PH142 Spring 2021 Final Exam - Gradescope Portion

Your name here

Question 1 [1 point total]

According to the National Organization for Rare Disorders, Acoustic Neuroma is a rare disease resulting a non-cancerous tumor in the vestibulocochlear nerve. Acoustic neuromas are estimated to affect about 1 in 100,000 people in the general population. Assume that acoustic neuroma are random and independent events. There is a local hospital which sees 12,000 patients a year. Let X be the count of patients with an acoustic neuroma in a year at that hospital.

What is an appropriate distribution for X?

a) Binomial Distribution
b) Poission Distribution
c) Normal Distribution
d) Student t distribution

```
# SOLUTION: b)
```

Question 2 [1 point total]

In order to calculate the 95% confidence interval for the sample mean, we use the r function qnorm to get the critical value z: `qnorm(____, lower.tail=____)`. Please choose the one that correctly fills in the function.

a) 0.05, FALSE
b) 0.975, FALSE
c) 0.95, TRUE
d) 0.025, FALSE
e) 0.05, TRUE

```
# SOLUTION: d)
# `qnorm(0.025, lower.tail = FALSE)`
```

Question 3 [5 points total]

We have been hired by lead NGOs to investigate into the severe acute malnutrition at the community-level. The outcome of interest Y is the average mid-upper arm circumference (MUAC) in cm of children aged 6-59 months in each community. We would like to see whether a community's socioeconomic status, measured on a scale from 1-5, affects the MUAC of children in that community. There are 5 communities of socio-economic status from 1-5 respectively in the study.

3.1 [1 point] What type of variable is the outcome?

a) Numerical, Discrete
b) Numerical, Continuous
c) Categorical, Nominal
d) Categorical, Ordinal

```
# SOLUTION: b, Numerical, Continuous
```

3.2. [1 point] What is the type of problem the study addresses?

a) Causative
b) Descriptive
c) Predictive

```
# SOLUTION: a, Causative
```

3.3. [1 point] Identify ONE appropriate visualization for the data.

```
# SOLUTION: There are a bunch, e.g. box plots, density plot, histogram with facets
```

3.4. [1 point] Which test is most appropriate to apply to the outcome variable and explanatory variable?

(a) Z-test
(b) T-test
(c) Two-Sample Paired T-test
(d) ANOVA
(e) Chi-square test

```
# SOLUTION: d, ANOVA
```

3.5 [1 point] What is the null hypothesis in the context of the question?

```
# SOLUTION: mu_1 = mu_2 = ... = mu_5;
# all the communities have the same MUAC.
```

Question 4 [2 points total]

Cystic fibrosis is a progressive, genetic disease that causes persistent lung infections and limits the ability to breathe over time. One key outcome for patients with cystic fibrosis is lung function (also known as FEV1). FEV1 is the volume of air that can forcibly be blown out in first 1 second, after full inspiration. A clinical study examined the effect of a bronchodilator treatment called Xoponex. At the start of the study, researchers measured FEV1 levels of 18 children diagnosed with cystic fibrosis. Here are the findings (in percents):

70.8 69.6 77.5 78.0 74.8 66.1 64.3 62.2 67.5 69.0 64.3 54.9 55.8 57.3 54.7 55.9 48.7 49.0

These data have mean $\bar{x} = 63.36$ and standard deviation $s = 9.13$. In healthy children, FEV1 values below 80% would be considered as abnormal. Consider patients in this study as a simple random sample of all children with cystic fibrosis. We expect cystic fibrosis to *decrease* FEV1 on average.

4.1 [1 point] We want to test if there is a difference from the population mean FEV1 $\mu_0 = 80$ in our sample. What test should we use to approach this data?

a) one sided z test
b) two sided z test
c) one sided t test
d) two sided t test

```
# Solution: C one sided t test because we don't know the true value
# for population sd. use sample sd from data.
# want to compare sample mean with population mean.
```

4.2 [1 point] What is the value of the test statistic for these data?

a) -7.7
b) 7.5
c) -32.81
d) 67.9

```
# Solution: A (63.36 - 80) / ((9.13) / sqrt(18)) = -7.7
```

Question 5 [6 points total]

Moderna, an American biotechnology company, released results from their phase 3 randomized control trial, which was conducted to determine the efficacy of their COVID-19 vaccine candidate. Suppose you are interested in knowing whether the vaccine and placebo groups have different proportions of COVID-19. Below is a table of the results.

```
##         COVID-19 no COVID-19
## vaccine       11       14989
## placebo      185       14815
```

5.1 [1 point] Select the correct $\hat{p}$ rounded to 4 decimal places.

