Name: Connor Raymond Stewart
ID: 101041125
Lab Section: STAT 2507-H4

Carleton University

School of Mathematics and Statistics

STAT 2507: Introduction to Statistical Modeling I

Assignment 4

INSTRUCTIONS:

I You must print this assignment and write your answers in the space provided by either printing this assignment or writing directly on this PDF with a tablet. If you print the assignment, you must use A4 or letter size paper. DO NOT CHANGE THE EXISTING SPACING OF THE QUESTION.

II Assignments are to be uploaded to the course website on cuLearn as a single PDF file by **Wednesday, April 1 at 8:35am**. No late assignments will be accepted. No other file types will be accepted. Technical issues are not an excuse so don't wait until the deadline to submit.

III The document should in the proper orientation so as not to require rotation. The document must be of sufficient resolution and writing must be legible. Pictures of your work are NOT an acceptable substitute for a scan.

IV You must show and explain all of your work. This includes explicitly defining random variables where necessary, writing out any formula that you use, and explaining your reasoning where applicable. No credit will be given for answers without justification.

V This assignment is intended to represent your individual knowledge. It is not a group assignment.

VI When you save your PDF file, save it with the format: **LastName.StudentNumber.A4.pdf** .

VII Failure to follow these instructions will result in a grade of zero.

Question 1. Data collected over a long period of time showed that 1 in 1000 high school students like mathematics. A random sample of 30,000 high school students was surveyed. Let X be the number of students in the sample who like mathematics

a) What is the probability distribution of X?

b) What distribution can be used to approximate the distribution of X? Explain.

c) Find the approximate probability of observing a value of X equal to 40 or more?

d) Find the approximate probability of observing a value of X between 35 and 40 inclusive ?

P (Probability of Success) $= \frac{1}{1000} = 0.001$, $n = 30\,000$

a) The distribution is binomial. This is because :

- The experiment has two possible outcomes, that a Student likes math or dosn't
- The experiment is repeated 30 000 times

Thus, X follows a $\boxed{\text{binomial distributions}}$

b) We Can Approximate P using the $\boxed{\text{Normal Approximation}}$ to the binomial probability distribution!

For the Normal Approximation to be Valid: $\left(Z = \frac{X-\mu}{\sigma}\right)$

$$\mu = np = (30\,000)(0.001) = 30 > 5$$

$$\sigma = \sqrt{npq} = \sqrt{(30000)(0.001)(1-0.001)} = \sqrt{29.97} \doteq 5.47449$$

Where :

$$np = 30 > 5 \;\; \& \;\; nq = (30\,000)(1-0.001) = 29970 > 5$$

thus, the Normal dist. Can Approximate this.

c) $X = 40$, $\mu = 30$, $\sigma = \sqrt{29.97}$

$P(X \geq 40) = 1 - P(X \geq 40)$

$\rightarrow Z = \frac{40-30}{\sqrt{29.97}} \doteq 1.8267 \doteq 1.83$ $= 1 - P(Z < 1.83) = 1 - 0.9664 = \boxed{0.0339}$

thus, the Approximate probability for $P(X \geq 40)$ is 0.0339.

d) $35 \leq X \leq 40$, $\mu = 30$, $\sigma = \sqrt{29.97}$

$$P(35 \leq X \leq 40) = P(X \leq 40) - P(X \leq 35)$$

$$= P\left(X \leq \frac{40-30}{\sqrt{29.97}}\right) - P\left(X \leq \frac{35-30}{\sqrt{29.97}}\right) \doteq P(X \leq 1.83) - P(X \leq 0.91)$$

$$= 0.9664 - 0.8186 = \boxed{0.1478}$$

thus, the Approximate Probability for $P(35 \leq X \leq 40)$ is 0.1478.

Question 2. Based on a survey, workers in Ontario earn an average of \$60,000 per year with a known standard deviation of \$6000. In an attempt to verify this salary level, a random sample of 36 workers in Ontario was selected. Let $\overline{X}$ represent the mean salary of these 36 workers.

a) Describe the sampling distribution of $\overline{X}$.

× b) Calculate the probability that $\overline{X}$ is between 58,500 and 63,000.

c) What is the 90th percentile for $\bar{X}$.

$\mu = 60\,000,\ \sigma = 6000,\ n = 36$

a) According to the Central limit theorm, the Sampling dist. is Approx. normal if the Sample's Sufficiently large (normally when $n \geq 30$)

$n = 36 > 30$ ∴ the mean's approximatly normal by the central limit theorm

the mean of the Sampling dist. is Approx. the population mean:

$$\mu_{\bar{X}} = \mu = \boxed{60\,000}$$

The Std. dev. of the Sampling dist. is:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{6000}{\sqrt{36}} = \frac{6000}{6} = \boxed{1000}$$

