Final Exam

Jackson Dial

12/8/2021

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.4 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.1 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(epitools)  
library(abind)

# Question 1

Standard error is a measure of spread regarding a sampling distribution. It is also referred to as the standard deviation of a sampling distribution, and is interpreted as how the values of a statistic vary about the mean.

# Question 2

The two populations are not independent. This is because a WSR is a test for one-sample or paired data. Given that we are wanting to compare two populations, this means it must be a WSR for paired-data. Paired data are not independent of each other.

# Question 3

An experimental study can be non-randomized. An example could be a situation where during assignment of treatment group, assignment is based entirely on the time that an individual ‘signed-up’ for the study. Though it could be argued that this method of treatment assignment is random because you do not know at what time individuals will register, I would disagree. Time is not random, it follows a strict pattern. If it were to be random, you would expect approximately the same number of individuals to register between, say, 1 AM & 2 AM as would between 1 PM and 2 PM, assuming large enough sample size. This would not be the case however, so it IS possible for an experimental study to be non-randomized.

# Question 4

## Part A

This is a case-control study.

## Part B

The most appropriate test in this situation would be a Chi-Squared test. This is because McNemar’s test is for data of two populations that are matched or paired in some way. Fisher’s exact test is used when the sample sizes or expected counts are small , and since the smallest expected count is 7.5, we should use the Chi-Squared test.

# Question 5

## Part A

I do have concerns when examining the QQ-Plot. The tail at the top right shows a sizable skew and does not follow the line very well, which suggests evidence of non-normality.

## Part B

Yes, there is a significant association between discrimination and depression, due to the p-value associated with the variable ‘dailyDiscr’, after controlling for age, gender, income, and BMI.

## Part C

Those with “Income Poor” tend to have 6.08 units higher depression on average compared to those categorized as “Affluent”

## Part D

The interpretation of the intercept here is not meaningful, because it represents what our depression score would be if all other values are 0. One of those values is age, which would mean we are interpreting the depression score of an unborn individual, which is not meaningful.

## Part E

No. The model is not good for predicting depression. We can gather that from the values. Our Multiple is only 0.1207, which means that only 12.07% of the variability in our data is explained by our model. That is very low, and means the model is not very useful for predicting depression.

## Part F

The model is significantly better than an intercept-only model. This is shown by the F-statistic and accompanying p-value at the bottom of the output. The null hypothesis for that F-test is that all the parameters except for the intercept are jointly 0. Since we have a significant p-value, we reject that null and conclude that all the parameters except the intercept 0. Since the parameters are not jointly 0, they do add to the predictive capability of our model, thus allowing us to conclude that this model is significantly better than an intercept-only model.

## Part G

The sample size was 1577, calculated by the residual standard error degrees of freedom plus the number of parameters being estimated.

# Question 6

## Part A

Oij <- matrix(c(23,35,19,23), nrow = 2, byrow = TRUE)  
Oij

## [,1] [,2]  
## [1,] 23 35  
## [2,] 19 23

oddsratio.wald(Oij)

## $data  
## Outcome  
## Predictor Disease1 Disease2 Total  
## Exposed1 23 35 58  
## Exposed2 19 23 42  
## Total 42 58 100  
##   
## $measure  
## odds ratio with 95% C.I.  
## Predictor estimate lower upper  
## Exposed1 1.0000000 NA NA  
## Exposed2 0.7954887 0.3561063 1.777004  
##   
## $p.value  
## two-sided  
## Predictor midp.exact fisher.exact chi.square  
## Exposed1 NA NA NA  
## Exposed2 0.5840246 0.682116 0.5766448  
##   
## $correction  
## [1] FALSE  
##   
## attr(,"method")  
## [1] "Unconditional MLE & normal approximation (Wald) CI"

(23\*23) / (19\*35)

## [1] 0.7954887

The odds ratio point estimate is 0.795, relevant to the novel regimen. This verifies the claim that the novel regimen has lower odds of SVR compared to the gold standard. This means the odds of SVR == Yes in the novel treatment group are approximately 21% lower than in the gold standard regimen group.

## Part B

#Try MH Test?  
# S1 <- matrix(c(13, 8, 17, 12), nrow=2, byrow = TRUE)  
# S2 <- matrix(c(10, 27, 2, 11), nrow=2, byrow = TRUE)  
# mantelhaen.test(abind(S1,S2, along=3))  
  
#It shouldn't be the MH test because it test uses separate data, not paired or correlated data  
  
#Pretty sure I am supposed to use mcnemar because this is a two sample test of proportions where populations are matched/paired, and the question said nothing relative to SVR  
  
#mcnemar.test(Oij)  
  
#I actually don't think so, because mcnemar's test is for paired, and we can't determine from the information given which observations are paired.  
  
# mymat <- matrix(c(29, 31, 13, 37), nrow = 2, byrow = TRUE)  
# oddsratio.wald(mymat)  
  
# This also doesn't make sesne in context of part c wanting to relate back to part b.... am I supposed to still consider the SVR in part b?  
  
# I think so. The one that makes the most sense to me is the MH test. BreslowDay would work but Tina never showed us how to do that and I am not allowed to look it up.  
  
S1 <- matrix(c(13, 8, 17, 12), nrow=2, byrow = TRUE)  
S2 <- matrix(c(10, 27, 2, 11), nrow=2, byrow = TRUE)  
mantelhaen.test(abind(S1,S2, along=3))

##   
## Mantel-Haenszel chi-squared test with continuity correction  
##   
## data: abind(S1, S2, along = 3)  
## Mantel-Haenszel X-squared = 0.21796, df = 1, p-value = 0.6406  
## alternative hypothesis: true common odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.549523 3.566729  
## sample estimates:  
## common odds ratio   
## 1.4

The common odds ratio here is 1.4, which means the odds of SVR == Yes in the novel treatment group is approximately 1.4 times the odds in the gold standard treatment group.

## Part C

A possible reason that the association is in one direction within physician but the opposite direction across physicians is that physician is a confounding variable.

# Question 7

## Part A

**TRUE**

## Part B

**TRUE**

## Part C

**TRUE**

## Part D

**TRUE**

## Part E

**TRUE**