Note # 6: Sampling Distributions and Statistical Inference

Problem 1. Suppose you are interested in knowing the average cholesterol level of all women between the ages of 21 and 30 who live in College Station. In this context, what type of distribution is described in each of the scenarios below?

a. You take a random sample of 50 females in the College Station area between the ages of 21and 30. You collect the cholesterol level of these 50 women and make a histogram of these 50 values.

b. You collect the cholesterol level of every 21-30 year old female in the College Station area and create a histogram of all of these values.

c. Everyone in the class (60 students) takes a random sample of 50 females in the College Station between the ages of 21 and 30. Each person calculates the average cholesterol level in their sample and we make a histogram of the 60 average cholesterol levels.

Solution:

a. Data distribution.

b. Population distribution.

c. Sampling distribution.

Problem 2. As part of a quality control process for computer chips, an engineer at a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

a. What population is under consideration in the data set?

b. What parameter is being estimated?

c. What is the point estimate for the parameter?

d. What is the name of the statistic can we use to measure the uncertainty of the point estimate?

e. Compute the value from part (**d**) for this context.

f. The historical rate of defects is 10%. Should the engineer be surprised by the observed rate of defects during the current week?

g. Suppose the true population value was found to be 10%. If we use this proportion to recompute the value in part (**e**) using $p = 0.1$ instead of $\hat{p}$, does the resulting value change much?

Solution:

a. All computer chips produced at this factory this particular week.

b. Estimate $p =$ proportion of computer chips manufactured at this factory this particular week that had defects.

c. Point estimate, $\hat{p}$, proportion in the sample $\hat{\boldsymbol{p}} = \frac{\mathbf{27}}{\mathbf{212}} = \mathbf{0.127}$.

d. Standard error (SE) → standard deviation of the sampling distribution.

e. $SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.127)(1-0.127)}{212}} = \mathbf{0.023}$

f. $\hat{p} = 0.127$ and $p = 0.10$ → difference: $\hat{p} - p = 0.127 - 0.10 = \mathbf{0.027}$. Our sample value is a little over 1 SE away from the true parameter, so this really is not too surprising.

g. $SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.10)(1-0.10)}{212}} = \mathbf{0.021}$. Using $\hat{p}$ to estimate p, $SE \approx \mathbf{0.023}$. There is not much difference at all.

Problem 3. In a random sample of 765 adults in the United States, 322 say they could not cover a $400 unexpected expense without borrowing money or going into debt.

a. What population is under consideration in the data set?

b. What parameter is being estimated?

c. What is the point estimate for the parameter?

d. What is the name of the statistic can we use to measure the uncertainty of the point estimate?

e. Compute the value from part (**d**) for this context.

f. A cable news expert thinks the value is actually 50%. Should she be surprised by the data?

g. Suppose the true population value was found to be 40%. If we use this proportion to recompute the value in part (**e**) using $p = 0.4$ instead of $\hat{p}$, does the resulting value change much?

Solution:

a. All adults in the United States.

b. Estimate p=proportion of all adults in the US who could not cover a $400 unexpected expense with out barrowing money or going into debt.

c. Point estimate, $\hat{p}$, sample proportion $\boldsymbol{\hat{p} = \frac{322}{765} = 0.421}$

d. Standard error (SE) → Standard deviation of the sampling distribution.

e. $SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.421)(1-0.421)}{765}} = \mathbf{0.0179}$

f. $\hat{p} = 0.421$ and $p = 0.5$ → The difference: $p - \hat{p} = 0.5 - 0.421 = \mathbf{0.079}$ is more than **4** SE away from the true parameter, so this would be surprising.

g. $SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.40)(1-0.40)}{765}} = \mathbf{0.0177}$. Using $\hat{p}$ to estimate p, $SE \approx \mathbf{0.0179}$, not very different.

