Problem Set 8

Michael Ghattas

Introduction

Collaboration

(1 point)

Other students who I worked with on this assignment (if any) : None.

Notes

These questions were rendered in R markdown through RStudio (https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf, http://rmarkdown.rstudio.com).

Please complete the following tasks regarding the data in R. Please generate a solution document in R markdown and upload the .Rmd document and a rendered .pdf document. Your solution document should have your answers to the questions and should display the requested plots.

Question 1

Let W be a random variable with expected value equal to μ_W and variance equal to σ^2_W.Let X be a random variable with expected value equal to μ_X and variance equal to σ^2_X. Suppose W and X are independent. Define the random variable Y by $Y = X + W$. Recall that if U and V are independent random variables then $E[UV] = E[U]E[V]$. (This is true provided $Var[U]$ and $Var[V]$ are defined.)

Question 1.1

(5 points)

Please compute the covariance of X and Y, $Cov[X, Y]$, in terms of μ_W, σ^2_W, μ_X, and σ^2_X.

Your answer here: $Cov[X, Y] = Cov[X, X + W] = Cov[X, X] + Cov[X, W]$ and since X and W are independent we know $Cov[X, W] = 0$. Thus, $Cov[X, Y] = Cov[X, X] = \sigma^2_X$.

Question 1.2

(5 points)

Let W, X, and Y be as defined above and suppose the correlation of X and Y equals $\rho > 0$. Find the value of the ratio $\frac{\sigma^2_W}{\sigma^2_X}$ in terms of ρ.

Your answer here: Given that the correlation ρ between two random variables X and Y is defined as:

$$\rho = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}}$$

and knowing that:

$$\mathrm{Cov}[X,Y] = \sigma_X^2, \quad \mathrm{Var}(Y) = \sigma_X^2 + \sigma_W^2,$$

we can write:

$$\rho = \frac{\sigma_X^2}{\sqrt{\sigma_X^4 + \sigma_X^2\sigma_W^2}} = \frac{\sigma_X}{\sqrt{\sigma_X^2 + \sigma_W^2}}.$$

Squaring both sides:

$$\rho^2 = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}.$$

Rearranging to solve for the ratio $\frac{\sigma_W^2}{\sigma_X^2}$:

$$\rho^2(\sigma_X^2 + \sigma_W^2) = \sigma_X^2,$$

$$\frac{\sigma_W^2}{\sigma_X^2} = \frac{1-\rho^2}{\rho^2}.$$

Question 2

This problem set uses 2019 data primarily for Denver county accessed through IPUMS-USA, University of Minnesota, www.ipums.org ,

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 7.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

The PUMA-to-county restriction was done using MABLE, https://mcdc.missouri.edu/applications/geocorr.html

This problem set uses a subsample of demographic data for the Denver area.

The sample was drawn according to the values in the variable "perwt". This is a weight value provided by the US Census Bureau to correct for differences between the sampled population and the target population. It is called a sample weight or an expansion weight.

It can be thought of as the number of people in Colorado that the one observation represents in terms of demographic characteristics. For example, if you add all the weights in the original sample for all of Colorado, you will get an approximation of the population of Colorado in the sample year. If you multiply the "age" variable by "perwt" then divide by the sum of the "perwt" values, you will get an approximation of the average age in the state, whether or not the ages of the cases are present in the same proportion in the sample as in the population.

Thus a sample drawn using "perwt" as the probability will be a better approximation of the population than a simple random sample in which each case has an equal likelihood of being selected.

The category "educ"=7 corresponds to 1 year of college. The category "educ"=10 corresponds to 4 years of college.

Samples of size 25 are drawn from the responses with "educ"=7 and with "educ"=10 according to the weights in the data set and saved in "dat_7_10.RData".

IPUMS Data

Read in the subsample of the IPUMS data.

