PH142 Spring 2021 Final Exam - Takehome Portion SOLUTIONS

Question 1 [7 points total]

A common belief among fans of Star Trek: The Original Series is that U.S.S. Enterprise crew members wearing red shirts are more likely to be killed off than those wearing other shirt colors, so much so that "Redshirt" is a commonly-used term to refer to any stock character in fiction who is considered "expendable" or dies soon after being introduced. The following is a table summarizing the deaths of crew members throughout the series:

Shirt Color	Fatalities	Survivors	Total
Blue	7	129	136
Gold	9	46	55
Red	24	215	239
Total	40	390	430

1.1. [0.5 points] What statistical test can we use to determine whether a crew member's shirt color is associated with their survival?

○ χ^2 Goodness-of-Fit Test
○ Two Sample Paired t-Test
○ One Sample Test of Proportion
○ χ^2 Test of Independence
○ Two Sample Test of Proportions

```
# Solution: $\chi^2$ Test of Independence
```

1.2. [0.5 points] Which of the following sets of null and alternative hypotheses are appropriate for this test? (Assume $p_{\textbf{color}}$ refers to the proportion of fatalities among crew members wearing that color) Select all that apply.

□ H_0 : Shirt color and fatality are independent
H_0 : Shirt color and fatality are not independent

□ $H_0 : p_{\text{blue}} = p_{\text{gold}} = p_{\text{red}}$
$H_a : p_{\text{blue}} \neq p_{\text{gold}} \neq p_{\text{red}}$

□ $H_0 : p_{\text{blue}} = p_{\text{gold}} = p_{\text{red}}$
$H_a : p_{\text{blue}} \neq p_{\text{gold}}$, or $p_{\text{blue}} \neq p_{\text{red}}$, or $p_{\text{gold}} \neq p_{\text{red}}$, or $p_{\text{blue}} \neq p_{\text{gold}} \neq p_{\text{red}}$

□ H_0 : Shirt color is not associated with fatality
H_a : Shirt color is associated with fatality

```
# Solution: All but $H_a:  p_{\text{blue}} \neq p_{\text{gold}} \neq p_{\text{red}}$
```

1.3. [1.5 points] Calculate and fill in the following table with the expected counts of fatalities and survivors among the different shirt colors. Round to 2 decimal places.

Shirt Color	Fatalities	Survivors
Blue	A	B
Gold	C	D
Red	E	F

A.

```
# Solution: 12.65
```

B.

```
# Solution: 123.35
```

C.

```
# Solution: 5.12
```

D.

```
# Solution: 49.88
```

E.

```
# Solution: 22.23
```

F.

```
# Solution: 216.77
```

1.4. [2.5 points] Using the *unrounded* expected counts from the previous question, calculate and fill in the following table with the contributions of each cell to the test statistic rounded to 3 decimal places. Then using the *unrounded* cell contributions, report the final test statistic rounded to 3 decimal places.

Shirt Color	Fatalities	Survivors
Blue	G	H
Gold	I	J
Red	K	L

G.

```
# Solution: 2.524
```

H.

```
# Solution: 0.259
```

I.

```
# Solution: 2.948
```

J.

```
# Solution: 0.302
```

K.

```
# Solution: 0.141
```

L.

```
# Solution: 0.014
```

Test Statistic:

```
# Solution: 6.189
```

1.5 [1 point] Write one line of `R` code to find the p-value for this test using the test statistic you just calculated. (Hint: Remember you can use `R` to check whether your code runs as expected!)

```
# Solution: pchisq(q = 6.189, df = 2, lower.tail = FALSE)
```

1.6. [1 point] In no more than two sentences, interpret the results of your test using $\alpha = 0.05$. Do your results confirm that red shirts specifically are more likely to die?

```
# Solution: Our p-value of 0.04530638 is less than $\alpha = 0.05$ so we find
# evidence to reject the null hypothesis that shirt color is independent of fatality.
# However, this test cannot tell us whether red shirts specifically
# are more likely to die.
```

Question 2 [3 points total]

Another theory is that it is not the red shirt, but rather working in security that determines a crew member's likelihood of death (Security, Engineering, and Operations all wear red shirts). To investigate this question we will use the below table summarizing the deaths of crew members wearing red shirts working in Security versus other departments:

Department	Fatalities	Survivors	Total
Security	18	72	90
Non-Security	6	143	149
Total	24	215	239