a) 0.0007
b) 0.1249
c) 0.0065
d) 0.0058

```
# Solution: C $\hat{p}$ = (11+185) / (30000) = 0.0065
```

5.2 [1 point] Compute the standard error for this hypothesis test. Round your answer to four decimal places and show your work.

```
# SOLUTION:
# sqrt(0.006533333 * (1 - 0.006533333) * (1 / 15000 + 1 / 15000))
# = 0.0009302794 = 0.0009
```

5.3 [1 point] Compute the test statistic using the appropriate statistical test. If you did not get a value for question 5.2, use 0.001. Round to one decimal place.

```
# SOLUTION:
# z = ((185/15000) - (11/15000)) / 0.0006521513 = 12.46937 = 12.5
```

5.4 [1 point] We save the test statistic above in an object called `q5stat`. Which of the following will calculate the correct p-value for this hypothesis test? Select all that apply.

a) `pnorm(q = q5stat, lower.tail = F)*2`
b) `pt(q = q5stat, df = 1, lower.tail = F)*2`
c) `prop.test(x = c(11,185), n = c(15000, 15000))`
d) `prop.test(x = c(11, 185), y = c(14989, 14815))`

```
# Solution: A, C
```

5.5 [1 point] Suppose the p-value is equal to 8.86e-71. Interpret the p-value in two sentences or less in the context of the question.

```
# Solution: The p-value is equal to approximately 0%.
# Under the null hypothesis of no difference between the proportions,
# there is an approximately 0% chance of observing the difference we saw
# or more extreme, which provides evidence in favor of
# the alternative hypothesis that these proportions are different.
```

Question 6 [3 points total]

Take Action is the common name for a contraception drug, levonorgestrel. Researchers enrolled a cohort of fertile women after unprotected intercourse during their fertile period. In this cohort, 97 women had unprotected intercourse before ovulation. Based on standard fertility rates, the researchers expected 14.4 pregnancies. Instead, they observed 0 pregnancies.

6.1 [1 point] We want to know if there is evidence that Take Action impacts the chance of pregnancy when taken before ovulation. Choose the appropriate statistical test for this data.

a) Kruskal Wallis test
b) Two-sided t-test
c) Wilcoxon rank sum test
d) Chi-squared test for goodness of fit

```
# Solution: D Chi-squared test
```

6.2 [1 point] Select the appropriate value of the test statistic rounded to one decimal place.

a) 16.9
b) -14.4
c) 0
d) 6.7

```
# SOLUTION: A (0-14.4)^2 / 14.4 + (97-82.6)^2 / 82.6 = 16.91041 = 16.9
```

6.3 [1 point] Let the test statistic calculated above be saved in an object called `q6stat`. Fill in the blanks in the following R code to calculate the appropriate p value. The first blank should go in the first box, second in the second box, and so on.

```
p_value <- _______(q6stat, _______ , lower.tail= ___)
```

First blank:
Second blank:
Third blank:

```
# SOLUTION: pchisq, df = 1 (or just 1), F
```

Question 7 [2 points total]

Title: Multiple Linear Regression Model of Meningococcal Disease in Ukraine: 1992–2015

Abstract: Estimating the rates of invasive meningococcal disease (IMD) from epidemiologic data remains critical for making public health decisions. In Ukraine, such estimations have not been performed. We used epidemiological data to develop a national database. These data were used to estimate the population susceptible to IMD and identify the prevalence of asymptomatic carriers of N. meningitidis using simple epidemiological models of meningococcal disease that may be used by the national policy makers. The goal was to create simple, easily understood analysis of patterns of the infection within Ukraine that would capture the major features of the infection dynamics. Studies used nationally reported data during 1992–2015. A logic model identified the prevalence of carriage and the proportion of the population susceptible to IMD as key drivers of IMD incidence. Multiple linear regression models for all ages (total population) and for children $\leq$ 14 years old were fit to national-level data. Linear models with the incidence of IMD as an outcome were highly associated with carriage (b = 0.84) and estimated susceptible population b = 455.58) in both total population and children (R2= 0.994 and R22= 0.978, respectively).

https://doi.org/10.1155/2020/5105120

7.1 [1 point] Interpret the slope of the linear model used for predicting IMD incidence by carriage prevalence among the total population outlined in the abstract.

```
# Solution: For every one percentage increase in carriage,
# there is a corresponding 0.84 case increase in IMD incidence
```

7.2 [1 point] Say we calculate two confidence intervals for this slope coefficient. The 95% confidence interval is (0.09, 1.59). The 99% confidence interval is (-0.51, 2.19). Which of the following *best* describes the p-value?

a) $p < 0.01$
b) $p > 0.01$
c) $p > 0.025$
d) $0.01 < p < 0.05$

```
# SOLUTION: D
```

END