STAA 552: HW 2

YOUR NAME HERE

See Canvas Calendar for due date.
54 points total.
Content for Q1-Q8 is from section 02.
Content for Q9-Q14 is from section 03.
Add or delete code chunks as needed.
For full credit, your numeric answers should be clearly labeled, outside of the R output.

Diabetes 2 (Q1 - Q6)

A researcher is interested in estimating the proportion of US adults who have been diagnosed with diabetes (π). Consider a study where n = 65 people are contacted and asked whether they have been diagnosed with diabetes. Suppose this is based on a random sample of US adults. From this sample, 7 people have been diagnosed with diabetes.

Q1 (2 pts)

Calculate $\hat{\pi}$.

Response

```
## [1] 0.1076923
```

Q2 (4 pts)

Discuss whether the large sample normal approximation is reasonable here. Use the rule of thumb from the course notes.

Response

This should be sufficient. We generally want to see two things:

1. n*pihat > 5 TRUE
2. n*(1-pihat) > 5 True

In our case both 7 and 58 are greater than 5 *****

Q3 (2 pts)

Use `prop.test( , correct = TRUE)` to calculate an (approximate) 95% confidence interval for π. The resulting interval is generated using using Wilson's score method. You can just show the R output for full credit.

```
##
##  1-sample proportions test with continuity correction
##
## data:  7 out of 65, null probability 0.5
## X-squared = 38.462, df = 1, p-value = 5.584e-10
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.04802157 0.21530910
## sample estimates:
##         p
## 0.1076923
```

Q4 (4 pts)

Using the confidence interval from the previous question, interpret the interval in context. Aim for an interpretation that could be used in a write-up: avoid abbreviations and statistical jargon.

Response

Based on the confidence interval shown above, we have 95% confidence that between 4.8 and 21.5% of adults in the US have been diagnosed with diabetes. *****

Q5 (4 pts)

Construct an (approximate) 95% Wald confidence interval for π. Do this "by hand" (using R as a calculator) and echo your R code.

Response

```r
#Q5

z <- qnorm(0.975)

pihat <- 7/65

lower <- pihat - z*sqrt((pihat*(1-pihat))/n)
upper <- pihat + z*sqrt((pihat*(1-pihat))/n)

cat("Wald confidence interval is", lower, "-", upper, "\n")
```

```
## Wald confidence interval is 0.03233228 - 0.1830523
```

Q6 (2 pts)

Use `binom.test()` to calculate the exact 95% confidence interval for π. You can just show the R output for full credit.

```
##
##  Exact binomial test
##
## data:  7 and 65
## number of successes = 7, number of trials = 65, p-value = 4.271e-11
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.04440866 0.20938160
## sample estimates:
## probability of success
##              0.1076923
```

Bomb Hits (Q7 - Q8)

This data is taken from Ott&Longnecker and originally from Hand (1993). During WWII, bomb hits were recorded in the n = 576 grids in a map of a region of South London. If the bomb hits were purely random, a Poisson model would produce the number of hits per grid.

# Bomb Hits	0	1	2	3	4+
# Grids	229	211	93	35	8

Q7 (2 pts)

Calculate $\hat{\mu}$, the sample mean hits per grid.
Notes: (1) To do this calculation find the total number of bomb hits and divide by the number of grids. (2) For the calculation, treat 4+ as 4. (3) Some example R code has been provided to get you started. But other approaches are allowed.

Response

```
## [1] 0.9270833
```

Q8 (8 pts)

Use the Pearson chi-square GOF test to test whether the number of bomb hits per grid follows the Poisson distribution. Calculate the test statistic, df, p-value and give a conclusion (in context) using α = 0.05.
Notes: (1) Do this "by hand" (using R as a calculator) but you can double check the test statistic using `chisq.test()`. (2) When calculating Poisson probabilities, you will use your $\hat{\mu}$ value from the previous question. (3) When calculating Poisson probabilities, the last value should represent P($Y \geq 4$). This can be done "manually". (4) When calculating df, we "lose" 1 df because estimated $\hat{\mu}$.

Response

```
## [1] 1.133167
```

```
## [1] 0.7690742
```

```
## Fail to reject the null hypothesis as our P value is:  0.7690742 .
##  There is no evidence against the bomb hits following a Poisson distribution.
```

Auto Accidents (Q9 - Q14)

CDA problem 2.4 concerns fatality results for drivers and passengers in auto accidents in Florida in 2008, according to whether the person was wearing a seat belt.

```
##             Nonfatal Fatal
## YesSeatBelt   441239   703
## NoSeatBelt     55623  1085
```

Q9 (3 pts)

Estimate (a) the marginal probability of **fatality**. Also estimate the probability of fatality, conditional on seat-belt use in category (b) yes and (c) no.

