COLLEGE OF CHARLESTON

Week 7

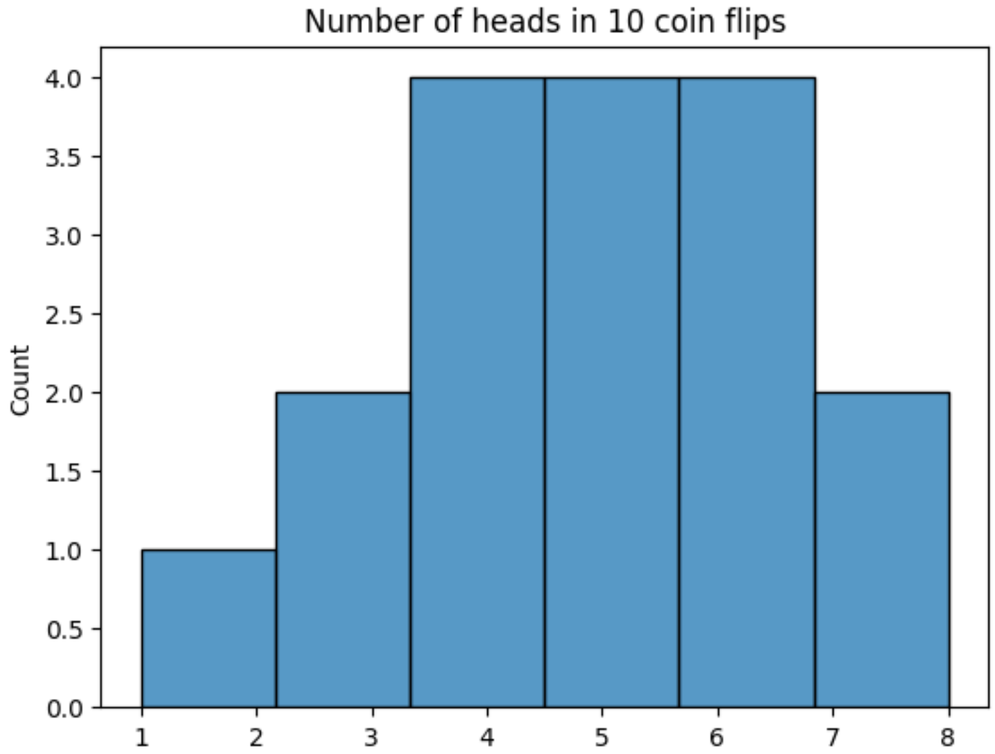
Math 104-03: Elementary Statistics

7.1 Sampling Distributions and the Central Limit theorem

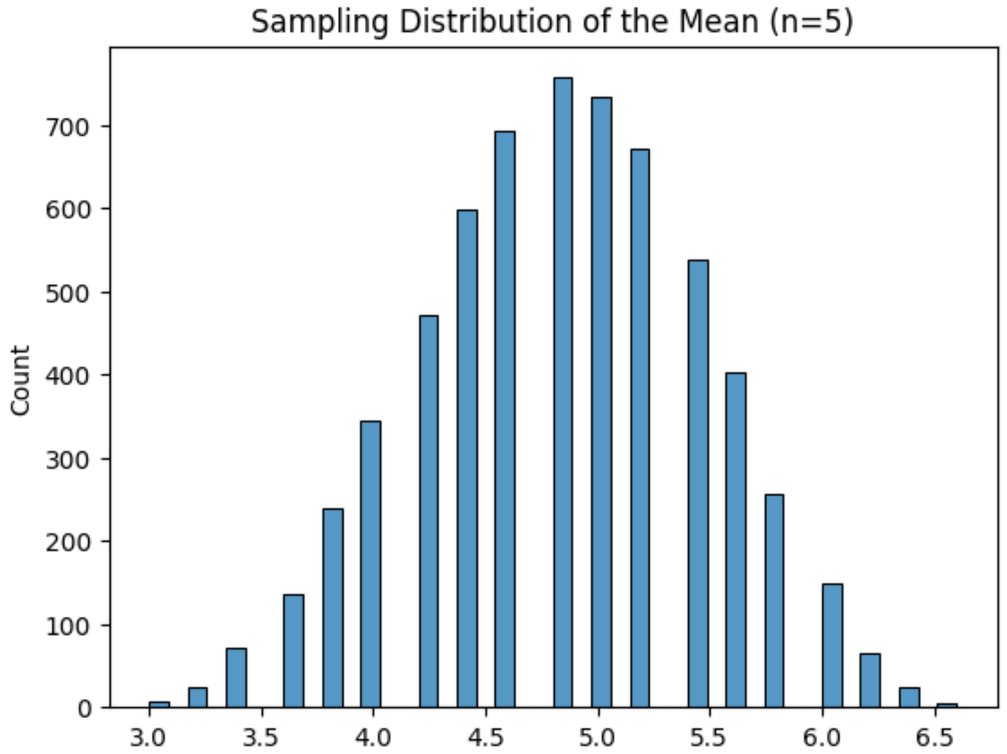
Sampling Distribution

- Problem: when we calculate a simple statistic from a sample of size n, it is possible that the sample statistic (e.g. sample mean) is a bad representation of a population parameter (e.g. population mean).
  - Suggests looking at a new distribution called a sampling distribution.
  - **Sampling distribution**: the distribution of values of a particular sample statistic for all possible samples of a given size n
  - **Sampling distribution of sample mean**: the distribution of sample means for all possible samples of a given size n

