Lecture 10

Sampling Distributions

Yen-Yi Ho

Department of Statistics
University of South Carolina

Outline

- 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

Random Sample

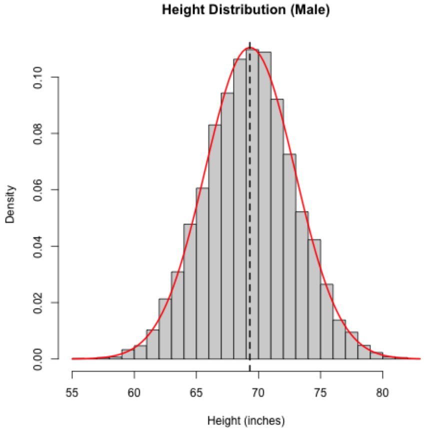
- When a sample is randomly selected from a population, it is called a random sample.
- In a simple random sample, each individual in the population has an equal chance of being chosen for the sample.
- But even with random sampling, there is still **sampling variability**.

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 answer each time), then it is fairly reliable (not a lot of variability)

Example: Self-Reported Heights of Male Students

- We have self-reported heights from 812 students from 4 data science courses.
- Assume the population distribution looks like this:
- Population mean(μ) $=$ 69.31 inches, sd(σ)$=$3.61 inches

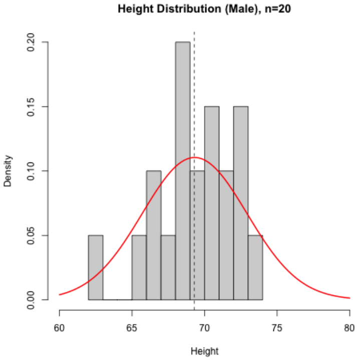


Example 1: Self-Reported Heights of Male Students

- Suppose we had all the time in the world
- We are going to take 500 separate random samples from this population of mean, each with **20** students
- For each of the 500 samples, we will plot a histogram of the height values of students and record the sample mean and sample sd.
- Ready, set, go ...

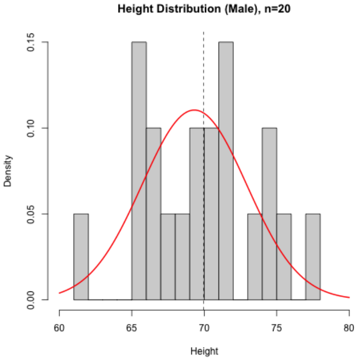
Sample Distributions

Sample 1: n=20,



$\overline{x}_H = 69.29$ inches,
sd_H=2.74 inches,

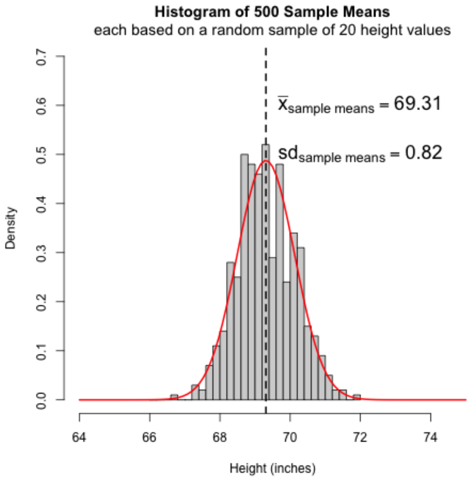
Sample 2: n=20,



$\overline{x}_H = 69.94$ inches
sd_H=3.97 inches

Sampling Distribution

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



Sampling Distributions — Definitions

Definition: Sampling Distribution

The distribution of the probability that a sample statistic takes each of its possible values (computed from samples of the same size, randomly drawn from the same population) is called the sampling distribution of that statistic.

- A **statistic** is any quantity computed from the data in a sample.
- A **parameter** is any quantity computed from the data of the entire population.