Problem 4. The American Heart Association reports that 4.9% of adolescents aged 12-17 are current smokers. Suppose you take a random sample of 32 adolescents between the ages of 12 and 17 and find the percent of them who are current smokers is 3.2%?

a. What is the value of the parameter?

b. What is the value of the statistic?

c. Describe the sampling distribution of $\hat{p}$.

Solution:

a. $p = \mathbf{0.049}$

b. $\hat{p} = \mathbf{0.032}$

c. Shape: $np = (32)(0.049) = \mathbf{1.568} < \mathbf{10}$ ×

$$n(1-p) = (32)(1-0.049) = \mathbf{30.432} \geq \mathbf{10}\ \surd$$

Shape is not Normal (will be skewed to the right)

Mean: $p = \mathbf{0.049}$ and Stdev: $= \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.049)(1-0.049)}{32}} = \mathbf{0.038}$

Problem 5. The distribution of weights of United States pennies is approximately normal with a mean of 2.5 grams and a standard deviation of 0.03 grams.

a. What is the probability that a randomly chosen penny weighs less than 2.48 grams?

b. Describe the sampling distribution of the mean weight of 10 randomly chosen pennies.

c. What is the probability that the mean weight of 10 pennies is less than 2.48 grams?

Solution:

a. $X = weight\ of\ a\ single\ penny$

$X \sim N(\mu = 2.5\ grams, \sigma = 0.03\ grams)$

$$P(X < 2.48) = P\left(Z < \frac{2.48 - 2.5}{0.03}\right) = P\left(Z < \frac{-0.02}{0.03}\right) = P(Z < -0.67) = \mathbf{0.2514}$$

b. $\bar{x} = mean\ weight\ of\ 10\ pennies$

Shape: approximately normal (because population distribution is approximately normal)

Mean: $\mu = 2.5$ and Stdev: $= \frac{\sigma}{\sqrt{n}} = \frac{0.03}{\sqrt{10}} = 0.009487$

$\boldsymbol{\bar{x} \sim N(\mu = 2.5\ grams, \sigma = 0.009487\ grams)}$

c. $P(\bar{x} < 2.48) = P\left(Z < \frac{2.48-2.5}{0.009487}\right) = P\left(Z < \frac{-0.02}{0.009487}\right) = P(Z < -2.11) = \mathbf{0.0174}$

Problem 6. Fourth-grade children are on average 54.5 inches tall with a standard deviation of 2.7 inches. Consider you took a sample of 36 students in a fourth-grade classroom and determined the average height of these students was 52.7 inches, with a standard deviation of 1.5 inches.

a. What is the mean of the population distribution?

b. What is the standard deviation of the population's distribution?

c. What is the mean of the data distribution?

d. What is the standard deviation of the data distribution?

e. What is the mean of the sampling distribution?

f. What is the standard deviation of the sampling distribution?

Solution:

a. $\mu = \mathbf{54.5}$

b. $\sigma = \mathbf{2.7}$

c. $\bar{x} = \mathbf{52.7\ inches}$

d. $s = \mathbf{1.5\ inches}$

e. $Mean = \mu = \mathbf{54.5}$

f. Stdev= SE $= \frac{\sigma}{\sqrt{n}} = \frac{2.7}{\sqrt{36}} = \mathbf{0.45}$

Problem 7. For a given confidence interval, what would happen to the width if we increased the sample size and decreased the confidence level?

a. It would increase.

b. It would decrease.

c. It would stay the same.

d. It is impossible to determine.

Solution:

b. It would decrease.

$$\boldsymbol{CI}: Point\ estimate\ \pm Margin\ of\ Error \begin{cases} Z^*(critical\ value) \\ Std\ error\ of\ Sampling\ Distribution \end{cases}$$

$$CL\ \uparrow,\ Z^*\ \uparrow, ME\ \uparrow, Width\ \uparrow$$

$$\boldsymbol{CL\ \downarrow, Z^*\ \downarrow, ME\ \downarrow, Width\ \downarrow}$$
$$\boldsymbol{n\ \uparrow, SE\ \downarrow, ME\ \downarrow, Width\ \downarrow}$$

$$n\ \downarrow, SE\ \uparrow, ME\ \uparrow, Width\ \uparrow$$

Problem 8. A random sample of 50 marriage records in Contra Costa County in California yields a 95% confidence interval of 21.5 to 23.0 years of age for the average age at first marriage for women. Which of the following is the correct interpretation of this interval?

