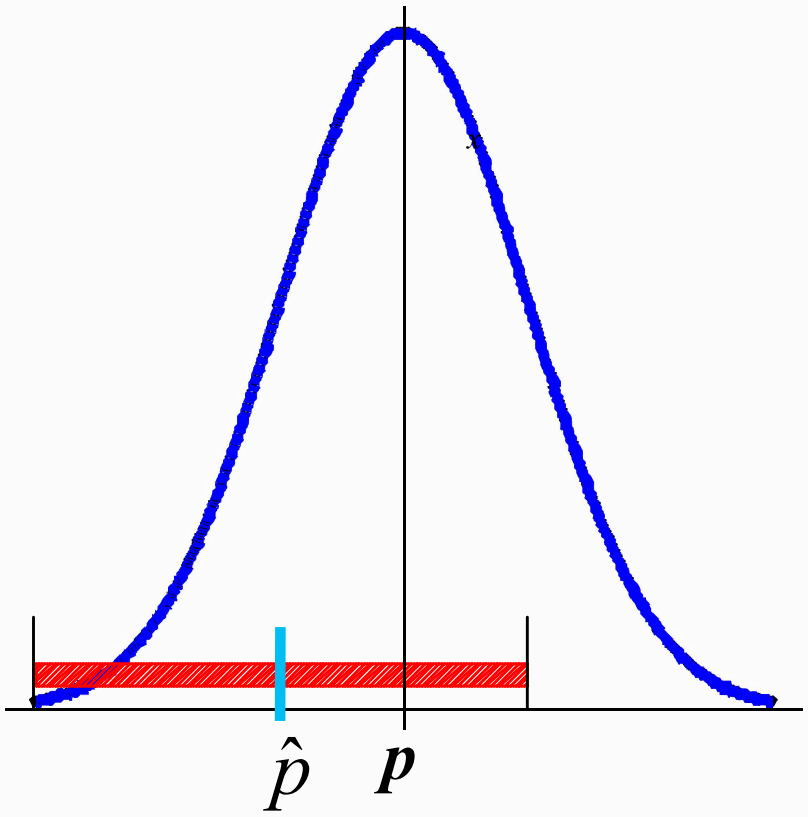
- But for most (95%) of the random samples we can get, our $\hat{p}$ will fall within +/- 2 SEs of *p*



CLT: So What?

- We are going to take a single sample of size *n* and get one $\hat{p}$

- So if we start at $\hat{p}$ and go 2 SEs in either direction, the interval created will contain *p* most (95 out of 100) of the time



Estimating a Confidence Interval

- Such and interval is a called a 95% confidence interval for the population proportion *p*
- Interval given by $\hat{p} \pm 2SE(\hat{p}) \rightarrow \bar{x} \pm 2 \times \sqrt{\frac{p \times (1 - p)}{n}}$
- Problem: we don't know *p*
  - Can estimate with $\hat{p}$, will detail this in next section
- What is interpretation of a confidence interval?

Interpretation of a 95%Confidence Interval (CI)

- Laypersons' range of "plausible" values for true proportion
  - Researcher never can observe true mean *p*
  - $\hat{p}$ is the best estimate based on a single sample
  - The 95% CI starts with this best estimate and additionally recognizes uncertainty in this quantity

- Technical
  - Were 100 random samples of size *n* taken from the same population, and 95% confidence intervals computed using each of these 100 samples, 95 of the 100 intervals would contain the values of true proportion *p* within the endpoints

Notes on Confidence Intervals

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

Notes on Confidence Intervals

- Are all CIs 95%?
  - No
  - It is the most commonly used
  - A 99% CI is wider
  - A 90% CI is narrower

- To change level of confidence adjust number of SE added and subtracted from $\hat{p}$
  - For a 99% CI, you need ± 2.6 SE
  - For a 95% CI, you need ± 2 SE
  - For a 90% CI, you need ± 1.65 SE

Summary

- What did we see with this set of examples

- A couple of trends:
  - Distribution of sample proportions tended to be approximately normal—even when original—and individual level data was not (binary outcome)
  - Variability in sample mean values decreased as the size of the sample each proportion was based upon increased

Clarification

- As with means for continuous data, variation in proportions values tied to the size of each sample selected in our exercise: NOT the number of samples

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section G

Estimating Confidence Intervals for the Proportion of a Population Based on a Single Sample of Size *n*: Some Examples

Estimating a 95% Confidence Interval

- In last section we defined a 95% confidence interval for the population proportion *p*
- Interval given by $\hat{p} \pm 2SE(\hat{p})$: $\hat{p} \pm 2^* \sqrt{\frac{p \times (1 - p)}{n}}$
- Problem: we don't know *p*
  - Can estimate with $\hat{p}$, such that our estimated SE is
  - $S\hat{E}(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$
- Estimated 95% CI for based on a single sample of size *n*
  - $\hat{p} \pm 2 \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$

Example 1

- Proportion of dialysis patients with national insurance in 12 countries (only six shown . . .)

