MATH 321 – QUIZ 5 & 6 SOLUTIONS – SPRING 2019 ‖ 50 POINTS

Instructions: Show all work. Collaboration and references are allowed. Approved computational devices allowed.

Print Name _____

1. (2 pts) Following are given two confidence intervals which were constructed from the same data set. One has 95% confidence level and the other has 99% confidence level. Which is which?

$$(19.44, 21.48) \qquad (19.11, 21.81)$$

Solution: $(19.44, 21.48)$ is the 95% CI and $(19.11, 21.81)$ is the 99% CI. Remember that increased confidence requires a wider interval (given a fixed data set).

2. (3 pts) Find the level of confidence associated with confidence interval $\overline{X} \pm 2.33\frac{\sigma}{\sqrt{n}}$. Write the R code and use R to calculate it.

Solution: It should be clear, that since σ is in the CI formula, that we are looking at a 'known variance' situation. This means that we are using a standard normal z-quantile in the CI formula.

We have that $2.33 =$`qnorm(1-`α`/2)` so $1-\alpha/2 =$`pnorm(2.33)` ≈ 0.9900969 and solving for $\alpha = 0.02$ thus $1-\alpha = 0.98$. So it is a 98% CI.

It's probably simpler just to think about the 99% meaning that we are chopping off an upper 1% tail but the CI also chops off the lower tail, so that the total middle would be 98%.

3. (4 pts) For a dataset of size $n = 120$ with sample mean $\overline{X} = 32$ taken from a population with known variance $\sigma^2 = 9$, construct a 95% confidence interval for the population mean μ. Write any R code used.

Solution: This is a known variance situation, so we'll use a standard normal quantile. Note that there is an underlying assumption of the data coming from a normal population, but since the sample size is so large $n = 120$, even if the population is not normal, this will generally be an acceptable approximation.

`xbar+c(-1,1)*qnorm(1-`α`/2)*`σ`/sqrt(`n`)`
`32+c(-1,1)*qnorm(0.975)*3/sqrt(120)` gives $(31.46324, 32.53676)$.

4. (5 pts) In a consumer preference survey of 1,500 individuals, it is found that 347 stated that they are willing buy brand A. Estimate the true population proportion that would buy brand A with a 95% confidence interval.

Solution:

```
a=0.05
x=347
n=1500
p=x/n
p+c(-1,1)*qnorm(1-a/2)*sqrt(p*(1-p)/n)
```

gives $(0.2099935, 0.2526731)$. So we are 95% confident that the true proportion of customers willing to buy brand A is between 21% and 25%.

5. (4 pts) Calculate the p-value, for a hypothesis test for a population mean. Use R and write the R code used.

(a) Two-sided test, unknown variance, sample size $n = 10$, test statistic $t = -2.5$.

Solution: $p =$ `2*pt(-2.5,9)` ≈ 0.03386183.

Note that whatever the hypotheses are, we would reject H_0 here if α was set to anything above our p value. So this example would result in a rejection of H_0 at $\alpha = 5\%$ definitely but not at $alpha = 1\%$ for example. We would reject H_0 at any level $\alpha \geq p$.

(b) One-sided test $H_a : \mu < \mu_0$, unknown variance, sample size $n = 10$, test statistic $t = -2.5$.

Solution: $p =$ `pt(-2.5,9)` ≈ 0.01693091. As noted above, we would reject H_0 at the 2% level here but not at 1% significance level.

6. (6 pts) A rope manufacturer claims that their ropes can handle a load of at least 6,000 lbs before failing. A sample of 15 rope segments are collected and tested until failure, and the average force required to cause failure was found to be $\overline{X}_{15} = 5,800$ lb with standard deviation $s = 500$ lb.
(a) Write the null and alternative hypotheses.
(b) Test the manufacturer's claim at significance level $\alpha = 0.1$.
(c) Clearly state your conclusion.

Solution: (a) $H_0 : \mu = 6,000$, $H_a : \mu < 6,000$.
(b) This is both a small sample size and unknown variance, so we'll use a t-test. $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} = \frac{5800-6000}{500/\sqrt{15}} \approx -1.549193$. The p-value is `pt(-1.549193,14)`=0.07182004.
(c) Since $p = 0.07182004 < \alpha = 0.1$, we reject H_0. This means that under the assumption that the null hypothesis is true, our data is sufficiently rare in our opinion that we don't want to risk trusting it.

7. (5 pts) If we are to perform a hypothesis, we need to choose the significance level α. If we assume the null hypothesis is true, what does α represent? What will happen if we increase or decrease α?