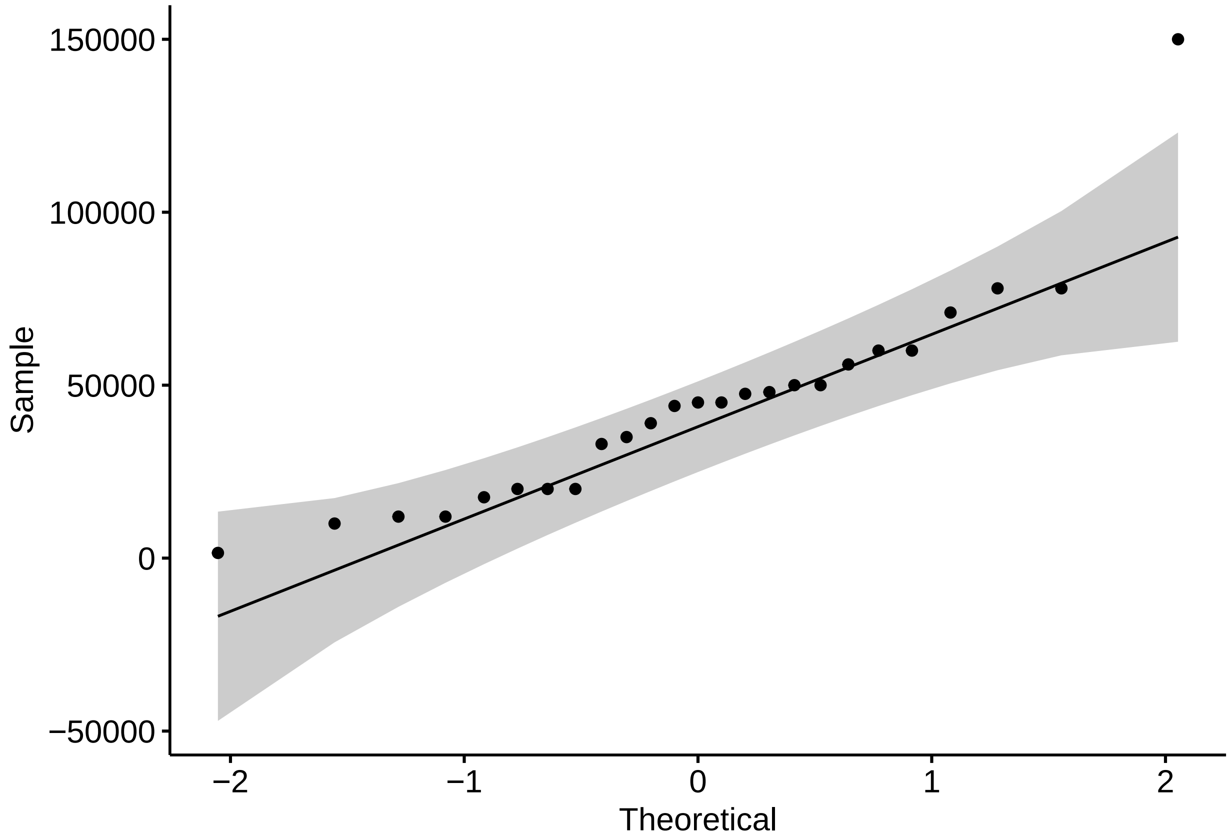
```
load("dat_7_10.RData")
```

Question 2.1

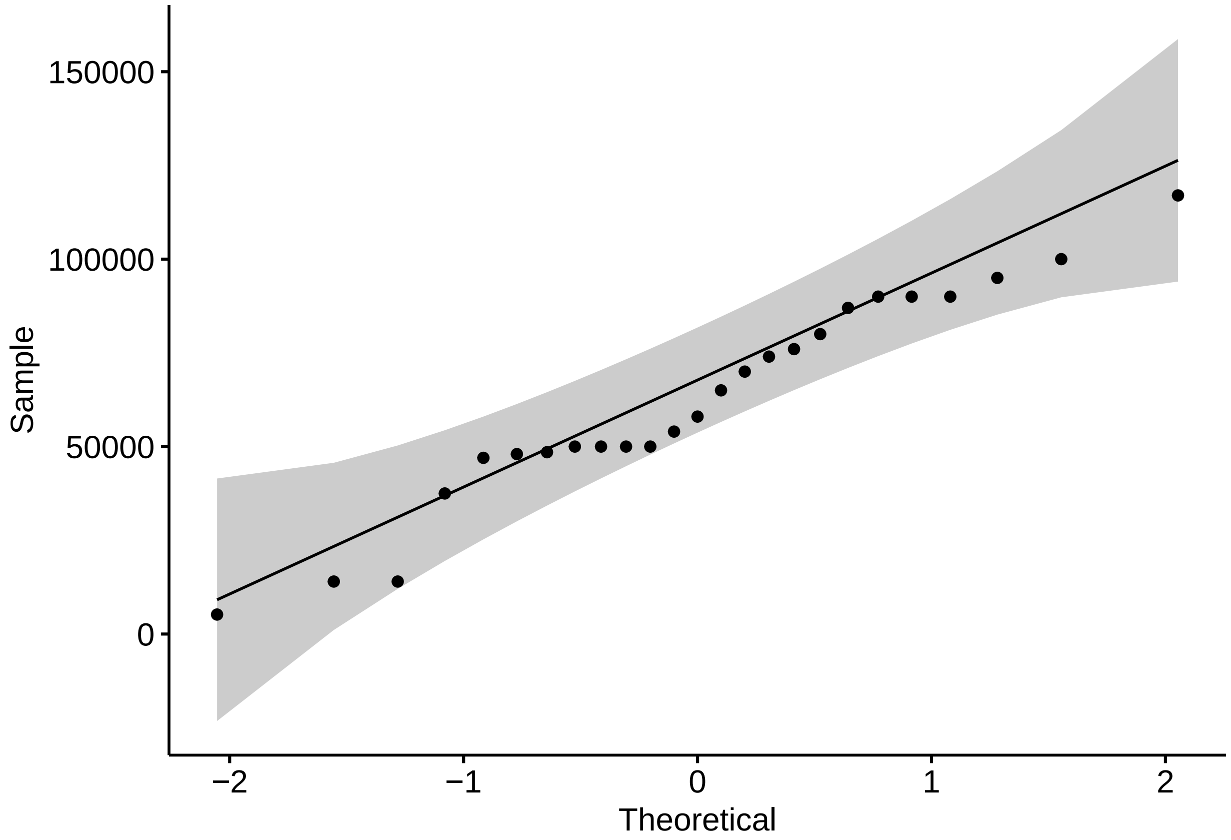
(5 points)

Are these samples consistent with Normality in the distributions of the two populations sampled? Please perform a visual check. Please perform Welch's test for the null hypothesis of equal means. Please interpret the results given the assessment of Normality.

```
ggqqplot(dat.7.10$incwage[dat.7.10$educ == 7])
```



```
ggqqplot(dat.7.10$incwage[dat.7.10$educ == 10])
```



Your code and answer here:

```r
# Perform Welch's T-test
welch_test <- t.test(incwage ~ educ, data = dat.7.10, var.equal = FALSE)
# Display the test results
welch_test
```

```
##
##  Welch Two Sample t-test
##
## data:  incwage by educ
## t = -2.1967, df = 47.617, p-value = 0.03294
## alternative hypothesis: true difference in means between group 7 and group 10 is not equal to 0
## 95 percent confidence interval:
##  -35061.182  -1546.818
## sample estimates:
##  mean in group 7 mean in group 10
##            44104            62408
```

Based on the Welch's T-test and assuming the distributions are approximately normal, there is evidence to suggest that individuals with 4 years of college education earn significantly more than those with just 1 year of college education in the Denver area. The p-value is 0.03294, which is less than the common significance level of 0.05. Therefore, we reject the null hypothesis that the means of the two groups (educ=7 and educ=10) are equal. The 95% confidence interval for the difference in means is between -35,061.182 and -1,546.818. Since this interval does not include 0, it further supports the conclusion that the difference in means is statistically significant.