∴ the Sampling dist. is Approx. normal w/ a mean of 60 000 & a Std. dev. of 1000

b) We must find the Z-score:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \text{ for } 58500 \leq \bar{X} \leq 63000$$

$$Z_1 = \frac{58500 - 60\,000}{1000} = -1.5$$

$$Z_2 = \frac{63000 - 60000}{1000} = \frac{3000}{1000} = 3$$

So: $P(58500 \leq \bar{X} \leq 63000) = P(-1.5 \leq Z \leq 3) = P(Z \leq 3) - P(Z \leq -1.5)$

$= 0.9987 - 0.0668 = \boxed{0.9319}$

So, the probability of $\bar{X}$ being between 58500 & 63000 is 0.9319.

c) We must first find the Z-score that results in the 90th Percentile:

$P(Z \leq Z_0) = 0.9$ So $Z_0 \doteq 1.28$

Since $Z_0 = \frac{\bar{X}_0 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$, we know $Z_0 \sigma_{\bar{X}} + \mu_{\bar{X}} = \bar{X}_0$

↳ $(1.28)(1000) + 60\,000 = \boxed{61280}$

So, the 90th Percentile value for $\bar{X}$ is $\boxed{61280}$

Question 3. Based on a random sample of 100 customers' responses, the average time to complete an online order at Amazon is 5 minutes. It is also known the population standard deviation of the time to complete an online order at Amazon is 2 minutes.

a) Construct an 90% confidence interval for the average time μ to complete an online order at Amazon and interpret it. Comment on the validity of the interval.

b) If the confidence level is changed to 80% in a), what z-value will you use for the 80% confidence interval.

c) What sample size is necessary to estimate true average time μ to within 0.17 minute with 99% confidence?

$n=100$, $\mu=5$, $\sigma=2$ ($n>30$ So $S\approx\sigma$)

a) for Confidence interval $1-\alpha=0.9$, find $Z_{a/2}=Z_{0.05}$

find 0.05 in the Z-Score table & reverse the Z-Score's Sign

↳ -1.645 Corresponds to 0.05 thus $Z_{a/2}=1.645$

$$E=(Z_{a/2})\left(\frac{\sigma}{\sqrt{n}}\right)=1.645\left(\frac{2}{\sqrt{100}}\right)=1.645(2/10)=\boxed{0.329}$$

thus, the Margin of error is $\bar{X}\pm E$:

$\bar{X}+E=5+0.329=\boxed{5.329}$ & $\bar{X}-E=\boxed{4.671}$

∴ We are 90% Confident that the population mean (μ) is between 5.329 & 4.671. (5 ± 0.329).

b) first, we look at the Confidence interval:

$1-\alpha=0.8$, we must thus find $Z_{a/2}=Z_{0.1}$

↳ According to the Z-Score table, 0.1 Corresponds to Approx. -1.28

↳ flip the Sign to get $\boxed{1.28}$, So $Z_{a/2}=1.28$

thus, the Z-Score Value used for an 80% Confidence interval is 1.28

c) Since the error bound is 0.17, $E=0.17$

We Know: $\sigma=2$, $\bar{X}=5$, & $1-\alpha=0.99$, So:

$Z_{a/2}=Z_{0.005}$, find 0.005 in the Z-Score table & reverse the Sign

↳ thus: $Z_{a/2}=2.575$

So, $E=(Z_{a/2})\left(\frac{\sigma}{\sqrt{n}}\right)\rightarrow\frac{1}{\sqrt{n}}=\frac{E}{Z_{a/2}(\sigma)}$ thus: $\frac{1}{n}=\left[\frac{E}{(Z_{a/2})(\sigma)}\right]^2$

$$\frac{1}{n}=\left[\frac{0.17}{(2.575)(2)}\right]^2\Rightarrow n=\left(\frac{515}{17}\right)^2\doteq 917.733564\doteq\boxed{918}$$

∴ A Sample Size of 918 is needed for the given Confidence interval

Question 4. In a pool of $n = 1000$ randomly selected teenagers, 450 indicated that they like Netflix.

a) Construct a 95% confidence interval for the proportion of teenagers who like Netflix and interpret it.

b) Why the result in a) is approximately valid.

c) If you wish to estimate the proportion of teenagers who like Netflix correct to within 0.025 with 95% confidence, how large should the sample size n be?

$n=1000,\ X=450$

a) $\hat{p}=\frac{X}{n}=\frac{450}{1000}=0.45$

for $1-\alpha=0.95$, find $Z_{\alpha/2}=Z_{0.025}$

the Z-score that results in 0.025 is -1.96

thus, $Z_{\alpha/2}=1.96$

thus, the margin of error is then:

$$E=Z_{\alpha/2}\left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)=1.96\left(\sqrt{\frac{(0.45)(1-0.45)}{1000}}\right)\doteq 0.03083498$$

the intervals boundaries are:

$\hat{p}\pm E$ thus: $\hat{p}+E\doteq 0.48083498\doteq\boxed{0.4808}$

$\hat{p}-E\doteq 0.419165019\doteq\boxed{0.4192}$

∴ We are 96% Confident that between $\boxed{0.4192 \ \& \ 0.4808}$ (0.45 ± 0.0308) lies the proportion of teenagers who like Netflix.

b) Since the problem is binomial (teenagers represent a true or false outcome w/ Netflix, & we check 1000 teens.) with a large n ($n>30$) of 1000, we can use the normal dist. as an approximation

In a, we use the normal dist. to find E by using Z values's Corresponding w/ 0.95

Since the Normal dist. is an Approx., & we use the Normal dist. in a, then a's Solⁿ is also an approx.