Example: Sampling Distribution

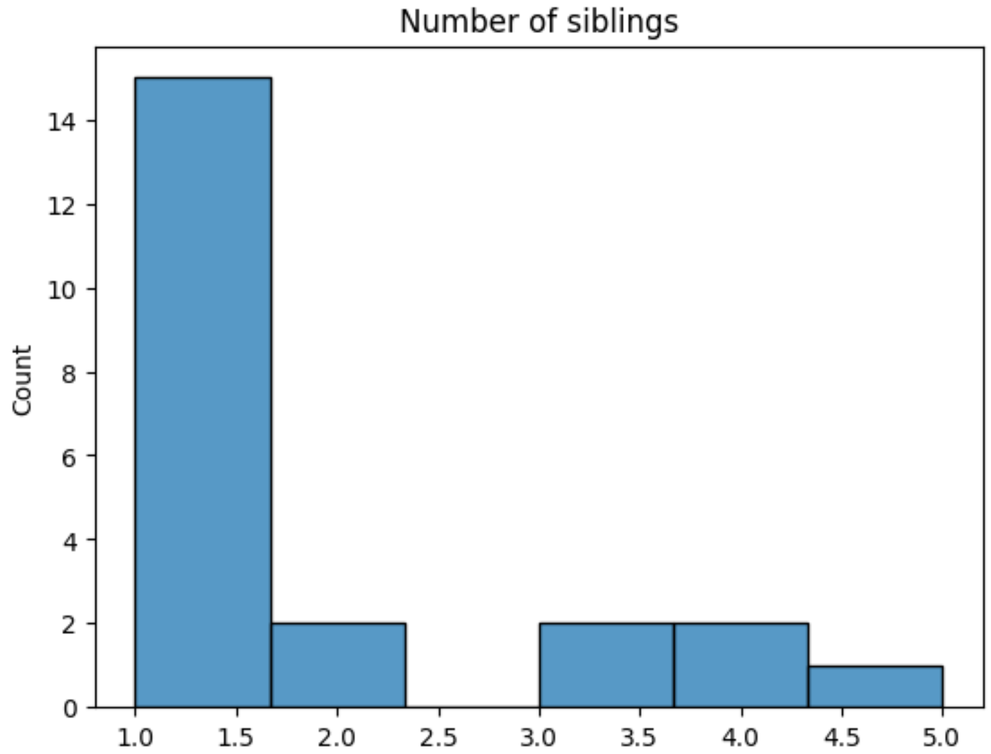


Mean 4.82, Standard Deviation: 1.617

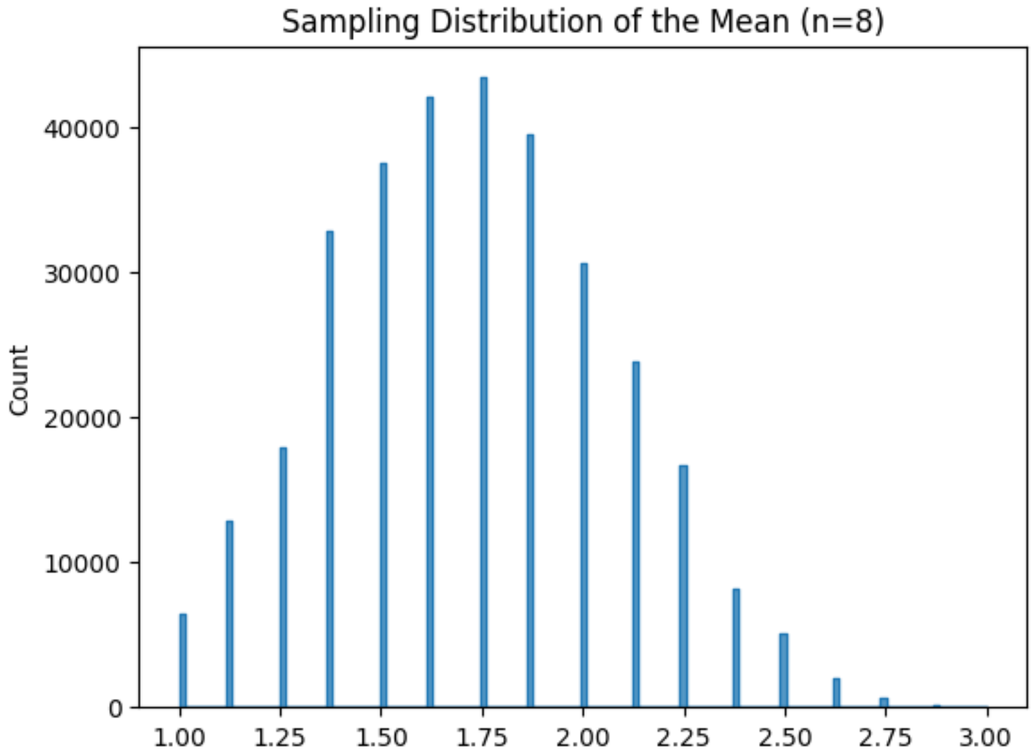


Mean: 4.82, Standard Deviation: 0.626

Example: Sampling Distribution



Mean: 1.72, Standard Deviation: 1.21



Mean: 1.72, Standard Deviation: 0.350

Sampling Distribution of Sample Means

- The mean of a sampling distribution of sample means is equal to the mean of the population
  - $\mu_{\bar{x}} = \mu$
  - Let's prove it.
- The standard deviation of a sampling distribution of sample means equals the standard deviation of the population divided by the square root of the sample size
  - $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$
  - Let's prove it!
- Example: suppose $n = 5$ and $\sigma = 1.61$. Then, $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{1.617}{\sqrt{5}} \approx 0.723$

