Lecture 20: Chapter 8, Section 2 Sampling Distributions: Means

- Typical Inference Problem for Means
- 3 Approaches to Understanding Dist. of Means
- Center, Spread, Shape of Dist. of Means
- 68-95-99.7 Rule; Checking Assumptions

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-4)
  - Displaying and Summarizing (Lectures 5-12)
  - Probability
    - Finding Probabilities (discussed in Lectures 13-14)
    - Random Variables (discussed in Lectures 15-18)
    - Sampling Distributions
      - Proportions (discussed in Lecture 19)
      - Means
  - Statistical Inference

Typical Inference Problem about Mean

The numbers 1 to 20 have mean 10.5, s.d. 5.8. If numbers picked "at random" by sample of 400 students have mean 11.6, does this suggest bias in favor of higher numbers?

Solution Method: Assume (temporarily) that population mean is 10.5, find probability of sample mean as high as 11.6. If it's too improbable, we won't believe population mean is 10.5; we'll conclude there *is* bias in favor of higher numbers.

Key to Solving Inference Problems

For a given population mean μ , standard deviation σ, and sample size *n*, need to find probability of sample mean $\bar{X}$ in a certain range:

Need to know **sampling distribution** of $\bar{X}$.

Notation: $\bar{x}$ denotes a single statistic.

$\bar{X}$ denotes the random variable.

Definition *(Review)*

Sampling distribution of sample statistic tells probability distribution of values taken by the statistic in repeated random samples of a given size.

Looking Back: *We summarized probability distribution of **sample proportion** by reporting its center, spread, shape. Now we will do the same for **sample mean**.*

Understanding Sample Mean

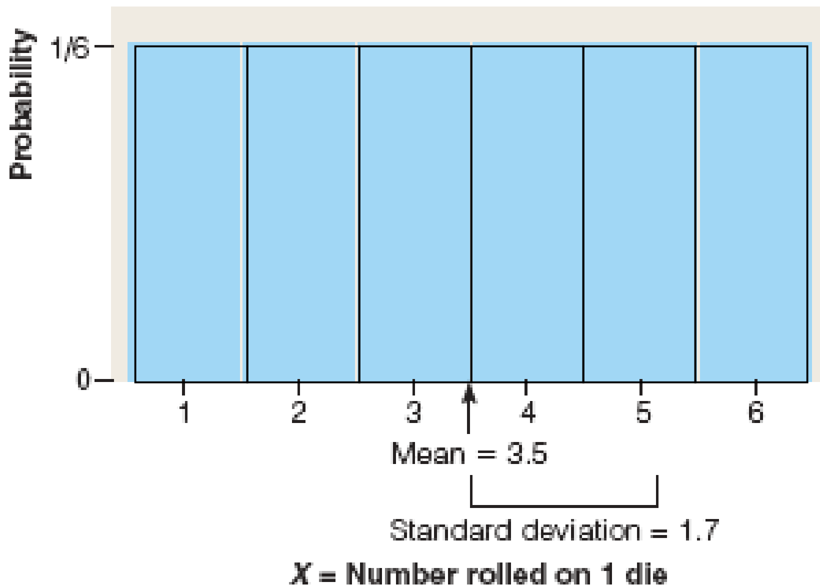
3 Approaches:

1. Intuition
2. Hands-on Experimentation
3. Theoretical Results

Looking Ahead: *We'll find that our* ***intuition*** *is consistent with* ***experimental*** *results, and both are confirmed by mathematical* ***theory****.*

Example: *Shape of Underlying Distribution (n=1)*

- **Background**: Population of possible dicerolls X are equally likely values {1,2,3,4,5,6}.
- **Question:** What is the probability histogram's shape?
- **Response:**



***Looking Ahead:** The shape of the underlying distribution will play a role in the shape of $\bar{X}$ for various sample sizes.*

Practice: 8.25a p.368

Example: *Sample Mean as Random Variable*

- **Background**: Population mean roll of dice is 3.5.
- **Questions:**
  - Is the underlying variable (dice roll) categorical or quantitative?
  - Consider the behavior of sample mean $\bar{X}$ for repeated rolls of a given number of dice. What type of variable is sample mean dice roll?
  - What 3 aspects of the distribution of sample mean should we report to summarize its behavior?
- **Responses:**
  - Underlying variable (number rolled) is
  - It's
  - Summarize with

Practice: 8.36 p.372

Example: *Center, Spread, Shape of Sample Mean*

- **Background**: Dice rolls X uniform with $\mu = 3.5$, $\sigma = 1.7$.
- **Question:** What are features of $\bar{X}$ for repeated rolls of 2 dice?
- **Response:**
  - **Center:** Some $\bar{X}$'s more than others less; they should balance out so mean of $\bar{X}$'s is μ= .
  - **Spread** of $\bar{X}$'s: (n=2 dice) easily range from to .
  - **Shape:**