2.1. [2 points] Calculate a 95% confidence interval for the difference in fatality proportion among security red shirts versus non-security red shirts. Round to 3 decimal places at the *very end* of your calculations.

```r
p_sec <- 18/90
p_nonsec <- 6/149
n_sec <- 90
n_nonsec <- 149
est <- p_sec - p_nonsec
se <- sqrt(((p_sec*(1 - p_sec))/n_sec) + ((p_nonsec*(1 - p_nonsec))/n_nonsec))
z <- qnorm(0.975)
me <- z*se
ci <- c(est - me, est + me)
# Solution: [0.071, 0.248]
```

2.2. [1 point] Based on your confidence interval, what do you expect will be the result of the two sample difference of proportions test at the $\alpha = 0.05$ level? Explain your answer in one sentence.

```r
# Solution: We expect to find statistically significant evidence
# that Security crewmembers have a higher proportion
# of fatalities than other red shirts since the
# confidence interval is all above zero.
```

Question 3 [4 points total]

We would like to use the South African heart disease data set to determine whether coronary heart disease (`chd`) and low density lipoprotein cholesterol (`ldl`) are related. `chd=0` means that no CHD event occurred and `chd=1` means CHD event occurred. Each row in this data set is a single independent observation from a South African man. Here are the first six rows:

```
##   chd  ldl
## 1   1 5.73
## 2   1 4.41
## 3   0 3.48
## 4   1 6.41
## 5   1 3.50
## 6   0 6.47
```

We perform an analysis in R. (Note: this is not the usual format of the output of this test. It has been modified).

```
## # A tibble: 1 x 8
##   mean_group0 mean_group1 statistic p.value    df conf.low conf.high alternative
##         <dbl>       <dbl>     <dbl>   <dbl> <dbl>    <dbl>     <dbl> <chr>
## 1        4.34        5.49     -5.55 6.77e-8  280.    -1.55    -0.738 two.sided
```

3.1 [2 points] Is this a paired t-test or a two-sample t-test? Justify your answer.

```
# SOLUTION: two-sample because the observations
# from the two groups (chd=0 and chd=1) are
# independent from one another.
```

3.2 [1 point] What are the null and alternative hypotheses? You may write them in words or notation (as well as you can).

```
# SOLUTION: H_0: mu1 - mu2 = 0 (1 point);
# H_A: mu1 - mu2 != 0 (1 point)
```

3.3 [1 point] Would you reject the null hypothesis based on these results based on an α of 0.05? Interpret the p-value in the context of the question.

```
# SOLUTION: Yes, the p-value is ___ meaning there is a ____% chance of
# observing a test statistic at least as extreme
# as this if the null hypothesis of
# no difference between cholesterol in CHD groups is true.
```

Question 4 [2 points total]

Now we are interested in investigating the relationship between family history of heart disease and presence/absence of coronary heart disease using the South African heart disease data. Below is the contingency table for these two variables:

```
##          chd
## famhist     0   1
##   Absent  206  64
##   Present  96  96
```

According to this table, the proportion of people who suffer from coronary heart disease given that they do not have family history of heart disease is 64/270 = 0.237. The proportion of people who suffer from coronary heart disease given that they do have family history of heart disease is 96/192 = 0.5.

4.1 [1 point] Calculate the risk ratio. Show your work and round your answer to 2 decimal places.

```
# SOLUTION: 0.5/0.237 = 2.11
# 0.5 point for showing correct-ish work. 0.5 point for correct answer.
```

4.2 [1 point] Calculate the odds ratio. Show your work and round your answer to 2 decimal places.

```
# SOLUTION: ((0.5)/(1-0.5))/((0.237)/(1-0.237)) = 3.22
# 1 point for showing correct-ish work. 0.5 point for correct answer.
```

Question 5 [8 points total]

In an effort to promote the Winter Olympic Games, Moscow city officials once led a "Russian health campaign", which allows train users in Moscow to pay in squats.

Suppose we took a random sample (sample size $n = 500$) of Moscow train users to do squats and grouped into 1 of 5 groups. These groups represent the range of squats they need to complete to "earn" a train ride.

Squats Group	Group Size
$A : 0 - 20$	88
$B : 20 - 30$	156
$C : 30 - 40$	143
$D : 40 - 50$	70
$E :> 50$	43

We also measured the happiness of the riders, which is a continuous score on a validated scale from 30-45. Here's a glance of the data (stored in `squats`):

```
##   group happiness
## 1     A  36.88208
## 2     A  37.55613
## 3     A  36.32048
## 4     A  34.12768
## 5     A  34.74622
## 6     A  36.59717
```

