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

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

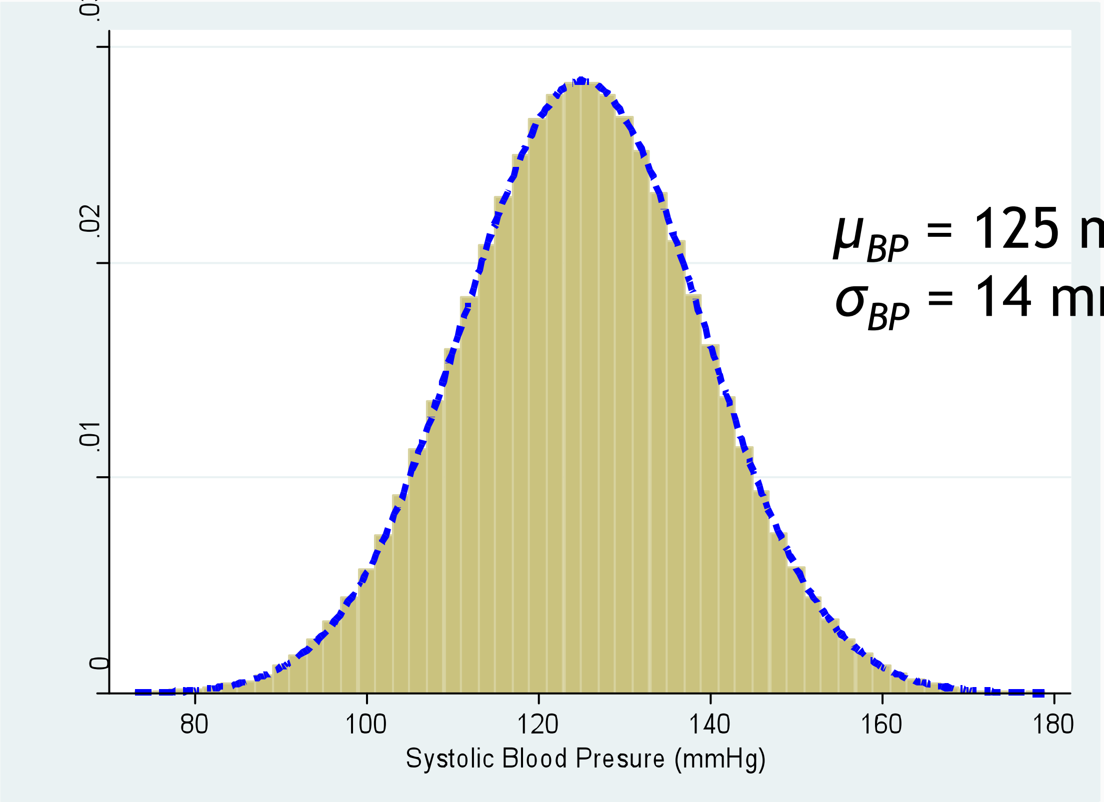
- If we repeatedly choose samples from the same population, a statistic will take different values in different samples

- 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:



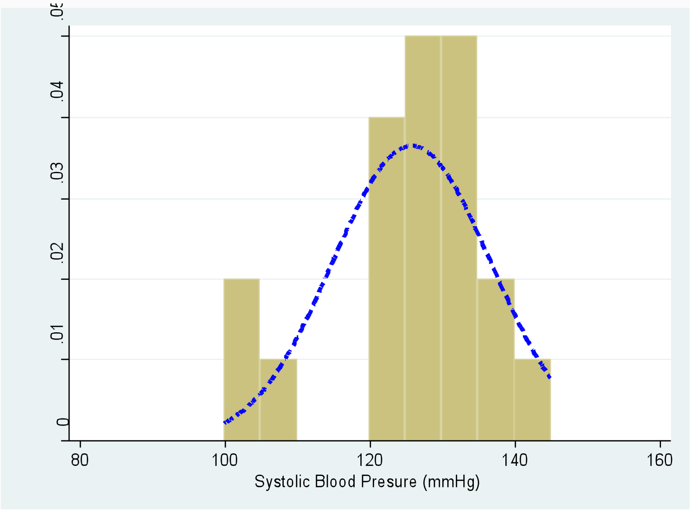
μ_{BP} = 125 mmHg
σ_{BP} = 14 mmHg

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

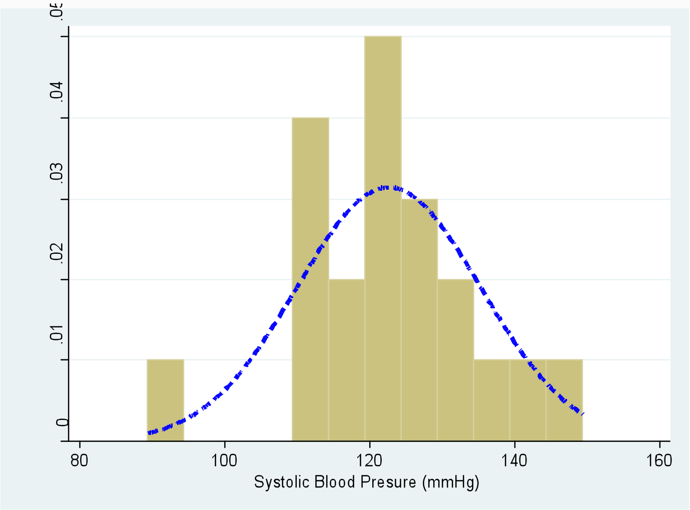
- Sample 1: n = 20



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

s_{BP} = 10.9 mmHg

- Sample 2: n = 20

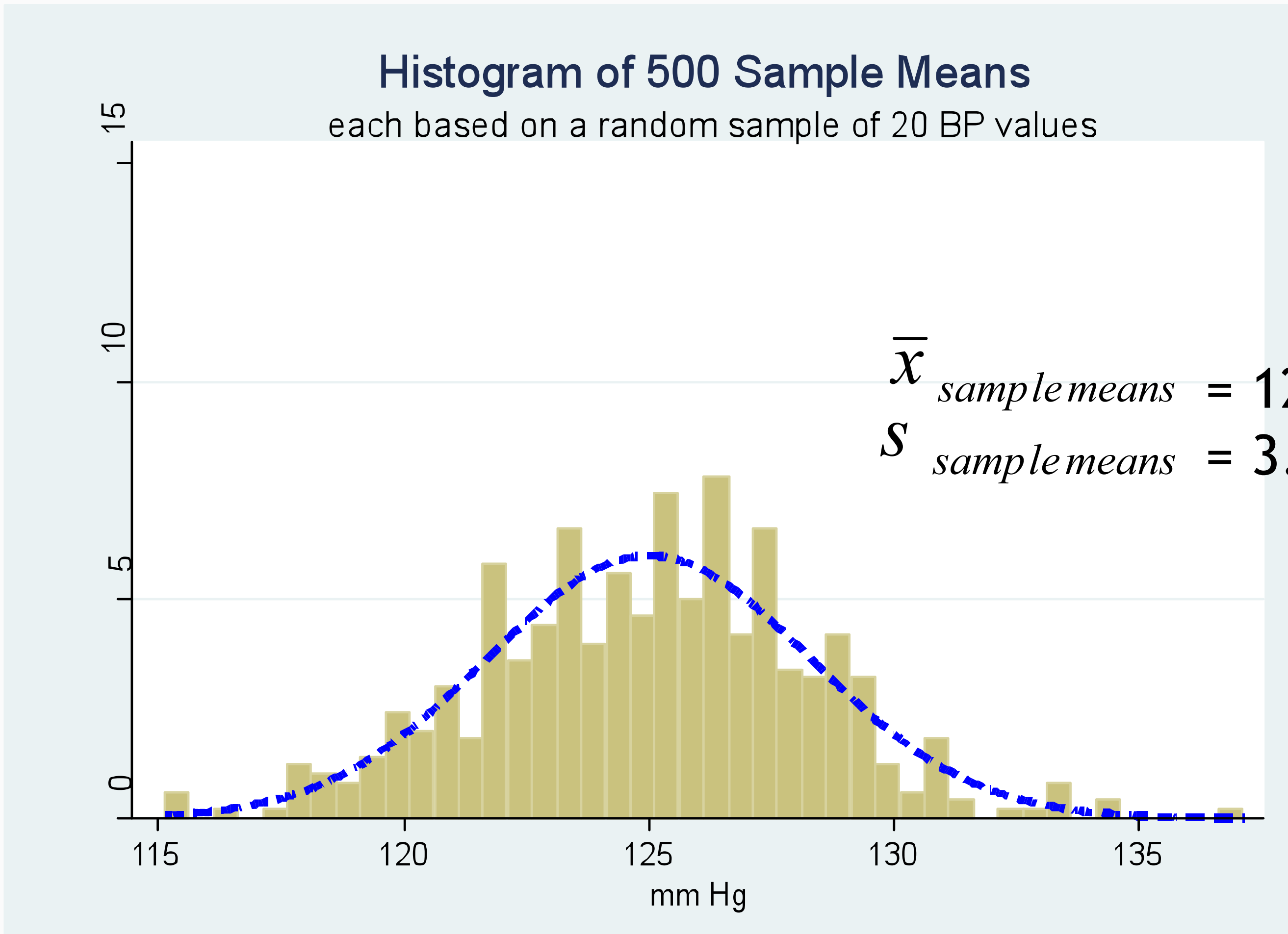


$\bar{x}_{BP}$ = 122.6 mmHg

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



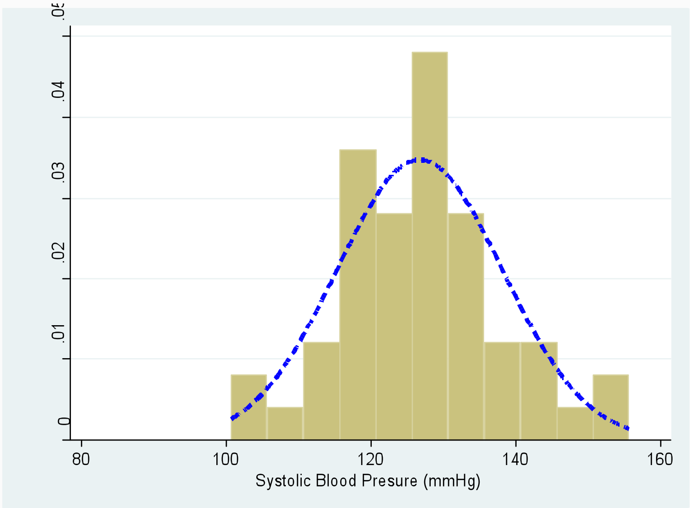
$$\bar{x}_{\text{sample means}} = 125 \text{ mmHg}$$
$$s_{\text{sample means}} = 3.3 \text{ mmHg}$$

Example: Blood Pressure of Males

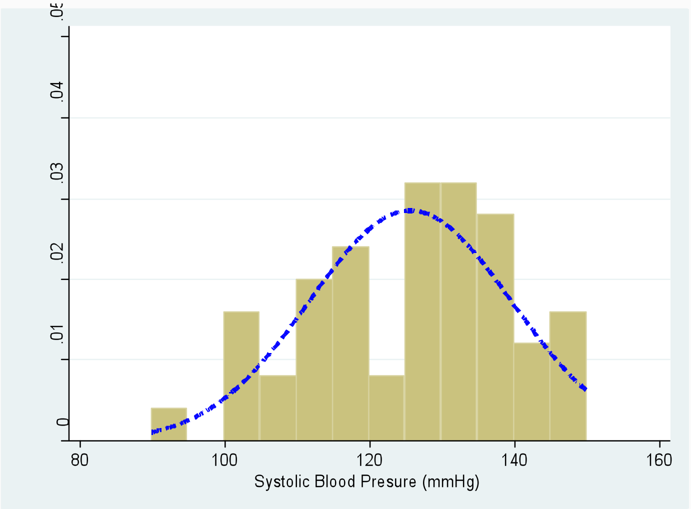
- We decide to do another experiment
- We are going to take 500 separate random samples from this population of me, each with 50 subjects
- For each of the 500 samples, we will plot a histogram of the sample BP values, and record the sample mean and sample standard deviation
- Ready, set, go . . .