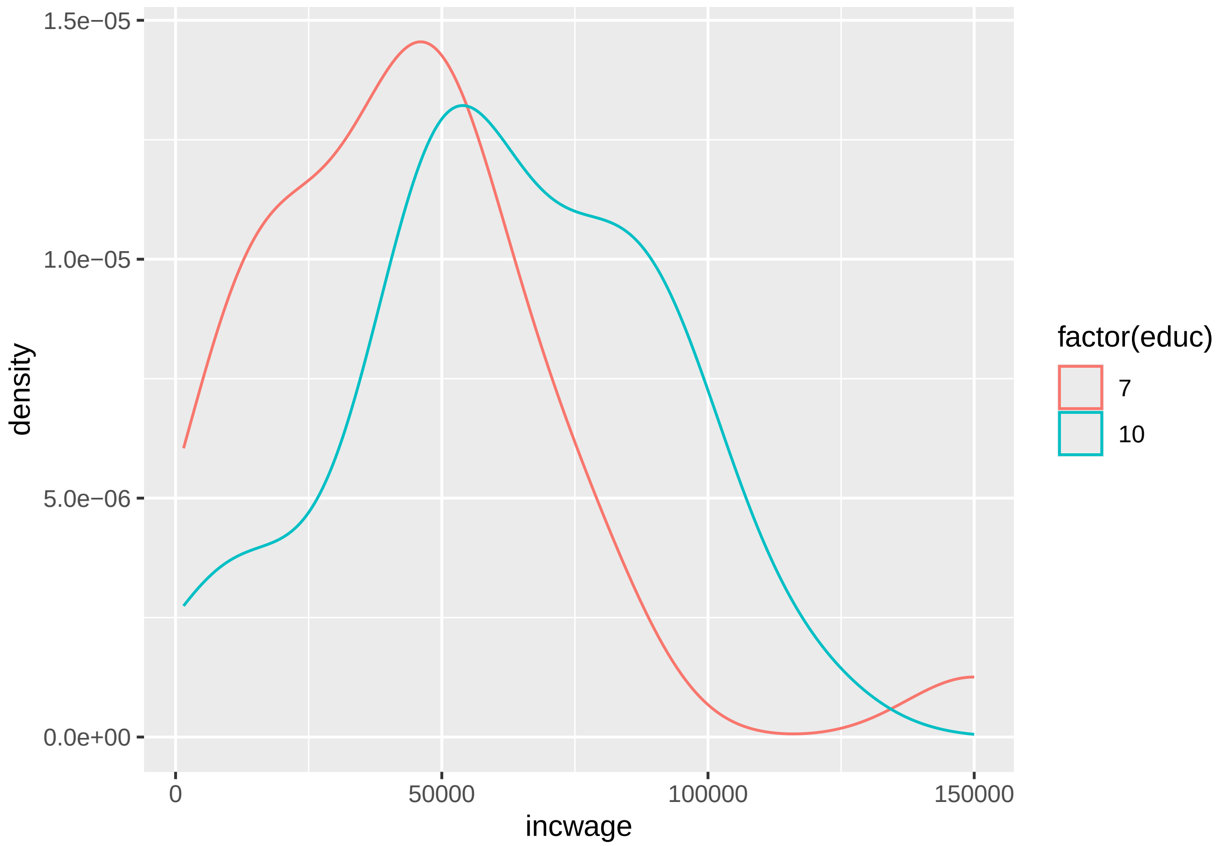
Question 2.2

(4 points)