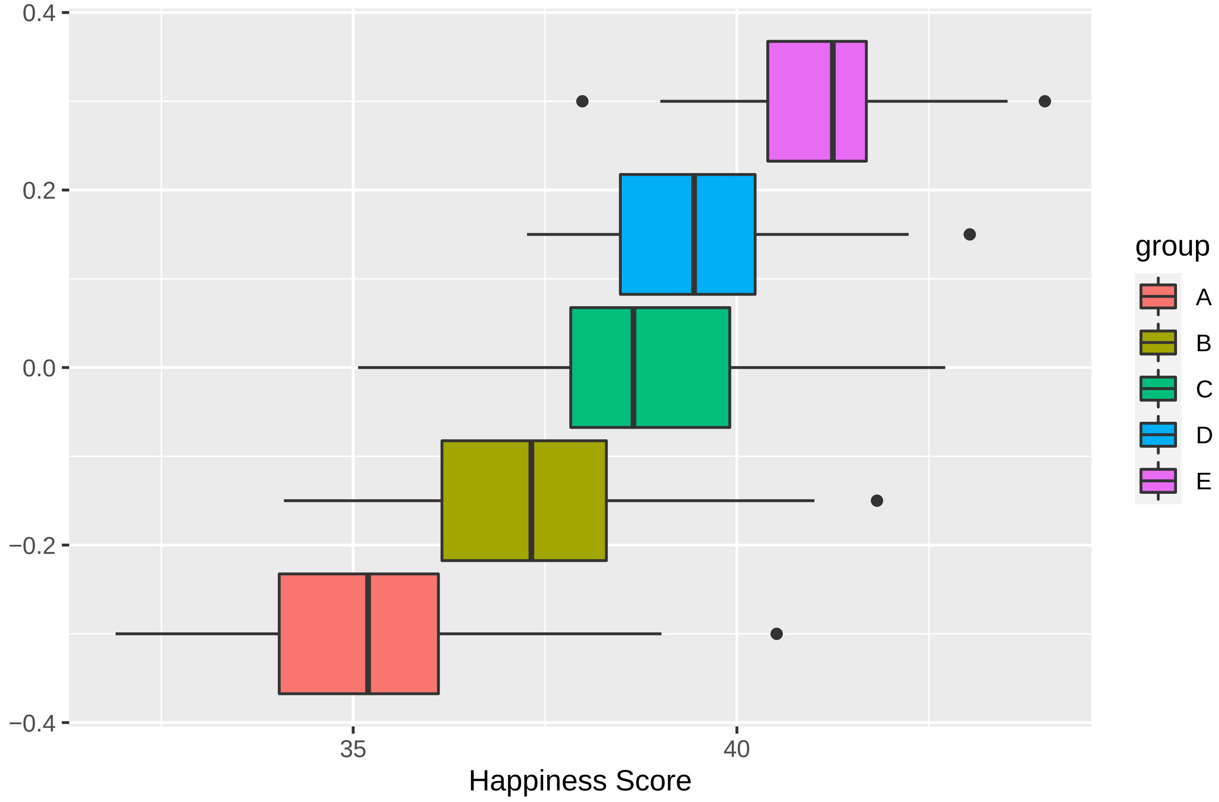
Now we will see if there is a difference in the average happiness score between the 5 groups.

5.1 [1 point] What is the most appropriate test for this research question?

```
#Solution: ANOVA.
```

5.2 [3 points] Using the visualization below and information in the question, are the assumptions met for your chosen test? Comment on each assumption for your chosen test. If you need more information, state what information you would need to assess that assumption.

Figure 1



```
# Solution: SRS, standard deviations, and normality of the groups.
# SRS is stated in the question, standard deviations look approximately equal,
# from the boxplots, the data appears approximately centered
# with no strong outliers - therefore approximately normal.

# -0.5 for each missed assumption
# -0.5 for missing/incorrext explanation
```

5.3 [1 point] Assume all conditions for this test are satisfied. State the alternative hypothesis

in words in the context of the question.

```
# Solution: HA: Not all average happiness scores for the squat groups are the same.
```

5.4 [1 point] **Write one line of `R` code to implement your test.**

```
# solution:
squats.test <- aov(happiness~group, data=squats)
```

5.5 [1 point] **The output from the statistical test is shown below. What distribution does the test statistic come from?**

term	df	sumsq	meansq	statistic	p.value
group	AAA	1468.53	367.13	174.33	4.70e-93
Residuals	BBB	1042.47	2.11	NA	NA

○ Student's t
○ Chi-Squared
○ ANOVA
○ F

```
# SOLUTION: D The null distribution is an F-distribution.
```

5.6 [1 point] **You might have noticed, two boxes are filled in with AAA and BBB instead of numbers. Fill in the numbers that belong in those boxes.**

AAA:

```
# SOLUTION: 4
```

BBB:

```
# SOLUTION: 195
```

Question 6 [2 points total]

6.1 [2 points] We know that TukeyHSD sets the overall error rate at 5% as an extension of an ANOVA test. In your own words, identify why this is useful and one other reason we use TukeyHSD.

```
# SOLUTION: Correct for multiple comparisons and identify which group(s) are different.
```

Question 7 [6 points total]

UK Biobank is a large-scale biomedical database containing health information from half a million UK participants. To study the statistical association between systolic blood pressure(in $mmHg$) and BMI(in kg/m^2) for humans, we took a sub-sample from the Biobank data, stored in `Biobank`.

