Lecture 13: Chapter 8, Sections 1-2 Sampling Distributions: Proportions; begin Means

- Typical Inference Problem
- Definition of Sampling Distribution
- 3 Approaches to Understanding Sampling Dist.
- Applying 68-95-99.7 Rule
- Means: Inference Problem, 3 Approaches
- Center, Spread, Shape of Sample Mean

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-3)
  - Displaying and Summarizing (Lectures 3-8)
  - Probability
    - Finding Probabilities (discussed in Lectures 9-10)
    - Random Variables (discussed in Lectures 10-12)
    - Sampling Distributions
      - Proportions
      - Means
  - Statistical Inference

Typical Inference Problem

If sample of 100 students has 0.13 left-handed, can you believe population proportion is 0.10?

Solution Method: Assume (temporarily) that population proportion is 0.10, find probability of sample proportion as high as 0.13. If it's too improbable, we won't believe population proportion is 0.10.

Key to Solving Inference Problems

For a given population proportion p and sample size n, need to find probability of sample proportion $\hat{p}$ in a certain range:

Need to know **sampling distribution** of $\hat{p}$.

Note: $\hat{p}$ can denote a single statistic or a random variable.

Definition

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 summarize a probability distribution by reporting its* ***center, spread, shape.***

Behavior of Sample Proportion *(Review)*

For random sample of size *n* from population with *p* in category of interest, sample proportion $\hat{p} = \frac{X}{n}$ has

- mean *p*
- standard deviation $\sqrt{\frac{p(1-p)}{n}}$
- shape approximately normal for large enough *n*

Looking Back: *Can find normal probabilities using 68-95-99.7 Rule, etc.*

Rules of Thumb *(Review)*

- Population at least 10 times sample size n

  (formula for standard deviation of $\hat{p}$ approximately correct even if sampled without replacement)

- np and $n(1-p)$ both at least 10

  (guarantees $\hat{p}$ approximately normal)

Understanding Dist. of Sample Proportion

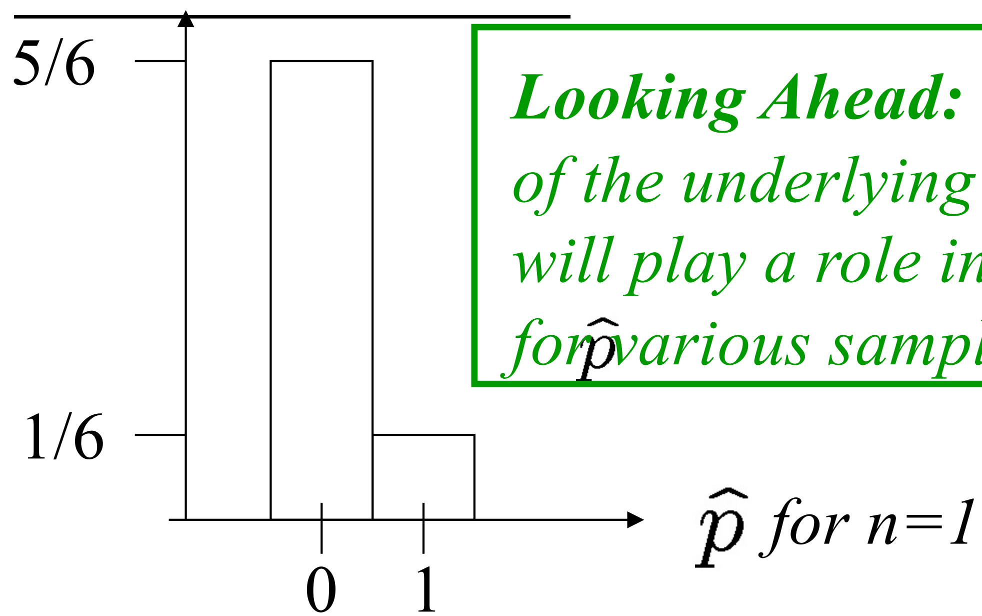
3 Approaches:

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 proportion of blue M&M's is $p=1/6=0.17$.
- **Question:** How does the probability histogram for sample proportions appear for samples of size 1?
- **Response:**



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

Practice: 8.53b p.375

Example: *Sample Proportion as Random Variable*

- **Background**: Population proportion of blue M&Ms is 0.17.
- **Questions:**
  - Is the underlying variable categorical or quantitative?
  - Consider the behavior of sample proportion $\hat{p}$ for repeated random samples of a given size. What type of variable is sample proportion?
  - What 3 aspects of the distribution of sample proportion should we report to summarize its behavior?
- **Responses:**
  - Underlying variable ___________________________________
  - ___________________________
  - Summarize with __________, __________, __________

Practice: 8.53a p.375

Example: *Center, Spread of Sample Proportion*

- **Background**: Population proportion of blue M&M's is p=1/6=0.17.
- **Question:** What can we say about center and spread of $\hat{p}$ for repeated random samples of size n = 25 (a teaspoon)?
- **Response:**
  - **Center:** Some $\hat{p}$'s more than ____, others less; should balance out so mean of $\hat{p}$'s is p = ________.
  - **Spread** of $\hat{p}$'s: s.d. depends on ____.
    - For n=6, could easily get $\hat{p}$ anywhere from ___to ___.
    - For n=25, spread of $\hat{p}$ will be ____than it is for n = 6.

Example: *Intuit Shape of Sample Proportion*

- **Background**: Population proportion of blue M&M's is p=1/6=0.17.
- **Question:** What can we say about the shape of $\hat{p}$ for repeated random samples of size n = 25 (a teaspoon)?
- **Response:**

  $\hat{p}$ close to____ most common, far from _____ in either direction increasingly less likely→

  ____

Practice: 8.53e p.375

Example: *Sample Proportion for Larger n*

- **Background**: Population proportion of blue M&M's is p=1/6=0.17.
- **Question:** What can we say about center, spread, shape of $\hat{p}$ for repeated random samples of size n = 75 (a Tablespoon)?
- **Response:**
  - **Center:** mean of $\hat{p}$'s should be p = _____ (for any n).
  - **Spread** of $\hat{p}$'s: compared to n=25, spread for n=75 is ____
  - **Shape:** $\hat{p}$'s clumped near 0.17, taper at tails→ _________

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

Understanding Sample Proportion

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****.*

Central Limit Theorem

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.

Center of Sample Proportion *(Implications)*

For random sample of size n from population with p in category of interest, sample proportion $\hat{p} = \frac{X}{n}$ has

- mean p

→ $\hat{p}$ is *unbiased estimator* of p

(sample must be random)

Spread of Sample Proportion *(Implications)*

For random sample of size *n* from population with *p* in category of interest, sample proportion $\hat{p} = \frac{X}{n}$ has

- mean *p*
- standard deviation $\sqrt{\frac{p(1-p)}{n}}$ ← *n* in denominator

→ $\hat{p}$ has less spread for larger samples

(population size must be at least 10n)

Shape of Sample Proportion *(Implications)*

For random sample of size *n* from population with *p* in category of interest, sample proportion $\hat{p} = \frac{X}{n}$ has

- mean *p*
- standard deviation $\sqrt{\frac{p(1-p)}{n}}$
- shape approx. normal for large enough *n*

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

Example: *Behavior of Sample Proportion*

- **Background**: Population proportion of blue M&M's is p=0.17.
- **Question:** For repeated random samples of n=**25**, how does $\hat{p}$ behave?
- **Response:** For n=25, $\hat{p}$ has
  - **Center:** mean ________
  - **Spread:** standard deviation __________________
  - **Shape:** not really normal because

    ____

Practice: 8.53c-e p.375

Example: *Sample Proportion for Larger n*

- **Background**: Population proportion of blue M&M's is p=0.17.
- **Question:** For repeated random samples of n=**75**, how does $\hat{p}$ behave?
- **Response:** For n=75, $\hat{p}$ has
  - **Center:** mean ________
  - **Spread:** standard deviation __________________
  - **Shape:** approximately normal because