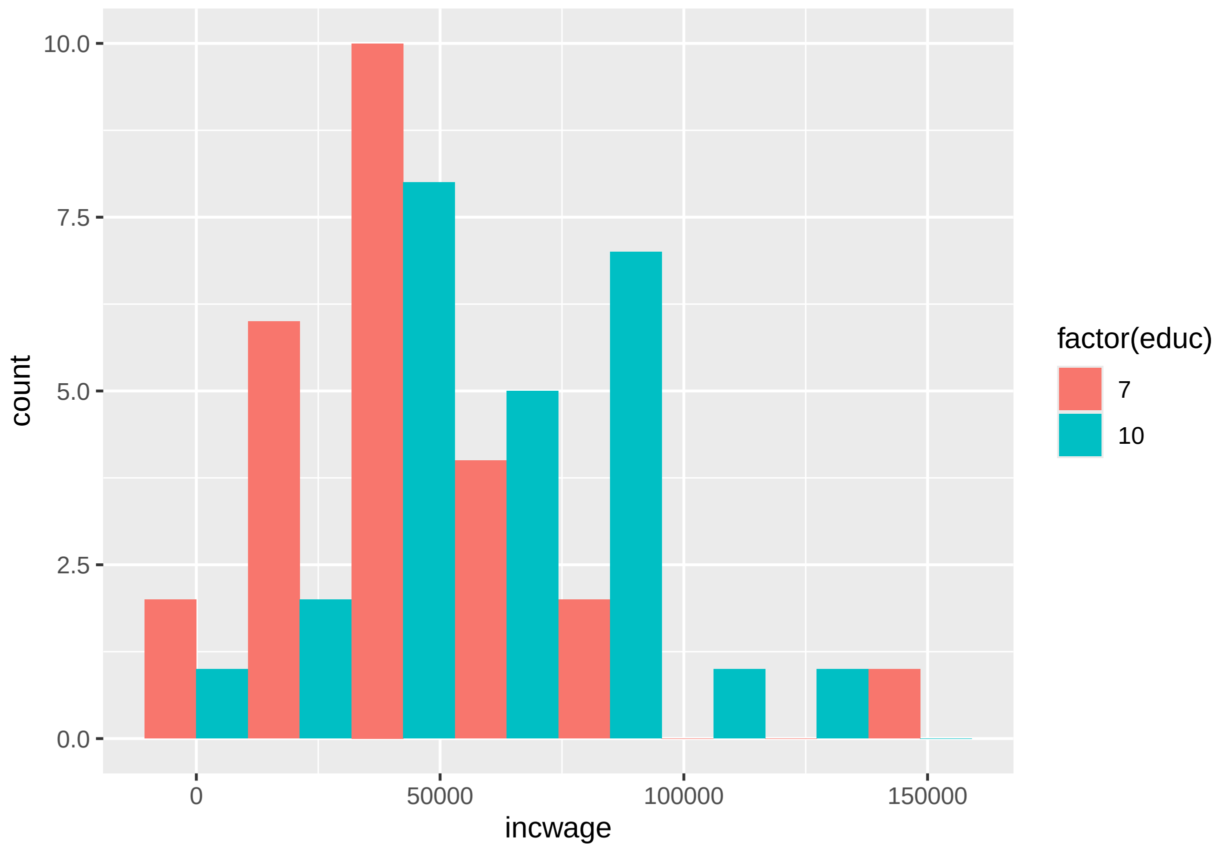
The interpretation of the Mann-Whitney U test depends on whether the distributions of the two populations sampled are viewed as related by translation (same distribution except shifted left or right by some constant). Thus a standard first step before applying the Mann-Whitney U test is to assess this assumption visually or on the basis of subject matter expertise. Given the plots below, which description best describes the relationship of the samples? Please justify/explain your choice.

- The plots clearly indicate that distributions of the two populations sampled are related by translation.
- The plots are somewhat consistent with the distributions of the two populations sampled being related by translation.
- The plots clearly indicate that distributions of the two populations sampled are not related by translation.

```
ggplot(dat.7.10,aes(x=incwage,color=factor(educ)))+geom_density()
```



```
ggplot(dat.7.10,aes(x=incwage,fill=factor(educ)))+geom_histogram(position="dodge",bins=8)
```



Your answer here: When distributions appear as a shift and mirror image of each other, this indicates that the distributions are not simply related by a horizontal translation. Instead, the mirroring suggests that the distributions have fundamentally different shapes or spreads. This visual evidence strongly suggests that the Mann-Whitney U test's underlying assumption of translation is violated in this case. The plots clearly indicate that distributions of the two populations sampled are not related by translation.