Sampling Distributions (cont'd)

Statistics of common interest are:

- Mean, median
- Variance
- Histogram

Important sampling distributions are those of:

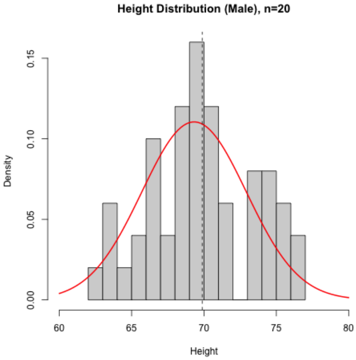
1. Sample mean
2. Difference between 2 sample means
3. Sample proportion
4. Difference between 2 sample proportions

Example 1: Self-Reported Heights of Male Students

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

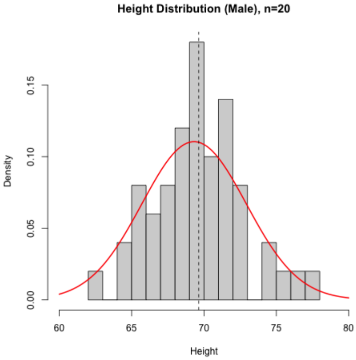
Sample Distributions

Sample 1: n=50,



$\overline{x}_H = 68.91$ inches,
sd_H=4.12 inches,

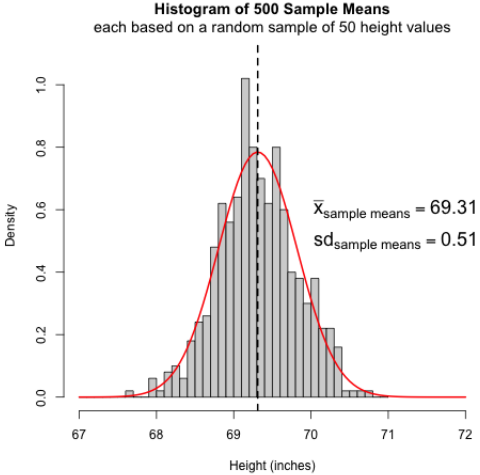
Sample 2: n=50,



$\overline{x}_H = 69.63$ inches
sd_H=3.18 inches

Sampling Distribution

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

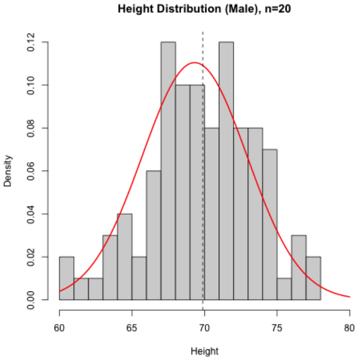


Example 1: Self-Reported Heights of Male Students

- We decide to do another experiment
- We are going to take 500 separate random samples from this population of mean, each with **100** students
- For each of the 500 samples, we will plot a histogram of the height values of students and record the sample mean and sample sd.
- Ready, set, go ...

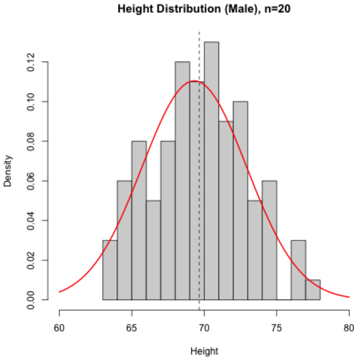
Sample Distributions

Sample 1: n=100,



$\overline{x}_H = 69.88$ inches,
sd_H=3.68 inches,

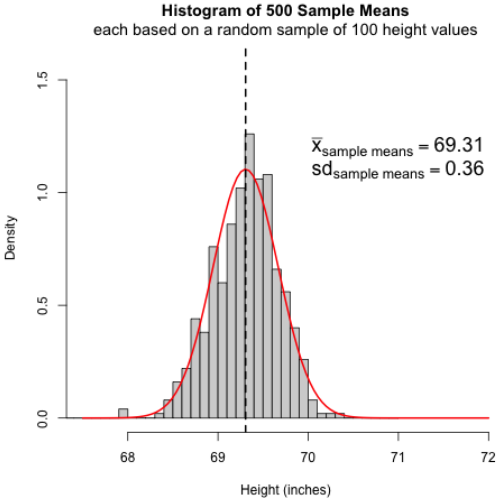
Sample 2: n=100,



$\overline{x}_H = 69.65$ inches
sd_H=3.23 inches

Sampling Distribution

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

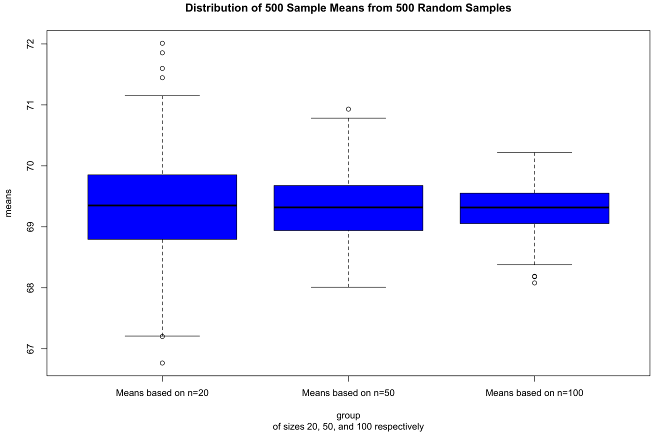


Example: Self-Reported Heights of Male Students

- Let's review the results
  - Population distribution of individual height measurements for male students: normal
  - True Population mean(μ) $=$ 69.31 inches, sd(σ)=3.61 inches
  - Results from 500 random samples:

	Means of 500 Sample Means	SD of 500 Sample Means	Shape of Distribution of 500 Sample Means
n=20	69.31	0.82	Approx Normal
n=50	69.31	0.51	Approx Normal
n=100	69.31	0.36	Approx Normal

Example: Self-Reported Heights of Male Students



Sampling Distribution of the Sample Mean

A. When observations are sampled independently at random from a normally distributed population:

1. The distribution of the values of the sample mean (the sample statistic of interest) will be **exactly** normal.
2. The mean of the sampling distribution will equal the true population mean.
3. The variance of the sampling distribution will be the population variance divided by the sample size.
4. The standard deviation of the sampling distribution (i.e., $\sqrt{\text{variance}}$) is referred to as the **standard error** of the statistic.

Sampling Distribution of the Sample Mean $\bar{X}$

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

Sample mean $\bar{X}$ is always an unbiased estimate of the population mean:

$$\mu_{\bar{X}} = E(\bar{X}) = \mu.$$

The standard deviation of $\bar{X}$ (the **standard error of the mean**) is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

Example 1: Self-Reported Heights of Male Students

- Let's review the results
  - Population distribution of individual height measurements for male students: normal
  - True Population mean(μ) = 69.31 inches, sd(σ)=3.61 inches
  - Results from 500 random samples:

	Means of 500 Sample Means	SD of 500 Sample Means	Shape of Distribution of 500 Sample Means
n=20	69.31	0.82	Approx Normal
n=50	69.31	0.51	Approx Normal
n=100	69.31	0.36	Approx Normal

Sampling Distribution of $\bar{X}$ from a Normal Population

If a random sample $(X_1, \ldots, X_n)$ is selected from a population with a **normal distribution** $N(\mu, \sigma^2)$, then the sample mean $\bar{x}$ is a normal random variable with mean μ and standard deviation $\sigma/\sqrt{n}$:

$$\bar{X} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$$

(CLT) Central Limit Theorem

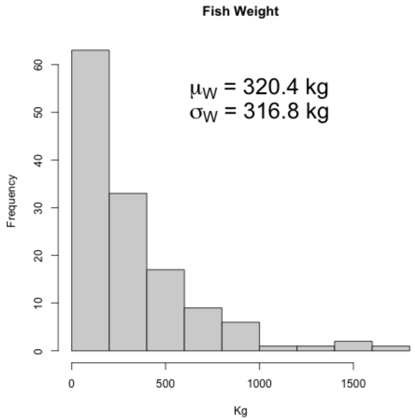
A random sample $(X_1, \ldots, X_n)$ selected from **any** population with mean μ and standard deviation σ. When n is sufficiently large ($n \geq 30$), the sampling distribution of $\bar{x}$ will be approximately

$$\bar{X} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right).$$

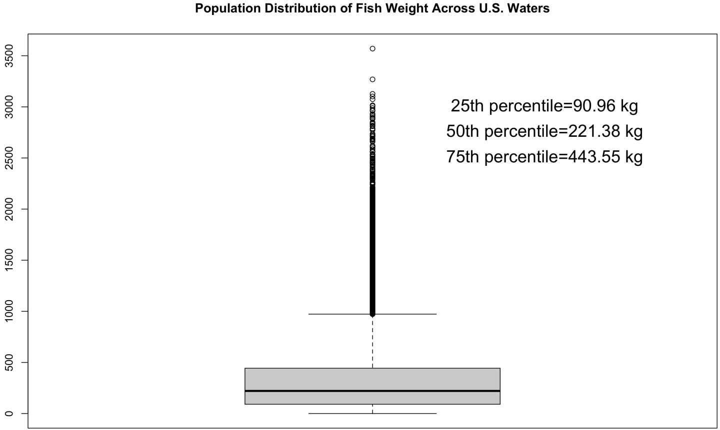
The larger the sample size n, the better the normal approximation.