Example: *Sample Mean for Larger n*

- **Background**: Dice rolls X uniform with $\mu = 3.5,\ \sigma = 1.7.$
- **Question:** What are features of $\bar{X}$ for repeated rolls of 8 dice?
- **Response:**
  - **Center:** Mean of $\bar{X}$'s is __________ (for any n).
  - **Spread:** (n=8 dice) __________ __________ spread than for n=2.
  - **Shape:** bulges more near 3.5, tapers at extremes 1 and 6→ shape close to __________

***Looking Ahead:** Sample size does **not** affect center but plays an important role in spread and shape of the distribution of sample mean (as it did for sample proportion).*

Practice: 8.19 p.367

Mean of Sample Mean (Theory)

For random samples of size *n* from population with mean μ, we can write sample mean as

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

where each X_i has mean μ. The Rules for constant multiples of means and for sums of means tell us that $\bar{X}$ has mean

$$\mu_{\bar{X}} = \frac{1}{n}(\mu + \mu + \cdots + \mu) = \frac{1}{n}(n\mu) = \mu$$

Standard Deviation of Sample Mean

For random samples of size *n* from population with mean μ, standard deviation σ, we write

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

where each X_i has s.d. σ. The Rules for constant multiples of s.d.s and for sums of **variances** tell us that $\bar{X}$ has s.d.

$$\frac{1}{n}\sqrt{\sigma^2 + \cdots + \sigma^2} = \frac{1}{n}\sqrt{n\sigma^2} = \frac{\sigma}{\sqrt{n}}$$

Rule of Thumb *(Review)*

- Need population size at least 10n

  (formula for s.d. of $\bar{X}$ approx. correct even if sampled without replacement)

Note: For means, there is no Rule of Thumb for approximate normality that is as simple as the one for proportions [np and $n(1-p)$ both at least 10].

Central Limit Theorem *(Review)*

Approximate normality of sample statistic for repeated random samples of a large enough size is cornerstone of inference theory.

- Makes intuitive sense.
- Can be verified with experimentation.
- Proof requires higher-level mathematics; result called Central Limit Theorem.

Shape of Sample Mean

For random samples of size n from population of quantitative values X, the shape of the distribution of sample mean $\bar{X}$ is approximately normal if

- X itself is normal; or
- X is fairly symmetric and n is at least 15; or
- X is moderately skewed and n is at least 30

Behavior of Sample Mean: Summary

For random sample of size n from population with mean μ, standard deviation σ, sample mean $\bar{x}$ has

- mean μ
- standard deviation $\frac{\sigma}{\sqrt{n}}$
- shape approximately normal for large enough n

Center of Sample Mean *(Implications)*

For random sample of size *n* from population with mean μ, sample mean $\bar{X}$ has

- mean μ

  → $\bar{X}$ is *unbiased estimator* of μ

(sample must be random)

Looking Ahead: *We'll rely heavily on this result when we perform inference. As long as the sample is random, sample mean is our "best guess" for unknown population mean.*

Spread of Sample Mean *(Implications)*

For random sample of size n from population with mean μ, s.d. σ, sample mean $\bar{X}$ has

- mean μ
- standard deviation $\frac{\sigma}{\sqrt{n}}$ ← n in denominator

→ $\bar{X}$ has *less spread for larger samples*

(population size must be at least 10n)

Looking Ahead: *This result also impacts inference conclusions to come. Sample mean from a larger sample gives us a better estimate for the unknown population mean.*

Shape of Sample Mean *(Implications)*

For random sample of size *n* from population with mean μ, s.d. σ , sample mean $\bar{X}$ has

- mean μ
- standard deviation $\frac{\sigma}{\sqrt{n}}$
- shape approx. normal for large enough *n*

→can find probability that sample mean takes value in given interval

Looking Ahead: *Finding probabilities about sample mean will enable us to solve inference problems.*

Example: *Behavior of Sample Mean, 2 Dice*

- **Background**: Population of dice rolls has $\mu = 3.5,\ \sigma = 1.7$
- **Question:** For repeated random samples of ***n*=2**, how does sample mean $\bar{X}$ behave?
- **Response:** For ***n*=2**, sample mean roll $\bar{X}$ has
  - **Center:** mean
  - **Spread:** standard deviation
  - **Shape** because the population is flat, not normal, and