EXHIBIT 1
Descriptive Measures Of The Prevalent Cross-Sectional Patient Sample, Dialysis Patients In Twelve Countries, 2002–2004

	A/NZ (n = 561)	BEL (n = 468)	CAN (n = 503	FRA (n = 481)	GER (n = 524)	ITA (n = 540)
Mean age (years)	59.9 (14.7)	66.2 (13.4)	62.1 (14.7)	64.1 (14.5)	61.7 (14.1)	64 (13.7)
Minority[1]	21.5%	5.3%	18.7%	7.1%	0.4%	0.4%
Income ($US)						
<$20,000	85.0%	73.4%	71.8%	67.0%	59.7%	78.3%
$20,000–$39,000	9.1	17.5	20.8	21.8	27.1	17.4
≥$40,000	5.9	9.1	7.4	11.2	13.1	4.2
Insurance type						
National only	69.8%	74.1%	79.6%	45.5%	95.4%	99.6%
Private only	5.4	0.4	0.2	0.2	2.9	0.0
Mean number of comorbid conditions[2]	3.7 (2)	3.9 (2.1)	4.1 (2.1)	3.1 (1.9)	3.4 (2.1)	2.7 (1.9)
Mean number of prescribed medications	8.7 (3.6)	9.9 (4.1)	12.6 (4.8)	7.7 (3.5)	9.7 (3.5)	6.4 (3.6)

- Example, France: $\hat{p} = \frac{219}{481} = .46$

Example 1

- Estimated confidence interval

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

$$.46 \pm 2 \times \sqrt{\frac{.46 \times (1 - .46)}{481}}$$

$$.46 \pm 2 \times .023$$

$$.46 \pm .046$$

$$(.414,.505) \approx (.41,.51)$$

$\rightarrow$ *41% to 51%*

Example 1 in Stata

- Can use *cii* command for binary outcomes to get CIs for *p*
- Syntax: *cii n y*
  - Where *n* is the total sample size, *y* is number of "yes" outcomes
- National health insurance in France

```
. cii 481 219, bin

                                                        -- Binomial Exact  --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |        481   .4553015    .0227068        .4101514    .5010042
```

Example 2

- Maternal/infant transmission of HIV
  - HIV-infection status was known for 363 births (180 in the zidovudine (AZT) group and 183 in the placebo group)
  - Thirteen infants in the zidovudine group and 40 in the placebo group were HIV-infected

$$\hat{p}_{AZT} = \frac{13}{180} = 0.07 = 7\%$$

$$\hat{p}_{PLAC} = \frac{40}{183} = 0.22 = 22\%$$

Example 2

- Estimated confidence interval for tranmission percentage in the placebo group

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

$$.22 \pm 2 \times \sqrt{\frac{.22 \times (1 - .22)}{183}}$$

$$.22 \pm 2 \times .031$$

$$.46 \pm .062$$

$$(.158,.282) \approx (.16,.28)$$

$$\rightarrow \textit{16\% to 28\%}$$

Example 2 in Stata

- Results from *cii* command

```
. cii 183 40

                                                    -- Binomial Exact  --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |        183   .2185792    .0305507         .160984    .2855248
```

Notes on 95% Confidence Interval for Proportion

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

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section H

Small Sample Considerations for Confidence Intervals for Population Proportions

The Central Limit Theorem (CLT)

- The Central Limit Theorem (CLT) is a powerful mathematical tool that gives several useful statistics including:
  - The sampling distribution of sample proportions based on all samples of same size *n* is *approximately* normal
  - Mother/infant transmission example, placebo group:

  - CLT 95% CI:
    (can be done by hand)

    $(.158,.282) \approx (.16,.28)$

    $\rightarrow$ *16% to 28%*

  - Exact 95% CI:
    (requires computer,
    always correct)

```
From Ôcii 183 40Ó command

 [95% Conf. Interval]

---------------------

.160984    .2855248
```

Notes on 95% Confidence Interval for Proportion

- The CLT based formula for a 95% CI is only *approximate*; it works very well if you have enough data in your sample

- The approximation works better the bigger $n \times \hat{p} \times (1 - \hat{p})$

- "Large sample" for binary outcomes is not only a function of total sample size *n*, but the split between "yes" and "no" outcomes

Mother/Infant Transmission: AZT Group

- Mother/infant transmission example, AZT group:
  - (n = 180, $\hat{p} = \frac{13}{180} = .07$)
- CLT 95% CI:
  (can be done by hand)

$$(.032,.108) \approx (.03,.11)$$

$$\rightarrow \textit{3\% to 11\%}$$

- Exact 95% CI:
  (requires computer,
  always correct)

```
From Òcii 180 13Ó command

 [95% Conf. Interval]

---------------------

 .0390137    .1203358
```

Mother/Infant Transmission CIs

- In the placebo sample

$$n \times \hat{p}_{plac} \times (1 - \hat{p}_{plac}) =$$

$$183 * .22 * .78 \approx 31$$

- In the AZT sample

$$n \times \hat{p}_{AZT} (1 - \hat{p}_{AZT}) =$$

$$180 * .07 * .93 \approx 12$$

Notes on 95% Confidence Interval for Proportion

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

- Interpretation of 95% CIs exactly the same with either method
  - In real life, using computer will always give valid result
  - CLT only breaks down with “small” sample sizes
  - In testing situations you will not be required to do exact CIs!

Really Small Sample Example for Illustration

- Random sample of 16 patients on drug A: two of sixteen patients experience drug failure in first month
  - CLT 95% CI: $\hat{p} \pm 2 \times S\hat{E}(\hat{p}) \rightarrow$

$$\frac{2}{16} \pm 2 \times \sqrt{\frac{(2/16) \times (1 - 2/16)}{16}} \rightarrow$$

$$(-0.05, 0.28)$$

  - Exact 95% CI: (0.02, 0.38)

```
. cii 16 2

                                                    -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |         16        .125    .0826797        .0155136    .3834762
```