# STAT 111

## Recitation 5

Mo Huang

Email: mohuang@wharton.upenn.edu

Office Hours: Wednesdays 3:00 - 4:00 pm, JMHH F96

Slides: `github.com/mohuangx/STAT111-Spring2019`

February 22, 2019

# Two-Standard-Deviation Rule

- From the chart:

$$P(Z < -1.96) = 0.025, \quad P(Z > 1.96) = 0.025.$$

# Two-Standard-Deviation Rule

- From the chart:

$$P(Z < -1.96) = 0.025, \quad P(Z > 1.96) = 0.025.$$

- Then:

$$P(-1.96 < Z < 1.96) = 0.95.$$

# Two-Standard-Deviation Rule

- From the chart:

$$P(Z < -1.96) = 0.025, \quad P(Z > 1.96) = 0.025.$$

- Then:

$$P(-1.96 < Z < 1.96) = 0.95.$$

- Approximate $1.96 \approx 2$ and "unstandardize":

$$P\left(-2 < \frac{X - \mu}{\sigma} < 2\right) = 0.95$$

# Two-Standard-Deviation Rule

- From the chart:

$$P(Z < -1.96) = 0.025, \quad P(Z > 1.96) = 0.025.$$

- Then:

$$P(-1.96 < Z < 1.96) = 0.95.$$

- Approximate $1.96 \approx 2$ and “unstandardize”:

$$P\left(-2 < \frac{X - \mu}{\sigma} < 2\right) = 0.95$$

$\Rightarrow$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95.$$

# Two-Standard-Deviation Rule

- From the chart:

$$P(Z < -1.96) = 0.025, \quad P(Z > 1.96) = 0.025.$$

- Then:

$$P(-1.96 < Z < 1.96) = 0.95.$$

- Approximate $1.96 \approx 2$ and "unstandardize":

$$P\left(-2 < \frac{X - \mu}{\sigma} < 2\right) = 0.95$$

$\Rightarrow$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95.$$

- The probability that a normal random variable is within 2 standard deviations of the mean is 95%.

# Normal Distribution: Sums and Averages

- Let $X_1, \ldots, X_n$ be independent and *normally distributed*. Let

$$T_n = X_1 + \cdots + X_n, \qquad \bar{X} = \frac{X_1 + \cdots X_n}{n}, \qquad D = X_2 - X_1.$$

# Normal Distribution: Sums and Averages

- Let $X_1, \dots, X_n$ be independent and *normally distributed*. Let

$$T_n = X_1 + \cdots + X_n, \qquad \bar{X} = \frac{X_1 + \cdots X_n}{n}, \qquad D = X_2 - X_1.$$

- Then $T_n$, $\bar{X}$ and $D$ are also normal random variables.

# Normal Distribution: Sums and Averages

- Let $X_1, \ldots, X_n$ be independent and *normally distributed*. Let

$$T_n = X_1 + \cdots + X_n, \qquad \bar{X} = \frac{X_1 + \cdots X_n}{n}, \qquad D = X_2 - X_1.$$

- Then $T_n$, $\bar{X}$ and $D$ are also normal random variables.

- Let $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$. Then:

$$T_n \sim N(n\mu, n\sigma^2)$$

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

$$D \sim N(0, 2\sigma^2)$$

# Example

- Suppose we know the weight $X$ of an adult man chosen at random is normally distributed with mean 160 pounds and variance 64 pounds$^2$.

a) Find $P(156 < X < 164)$.

# Example

- Suppose we know the weight $X$ of an adult man chosen at random is normally distributed with mean 160 pounds and variance 64 pounds$^2$.

a) Find $P(156 < X < 164)$.

$$\begin{aligned} P(156 < X < 164) &= P\left(\frac{156-160}{8} < \frac{X-160}{8} < \frac{164-160}{8}\right) \\ &= P(-0.5 < Z < 0.5) \\ &= 0.3830 \end{aligned}$$

# Example

- Suppose we know the weight $X$ of an adult man chosen at random is normally distributed with mean 160 pounds and variance 64 pounds$^2$.

a) Find $P(156 < X < 164)$.

$$\begin{aligned} P(156 < X < 164) &= P\left(\frac{156-160}{8} < \frac{X-160}{8} < \frac{164-160}{8}\right) \\ &= P(-0.5 < Z < 0.5) \\ &= 0.3830 \end{aligned}$$

b) Find the probability that the average weight of 16 men chosen at random is between 156 and 164 pounds.