Random Samples

- Sample 1: n = 50



- Sample 2: n = 50

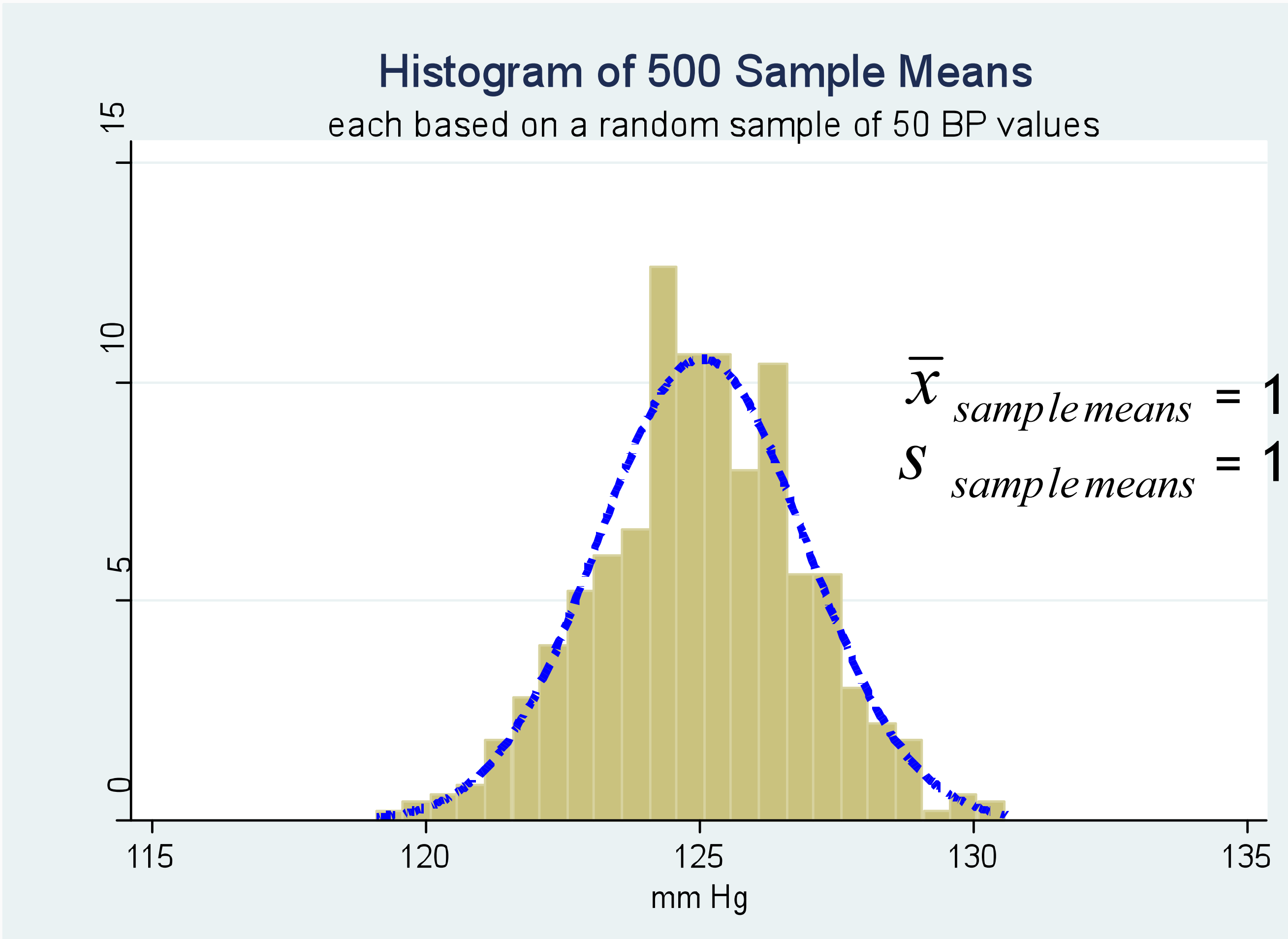


$\bar{x}_{BP}$ = 126.7 mmHg
s_{BP} = 11.5 mmHg

$\bar{x}_{BP}$ = 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



$\bar{x}_{sample\ means}$ = 125 mmHg

$s_{sample\ means}$ = 1.9 mmHg

Example: Blood Pressure of Males

- We decide to do one more experiment

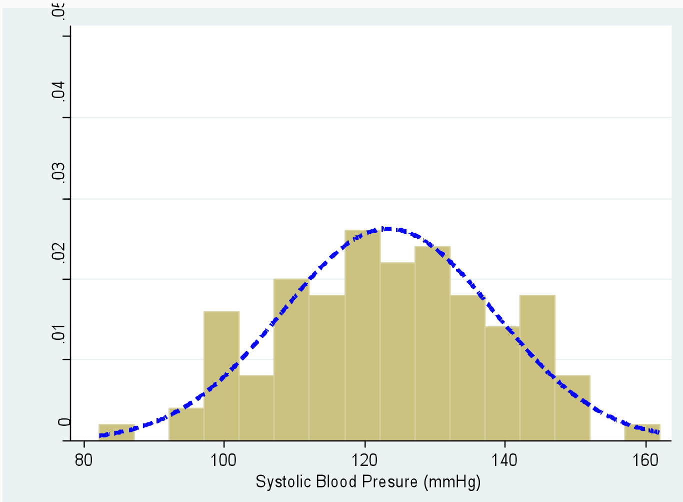
- We are going to take 500 separate random samples from this population of men, each with 100 subjects

- 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 . . .