Response

```
## The Marginal probability of Fatality is,  0.003585681
```

```
## The Marginal probability of Fatality given that you are wearing a seatbelt is, 0.001590706
```

```
## The Marginal probability of Fatality given that you are not wearing a seatbelt is, 0.0191331
```

Q10 (3 pt)

Estimate (a) the marginal probability of **wearing a seatbelt**. Also estimate the probability of wearing a seatbelt, conditional on the injury being (b) Nonfatal and (c) Fatal.

Response

```
## Probability of wearing a seatbelt is,  0.8862769
```

```
## Probability of wearing a seatbelt given that you were in a non fatal crash is, 0.8880514
```

```
## Probability of wearing a seatbelt given that you were in a Fatal crash is, 0.3931767
```

Auto Accidents continued (Q11 - Q13)

Use injury as the response variable (with fatal injury being the "event of interest") and compare those not wearing seatbelts vs those wearing seatbelts.

Q11 (6 pts)

Provide an estimate of the **difference in proportions** and the corresponding (Wald) 95% confidence interval. Use the interval to make a conclusion about association between seat belt use and injury. Be sure to mention the direction of association. (By direction of association, I just mean which group has the higher probability (risk) of fatal injury.)

Response

```
## Difference in proportions: 0.0175424
```

```
## 95% Confidence interval: ( 0.01640877 , 0.01867602 )
```

Q12 (6 pts)

Provide an estimate of the **relative risk** and the corresponding (Wald) 95% confidence interval. Use the interval to make a conclusion about association between seat belt use and injury. Be sure to mention the direction of association. (By direction of association, I just mean which group has the higher probability (risk) of fatal injury.)

Response

```
## Relative Risk: 12.02805
```

```
## 95% Confidence interval: ( 10.93914 , 13.22536 )
```

Q13 (6 pts)

Provide an estimate of the **odds ratio** and the corresponding (Wald) 95% confidence interval. Use the interval to make a conclusion about association between seat belt use and injury. Be sure to mention the direction of association. (By direction of association, I just mean which group has the higher odds of fatal injury.)

Response

```
## Odds Ratio: 12.24317
```

```
## 95% Confidence interval: ( 11.13023 , 13.46739 )
```

Q14 (2 pts)

Considering the results of the previous two questions, why are the relative risk and odds ratio approximated equal?

Response

The relative risk and odds ratio are approximately equal because fatal injury is rare in both groups. When the outcome is rare, the odds and the probabilities of the event are similar, making the odds ratio a close approximation of the relative risk.