Practice: 8.75a p.381

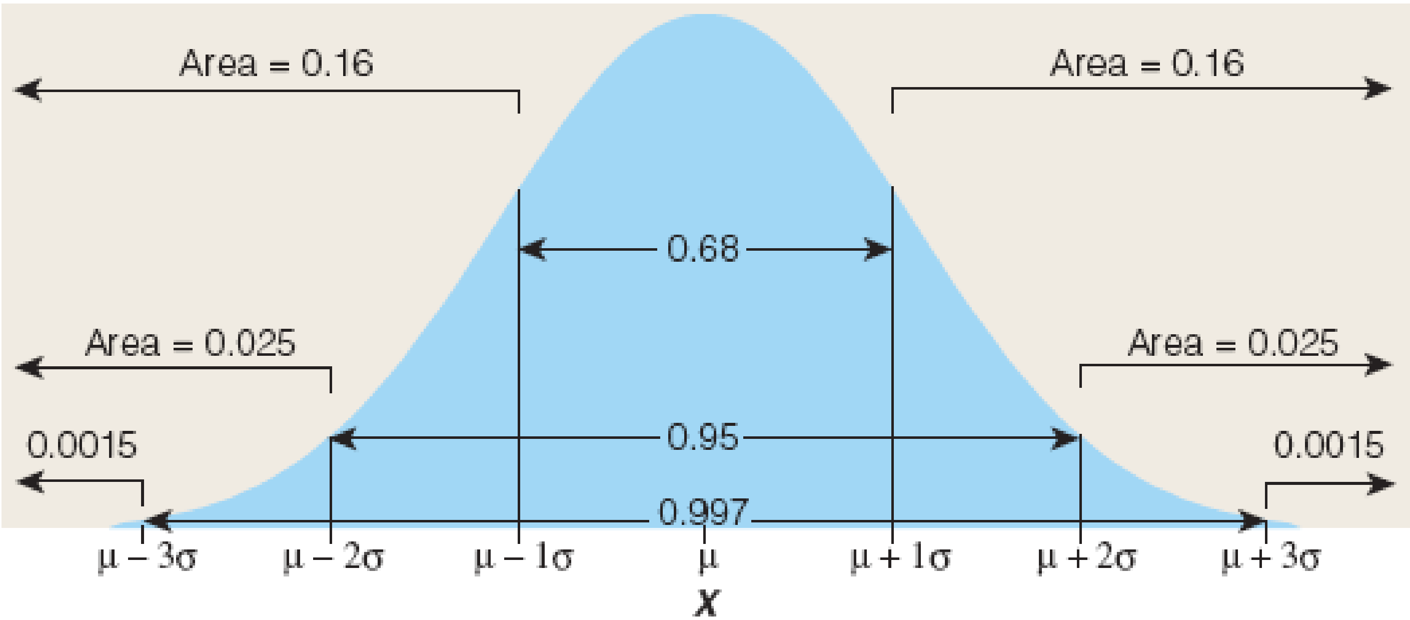
Example: *Behavior of Sample Mean, 8 Dice*

- **Background**: Population of dice rolls has $\mu = 3.5, \sigma = 1.7$
- **Question:** For repeated random samples of ***n*=8**, how does sample mean $\bar{X}$ behave?
- **Response:** For ***n*=8**, sample mean roll $\bar{X}$ has
  - **Center:** mean
  - **Spread:** standard deviation
  - **Shape:** normal than for *n*=2 (Central Limit Theorem)

68-95-99.7 Rule for Normal R.V. *(Review)*

Sample at random from normal population; for sampled value *X* (a R.V.), probability is

- 68% that *X* is within 1 standard deviation of mean
- 95% that *X* is within 2 standard deviations of mean
- 99.7% that *X* is within 3 standard deviations of mean



68-95-99.7 Rule for Sample Mean

For sample means $\bar{X}$ taken at random from large population with mean μ, s.d. σ, probability is

- 68% that $\bar{X}$ is within $1\frac{\sigma}{\sqrt{n}}$ of μ
- 95% that $\bar{X}$ is within $2\frac{\sigma}{\sqrt{n}}$ of μ
- 99.7% that $\bar{X}$ is within $3\frac{\sigma}{\sqrt{n}}$ of μ

These results hold only if n is large enough.

Example: *68-95-99.7 Rule for 8 Dice*

- **Background**: Population of dice rolls has $\mu = 3.5,\ \sigma = 1.7$. For random samples of size 8, sample mean roll $\bar{X}$ has mean 3.5, standard deviation 0.6, and shape fairly normal.
- **Question:** What does 68-95-99.7 Rule tell us about the behavior of $\bar{X}$?
- **Response:** The probability is approximately
  - 0.68 that $\bar{X}$ is within ______ of ______: in (2.9, 4.1)
  - 0.95 that $\bar{X}$ is within ______ of ______: in (2.3, 4.7)
  - 0.997 that $\bar{X}$ is within ______ of ______: in (1.7, 5.3)

Typical Problem about Mean *(Review)*

The numbers 1 to 20 have mean 10.5, s.d. 5.8. If numbers picked "at random" by sample of 400 students has mean 11.6, does this suggest bias in favor of higher numbers?

Solution Method: Assume (temporarily) that population mean is 10.5, find probability of sample mean as high as 11.6. If it's too improbable, we won't believe population mean is 10.5; we'll conclude there *is* bias in favor of higher numbers.

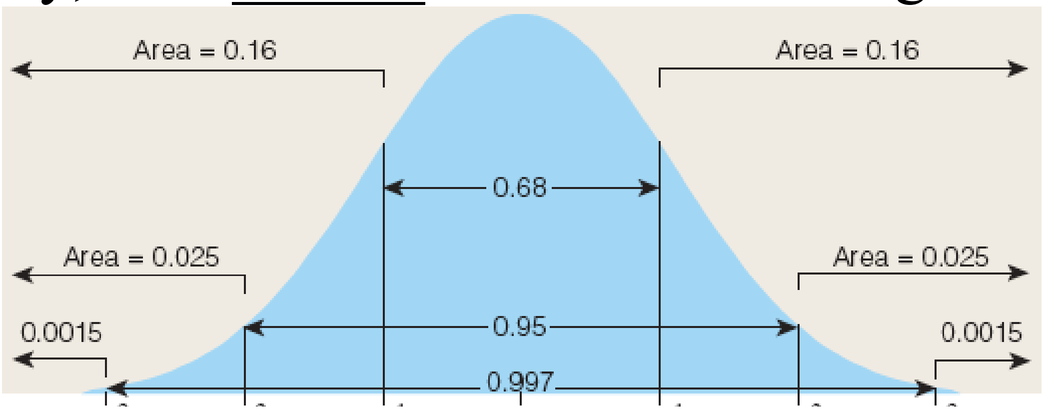
Example: *Establishing Behavior of* $\bar{X}$

- **Background**: We asked the following: "*The numbers 1 to 20 have mean 10.5, s.d. 5.8. If numbers picked 'at random' by 400 students have mean* $\bar{x}$ *=11.6, does this suggest bias in favor of higher numbers?*"
- **Question:** What are the mean, standard deviation, and shape of the R.V. $\bar{X}$ in this situation?
- **Response:** For μ= 10.5, σ=5.8, and n=400, $\bar{X}$ has
  - mean
  - standard deviation
  - shape

Practice: 8.17 p.366

Example: *Testing Assumption About μ*

- **Background**: Sample mean number picked at random from 1 to 20 by 400 students should have mean 10.5, s.d. 0.3.
- **Questions:** Is 11.6 improbably high for $\bar{X}$? Does a sample mean of 11.6 convince us of bias in favor of higher numbers?
- **Responses:** 11.6 is above more than 3 s.d.s. The probability of being this high (or higher) is Since this is extremely improbable, we ______ believe μ =10.5. Apparently, there_______ bias in favor of higher numbers.



9.6 9.9 10.2 10.5 10.8 11.1 11.4

Practice: 8.32 p.369

Example: *Behavior of Individual vs. Mean*

- **Background**: IQ scores are normal with mean 100, s.d. 15.
- **Question:** Is 88 unusually low for…
  - IQ of a randomly chosen individual?
  - Mean IQ of 9 randomly chosen individuals?
- **Response:**
  - IQ X of a randomly chosen individual has mean 100, s.d. 15. For x=88, $z =$
    not even 1 s.d. below the mean→
  - Mean IQ $\bar{X}$ of 9 randomly chosen individuals has mean 100, s.d. . For $\bar{x}$=88, $z =$
    unusually low (happens less than of the time, since
).

Example: *Checking Assumptions*

- **Background**: Household size X in the U.S. has mean 2.5, s.d. 1.4.
- **Question:** Is 3 unusually high for…
  - Size of a randomly chosen household?
  - Mean size of 10 randomly chosen households?
  - Mean size of 100 randomly chosen households?
- **Response:**
  -
  -
  - n=100 large→ $\bar{X}$ normal; mean 2.5, s.d. $\frac{1.4}{\sqrt{100}} = 0.14$
    so $\bar{x} = 3$ has z = (3-2.5)/0.14 = +3.57: unusually high.

Lecture Summary
(Sampling Distributions; Means)

- Typical inference problem for means
- 3 approaches to understanding dist. of sample mean
  - Intuit
  - Hands-on
  - Theory
- Center, spread, shape of dist. of sample mean
- 68-95-99.7 Rule for sample mean
  - Revisit typical problem
  - Checking assumptions for use of Rule