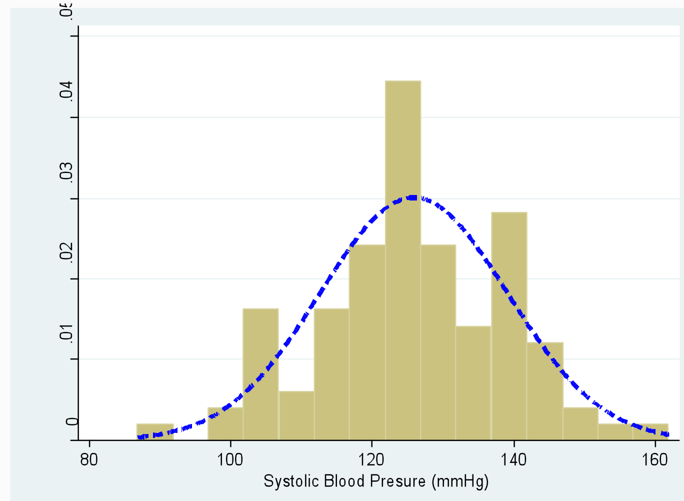
Random Samples

- Sample 1: n = 100



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

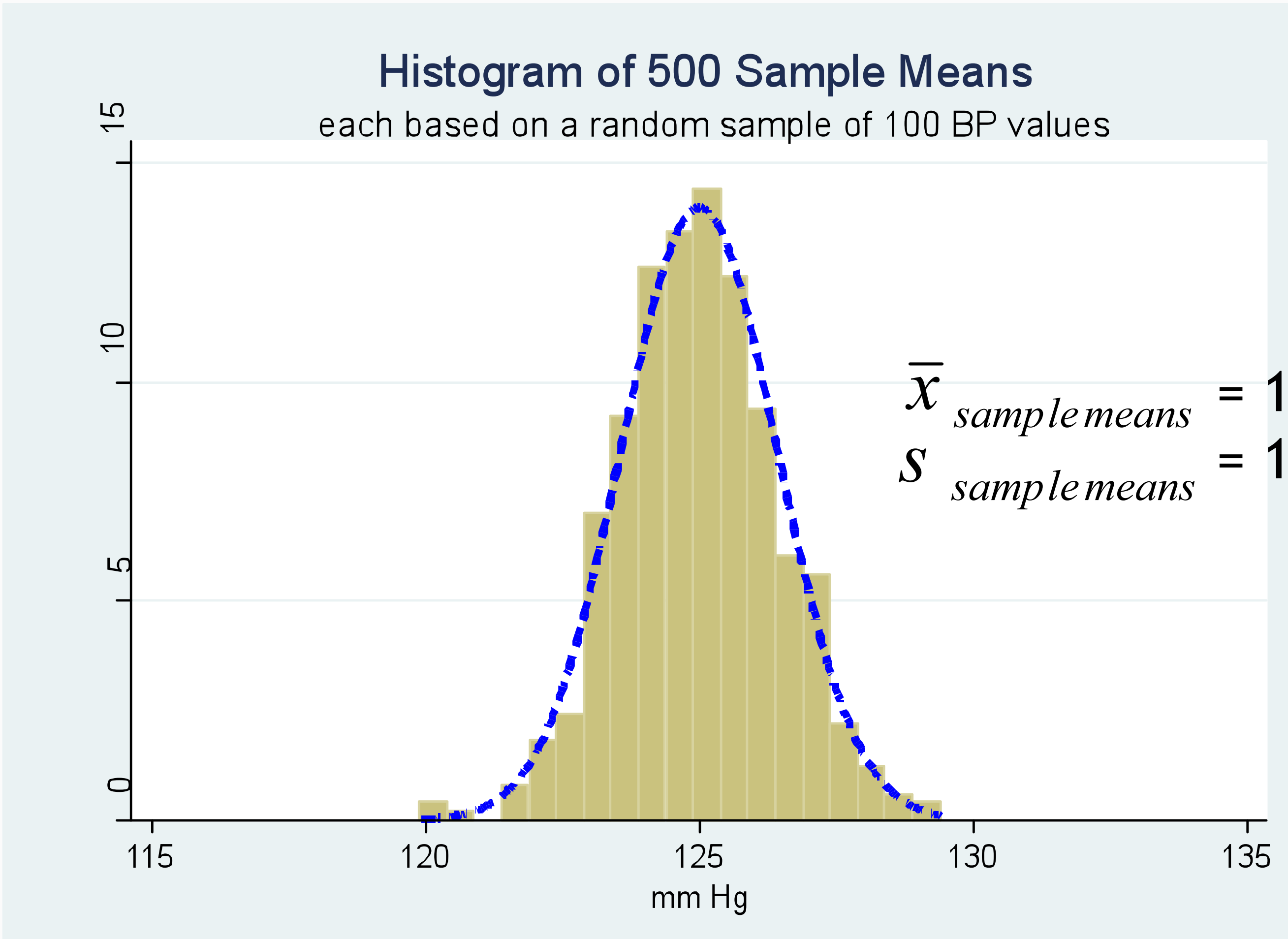
- Sample 2: n = 100



$\bar{x}_{BP}$ = 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



$\bar{x}_{sample\,means}$ = 125 mmHg
$s_{sample\,means}$ = 1.4 mmHg

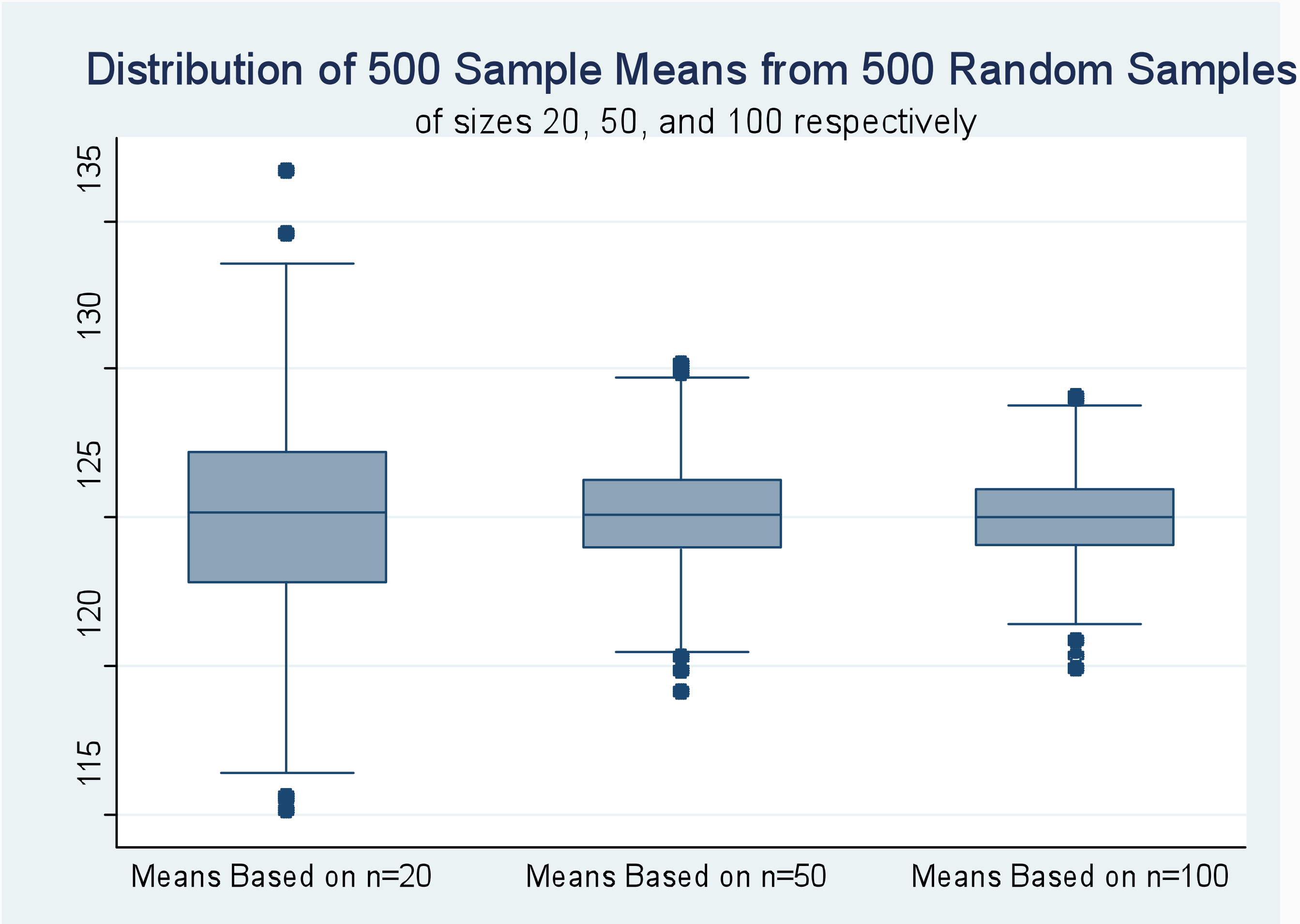
Example: Blood Pressure of Males

- 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