Appendix

```r
#Retain this code chunk!!!
#install.packages("tinytex")
#install.packages("kableExtra")

library(knitr)
library(kableExtra)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
#Q1
n <- 65
t <- 7

pihat <- t/n

print(pihat)
#Q3

prop.test(x = 7, n = 65, correct = TRUE)

#Q5

z <- qnorm(0.975)

pihat <- 7/65

lower <- pihat - z*sqrt((pihat*(1-pihat))/n)
upper <- pihat + z*sqrt((pihat*(1-pihat))/n)

cat("Wald confidence interval is", lower, "-", upper, "\n")


#Q6

binom.test(x = 7, n = 65)


#Q7
Bombs <- data.frame(Y = seq(from = 0, to = 4),
                    Obs = c(229, 211, 93, 35, 8))

total <- sum(Bombs$Y*Bombs$Obs)
n <- sum(Bombs$Obs)

mu <- total/n

print(mu)

#Q8
O <- Bombs$Obs

muhat <- total / n

E0 <- dpois(0, lambda<- muhat) * n
```

```r
E1 <- dpois(1, lambda<- muhat) * n
E2 <- dpois(2, lambda<- muhat) * n
E3 <- dpois(3, lambda<- muhat) * n
E4 <- dpois(4, lambda<- muhat) * n

E <- c(E0, E1, E2, E3, E4)

Chisq <- sum((O-E)^2/E)
Chisq

df <- length(O) - 2
p <- 1 - pchisq(Chisq, df)
p

if (p < 0.05) {
  cat("Reject the null hypothesis as our P value is: ", p, ".\n There is evidence that the bomb hits do
} else {
  cat("Fail to reject the null hypothesis as our P value is: ", p, ".\n There is no evidence against th
}

#Q9-Q14
Auto <- matrix(c(441239, 703,
                 55623, 1085), nrow = 2, byrow = TRUE)
rownames(Auto) <- c("YesSeatBelt", "NoSeatBelt")
colnames(Auto) <- c("Nonfatal", "Fatal")
Auto
#Q9

Total <- sum(Auto)

TotalFatal <- sum(Auto[,"Fatal"])
marginalFatality <- TotalFatal/Total
cat("The Marginal probability of Fatality is, ", marginalFatality, "\n")

TotalSeatBelt <- sum(Auto["YesSeatBelt", ])
FatalSeatBelt <- Auto["YesSeatBelt", "Fatal"]
ProbFatalGivenYes <- FatalSeatBelt / TotalSeatBelt
cat("The Marginal probability of Fatality given that you are wearing a seatbelt is,", ProbFatalGivenYes

TotalSeatBelt <- sum(Auto["NoSeatBelt", ])
FatalNoSeatBelt <- Auto["NoSeatBelt", "Fatal"]
ProbFatalGivenNo <- FatalNoSeatBelt / TotalSeatBelt
cat("The Marginal probability of Fatality given that you are not wearing a seatbelt is,", ProbFatalGiver

#Q10

TotalSeatBelt <- sum(Auto["YesSeatBelt", ])
MarginalProbSeatBelt <- TotalSeatBelt / Total
cat("Probability of wearing a seatbelt is, ", MarginalProbSeatBelt, "\n")


TotalNonfatal <- sum(Auto[, "Nonfatal"])
SeatBeltNonfatal <- Auto["YesSeatBelt", "Nonfatal"]
```

```r
ProbSeatBeltNonfatal <- SeatBeltNonfatal / TotalNonfatal
cat("Probability of wearing a seatbelt given that you were in a non fatal crash is,", ProbSeatBeltNonfa


TotalFatal <- sum(Auto[, "Fatal"])
SeatBeltFatal <- Auto["YesSeatBelt", "Fatal"]
ProbSeatBeltFatal <- SeatBeltFatal / TotalFatal
cat("Probability of wearing a seatbelt given that you were in a Fatal crash is,", ProbSeatBeltFatal, "\n
#Q11

fatal_no_seatbelt <- Auto["NoSeatBelt", "Fatal"]
nonfatal_no_seatbelt <- Auto["NoSeatBelt", "Nonfatal"]
n_no_seatbelt <- fatal_no_seatbelt + nonfatal_no_seatbelt

fatal_yes_seatbelt <- Auto["YesSeatBelt", "Fatal"]
nonfatal_yes_seatbelt <- Auto["YesSeatBelt", "Nonfatal"]
n_yes_seatbelt <- fatal_yes_seatbelt + nonfatal_yes_seatbelt

p_no_seatbelt <- fatal_no_seatbelt / n_no_seatbelt
p_yes_seatbelt <- fatal_yes_seatbelt / n_yes_seatbelt

D <- p_no_seatbelt - p_yes_seatbelt

SE <- sqrt(
  (p_no_seatbelt * (1 - p_no_seatbelt) / n_no_seatbelt) +
  (p_yes_seatbelt * (1 - p_yes_seatbelt) / n_yes_seatbelt)
)
z <- qnorm(0.975)

ME <- z * SE


Lower <- D - ME
Upper <- D + ME

cat("Difference in proportions:", D, "\n")
cat("95% Confidence interval: (", Lower, ",", Upper, ")\n")

#Q12
RR <- p_no_seatbelt / p_yes_seatbelt

SE_lnRR <- sqrt((1 / fatal_no_seatbelt) + (1 / fatal_yes_seatbelt))
z <- qnorm(0.975)
ME_lnRR <- z * SE_lnRR

lnRR <- log(RR)
Lower_lnRR <- lnRR - ME_lnRR
Upper_lnRR <- lnRR + ME_lnRR


Lower_RR <- exp(Lower_lnRR)
Upper_RR <- exp(Upper_lnRR)
```

```r
cat("Relative Risk:", RR, "\n")
cat("95% Confidence interval: (", Lower_RR, ",", Upper_RR, ")\n")

#Q13

a <- Auto["NoSeatBelt", "Fatal"]
b <- Auto["YesSeatBelt", "Fatal"]
c <- Auto["NoSeatBelt", "Nonfatal"]
d <- Auto["YesSeatBelt", "Nonfatal"]

# Odds Ratio
OR <- (a * d) / (b * c)

# Standard error of ln(OR)
SE_lnOR <- sqrt(1/a + 1/b + 1/c + 1/d)

# 95% Confidence interval for ln(OR)
z <- qnorm(0.975)
ME_lnOR <- z * SE_lnOR

lnOR <- log(OR)
Lower_lnOR <- lnOR - ME_lnOR
Upper_lnOR <- lnOR + ME_lnOR

# Transform back to OR
Lower_OR <- exp(Lower_lnOR)
Upper_OR <- exp(Upper_lnOR)

# Output
cat("Odds Ratio:", OR, "\n")
cat("95% Confidence interval: (", Lower_OR, ",", Upper_OR, ")\n")
```