Sampling Distribution of Sample Means

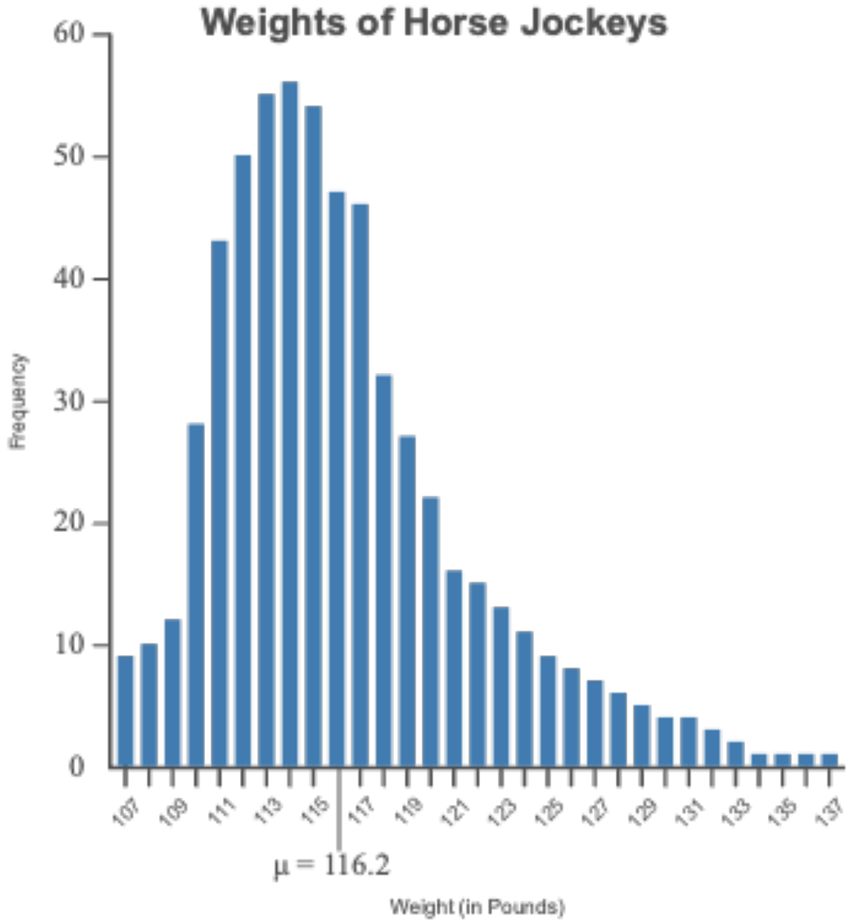
- The mean of a sampling distribution of sample means is equal to the mean of the population
  - $\mu_{\bar{x}} = \mu$
- The standard deviation of a sampling distribution of sample means equals the standard deviation of the population divided by the square root of the sample size
  - $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$
- **Example**: Suppose that movie ticket prices in one area have a mean of $8.32 and a standard deviation of $0.72. Ticket prices are recorded for samples of 52 theaters.
  - Calculate the mean of the resulting sampling distribution.
  - Calculate the standard deviation of the resulting sampling distribution.

The Central Limit Theorem (CLT)

- For any given population with mean μ and standard deviation σ, the shape of the sampling distribution of sample means will approach that of a normal distribution as the sample size increases.
  - We will consider a sample size to be large enough if $n \geq 30$
  - The larger the sample size, the better the normal distribution approximation will be
- The population does NOT need to be normally distributed
  - If the population is normally distributed, then we do not need n to be large