Question 2.3

(5 points)

Please run and interpret a Mann-Whitney U test comparing "incwage" for the observations with "educ" equal to 7 and "incwage" for the observations with "educ" equal to 10. In your interpretation, please address both the case in which you treat the distributions of the two populations as related by translation and the case in which you don't make this assumption.

Your code and answer here:

```r
# Perform Mann-Whitney U test (also known as Wilcoxon rank-sum test in R)
mann_whitney_test <- wilcox.test(incwage ~ educ, data = dat.7.10)
```

```
## Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
## compute exact p-value with ties
```

```r
# Display the test results
mann_whitney_test
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  incwage by educ
## W = 177.5, p-value = 0.008981
## alternative hypothesis: true location shift is not equal to 0
```

- Translation Assumption: The significant p-value suggests that the income distributions are different, likely due to a shift in median income between the two education levels.
- No Translation Assumption: The significant p-value indicates that the income distributions differ, potentially due to differences in their shapes, spreads, or other characteristics, not just a simple shift.

Question 2.4

(5 points)

Please run a Mann-Whitney U test comparing log(incwage) for the observations with "educ" equal to 7 and with "educ" equal to 10 and compare to the result in part a. Please explain what you observe about the two tests.

Your code and answer here:

```
# Log-transform the income variable
dat.7.10$log_incwage <- log(dat.7.10$incwage)
# Perform Mann-Whitney U test on the log-transformed income
mann_whitney_log_test <- wilcox.test(log_incwage ~ educ, data = dat.7.10)
```

```
## Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
## compute exact p-value with ties
```

```
# Display the test results
mann_whitney_log_test
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  log_incwage by educ
## W = 177.5, p-value = 0.008981
## alternative hypothesis: true location shift is not equal to 0
```

Both the original and log-transformed Mann-Whitney U tests indicate a statistically significant difference in incomes between the two education groups. The similarity in results suggests that the observed difference is consistent and not driven by skewness or outliers. This reinforces the conclusion that education level (1 year vs. 4 years of college) is associated with significantly different income levels in the Denver area, regardless of whether the income data is considered on its original scale or after log transformation.

Question 3

The raw data in this question is the "Pew Research Center's American Trends Panel" Wave 69 Field dates: June 16 – June 22, 2020 Topics: Coronavirus tracking, politics, 2020 Census data and questionnaire downloaded 3/4/2021 from https://www.pewresearch.org/politics/dataset/american-trends-panel-wave-69/

The codebook was downloaded 3/5/2021 from https://www.pewresearch.org/wp-content/uploads/2018/05/Codebook-and-instructions-for-working-with-ATP-data.pdf

The Pew Research Center provides sample weights in the variable "WEIGHT_W69". These serve a similar purpose to the "perwt" variable in the IPUMS data, though these weights have the effect of readjusting the proportions of demographic groups in the sample to be approximately the proportions in the target population when the responses are viewed as representing the number of people given by the weight. The weights add up to the number of responses in the study.

The code below draws a sample from the full response set with probability based on the weight. Please use "dat.sub" in the questions below. The data frame "dat.sub" is provided with the assignment.

```
# The file path will need to be adjusted for the local system's directory structure:
# dat.pew<-data.frame(read.spss("W69_Jun20/ATP W69.sav"))
# sum(dat.pew$WEIGHT_W69)
# set.seed(1234)
# sub.index<-sample(1:nrow(dat.pew),200,prob = dat.pew$WEIGHT_W69,replace=TRUE)
# dat.sub<-dat.pew[sub.index,]
# save(dat.sub,file="dat_sub.RData")
load("dat_sub.RData")
```

The code below generates a contingency table for the answers to the question "How much of a problem do you think each of the following are in the country today?" applied to the coronavirus outbreak by the age category of the respondent. The respondents who refused to supply their age or an answer to the question are omitted.