Example: Fish Weight Across U.S. Waters

- A sample of 133 Fish across the United States
- Assume the population distribution is given by the following:
- True population mean ($\mu_W = 320.4$ kg); standard deviation (σ_W=316.8 kg)



Example 2: Fish Weight Across U.S. Waters

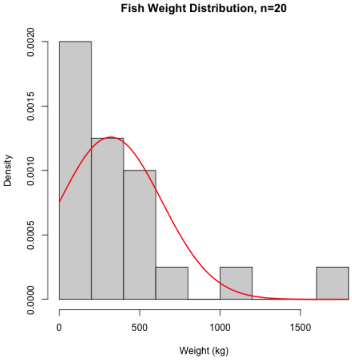


Example 2: Fish Weight Across U.S. Waters

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

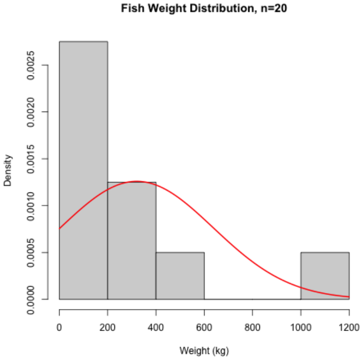
Fish Weight Across U.S. Waters

Sample 1: n=20,



$\overline{x}_W = 391.20$ Kg,
sd_W=385.66 Kg,

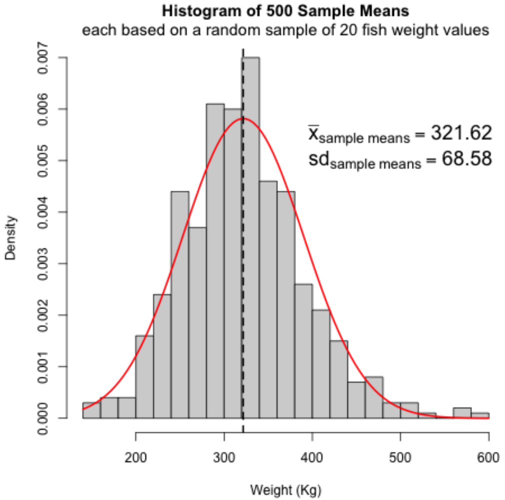
Sample 2: n=20,



$\overline{x}_W = 287.98.29$ Kg
sd_W=308.53 Kg

Example 2: Fish Weight Across U.S. Waters

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

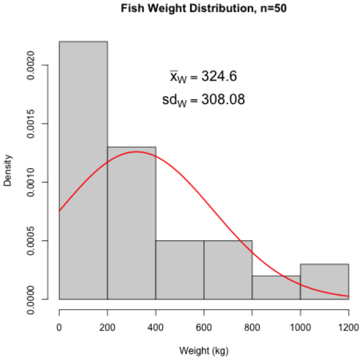


Example 2: Fish Weight Across U.S. Waters

- Suppose we had all the time in the world
- We decide to do another set of experiments
- We are going to take 500 separate random samples from this population of mean, each with **50** fish
- For each of the 500 samples, we will plot a histogram of sampled fish weight values and record the sample mean and sample sd.
- Ready, set, go ...