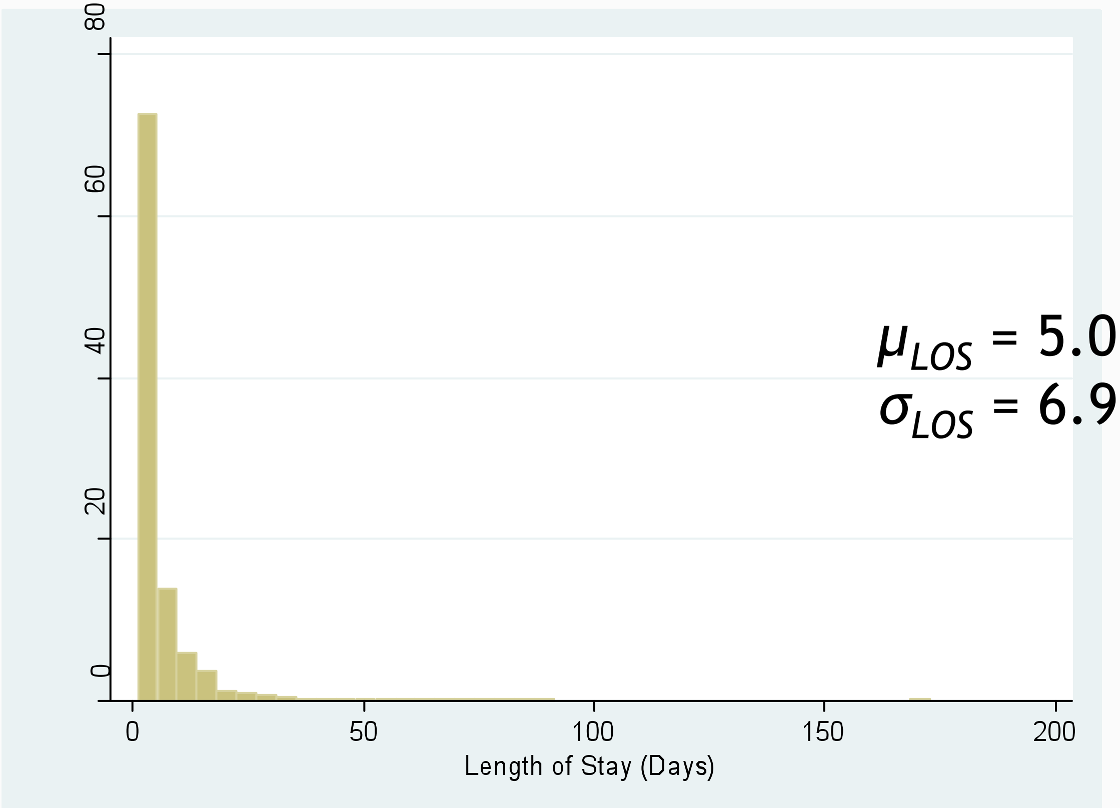
- Let's review the results



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

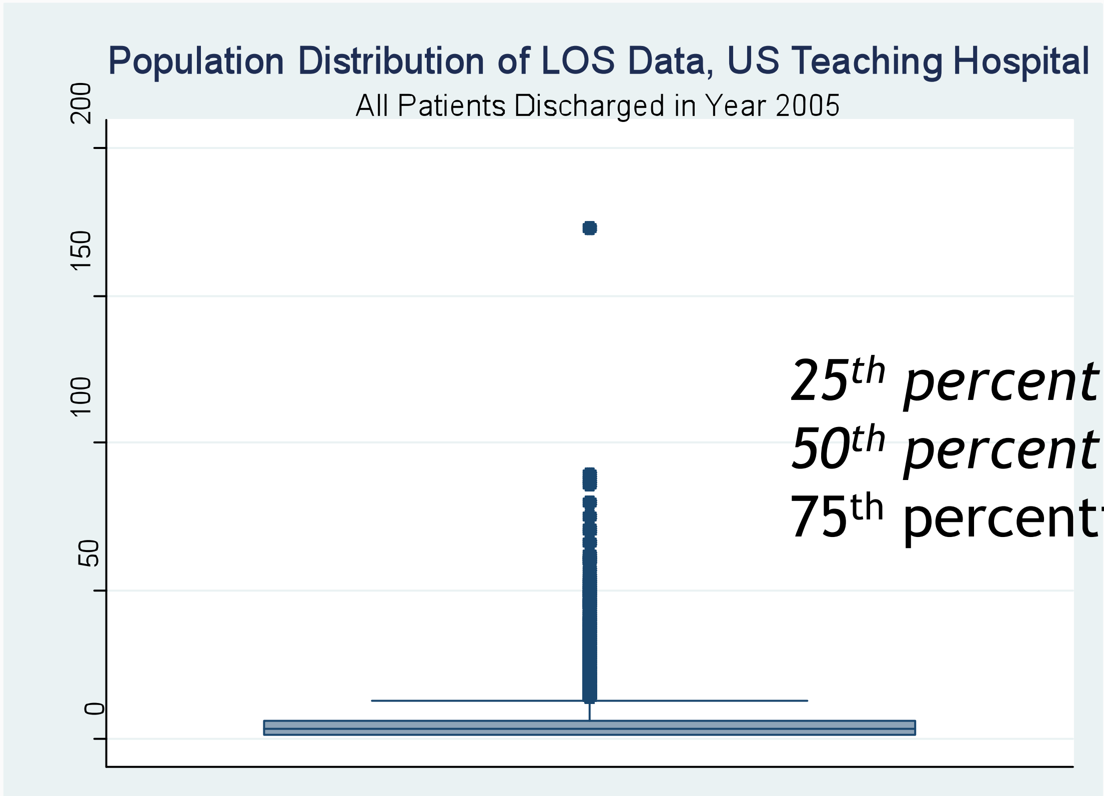
- Assume the population distribution is given by the following:



μ_{LOS} = 5.0 days
σ_{LOS} = 6.9 days

Example 2: Hospital Length of Stay

- Boxplot presentation



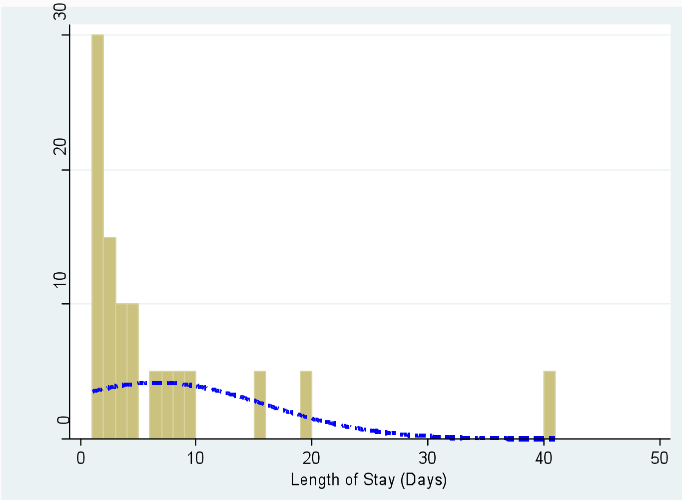
25th percentile: 1.0 days
50th percentile: 3.0 days
75th percentile: 6.0 days