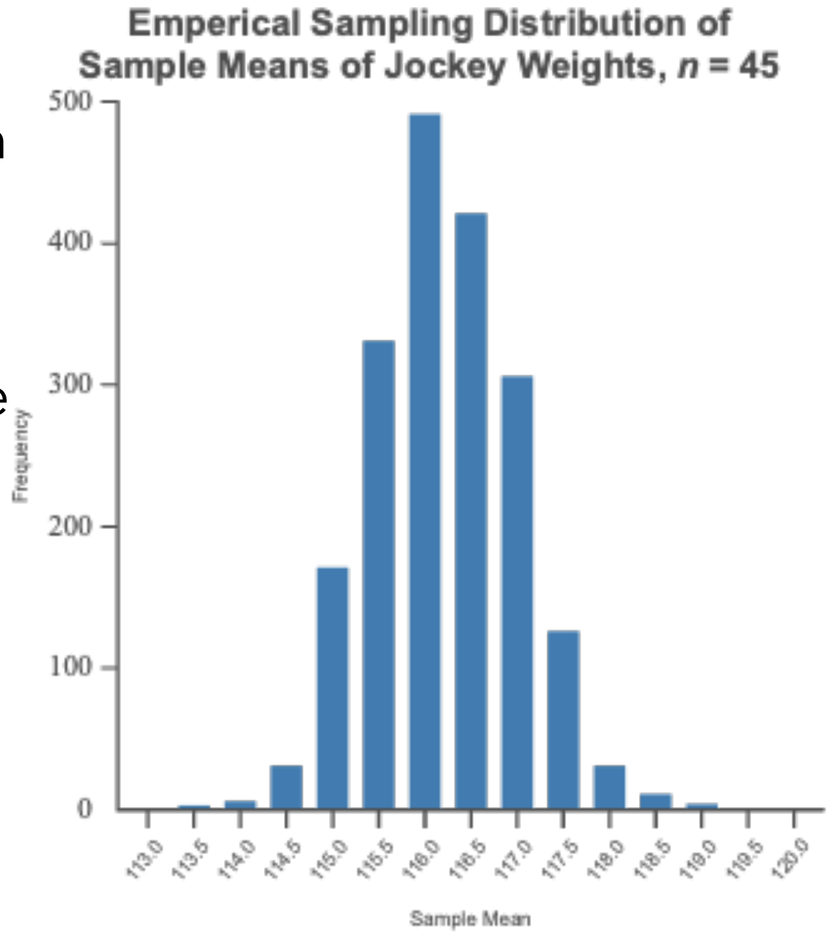
Example 1

- The following histogram represents the population distribution of the weights of 600 horse jockeys. The mean of the population is 116.2 pounds and the standard deviation is 3.9 pounds.
- Let's now consider the sampling distribution of sample means for samples of size n=45.
  - What is the sampling distribution's mean?
  - What is the sampling distribution's standard deviation?
  - Can a normal approximation be used for this sampling distribution?
  - What is the sampling distribution's shape?



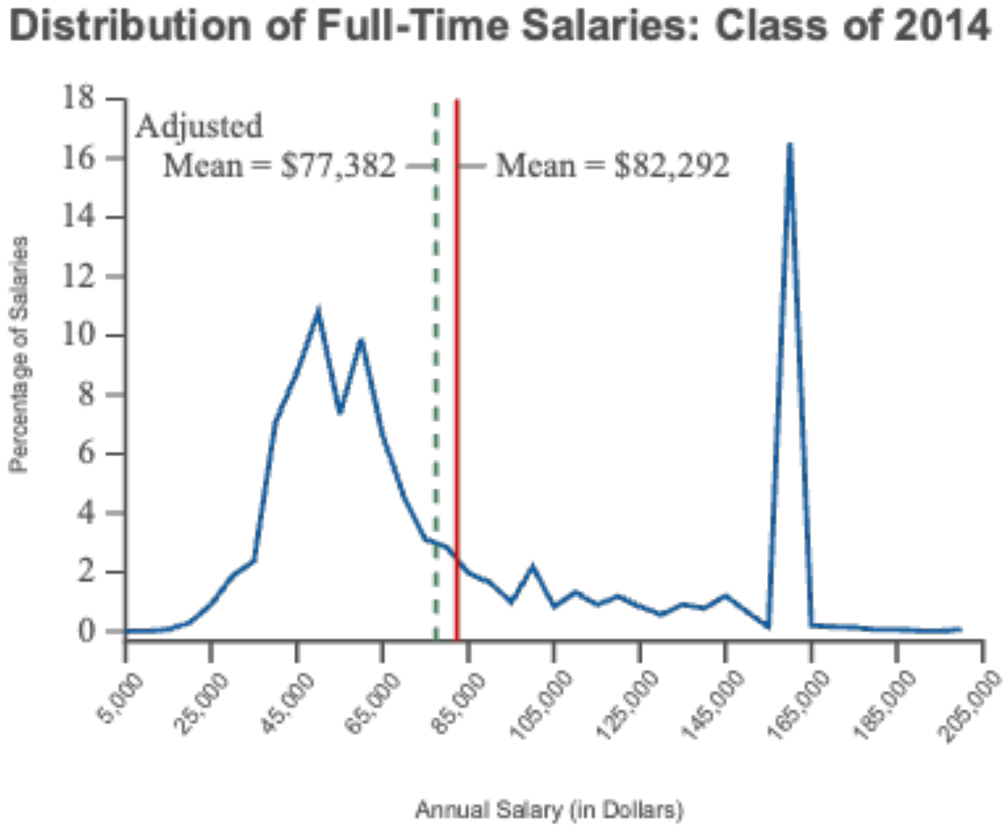
Example 1 (Answer)

- The sampling populations mean is equal to the mean of the population: $\mu_{\bar{x}} = \mu = 116.2$
- To find the standard deviation of the sampling distribution we will have to do a small calculation. We know that $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$. Hence, $\sigma_{\bar{x}} = \dfrac{3.9}{\sqrt{45}} \approx 0.58.$
- Yes, a normal approximation can be used because $n = 45$ satisfies the criteria $n \geq 30$
- Although the original distribution was skewed to the right, CLT tells us that the sampling distribution will be normal, so it will be bell-shaped



Example 2

- The following graph displays the starting salaries for law school graduates entering the workforce in 2014
  - Describe the shape of this distribution.
  - Consider the sampling distribution of the sample means created from this population for samples of size 25. Can a normal approximation be used for this sampling distribution?
  - Consider the sampling distribution of the sample means created from this population for samples of size 50. Can the Central Limit Theorem be applied in this situation?
  - Consider the sampling distribution of the sample means created from this population for samples of size 200. What would you expect the shape of this distribution to be? How would it compare to the sampling distribution described in part c.?



Example 2 (answer)

- The distribution is irregularly shaped, with a mode of approximately $160,000
- No, a normal approximation is not applicable in this situation because neither condition is met. The sample size is not large enough (25<30), and the population data are not normally distributed.
- Yes, the Central Limit Theorem is applicable here because the sample size is sufficiently large (50>30).
- We would expect the sampling distribution created from samples of size 200 to resemble a normal distribution. In fact, the larger the value of the sample size, n, the more "normal" the graph will appear. So, it follows that the sampling distribution described in part d. will more closely resemble a true normal distribution than the sampling distribution described in part c.

7.2 Central Limit Theorem with Means

Standard Score for a Sample Mean

- If either the sample size $n \geq 30$ or the population is normally distributed, then the standard score for a sample mean in a sampling distribution is given by

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

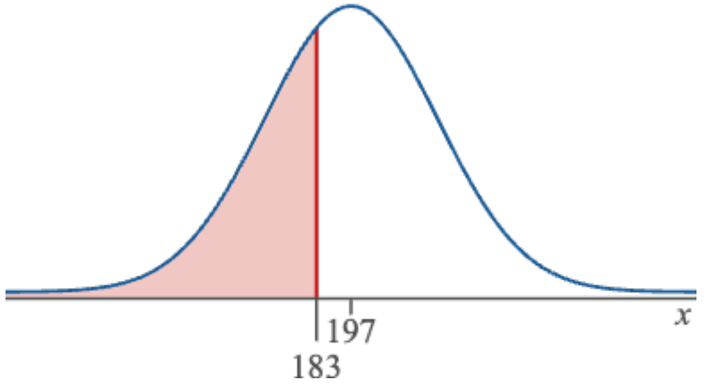
  - $\bar{x}$ is the sample mean
  - μ is the population mean
  - σ is the population standard deviation
  - n is the sample size used to create the sampling distribution

Example 3

- According to the Centers for Disease Control and Prevention (CDC), the mean total cholesterol level for persons 20 years of age and older in the United States from 2007–2010 was 197 mg⁄dL.
  - What is the probability that a randomly selected adult from the population will have a total cholesterol level less than 183 mg⁄dL? Use a standard deviation of 35 mg⁄dL and assume that the cholesterol levels for the population are normally distributed.
  - What is the probability that a randomly selected sample of 150 adults from the population will have a mean total cholesterol level of less than 183 mg⁄dL? Use a standard deviation of 35 mg⁄dL.

Example 3 (Answer)

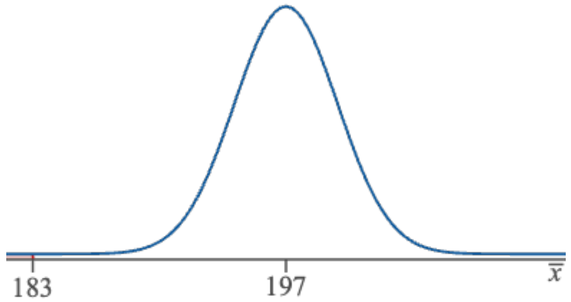
- We are looking to find the probability that an individual person selected at random has a cholesterol level less than 183. The graph of this area is shown below.



- First, we need to convert the value of 183 to a standard score. It is important to note that in the first part of the question we are referring to an individual and not a sample. Thus, $z = \frac{x - \mu}{\sigma} = \frac{183 - 197}{35} = -0.40$. Using the table, we find the area under the normal curve to the left of $z = -0.40$ is approximately 0.3446

Example 3 (Answer)

- The sketch of the normal curve will be constructed in a similar fashion as part a., with a mean of 197 and the area to the left of 183 shaded, except now we are considering the sampling distribution of sample means rather than the distribution of individual total cholesterol levels. Hence we now wish to find $\mathbb{P}(\bar{X} < 183)$.



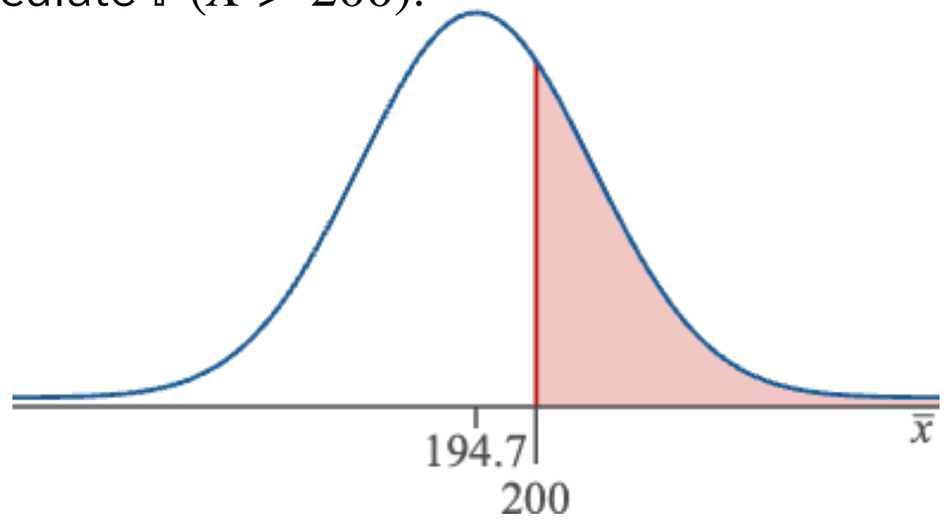
- We first need to convert the value of 183 into a standard score. Since we are referring to sample mean and not an individual, we use the following z-score formula:
  - $$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{183 - 197}{\frac{35}{\sqrt{150}}} \approx -4.90$$
- Using the normal distribution tables, we find that the area under the standard normal curve to the left of $z = -4.90$ is approximately 0.0000
- Thus, the probability of a randomly selected sample of 150 US adults having a mean total cholesterol level less than 183 mg/dL is approximately 0.0000, which means it is very unlikely, although not impossible.

Example 4

- Elevators are required to have weight capacities posted. If the weights of US men are normally distributed with a mean of 194.7 pounds and a standard deviation of 32.0 pounds, what is the probability that a random group of 10 men who enter an elevator will have a combined weight greater than the weight capacity of 2000" " pounds that is posted in that elevator?

Example 4 (Answer)

- Although the sample size is small $n = 10$ a normal approximation can be applied because the population is normally distributed. This problem is worded differently because the statistic is a sum not a mean. If there are 10 mean, and we are interested in a combined total of 2000 pounds, then the mean weight is $2000/10 = 200$ pounds.
- So, we can restate the problem as: "What is the probability that a randomly selected group of 10 men have a mean weight greater than 200 pounds?"
- Begin by drawing a normal curve with a mean of 194.7 and shade the area to the right of 200 to illustrate that we wish to calculate $\mathbb{P}(\bar{X} > 200)$.



Example 4 (Answer)

- We must first convert the value of 200 to a standard score for sample means.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{n}} = \frac{200 - 194.7}{\frac{32.0}{\sqrt{10}}} \approx 0.52$$

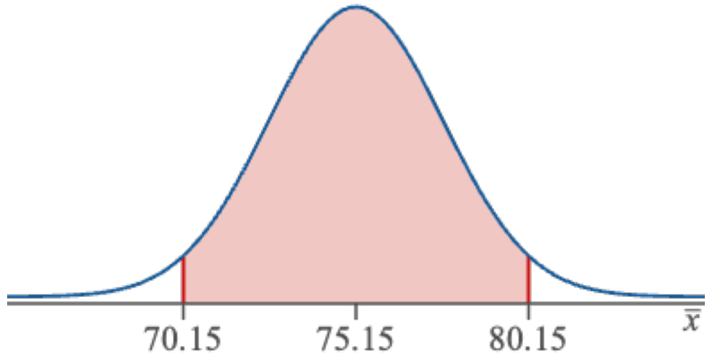
- Next, we use the standard normal distribution tables to find the area under the standard normal curve to the right of $z = 0.52$. The area is approximately 0.3015.
- This means that the probability that a random group of 10 men who enter the elevator will exceed the posted weight capacity is approximately 30%.

Example 5

- Suppose that prices of athletic shoes have a mean of $75.15 and a standard deviation of $17.89. What is the probability that the mean price of a random sample of 50 pairs of athletic shoes will differ from the population mean by less than $5.00?

Example 5 (Answer)

- By subtracting \$5.00 from and adding \$5.00 to the population mean, we find that the area under the normal curve in which we are interested is between \$70.15 and \$80.15, i.e., we wish to calculate $\mathbb{P}(70.15 < \bar{X} < 80.15)$



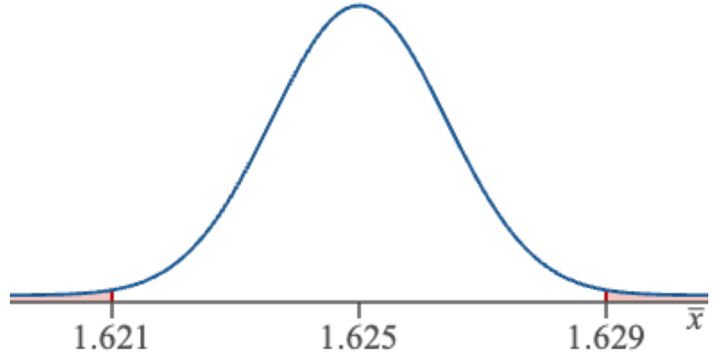
- We begin by converting the endpoints of the shaded region to z-scores. By symmetry, we will only need to find one z-score: $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{5.00}{\dfrac{17.89}{\sqrt{50}}} \approx 1.98$
- Thus, we need to find the area between the z-scores -1.98 and 1.98. Using the normal distribution tables, the area is 0.9522. This means that the probability of a random sample of 50 pairs of athletic shoes having a mean price that differs from the population mean by less than \$5.00 is approximately 95%.

Example 6

- At a local hardware manufacturing plant, the screws being manufactured have a mean length of 1.625 inches, with a standard deviation of 0.010 inches. If the quality control director randomly chooses a batch of 50 screws, what is the probability that their mean length differs from the mean of the population by more than the allowed 0.004 inches?

Example 6 (Answer)

- By subtracting 0.004 inches from and adding 0.004 inches to the mean of 1.625 inches, we find that the area under the normal curve in which we are interested is the sum of the areas to the left of 1.621inches and to the right of 1.629 inches.

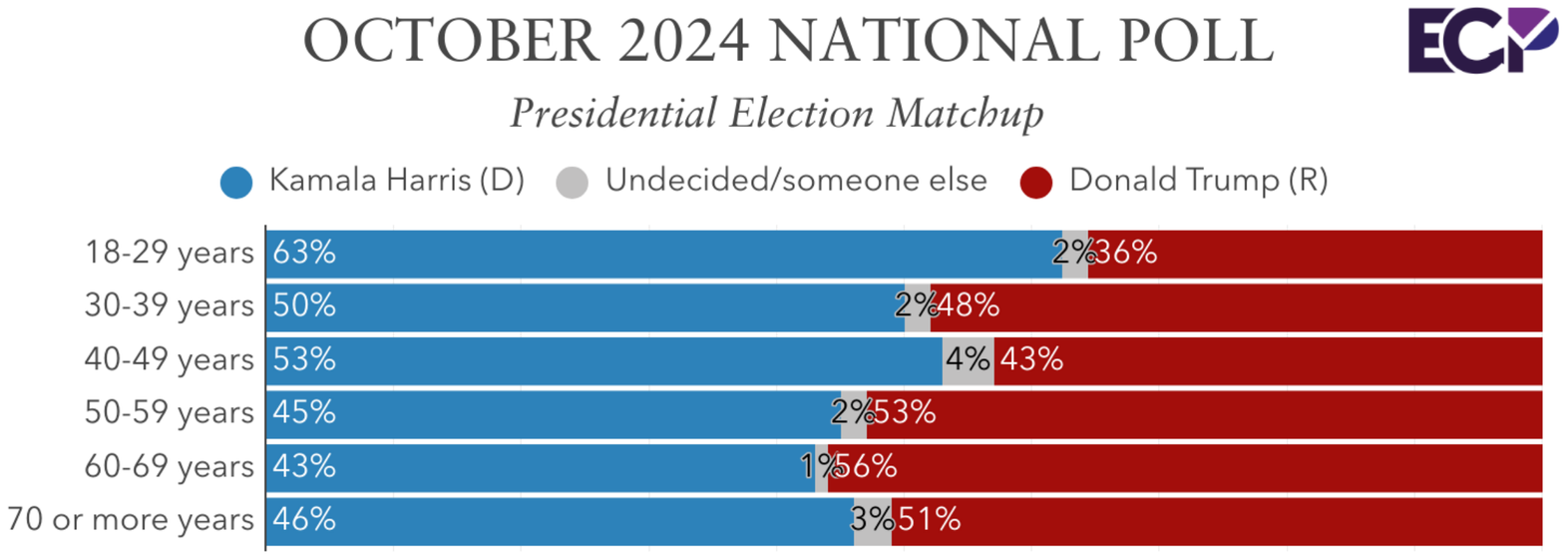


- As before we calculate the z-score: $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{0.004}{\dfrac{0.010}{\sqrt{50}}} \approx 2.83$
- Thus, the two z-scores are -2.83 and 2.83. Because we want to find the total area in the two tails, using the symmetry property of the normal distribution, we can find the area to the left of -2.83 and double it. Using the tables, we find that the area to the left of 2.83 is 0.0023, so the total area in the two tails is (0.0023)(2)=0.0046.
- This means that the probability of the sample batch not meeting the requirements is very small, approximately 0.5%.

7.3 Central Limit Theorem with Proportions

Proportion

- A **population proportion** is given by $p = \frac{x}{N}$ where x is the number of individuals in the population that have a certain characteristic and N is the size of the population.
- A **sample proportion** is given by $\hat{p} = \frac{x}{n}$ where x is the number of individuals in the sample that have a certain characteristic and p is the sample size.



Sampling Distribution of Sample Proportions

- Sampling distribution of sample proportions has a binomial distribution, not a normal distribution.
  - Can approximate binomial with a normal distribution!
- When samples taken are simple random samples, the conditions for a binomial distribution are met, and the sample size is large enough ($np \geq 10$ and $n(1-p) \geq 10$), a normal distribution can be used to approximate the binomial sampling distribution of sample proportions with mean and standard deviation
  - $\mu_{\hat{p}} = p; \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

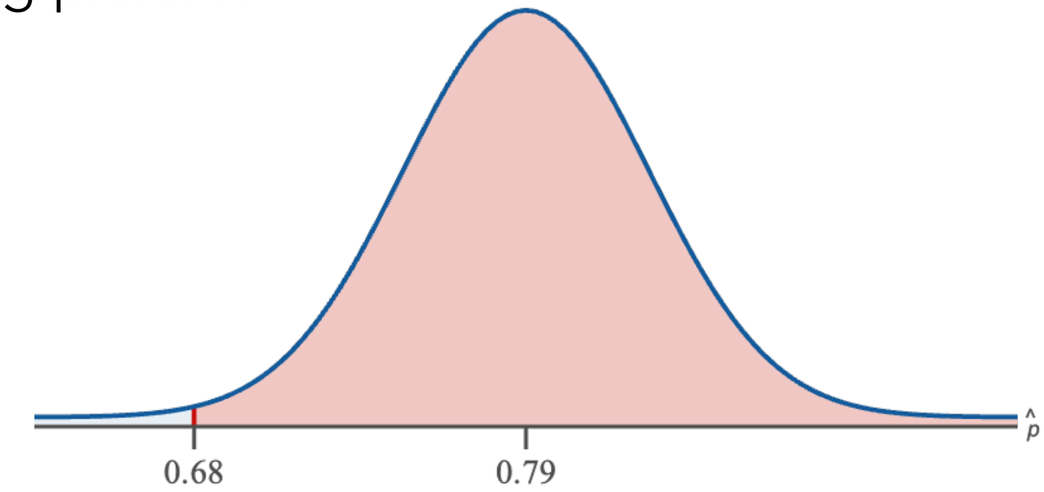
  where p is the population proportion and n is the population size.
- z-scores are: $z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p}{n}}}$

Example 1

- Question: In a certain conservative precinct, 71% of the voters are registered Republicans. What is the probability that in a random sample of 100 voters from this precinct, at least 68 of the voters would be registered Republicans?