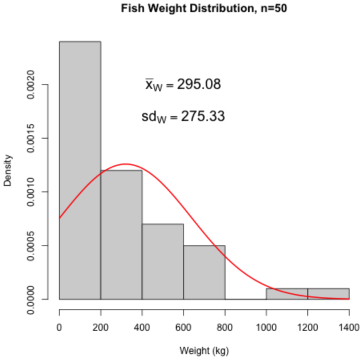
Fish Weight Across U.S. Waters

Sample 1: n=50,



$\bar{x}_W = 324.60$ Kg,
sd_W=308.08 Kg,

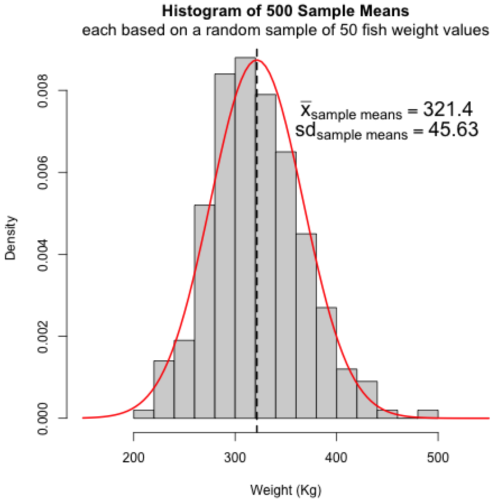
Sample 2: n=50,



$\bar{x}_W = 295.08$ Kg
sd_W=275.33 Kg

Example 2: Fish Weight Across U.S. Waters

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

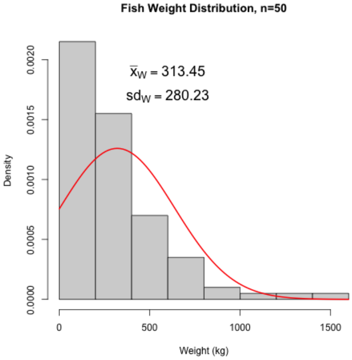


Example 2: Fish Weight Across U.S. Waters

- Suppose we had all the time in the world
- We decide to do another set of experiments
- We are going to take 500 separate random samples from this population of mean, each with $\mathbf{100}$ fish
- For each of the 500 samples, we will plot a histogram of sampled fish weight values and record the sample mean and sample sd.
- Ready, set, go ...

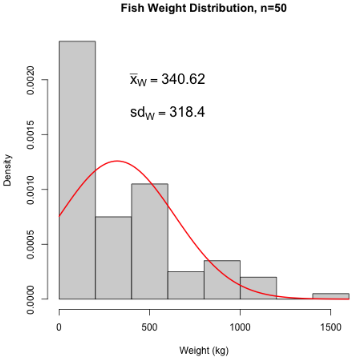
Fish Weight Across U.S. Waters

Sample 1: n=50,



$\overline{x}_W$ = 313.45 Kg,
sd_W=280.23 Kg,

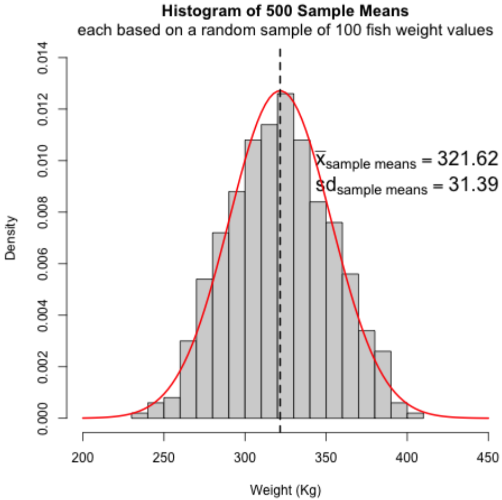
Sample 2: n=50,



$\overline{x}_W$ = 340.62 Kg
sd_W=318.4 Kg

Example 2: Fish Weight Across U.S. Waters

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

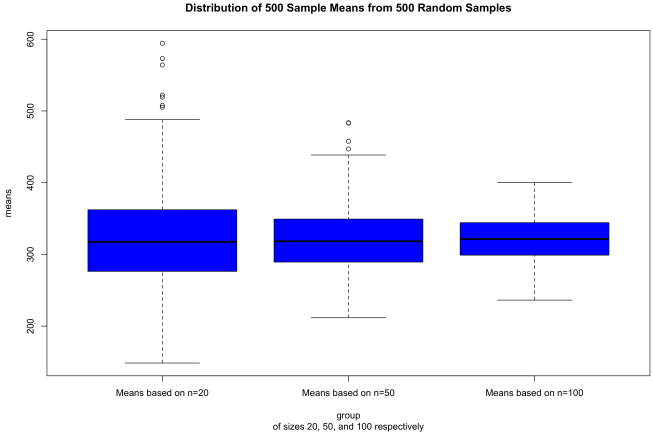


Example 2: Fish Weight Across U.S. Waters

- Let's review the results
  - Population distribution of fish weight Across U.S. waters: right skew
  - True population mean ($\mu_W = 320.4$ kg); standard deviation (σ_W=316.8 kg)
  - Results from 500 random samples:

	Means of 500 Sample Means	SD of 500 Sample Means	Shape of Distribution of 500 Sample Means
n=20	321.62	68.58	Approx Normal
n=50	321.40	45.63	Approx Normal
n=100	321.62	31.39	Approx Normal

Example 2: Fish Weight Across U.S. Waters



Summary

- What did we see across the two examples (Height of male students, Fish weight across U.S waters)?
- A couple of trends:
    - Distributions of sample means tended to be approximately normal even when original, individual level data was not (Fish Weight).
    - Variability in sample mean values decreased as size of sample of each mean based upon increased.
    - Distribution of sample means centered at true, population mean.

Central Limit Theorem (CLT)

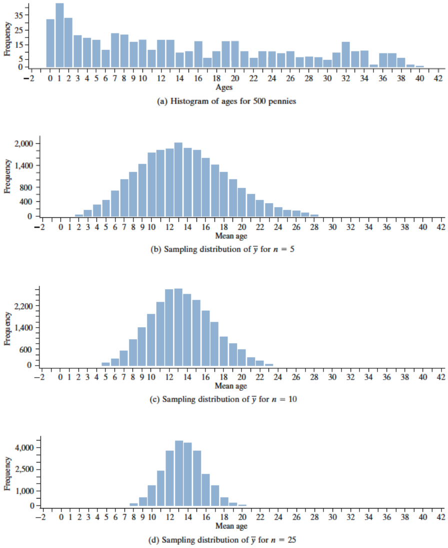
Central Limit Theorem

Given a population of any non-normal distribution with mean μ and finite variance σ^2, the sampling distribution of the sample mean computed from all possible samples of size n from this population will be approximately normal with mean μ and variance σ^2/n when the sample size is large.

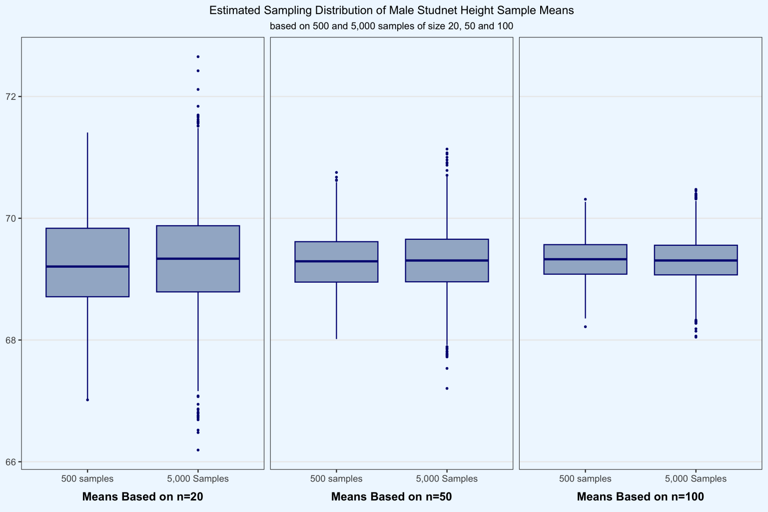
- In most practical situations, the CLT applies when $n > 25$.
- The approximation to normality becomes better as n increases and the population distribution is closer to normal.

3. Central Limit Theorem (cont'd)

FIGURE 4.20 Sampling distribution of $\bar{y}$ for $n = 1, 5, 10, 25$



Clarification



Example: Self-Reported Heights of Male Students

- Population distribution of individual height measurements for male students: normal
- True Population mean(μ) $=$ 69.31 inches, sd(σ)=3.61 inches

	Means of 500 Sample Means	Means of 5000 Sample Means	SD of 500 Sample Means	SD of 5000 Sample Means	CLT
n=20	69.25	69.32	0.84	0.81	$\frac{3.61}{\sqrt{20}} \approx 0.81$
n=50	69.30	69.30	0.52	0.52	$\frac{3.61}{\sqrt{50}} \approx 0.51$
n=100	69.33	69.31	0.36	0.36	$\frac{3.61}{\sqrt{100}} \approx 0.36$

Exercise: Soda Drinks

Suppose the mean and the standard deviation of the amount of soda in a large drink is 27.5oz and 1.5oz, respectively. If a random sample of 30 drinks is collected, what is the probability that they will average between 27 and 28 oz?