∴ a's Solⁿ uses an Approx. for the binomial dist. & thus is only Approx. valid

c) $E=0.025$ & $C=0.95$, $\hat{p}=0.45$

$$E=Z_{\alpha/2}\left(\sqrt{\hat{p}\hat{q}/n}\right)\rightarrow\frac{E^2}{Z^2_{\alpha/2}}=\frac{\hat{p}\hat{q}}{n}\quad\text{So}\quad n=\frac{(Z^2_{\alpha/2})(\hat{p})(1-\hat{p})}{E^2}$$

the Z-Score resulting in 0.025 is -1.96 So $Z_{\alpha/2}$ is 1.96

the Sample Size is thus:

$$n=\frac{(1.96)^2(0.45)(1-0.45)}{0.025^2}\doteq 1521.2736$$

→ Since we can't have half teens, we must round up for this to fall "w/in" 0.025 ∴ $\boxed{1522}$

thus, we need Approx. $\boxed{1522}$ (1521 w/ lower rounding) teenagers to estimate the proportion of teenagers who like netflix w/ the Given Constraints.

Question 5. One local hockey team conducts a study to compare the amount of money spent on refreshments at the Bell Centre. Two simple random samples of 100 men and 100 women are collected. For men, the average expenditure was \$20, with a known population standard deviation of \$5. For women, it was \$30, with a known population standard deviation of \$4.

a) Estimate the difference in the expenditure between man and women using a 99% confidence interval.

b) Based on your confidence interval, do you think there is a significant difference between the expenditure of men and women? Explain.

$\bar{X}_m = 20,\ S_m = 5,\ \bar{X}_w = 30,\ S_w = 4,\ n_m = 100,\ n_w = 100$

a) $C = 0.99$

the Samples are both $n > 30$ So we can Say:

$\sigma_m \doteq S_m$ & $\sigma_w \doteq S_w$

For Confidence interval $1-\alpha = 0.99$, find $Z_{\alpha/2} = Z_{0.005}$

$Z_{0.005} = -2.575$ So we flip the Sign to find $\boxed{Z_{0.005} = 2.575}$

$$E = Z_{\alpha/2}\sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}} = 2.575\sqrt{\frac{5^2}{100} + \frac{4^2}{100}}$$

$$= 2.575\sqrt{41/100} \doteq 1.6488$$

for the difference:

$(\bar{X}_w - \bar{X}_m) - E \doteq \boxed{8.3512}$

$(\bar{X}_w - \bar{X}_m) + E = \boxed{11.6488}$

We are ∴ 99% Confident that the difference in expendature between men & women lies between $\boxed{8.3512 \ \& \ 11.6488}$

b) Based off the Confidence interval, there is a Significant difference in the expendature between men and Women.

the difference interval is $(8.3512, 11.6488)$, which is far from 0 at the lower-bound, So we can Clearly See that there's a Significent difference between men & Women's expend.

thus, we can Conclude (W/ a 99% Confidence interval) Women Spend a non-trivial amount more (Roughly 1.5 times more) then men.

Question 6. According to a recent study, 75 out of 100 randomly selected teenagers supported the liberal government and 48 out of 80 randomly selected adults supported the liberal government. Assume these two samples are independent. Find a 90% confidence interval for the true difference in two population proportions and interpret it.

$n_1 = 100$, $n_2 = 100$, $X_1 = 75$, $X_2 = 48$

We can find the sample proportion when counting X_1 & X_2 as successes:

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{75}{100} = 0.75$$

$$\hat{p}_2 = \frac{X_2}{n_2} = \frac{48}{80} = 0.6$$

for the confidence interval $1-\alpha = 0.9$, find $Z_{a/2}$:

$Z_{0.05} = -1.645$, so if we reverse the sign we find $Z_{a/2} = 1.645$

thus, the margin of error is:

$$E = Z_{a/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 1.645\sqrt{\frac{(0.75)(1-0.75)}{100} + \frac{(0.6)(1-0.6)}{80}}$$

$$\boxed{\doteq 0.1149}$$

the endpoints of the confidence interval for $p_1 - p_2$ are then:

$$(\hat{p}_1 - \hat{p}_2) - E = (0.75 - 0.6) - 0.1149 \doteq \boxed{0.03514}$$

$$(\hat{p}_1 - \hat{p}_2) + E = (0.75 - 0.6) + 0.1149 \doteq \boxed{0.2649}$$

thus, we are 90% confident that the true difference in the two population proportions is between 0.03514 & 0.2649.

We can conclude that there is certainly a difference in the proportion of teenagers who support the liberal government vs. adults.