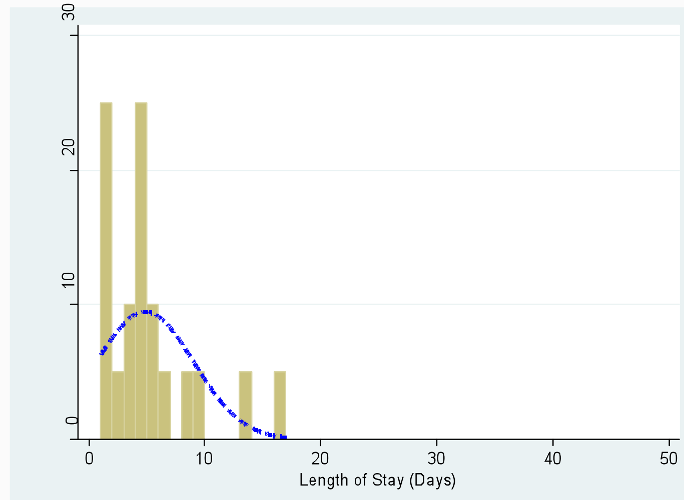
Example 2: Hospital Length of Stay

- 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
- Sample 2: n = 20



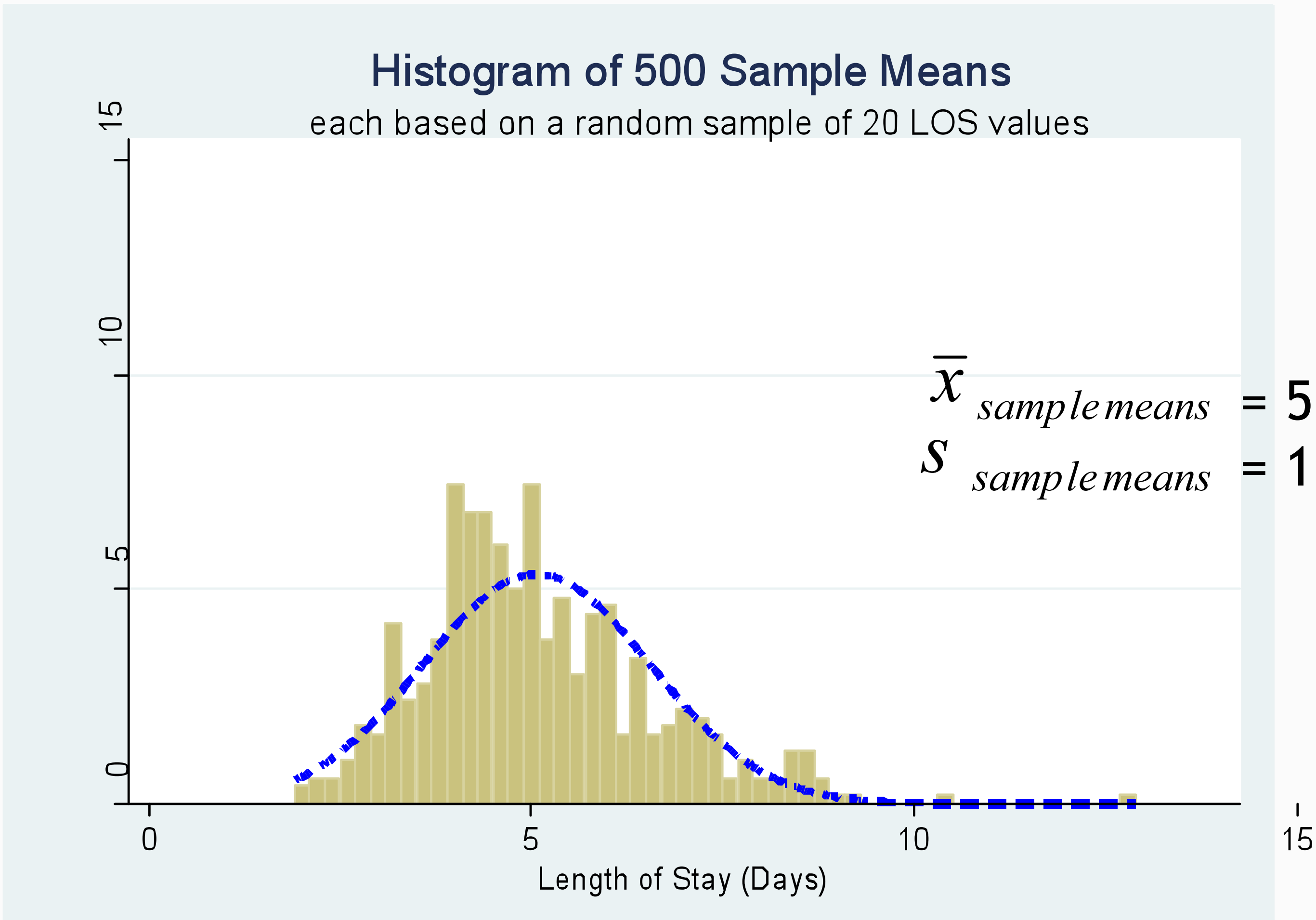


$\bar{x}_{LOS}$ = 6.6 days
s_{LOS} = 9.5 days

$\bar{x}_{LOS}$ = 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



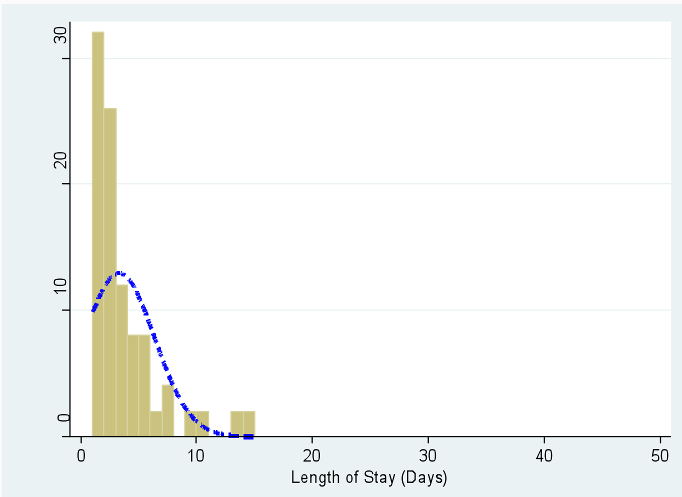
$$\bar{x}_{sample\,means} = 5.05 \text{ days}$$
$$s_{sample\,means} = 1.49 \text{ days}$$

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 . . .