- **a.** If random samples of 50 records were repeatedly selected, then 95% of the time, the sample mean age at the first marriage for women would be between 21.5 and 23.0 years.
- **b.** 95% of the ages at first marriage for women in Contra Costa County are between 21.5 and 23.0 years.
- **c.** We can be 95% confident the sample mean is between 21.5 and 23.0 years.
- **d.** If we repeatedly sampled the entire population, then 95% of the time the population means would be between 21.5 and 23.0 years.
- **e.** None of the above.

Solution:

None of the above. The correct interpretation is: We are 95% confident that the average age of first marriage for all women in Contra Costa County in California is between 21.5 and 23.0 years.

Problem 9. A Gallop poll wants to determine the proportion of people who plan on voting for neither the republican nor the democratic candidate in the upcoming election (those who will vote for the third party). They conduct a random phone poll, where they contact 827 individuals and ask them whether or not they plan on voting for a third-party candidate. Of these 827 respondents, 33 people say they plan on voting for a third-party candidate. **Create a 99% confidence interval for the true proportion of all people who plan on voting third party.**

a. What is the 99% confidence interval?

b. What is the correct interpretation of the confidence interval?

c. Are the assumptions met? Explain.

Solution:

a. $\because \hat{p} \pm Z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, $\hat{p} = \frac{33}{827} = 0.0399$, $n = 827$, $Z^* = 2.576$

$$\therefore 0.0399 \pm 2.576 \sqrt{\frac{(0.0399)(0.9601)}{827}} = 0.0399 \pm (2.576)(0.006806) = 0.0399 \pm 0.01753$$

$$\mathbf{99\%\ CI: (0.02237, 0.05743)}$$

b. We are 99% confident that the true proportion of all Americans who are planning to vote 3rd party is between 0.02237 and 0.05743.

c.

i. **Independence:**

- Random
- $n < 10\%\ of\ the\ population \rightarrow \mathbf{n = 827}$

ii. **Sample size:**

$n\hat{p} \geq 10 \rightarrow (827)(0.0399) = \mathbf{33 \geq 10}$

$n(1-\hat{p}) \geq 10 \rightarrow (827)(0.9601) = \mathbf{794 \geq 10}$

Problem 10. A Gallop poll wants to determine the proportion of people who plan on voting for neither the republican nor the democratic candidate in the upcoming election (those who will vote third party). They conduct a random phone poll, where they contact 827 individuals and ask them whether or not they plan on voting for a third-party candidate. Of these 827 respondents, 33 people say they plan on voting for a third-party candidate. **Gallop wants to determine whether or not the data supports the idea that more than 3% of people plan on voting third party. Conduct a hypothesis test at the 0.10 significance level to test this claim.**

a. What are the hypotheses?
b. What is the significance level?
c. What is the value of the test statistic?
d. What is the p-value?
e. What is the correct decision?
f. What is the appropriate conclusion/interpretation?
g. Are the assumptions met? Explain.

Solution:

a. $H_0: p = 0.03$ VS. $H_A: p > 0.03$

b. $\alpha = 0.10$

c. $TS = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.0399 - 0.03}{\sqrt{\frac{(0.03)(0.97)}{827}}} = \frac{0.0099}{0.005932} = \mathbf{1.67}$

d. $p - value = P(get\ our\ results\ or\ more\ extereme \mid H_0\ is\ true)$

$p - value = P(\hat{p} > 0.0399 \mid p = 0.03) = P(Z > 1.67) = 1 - P(Z < 1.67) = 1 - 0.9525 = \mathbf{0.0475}$

e. $\because 0.0475 < \alpha = 0.10, \therefore$ Reject H_0

f. The data does provide statistically significant evidence that the true proportion of all Americans who are planning to vote to the 3rd party is greater than 0.03.