For intuition, the percent within each age range selecting each response is shown.

If you would prefer to investigate the independence of another pair of variables, you may generate your own contingency table and base your answers to part 1 and part 2 on your own table.

```
t<-table(dat.sub$F_AGECAT,dat.sub$NATPROBS_b_W69)
(t<-t[1:4,1:4]) # Drop the "Refused" row and column
```

```
##
##         A very big problem A moderately big problem A small problem
##   18-29                 21                        7               5
##   30-49                 31                       13              13
##   50-64                 29                       22               9
##   65+                   24                       12               4
##
##         Not a problem at all
##   18-29                    2
##   30-49                    4
##   50-64                    2
##   65+                      2
```

```
# percent of each row that lies in the corresponding column
round(100*t/rowSums(t),0)
```

```
##
##         A very big problem A moderately big problem A small problem
##   18-29                 60                       20              14
##   30-49                 51                       21              21
##   50-64                 47                         35               15
##   65+                   57                         29               10
##
##         Not a problem at all
##   18-29                    6
##   30-49                    7
##   50-64                    3
##   65+                      5
```

Question 3.1

(10 points)

Please use the χ^2 test to test the independence of the probability distribution with the outcomes in the rows and the probability distribution with the outcomes in the columns for the table you chose. Is this an appropriate test for your table? Explain why or why not. If it an appropriate test, are the data consistent with the null hypothesis that the row distribution and the column distribution are independent?

Your code and answer here:

```r
# Perform the Chi-Square test
chisq.test(t)
```

```
## Warning in chisq.test(t): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = 7.1205, df = 9, p-value = 0.6246
```

The data are consistent with the null hypothesis that the age distribution and the perceived problem level of the coronavirus outbreak are independent. In other words, there is no strong evidence from this test to suggest that the perception of the coronavirus outbreak as a problem varies significantly with age category. The chi-square test is typically appropriate for large samples with sufficient expected cell counts. If the expected counts in any cell are too low, the test's assumptions may be violated, and results may not be reliable. The expected frequency in each cell should generally be 5 or more. If some expected frequencies are below 5, the test might not be valid, and a different test (like Fisher's exact test) might be necessary.

Question 3.2

(10 points)

Please use Fisher's exact test to test the independence of the probability distribution with the outcomes in the rows and the probability distribution with the outcomes in the columns for the table you chose. Is this an appropriate test for your table? Please explain why or why not. If it is an appropriate test, are the data consistent with the null hypothesis that the row distribution and the column distributions are independent?

Note: Setting cache=TRUE means that the code in the block will not be reevaluated on subsequent "knit" applications unless the code in the block is changed. This speeds up text editing once the calculations are in place, but shouldn't be used before then because the calculations won't be updated to reflect changes elsewhere.

Your code and answer here:

```r
# Perform Fisher's Exact Test (Increase the workspace size)
fisher.test(t, workspace = 2e7)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t
## p-value = 0.6345
## alternative hypothesis: two.sided
```

The chi-square test was initially questionable due to the low expected frequencies in the contingency table, as indicated by the warning. Fisher's Exact Test is more appropriate in such cases because it does not rely on large-sample approximations and can handle small expected counts more effectively. Both tests (chi-square and Fisher's) indicate that there is no significant relationship between the respondents' age categories and their views on the severity of the coronavirus outbreak. The high p-values from both tests suggest that any observed differences in the table are likely due to random variation rather than a true association. Given these results, the data are consistent with the null hypothesis that the row distribution (age categories) and the column distribution (perceptions of the coronavirus problem) are independent. There is no strong evidence to suggest a significant relationship between age and perceptions of the coronavirus outbreak in this dataset.

End.