# Example

- Suppose we know the weight $X$ of an adult man chosen at random is normally distributed with mean 160 pounds and variance 64 pounds$^2$.

a) Find $P(156 < X < 164)$.

$$\begin{aligned} P(156 < X < 164) &= P\left(\frac{156-160}{8} < \frac{X-160}{8} < \frac{164-160}{8}\right) \\ &= P(-0.5 < Z < 0.5) \\ &= 0.3830 \end{aligned}$$

b) Find the probability that the average weight of 16 men chosen at random is between 156 and 164 pounds.

$$\begin{aligned} \overline{X} &\sim N(160, 64/16 = 4) \\ P(156 < \overline{X} < 164) &= P(-2 < Z < 2) \\ &\approx 0.95 \end{aligned}$$

# Example

- Suppose we know the weight $X$ of an adult man chosen at random is normally distributed with mean 160 pounds and variance 64 pounds$^2$.

c) Calculate the numbers $A$ and $B$ such that $P(A < X < B) \approx 0.95$.

# Example

- Suppose we know the weight $X$ of an adult man chosen at random is normally distributed with mean 160 pounds and variance 64 pounds$^2$.

c) Calculate the numbers $A$ and $B$ such that $P(A < X < B) \approx 0.95$.

$$0.95 \approx P\left( -2 < \frac{X-\mu}{\sigma} < 2 \right) = P(-2\sigma + \mu < X < 2\sigma + \mu)$$

$$A = -2(8) + 160 = 144, \quad B = 2(8) + 160 = 176$$

# Example

- Suppose we know the weight $X$ of an adult man chosen at random is normally distributed with mean 160 pounds and variance 64 pounds$^2$.

c) Calculate the numbers $A$ and $B$ such that $P(A < X < B) \approx 0.95$.

$$0.95 \approx P\left(-2 < \frac{X-\mu}{\sigma} < 2\right) = P(-2\sigma + \mu < X < 2\sigma + \mu)$$

$$A = -2(8) + 160 = 144, \quad B = 2(8) + 160 = 176$$

d) Calculate the numbers $C$ and $D$ such that the average of 256 randomly chosen adults is between $C$ and $D$ with probability approximately 0.95.

# Example

- Suppose we know the weight $X$ of an adult man chosen at random is normally distributed with mean 160 pounds and variance 64 pounds$^2$.

c) Calculate the numbers $A$ and $B$ such that $P(A < X < B) \approx 0.95$.

$$0.95 \approx P\left(-2 < \frac{X-\mu}{\sigma} < 2\right) = P(-2\sigma + \mu < X < 2\sigma + \mu)$$
$$A = -2(8) + 160 = 144, \quad B = 2(8) + 160 = 176$$

d) Calculate the numbers $C$ and $D$ such that the average of 256 randomly chosen adults is between $C$ and $D$ with probability approximately 0.95.

$$\overline{X}_{256} \sim N(160, 64/256 = 1/4)$$
$$C = -2\sigma + \mu = -2(1/2) + 160 = 159$$
$$D = 2\sigma + \mu = 2(1/2) + 160 = 161$$

# Central Limit Theorem

The Central Limit Theorem:

- Suppose $X_1, X_2, \ldots, X_n$ are *iid* with mean $\mu$ and variance $\sigma^2$.
- Then, for large $n$

$$T_n \sim N(n\mu, n\sigma^2) \quad \text{and} \quad \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

*no matter the distribution of the individual* $X_i$

- Allows approximation of all distributions using the normal distribution if you know the mean and variance.

# Central Limit Theorem

The Central Limit Theorem:

- Suppose $X_1, X_2, \ldots, X_n$ are *iid* with mean $\mu$ and variance $\sigma^2$.
- Then, for large $n$

$$T_n \sim N(n\mu, n\sigma^2) \quad \text{and} \quad \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

*no matter the distribution of the individual $X_i$*

- Allows approximation of all distributions using the normal distribution if you know the mean and variance.

Note: if $X_1, \ldots, X_n$ are normally distributed, then this applies for *all* $n$, not just large $n$.

# Central Limit Theorem: Example

- Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Binomial(1, \theta)$.

# Central Limit Theorem: Example

- Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Binomial(1, \theta)$.

- For each $X_i$, $Mean(X_i) = \theta$ and $Var(X_i) = \theta(1 - \theta)$.

# Central Limit Theorem: Example

- Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Binomial(1, \theta)$.
- For each $X_i$, $Mean(X_i) = \theta$ and $Var(X_i) = \theta(1 - \theta)$.
- The sum is: $T_n = X_1 + X_2 + \cdots + X_n$.

# Central Limit Theorem: Example

- Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Binomial(1, \theta)$.
- For each $X_i$, $Mean(X_i) = \theta$ and $Var(X_i) = \theta(1 - \theta)$.
- The sum is: $T_n = X_1 + X_2 + \cdots + X_n$.
- The proportion is:

$$P = \frac{X_1 + \cdots + X_n}{n}.$$

# Central Limit Theorem: Example

- Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Binomial(1, \theta)$.
- For each $X_i$, $Mean(X_i) = \theta$ and $Var(X_i) = \theta(1-\theta)$.
- The sum is: $T_n = X_1 + X_2 + \cdots + X_n$.
- The proportion is:

$$P = \frac{X_1 + \cdots + X_n}{n}.$$

# Central Limit Theorem: Example

- Let $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim} Binomial(1, \theta)$.

- For each $X_i$, $Mean(X_i) = \theta$ and $Var(X_i) = \theta(1-\theta)$.

- The sum is: $T_n = X_1 + X_2 + \cdots + X_n$.

- The proportion is:
$$P = \frac{X_1 + \cdots + X_n}{n}.$$

- For large $n$,
$$T_n \sim N\left(n\theta, n\theta[1-\theta]\right)$$
$$P \sim N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$$

# Central Limit Theorem: Problem

- Suppose you are rolling a fair die 1000 times. Calculate the numbers $A$ and $B$ such that the average of the 1000 rolls is between $A$ and $B$ with probability approximately 0.95. You may assume the mean of one roll is 3.5 and the variance is $35/12$.

# Central Limit Theorem: Problem

- Suppose you are rolling a fair die 1000 times. Calculate the numbers $A$ and $B$ such that the average of the 1000 rolls is between $A$ and $B$ with probability approximately 0.95. You may assume the mean of one roll is 3.5 and the variance is 35/12.

$$Mean(X_i) = 3.5, \quad Var(X_i) = 35/12$$

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(3.5, \frac{35}{12000}\right) \quad \text{by CLT}$$

$$A = -2\sigma + \mu = -2\sqrt{35/12000} + 3.5 \approx 3.392$$

$$B = 2\sigma + \mu = 2\sqrt{35/12000} + 3.5 \approx 3.608$$

# Statistics

- We have finished the first half of the course on probability. Now, we move on to statistics.

# Statistics

- We have finished the first half of the course on probability. Now, we move on to statistics.
- Statistics is used to make inductive statements about some phenomenon (coin-flipping, dice rolling) **after** observing data.

# Statistics

- We have finished the first half of the course on probability. Now, we move on to statistics.
- Statistics is used to make inductive statements about some phenomenon (coin-flipping, dice rolling) **after** observing data.
- Three main activities of statistics:
  1. Estimating numerical values of a parameter or parameters.
  2. Assessing accuracy of these estimates.
  3. Testing hypotheses about the numerical values of parameters.

# Statistics

- We have finished the first half of the course on probability. Now, we move on to statistics.
- Statistics is used to make inductive statements about some phenomenon (coin-flipping, dice rolling) **after** observing data.
- Three main activities of statistics:
  1. Estimating numerical values of a parameter or parameters.
  2. Assessing accuracy of these estimates.
  3. Testing hypotheses about the numerical values of parameters.
- Example: Suppose flip a coin 1,000 times and observe 700 heads.
  1. How do I estimate the probability $\theta$ of achieving a head?
  2. How accurate is my estimate of $\theta$?
  3. Is this a fair coin ($\theta = 0.5$)?

# Estimation of a parameter: Binomial parameter $\theta$

- Recall a binomial random variable $X \sim Bin(n, \theta)$. How do we estimate the probability of success $\theta$?

# Estimation of a parameter: Binomial parameter $\theta$

- Recall a binomial random variable $X \sim Bin(n, \theta)$. How do we estimate the probability of success $\theta$?

- An intuitive estimator for $\theta$ is $p = x/n$, the **observed** proportion of successes.

# Estimation of a parameter: Binomial parameter $\theta$

- Recall a binomial random variable $X \sim Bin(n, \theta)$. How do we estimate the probability of success $\theta$?

- An intuitive estimator for $\theta$ is $p = x/n$, the **observed** proportion of successes.

- Consider the random variable $P$, the proportion of successes **prior** to performing the experiment.

# Estimation of a parameter: Binomial parameter $\theta$

- Recall a binomial random variable $X \sim Bin(n, \theta)$. How do we estimate the probability of success $\theta$?
- An intuitive estimator for $\theta$ is $p = x/n$, the **observed** proportion of successes.
- Consider the random variable $P$, the proportion of successes **prior** to performing the experiment.
  - $Mean(P) = \theta$ so $p$ is "shooting at the right target". $p$ is then referred to as an **unbiased** estimate of $\theta$.

# Estimation of a parameter: Binomial parameter $\theta$

- Recall a binomial random variable $X \sim Bin(n, \theta)$. How do we estimate the probability of success $\theta$?

- An intuitive estimator for $\theta$ is $p = x/n$, the **observed** proportion of successes.

- Consider the random variable $P$, the proportion of successes **prior** to performing the experiment.
  - $Mean(P) = \theta$ so $p$ is "shooting at the right target". $p$ is then referred to as an **unbiased** estimate of $\theta$.

- Difference between estimate and estimator:
  - **Estimate:** A function of the observed data used to estimate a given parameter. Ex: $p$.
  - **Estimator:** The random variable whose realization is the estimate. Ex: $P$.

# Estimation of a parameter: Binomial parameter $\theta$

- Recall a binomial random variable $X \sim Bin(n, \theta)$. How do we estimate the probability of success $\theta$?
- An intuitive estimator for $\theta$ is $p = x/n$, the **observed** proportion of successes.
- Consider the random variable $P$, the proportion of successes **prior** to performing the experiment.
  - $Mean(P) = \theta$ so $p$ is "shooting at the right target". $p$ is then referred to as an **unbiased** estimate of $\theta$.
- Difference between estimate and estimator:
  - **Estimate:** A function of the observed data used to estimate a given parameter. Ex: $p$.
  - **Estimator:** The random variable whose realization is the estimate. Ex: $P$.
- To investigate the precision of an estimate, we need to consider the random variable estimator.

# Precision of an estimate: Binomial parameter $\theta$

- Precision of $p$ as an estimate of $\theta$ depends on the variance of random variable $P$.

# Precision of an estimate: Binomial parameter $\theta$

- Precision of $p$ as an estimate of $\theta$ depends on the variance of random variable $P$.

- By the CLT, two standard deviation rule, and approximations (see pg. 40-41), we get the approximate 95% confidence interval for $\theta$ as

$$p \pm 2\sqrt{p(1-p)/n}$$

# Precision of an estimate: Binomial parameter $\theta$

- Precision of $p$ as an estimate of $\theta$ depends on the variance of random variable $P$.

- By the CLT, two standard deviation rule, and approximations (see pg. 40-41), we get the approximate 95% confidence interval for $\theta$ as

$$p \pm 2\sqrt{p(1-p)/n}$$

- We can further approximate the 95% confidence interval with $p(1-p) \leq 1/4$ to get

$$p \pm \sqrt{1/n} \tag{66}$$

# Precision of an estimate: Binomial parameter $\theta$

- Precision of $p$ as an estimate of $\theta$ depends on the variance of random variable $P$.

- By the CLT, two standard deviation rule, and approximations (see pg. 40-41), we get the approximate 95% confidence interval for $\theta$ as

$$p \pm 2\sqrt{p(1-p)/n}$$

- We can further approximate the 95% confidence interval with $p(1-p) \leq 1/4$ to get

$$p \pm \sqrt{1/n} \tag{66}$$

- Correspondingly, the 99% confidence interval is

$$p \pm 2.576\sqrt{p(1-p)/n} \approx p \pm 1.288\sqrt{1/n}$$

# Example

In the 2017-2018 NBA season, Lebron James shot 531 free throws and made 388. We want to estimate the probability $\theta$ that Lebron James makes a free throw.

# Example

In the 2017-2018 NBA season, Lebron James shot 531 free throws and made 388. We want to estimate the probability $\theta$ that Lebron James makes a free throw.

Q1: What is the estimate for $\theta$?

# Example

In the 2017-2018 NBA season, Lebron James shot 531 free throws and made 388. We want to estimate the probability $\theta$ that Lebron James makes a free throw.

Q1: What is the estimate for $\theta$?

$$p = x/n = 388/531 = 0.7307$$

# Example

In the 2017-2018 NBA season, Lebron James shot 531 free throws and made 388. We want to estimate the probability $\theta$ that Lebron James makes a free throw.

Q1: What is the estimate for $\theta$?

$$p = x/n = 388/531 = 0.7307$$

Q2: Calculate the 95% confidence interval for $\theta$ using the approximate 95% interval formula (66):

# Example

In the 2017-2018 NBA season, Lebron James shot 531 free throws and made 388. We want to estimate the probability $\theta$ that Lebron James makes a free throw.

Q1: What is the estimate for $\theta$?

$$p = x/n = 388/531 = 0.7307$$

Q2: Calculate the 95% confidence interval for $\theta$ using the approximate 95% interval formula (66):

$$p \pm \sqrt{1/n} = 0.7307 \pm \sqrt{1/531} = 0.7307 \pm 0.0434$$

# Example

In the 2017-2018 NBA season, Lebron James shot 531 free throws and made 388. We want to estimate the probability $\theta$ that Lebron James makes a free throw.

Q1: What is the estimate for $\theta$?

$$p = x/n = 388/531 = 0.7307$$

Q2: Calculate the 95% confidence interval for $\theta$ using the approximate 95% interval formula (66):

$$p \pm \sqrt{1/n} = 0.7307 \pm \sqrt{1/531} = 0.7307 \pm 0.0434$$

Q3: What is the sample size if we want the width of the confidence interval to be 0.02?

# Example

In the 2017-2018 NBA season, Lebron James shot 531 free throws and made 388. We want to estimate the probability $\theta$ that Lebron James makes a free throw.

Q1: What is the estimate for $\theta$?

$$p = x/n = 388/531 = 0.7307$$

Q2: Calculate the 95% confidence interval for $\theta$ using the approximate 95% interval formula (66):

$$p \pm \sqrt{1/n} = 0.7307 \pm \sqrt{1/531} = 0.7307 \pm 0.0434$$

Q3: What is the sample size if we want the width of the confidence interval to be 0.02?

$$\begin{aligned} \text{We want } \sqrt{1/n} &= 0.01 = 0.02/2. \\ 1/n &= 0.01^2 \\ n &= 1/0.01^2 = 10000 \end{aligned}$$