g.

i. **Independence:**
- Random
- $n < 10\%\ of\ the\ population \rightarrow \boldsymbol{n = 827}$

ii. **Sample size:**

$np_0 \geq 10 \rightarrow (827)(0.03) = \mathbf{24.81 \geq 10}$

$n(1 - p_0) \geq 10 \rightarrow (827)(0.97) = \mathbf{802.19 \geq 10}$

Problem 11. The National Center for Health Statistics reports that the average systolic blood pressure for males 35-44 years of age has a mean of 122. The medical director of a large company believes the average systolic blood pressure for male executives 35-44 years of age at his company is different from 122. He looks at the medical records of 81 randomly selected male executives in this age group and finds that the mean systolic blood pressure in this sample is $\bar{x} = 128.4$ and the standard deviation is 30. **Create a 95% confidence interval for the true average systolic blood pressure.**

a. What is the 95% confidence interval?

b. What is the correct interpretation of the confidence interval?

c. Are the assumptions met? Explain.

Solution:

a. $\because \bar{x} \pm Z^* \frac{s}{\sqrt{n}}$, $\bar{x} = 128.4$, $s = 30,\ n = 81$, $Z^* = 1.96$

$$\therefore 128.4 \pm 1.96\left(\frac{30}{\sqrt{81}}\right) = 128.4 \pm (1.96)(3.3333) = 128.4 \pm 6.5333$$

$$\mathbf{95\%\ CI}: (\mathbf{121.8667}, \mathbf{134.9333})$$

b. We are 95% confident that the true average systolic blood pressure for all 35- 44 year-old male executives at this company is between 121.8667 and 134.9333.

c.

i. **Independence:**
- Random
- $n < 10\%\ of\ the\ population$ → we do not know.

ii. Population Distribution is approximately normal, we don't know, but since n is large ($81 > 30$), so our sampling distribution should be approximately normal.

Problem 12. The National Center for Health Statistics reports that the average systolic blood pressure for males 35-44 years of age has a mean of 122. The medical director of a large company believes the average systolic blood pressure for male executives 35-44 years of age at his company is different from 122. He looks at the medical records of 81 randomly selected male executives in this age group and finds that the mean systolic blood pressure in this sample is $\bar{x} = 128.4$ and the standard deviation is 30. **Conduct a hypothesis test at the 0.05 significance level to test the medical director's claim.**

a. What are the hypotheses?
b. What is the significance level?
c. What is the value of the test statistic?
d. What is the p-value?
e. What is the correct decision?
f. What is the appropriate conclusion/interpretation?
g. Are the assumptions met? Explain.

Solution:

a. $H_0: \mu = 122$ VS. $H_A: \mu \neq 122$

b. $\alpha = 0.05$

c. $TS = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{128.4-122}{30/\sqrt{81}} = \frac{6.4}{3.333} = \mathbf{1.92}$

d. $p - value = P(get\ our\ results\ or\ more\ extereme \mid H_0\ is\ true)$

$p - value = P(\bar{x} > 128.4\ or\ \bar{x} < 115.6 \mid \mu = 122) = P(Z < -1.92) + P(Z > 1.92) = 0.0274 + 0.0274 = (2)(0.274) = 0.0548.$ Or simply: $= 2 \times P(Z < -1.92) = 0.0548.$

e. $\because 0.0548 > \alpha = 0.05, \therefore$ Fail to reject H_0

f. The data **does not** provide statistically significant evidence that the true average systolic blood pressure for all 35- 44 year-old male executives at this company is different from 122.

g.

i. **Independence:**
- Random
- $n < 10\%\ of\ the\ population \rightarrow \boldsymbol{n = 827}$

ii. Population Distribution is approximately normal, we don't know, but since n is large $(81 > 30)$, so our sampling distribution should be approximately normal.