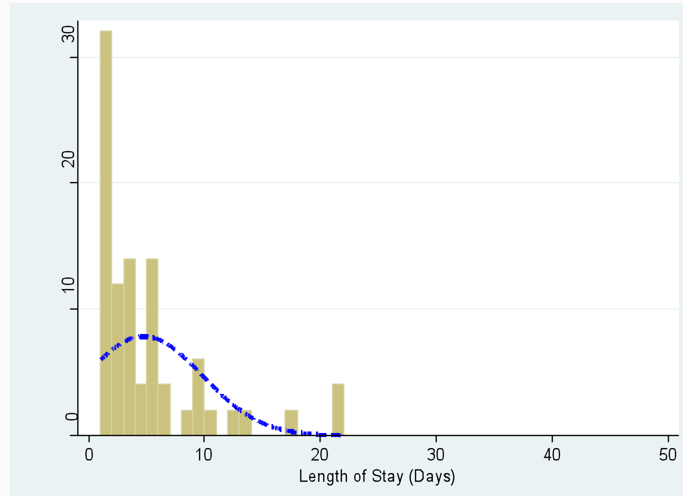
Random Samples

- Sample 1: n = 50



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

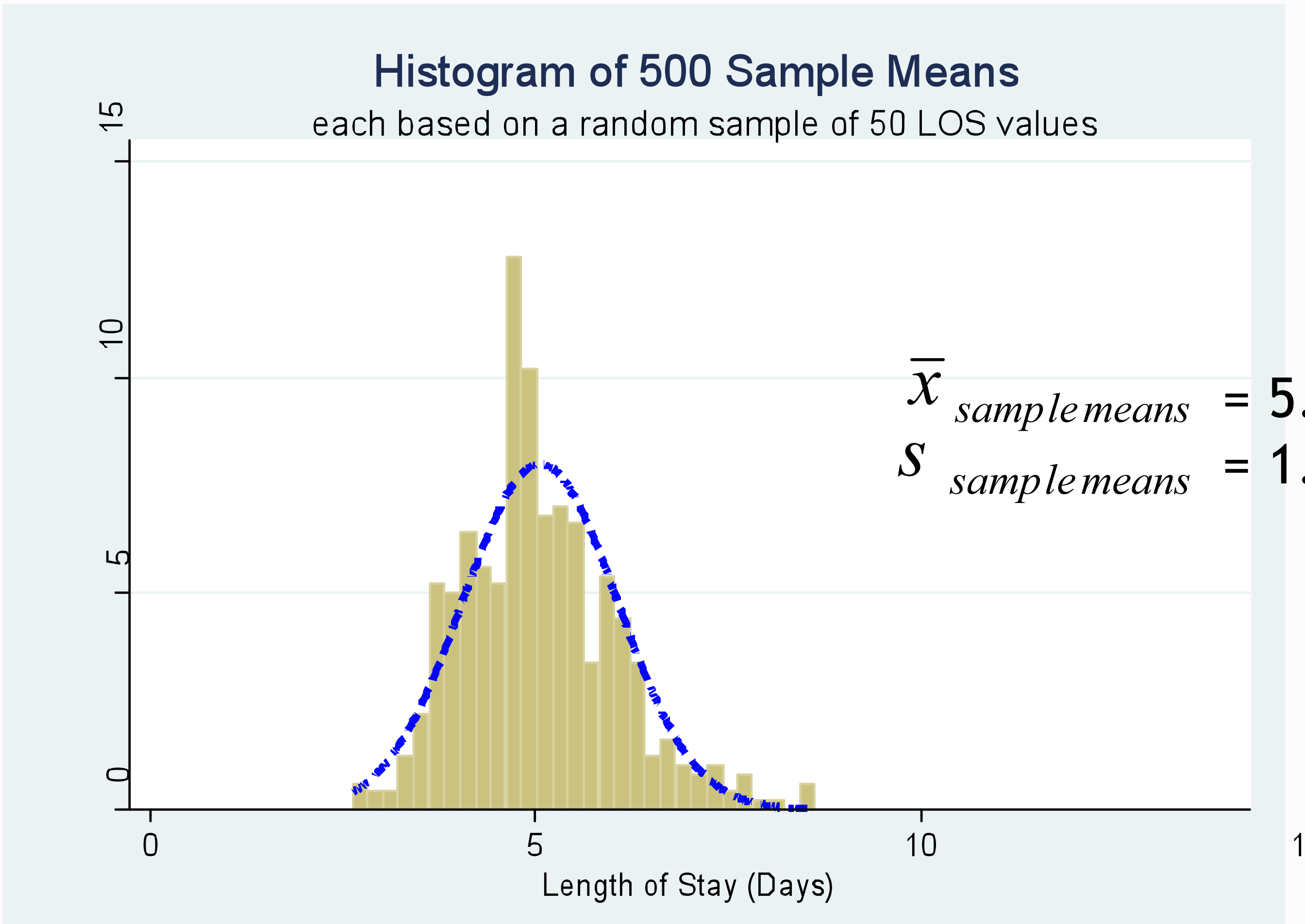
- Sample 2: n = 50



$\bar{x}_{LOS}$ = 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



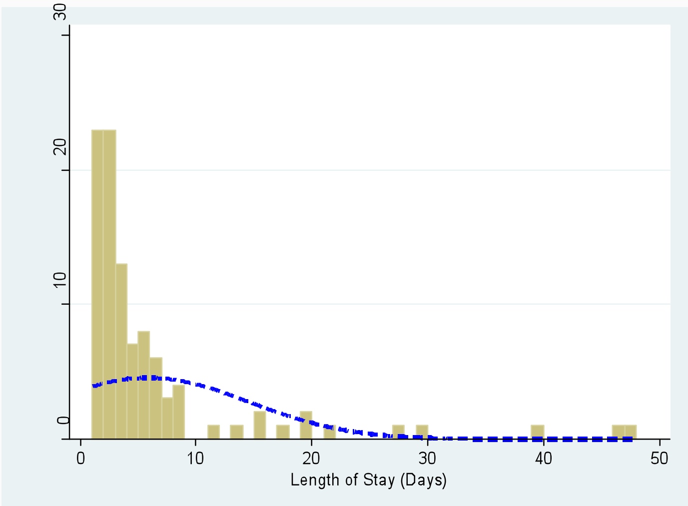
$$\bar{x}_{sample\,means} = 5.04 \text{ days}$$
$$s_{sample\,means} = 1.00 \text{ days}$$

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 . . .