Example 1 (answer)

- We first see that $\hat{p} = \frac{68}{100} = 0.68$. Because we're looking for the probability that at least 68% of voters in the sample are registered Republicans, we need to find the area under the normal curve of the sampling distribution to the right of 68%, which is denoted $\mathbb{P}(\hat{p} > 0.68)$. Since this value is smaller than the population proportion of 79%, we can mark a value to the left of the center, and shade from there to the right, as shown in the following picture.



Example 1 (answer)

- First we must calculate the z-score for sample proportions. To do this, we need to use $\hat{p} = 0.68, p = 0.68, n = 100$ to calculate z as follows:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.68 - 0.79}{\sqrt{\frac{0.79(1-0.79)}{100}}} \approx -2.70$$

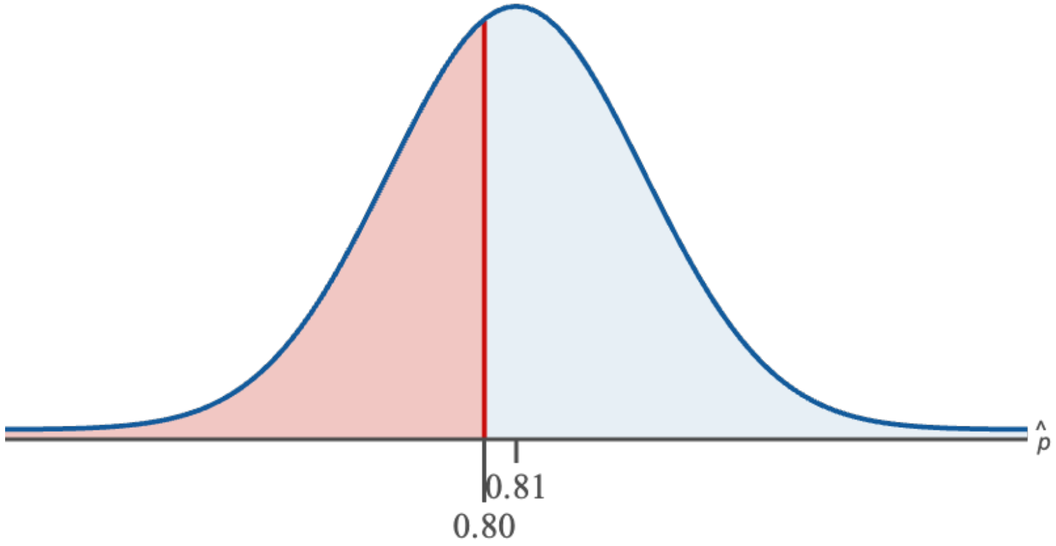
- Using the normal tables, we find that the area under the standard normal to the right of z = -2.70 us 0.9965
- Thus, the probability of at least 68 voters in a sample of 100 being registered Republicans is very high, approximately 99.65%

Example 2

- In another precinct across town, the population is very different. In this precinct, 81% of the voters are registered Democrats. What is the probability that, in a random sample of 100 voters from this precinct, no more than 80 of the voters would be registered Democrats?

Example 2 (answer)

- From the information given, we see that $\hat{p} = 80/100 = 0.8$. Now let's sketch the normal curve. This time we're interested in the probability that no more than 80% of voters in the sample are registered Democrats, so we need to find the area under the normal curve of the sampling distribution to the left of 80%, denoted $\mathbb{P}(\hat{p} < 0.8)$. Since this value is smaller than the population proportion of 81%, we can mark a value to the left of the center, and shade from there to the left, as shown in the following picture.



Example 2 (answer)

- We first need to calculate the z-score. To do this, we need to use $\hat{p} = 0.80, p = 0.81,$ and $n = 100$ to calculate z as follows

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.8 - 0.81}{\sqrt{\frac{0.81(1-0.81)}{100}}} \approx -0.25$$

- Using the normal tables, we find that the area under the standard normal curve to the left of z = -0.25 is 0.4013. Hence, the probability of no more than 80 voters in a randomly selected sample of 100 voters being registered Democrats is approximately 40%.

“How to Lie with Statistics”

Discussion Question

- Please answer the following question on OAKS:
  - Darrell Huff wrote How to Lie with Statistics in 1954. It is arguable that the amount of misinformation is even greater nowadays. Here is your assignment: 1) pick a specific technique that Huff identifies that can be used to lie with statistics, 2) briefly explain the technique, and 3) describe a modern-day example of that technique being used.
- A hypothetical/made-up example is fine, but compelling real-life examples will be featured in class.
- Responses should be around 100 words.