Solution: $\alpha = P(\text{type I error}) = P(\text{rejecting } H_0 \text{ under the assumption that it is true})$. Type I error can be thought of as a "false positive," rejecting a true hypothesis due to potentially faulty or rare data. If we increase α, we increase the chance of this happening by making our rejection region larger. Decreasing α makes th probability of type I error smaller.

8. (6 pts) A company is interested in testing whether or not any particular day of the week tends to have more or less customers. Here are the number of customers observed on each day for one week:

Day	Mon	Tues	Weds	Thurs	Fri
# customers	123	107	92	118	105

Does this data call into question the claim that each day should on average have the same number of customers? Write any R code used. Clearly state your conclusion. *(Hint: Use chi-squared goodness of fit test.)*

Solution:

$H_0 : p_i = \frac{1}{5}$ for $i = 1, 2, 3, 4, 5$
H_a : at least one $p_i \neq \frac{1}{5}$
or in words:
H_0 : the proportion of customers for every day of the week is the same
H_a : at least two days have different proportions of customers

$n = 123 + 107 + 92 + 118 + 105 = 545$ thus $p = 545/5 = 109$. This is our expected count for each day under the assumption that customers do not tend to arrive on some days more than others.

Our test statistics is $c = \sum \frac{(obs-exp)^2}{exp}$ and our p-value is $P(\chi^2_{k-1} > c)$.

```
obs=c(123,107,92,118,105)
p=rep(1/5,5)    or p=c(1,1,1,1,1)/5
n=sum(obs)
exp=sum(obs)*p
c=sum((obs-exp)^2/exp)
p=1-pchisq(c,length(obs)-1)
```

gives $p \approx 0.2508327$ thus we would reject H_0 only as a significance level greater than 25%. For the commonly used value of $\alpha = 0.05$, we do not reject H_0.

9. (7 pts) The current US House of Representatives is composed of 235 Democrats and 197 Republicans (there are 3 vacant seats). If 141 Demcorats support a particular bill, what is the maximum number of Republicans that can support it and still see a statistically significant difference at the 5% level? Write all R code used.

(Hint: Start at the same level of support among Republicans, perform a two proportion hypothesis test. Then decrease the number of republicans supporting the bill and re-do the hypothesis test. Continue decreasing the number of Republicans supporting the bill until the outcome of the test is to reject the null hypothesis that both parties support the bill equally. Carefully think whether this should be one- or two-sided.)

Solution:

$H_0 : p_D = p_R$ or $H_0 : p_D \leq p_R$
$H_a : p_D > p_R$.

141/235=0.6, so we start with close to 60% support among Republicans and will decrease that support level until we see a statistically significant difference, i.e. until $p < 0.05$. Run the following code starting with $x = 118$ and decreasing x until we get $p < 0.05$.

At $x = 103$, $p = 0.05335332$ and at $x = 102$, $p = 0.04281528$. Thus we need Republican support to be at $102/197 \approx 52\%$ or lower in order to show a statistically significant difference.

A 2-sided test is also reasonable here, but you would need to calculate the high and low levels of support among republicans that give a statistically significant difference. These occur at $x = 99$ and $x = 136$ Republicans supporting the bill.

```
x1=141
n1=235
x2=x
n2=197
p1=x1/n1
p2=x2/n2
z=(p1-p2)/sqrt(p1*(1-p1)/n1+p2*(1-p2)/n2)
pnorm(-abs(z))
```

10. (8 pts) The following table show real gross domestic product (GDP) and real personal expenditures (PE), both in trillions of US $, for the years 2009 to 2018. Write any R code used in answering the questions below.

(a) Find the covariance of GDP and PE.
(b) Find the correlation coefficient between GDP and PE.
(c) Construct a linear regression line for GDP as a linear function of PE.
(d) Find the SSE, sum of the squared errors.

Solution:

We will let GDP be the Y variable and PE be the X variable.

```
gdp=c(15.2,15.6,15.8,16.2,16.5,16.9,17.4,17.7,18.1,18.6)
pe=c(10.5,10.6,10.8,11,11.2,11.5,11.9,12.2,12.6,12.9)
cov(gdp,pe)
cor(gdp,pe)
b=cov(gdp,pe)/var(pe)
a=mean(gdp)-b*mean(pe)
gdpfit=a+b*pe
SSE=sum((gdpfit-gdp)^2)
```

(a) Cov(X,Y)=0.9544444
(b) Cor(X,Y)=0.9943344
(c) Intercept: $a = 1.472119$, slope $b = 1.330545$
(d) SSE = 0.1306165.