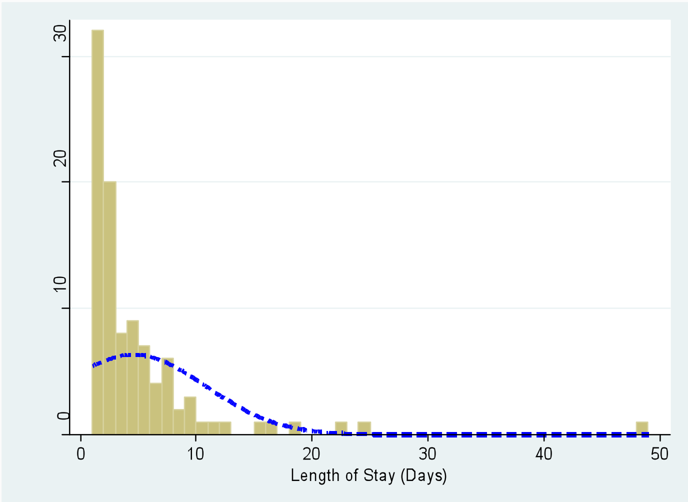
Random Samples

- Sample 1: n = 100



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

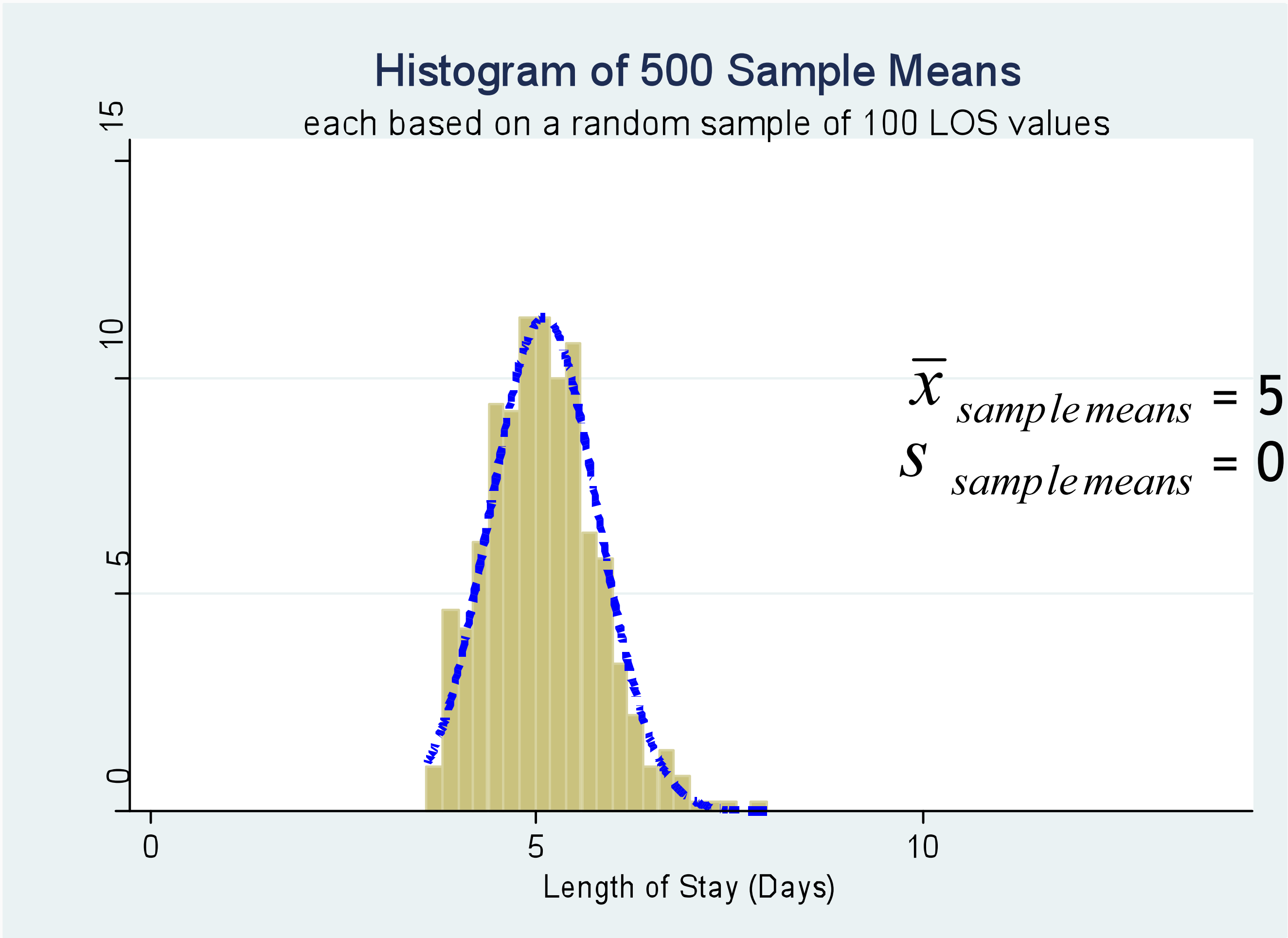
- Sample 2: n = 100



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

Distribution of Sample Means

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



$\bar{x}_{sample\,means}$ = 5.08 days

$s_{sample\,means}$ = 0.78 days

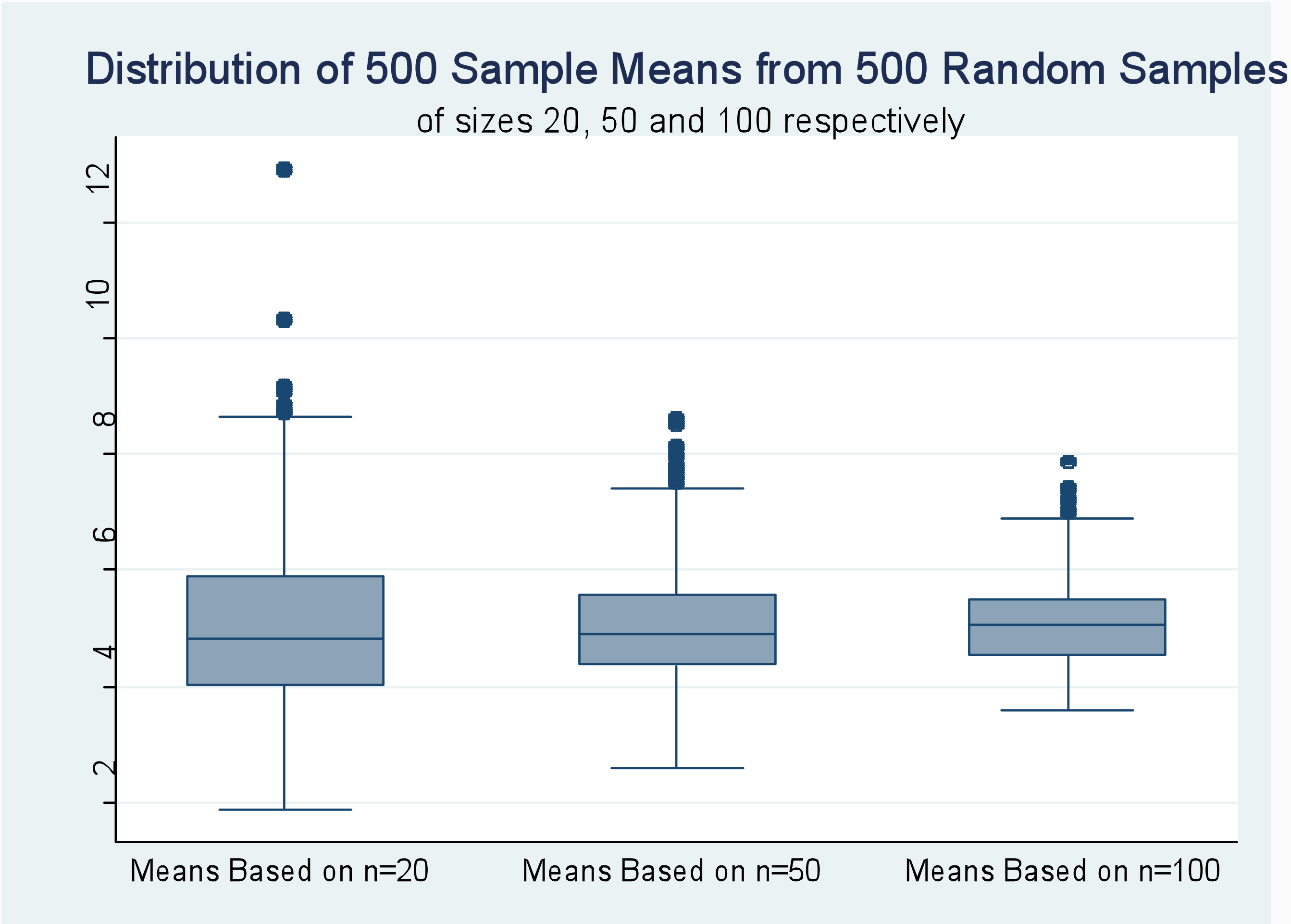
Example 2: Hospital Length of Stay

- Let's review the results
  - Population distribution of individual LOS values for population of patients: right skewed
  - True mean μ = 5.05 days: σ = 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