Let X = the amount of soda in a large drink; $\mu = 27.5$, $\sigma = 1.5$. $(X_1, \ldots, X_{30})$ is a random sample of 30 drinks ($n = 30$, large sample). By CLT, $\bar{X} \sim N(27.5, 0.2739^2)$ where $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 1.5/\sqrt{30} = 0.2739$.

$$P(27 < \bar{X} < 28) = P\left(\frac{27 - 27.5}{0.2739} < Z < \frac{28 - 27.5}{0.2739}\right)$$

$$= P(-1.83 < Z < 1.83) = 2 \cdot P(0 < Z < 1.83) = 2(0.4664) = 0.9328$$

Revisit the Central Limit Theorem

Sampling Distribution for the Mean

If $X_1, X_2, \ldots$ are an iid sequence of random variables from a population with mean μ and standard deviation σ, then the CLT says:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \mathrel{\dot\sim} N(0,1) \qquad \text{for large enough } n.$$

Sampling Distribution for the Sum

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} \mathrel{\dot\sim} N(0,1) \qquad \text{for large enough } n.$$

Sampling Distribution for the Sample Percentage

Let X be a binomial random variable with sample size n and probability p, and $\hat{p} = X/n$. Then

$$\frac{X - np}{\sqrt{np(1-p)}} \dot{\sim} N(0,1) \qquad \text{and} \qquad \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \dot{\sim} N(0,1)$$

for large enough n, which is equivalent to

$$\hat{p} \sim N\left(p, \left(\sqrt{\frac{pq}{n}}\right)^2\right).$$

Large Sample Rule

A sample is considered large if $n\hat{p} \geq 15$ and $n(1-\hat{p}) \geq 15$.

Example: Downloading Music

According to a *Pew Internet & American Life Project Survey* (Oct. 2010), 67% of adults who use the Internet have paid to download music. In a random sample of $n = 1000$ adults who use the Internet, let $\hat{p}$ represent the proportion who have paid to download music.

Example: (a)–(c)

(a) Mean of the sampling distribution of $\hat{p}$: $\mu_{\hat{p}} = E(\hat{p}) = p = 0.67$.
(b) SD of the **sampling distribution of** $\hat{p}$:

Example: Downloading Music

According to a *Pew Internet & American Life Project Survey* (Oct. 2010), 67% of adults who use the Internet have paid to download music. In a random sample of $n = 1000$ adults who use the Internet, let $\hat{p}$ represent the proportion who have paid to download music.

Example: (a)–(c)

(a) Mean of the sampling distribution of $\hat{p}$: $\mu_{\hat{p}} = E(\hat{p}) = p = 0.67$.
(b) SD of the **sampling distribution of** $\hat{p}$:

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.67(1-0.67)}{1000}} = 0.0149$$

(c) By the CLT: $\hat{p} \sim N(0.67, 0.0149^2)$

Example: Downloading Music (cont'd)

Example: (d) $P(\hat{p} < 0.75)$

Example: Downloading Music (cont'd)

Example: (d) $P(\hat{p} < 0.75)$

$$P(\hat{p} < 0.75) = P\left(\frac{\hat{p} - 0.67}{0.0149} < \frac{0.75 - 0.67}{0.0149}\right) = P(Z < 5.37) \approx 1$$

Example: (e) $P(\hat{p} > 0.50)$

Example: Downloading Music (cont'd)

Example: (d) $P(\hat{p} < 0.75)$

$$P(\hat{p} < 0.75) = P\left(\frac{\hat{p} - 0.67}{0.0149} < \frac{0.75 - 0.67}{0.0149}\right) = P(Z < 5.37) \approx 1$$

Example: (e) $P(\hat{p} > 0.50)$

$$P(\hat{p} > 0.50) = P\left(\frac{\hat{p} - 0.67}{0.0149} > \frac{0.50 - 0.67}{0.0149}\right) = P(Z > -11.41) \approx 1$$