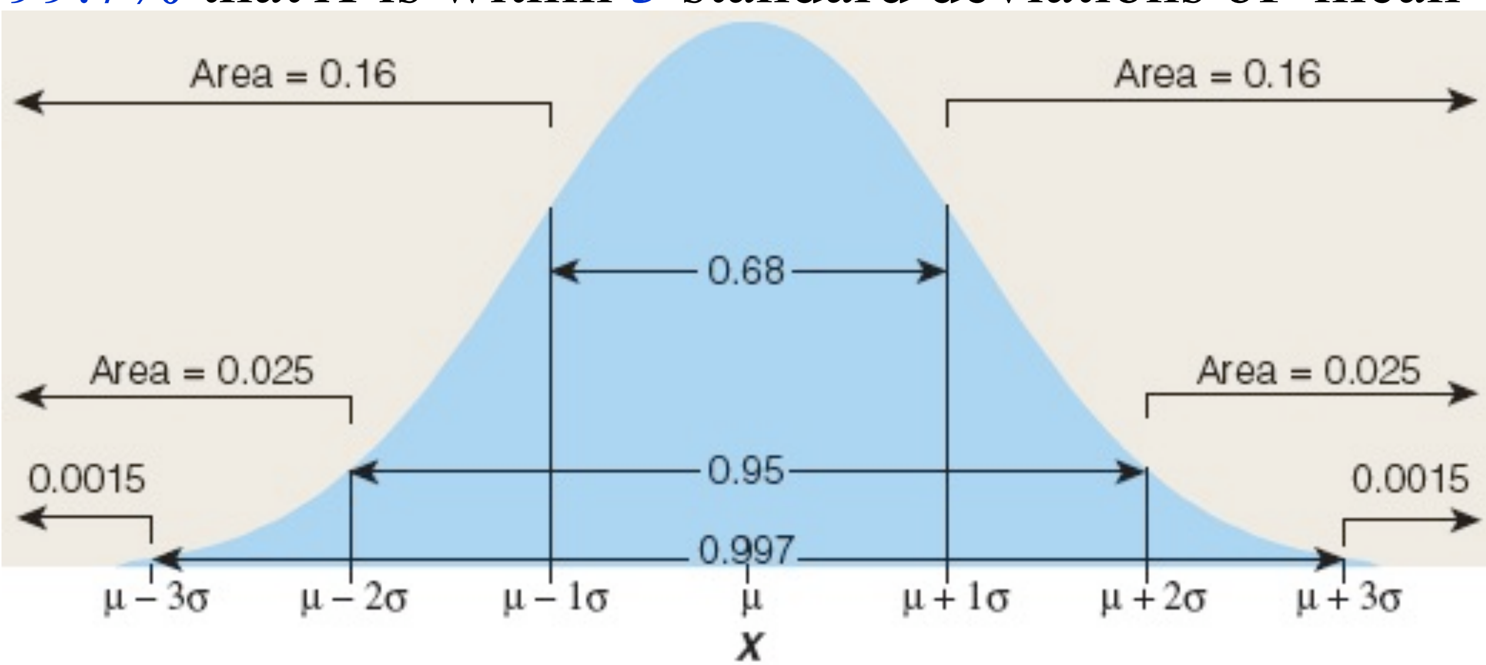
____

Practice: 8.53f-h p.375

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 Proportion

For sample proportions $\hat{p}$ taken at random from a large population with underlying *p*, probability is

- 68% that $\hat{p}$ is within 1 $\sqrt{\frac{p(1-p)}{n}}$ of *p*
- 95% that $\hat{p}$ is within 2 $\sqrt{\frac{p(1-p)}{n}}$ of *p*
- 99.7% that $\hat{p}$ is within $3\sqrt{\frac{p(1-p)}{n}}$ of *p*

Example: *Sample Proportion for n=75, p=0.17*

- **Background**: Population proportion of blue M&Ms is $p=0.17$. For random samples of $n=75$, $\hat{p}$ approx. normal with mean 0.17, s.d. $\sqrt{\frac{0.17(1-0.17)}{75}} = 0.043$
- **Question:**

What does 68-95-99.7 Rule tell us about behavior of $\hat{p}$?