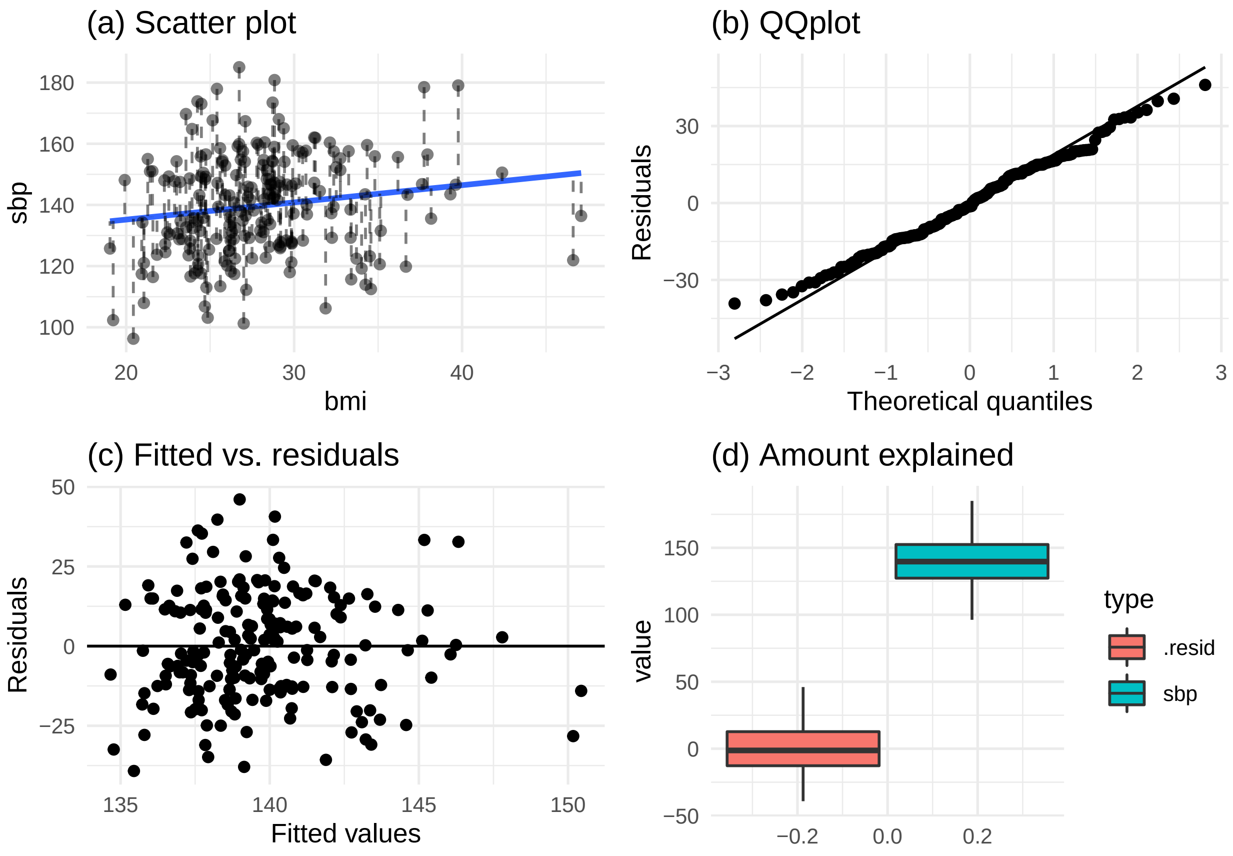
```
##        sbp      bmi
## 1 121.5518 25.86241
## 2 131.5333 35.16031
## 3 127.3799 29.86696
## 4 131.4940 24.41427
## 5 154.2746 23.00106
## 6 147.2640 31.19810
```

We fit a linear regression model using the systolic blood pressure(stored in `sbp`) as the response variable and BMI(stored in `bmi`) as the explanatory variable. The model is stored in `Biobank.model`.

```
Biobank.model <- lm(sbp~bmi, data=Biobank)
```

In this question we will perform an regression inference analysis.