- **Response:** The probability is approximately
  - 0.68 that $\hat{p}$ is within ________ of ____: in (0.13, 0.21)
  - 0.95 that $\hat{p}$ is within ________ of ____: in (0.08, 0.26)
  - 0.997 that $\hat{p}$ is within ________ of ____: in (0.04, 0.30)

Looking Back: *We don't use the Rule for n=25 because* ____________________

Practice: 8.57 p.376

Typical Inference Problem *(Review)*

If sample of 100 students has 0.13 left-handed, can you believe population proportion is 0.10?

Solution Method: Assume (temporarily) that population proportion is 0.10, find probability of sample proportion as high as 0.13. If it's too improbable, we won't believe population proportion is 0.10.

Example: *Testing Assumption About p*

- **Background**: We asked, "*If sample of 100 students has 0.13 left-handed, can you believe population proportion is 0.10?*"
- **Questions:**
  - What are the mean, standard deviation, and shape of $\hat{p}$?
  - Is 0.13 improbably high under the circumstances?
  - Can we believe p = 0.10?
- **Response:**
  - For p=0.10 and n=100, $\hat{p}$ has mean ____, s.d. ___________________ ; shape approx. normal since ____.
  - According to Rule, the probability is __________________ that $\hat{p}$ would take a value of 0.13 (1 s.d. above mean) or more.
  - Since this isn't so improbable, ____________________________________.

Practice: 8.11d-f p.355

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-3)
  - Displaying and Summarizing (Lectures 3-8)
  - Probability
    - Finding Probabilities (discussed in Lectures 9-10)
    - Random Variables (discussed in Lectures 10-12)
    - Sampling Distributions
      - Proportions (discussed just now in Lecture 13)
      - 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

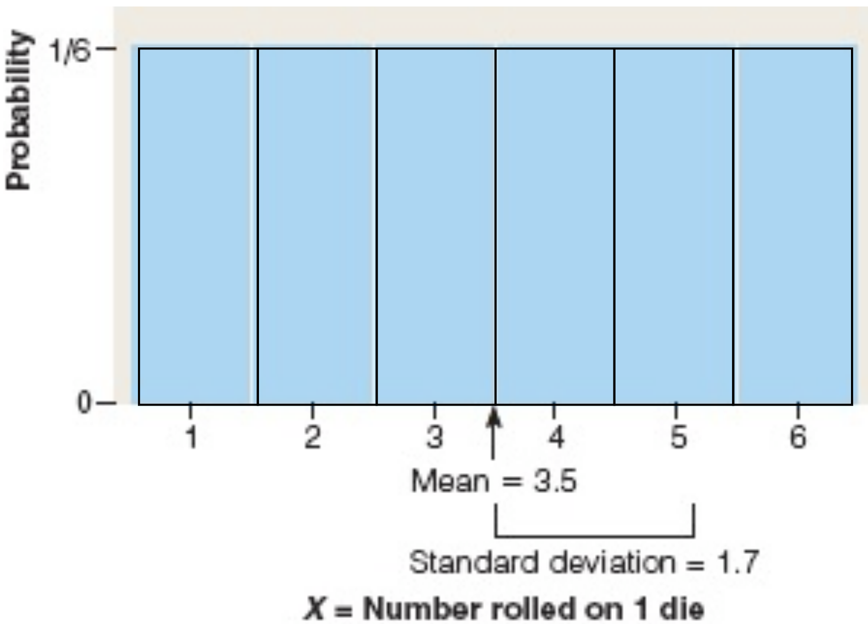
3 Approaches:

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.*

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)

Practice: 8.15 p.366

Lecture Summary
(Distribution of Sample Proportion)

- Typical inference problem
- Sampling distribution; definition
- 3 approaches to understanding sampling dist.
  - Intuition
  - Hands-on experiment
  - Theory
- Center, spread, shape of sampling distribution
  - Central Limit Theorem
- Role of sample size
- Applying 68-95-99.7 Rule

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