7.1 [4 points] Before we run the analysis, we must first check the assumptions of the linear regression test. Using the output below, identify and comment on each of the assumptions. Hint: There are 4.



Assumption 1:

Assumption 2:

Assumption 3:

Assumption 4:

```
# SOLUTION:
#1. linear relationship between x and y.
#2. Residuals are normally distributed.
#3. Variation is constant for all values of x's
#4. observations are independent
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  124.        7.29      17.0  1.50e-40
## 2 bmi            0.562     0.257      2.19 2.97e- 2
```

7.2 [1 point] State the null and alternative hypothesis in the context of this research question.

Null:

Alternative:

```
# SOLUTION:
# Null: There is no association between sbp and bmi (b = 0)
# Alternative: There is an association between sbp and bmi (b != 0)
```

7.3. [1 point] Using the output above, construct a 95% confidence interval for the slope parameter.

```
# SOLUTION: (0.0504, 1.0696)
```

Question 8 [7 points total]

8.1 [2 points] Select all of the scenarios below for which a nonparametric test is more appropriate than the parametric equivalent.

We have a random sample of 36 individuals who will be doing randomly assigned to either a diet plan or an exercise plan. We are interested in whether there is a difference in average weight loss between the two groups. 5 of 18 people drop out of Group Diet due to non-adherence and 3 of 18 drop out of Group Exercise. The median weight loss after 3 months of Group Diet is 3 with an average of 7 and the median weight loss of Group Exercise is 4 with an average of 5. □

We are interested in if there is a difference in the average lifespan of strictly indoor cats, strictly outdoor cats, and indoor/outdoor (aka hybrid) cats. We take a random sample of 300 cats in the city of Berkeley, 150 of which are indoor, 50 of which are strictly outdoor and 100 of which are hybrid. All of the groups are approximately normally distributed. The average lifespan of Group Indoor is 13 years with a standard deviation of 1, Group Outdoor is 8 years with a standard deviation of 2, Group Hybrid is 10 years with a standard deviation of 1.5. □

The PI of your research group is interested in understanding if there is a difference in the resting heart rate of men and women. To control for excess variation, you sample 10 heterosexual couples living in the same household and measure their heart rates. Both groups are approximately normally distributed but there is one household where each member of the couple has significantly higher than normal resting heart rate. □

You are interested in understanding if there is a difference in average screentime between undergraduates and graduate students among students interested in public health at Berkeley. You sample every student in your PH142 class and find that among the 47 graduate students, the screentime is slightly right skewed with no strong outliers and an average of 3.5 hours and sd of 0.56. Among the 253 undergraduate students, the sample is roughly normally distributed with an average of 4.2 hours and an sd of 0.32. □

```
# SOLUTION: A, B
```

8.2 [1 point] Whether or not you chose it as a nonparametric example, let's proceed with the study of resting heart rate among men and women and apply a nonparametric test. What is the nonparametric test that is appropriate for this problem?

○ Wilcoxon Rank Sum
○ Wilcoxon Sign Rank
○ Kruskal Wallis
○ Chi-Squared Test

```
# SOLUTION C
```

8.3 [3 points] The data you have collected from this study sample is below. Calculate the following values:

Household	Women	Men
1	54	60
2	61	52
3	51	58
4	50	45
5	71	72
6	63	60
7	45	47
8	65	68
9	47	45
10	87	85

T :

μ_T

σ_T

```
# SOLUTION:
# diff: 6, -9, 7, -5, 1, -3, 2, 3, -2, -2
# order: 1 (+), 2 (+), 2 (-), 2 (-), 3 (+), 3 (-), 5 (-), 6 (+), 7 (+), 9 (-)
# rank: 1, 3, 3, 3, 5.5, 5.5, 7, 8, 9, 10
# pos sum: 22.5
# neg sum: 27.5

#T: 27.5
#muT: 30.25
#sigmaT: 13.87
```

8.4 [1 point] Write one line of `R` code to calculate the p-value for this hypothesis test. ***Note: no points will be awarded for using a function that ends in*** **`.test`.**

```
# SOLUTION: pnorm(-0.19)*2
```

Extra Credit 1 [1 point total]

In your own words how does a permutation test construct a distribution?

Extra Credit 2 [4 points total]

Identify a peer reviewed scientific study published in the last 5 years that has results that are compelling enough to you that you would be willing to change your behavior as a result. This study must use a method we have explored in part III of this class to test an association. This study should be attempting to answer a causal question. Briefly describe the hypothesis of the study, what method discussed in PH142 was used in the study and why you found the study compelling. Include a citation for the published study.

Exam feedback:

If you experienced any issues with your exam please describe them here: