Stat511 Hw10

Amy Fox

11/19/2019

Question 1

Bacillus Calmette-Guerin (BCG) is a vaccine for preventing tuberculosis. For this question, we will examine data from 3 studies (Vandiviere et al 1973, TPT Madras 1980, Coetzee & Berjak 1968). The data is summarized below.

A. Calculate the odds ratio (corresponding to TB_pos for Trt vs Ctrl) for each study separately. (4 pts)

```
library(lawstat)
BCG_response <- array(c(619, 2537, 10, 8,
        87892, 87886, 499, 505,
        7232, 7470, 45, 29),
      dim = c(2, 2, 3),
      dimnames = list(Treatment = c("Ctrl", "Trt"),
                      Response = c("TB_neg", "TB_pos"),
                      Study = c("1", "2", "3")))

cmh.test(BCG_response)
```

```
##
##  Cochran-Mantel-Haenszel Chi-square Test
##
## data:  BCG_response
## CMH statistic = 0.53072, df = 1.00000, p-value = 0.46631, MH
## Estimate = 0.95700, Pooled Odd Ratio = 0.95685, Odd Ratio of level
## 1 = 0.19519, Odd Ratio of level 2 = 1.01210, Odd Ratio of level 3
## = 0.62391
```

OR Study 1: 0.19519

OR Study 2: 1.01210

OR Study 3: 0.62391

B. Use the Breslow-Day test to test for equality of odds ratios across the 3 studies. State your p-value and conclusion. Can we conclude that the odds ratios are equal across the 3 studies? Based on this test, should be combine information across studies? (4 pts)

A note about the BCG vaccine from Wikipedia:

The most controversial aspect of BCG is the variable efficacy found in different clinical trials that appears to depend on geography. Trials conducted in the UK have consistently shown a protective effect of 60 to 80%, but those conducted elsewhere have shown no protective effect, and efficacy appears to fall the closer one gets to the equator.

```
library(metafor)
cmh_bd_BCG <- rma.mh(ai = BCG_response[1, 1, ], bi =  BCG_response[1, 2, ],
       ci = BCG_response[2, 1, ], di = BCG_response[2, 2, ])

p_value <- cmh_bd_BCG$BDp
```

p-value $= 1.4567539 \times 10^{-4}$

Because the p-value is less than $\alpha = 0.05$, we reject H0 and assume that the odds ratios are not equal across the 3 studies. Therefore, we should not combine information across studies.

Question 2

Problem 10.36 involves bomb hits during WWII. Bomb hits were recorded in n = 576 grids in a map of a region of South London.

A. Find the sample mean ($\hat{\mu}$) bomb hits per grid.

```
bomb_hits = c(0, 1, 2, 3, 4) # Y aka events
grids = c(229, 211, 93, 35, 8) # obs aka number of events aka units

cbind(bomb_hits, grids)
```

```
##      bomb_hits grids
## [1,]         0   229
## [2,]         1   211
## [3,]         2    93
## [4,]         3    35
## [5,]         4     8
```

```
mu_hat <- sum(bomb_hits*grids)/sum(grids)
```

Sample mean: 0.9270833

B. Use the GOF test to test whether the number of bomb hits per grid follows the Poisson distribution. Calculate the GOF test statistic, df, p-value and give a conclusion using $\alpha = 0.05$. (6 pts)

```
# calculate Poisson probability
pois_dist <- dpois(bomb_hits, mu_hat)

# change so that the poisson adds up to 1
sum_pois <- sum(pois_dist)
pois_dist[5] <- 1-sum(pois_dist[1:4])

# expected values
Exp <- pois_dist*576

# contributions of Chi-squared
X2 <- (grids-Exp)^2/Exp

cbind(bomb_hits, grids, pois_dist, Exp, X2)
```

```
##      bomb_hits grids  pois_dist        Exp           X2
## [1,]         0   229 0.39570617 227.926755 0.0050536183
## [2,]         1   211 0.36685260 211.307096 0.0004463069
## [3,]         2    93 0.17005146  97.949643 0.2501180049
## [4,]         3    35 0.05255063  30.269161 0.7393941810
## [5,]         4     8 0.01483914   8.547345 0.0350502750
```

```
Chi_Sq_TS <- sum(X2)


df = 5-2
```

```r
pval <- 1-pchisq(Chi_Sq_TS, df)
pval
```

```
## [1] 0.7939783
```

GOF Test statistic: 1.0300624

df: 3

p-value: 0.7939783

Because the p-value $> \alpha = 0.05$, we fail to reject H0 and conclude that there is no evidence against Poisson aka it follows a poisson distribution.

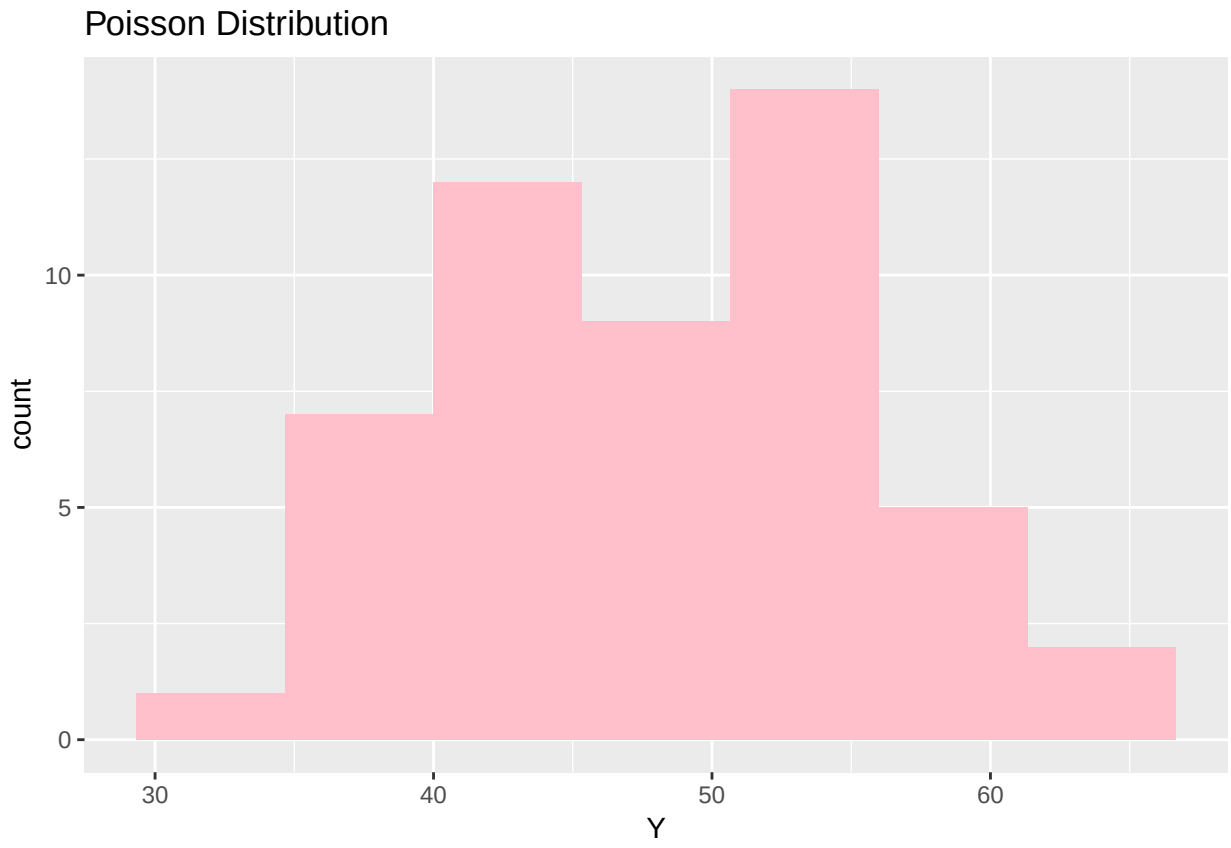
Question 3

The data "PoissonData.csv" gives observations Y (counts or events) for n = 50 (units) generated from the Poisson distribution (using the rpoiss() function).

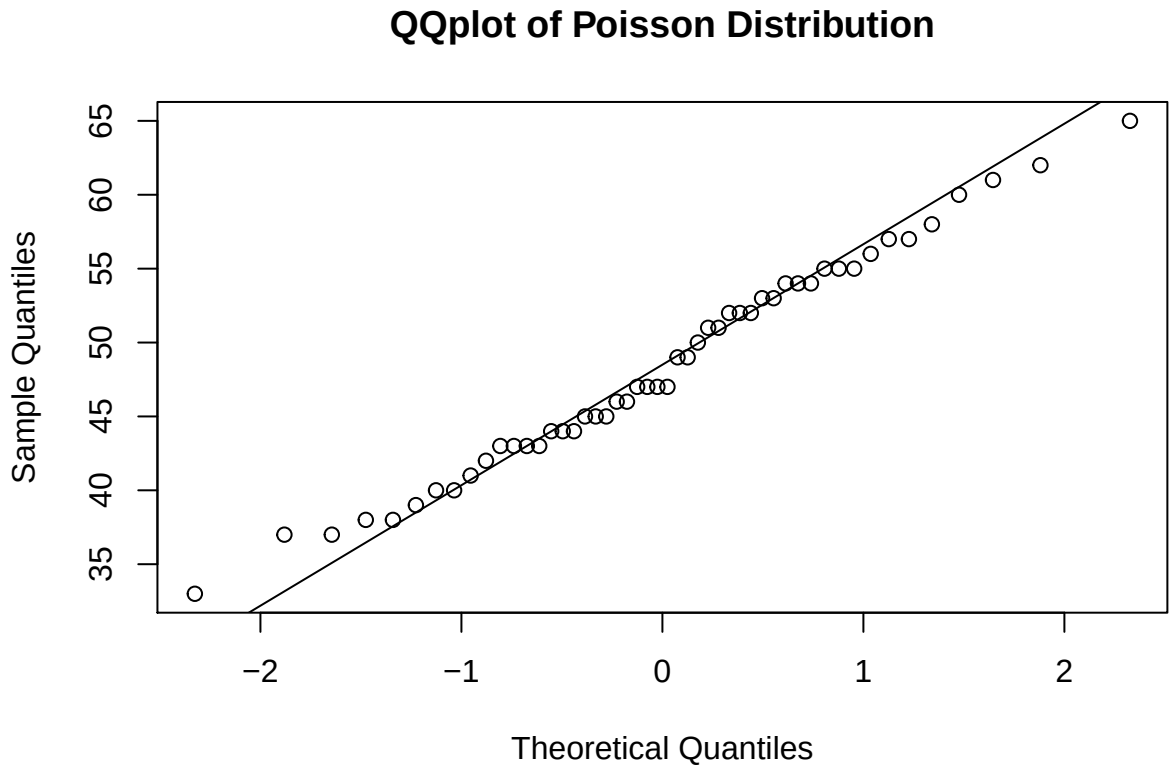
A. Calculate the sample mean and sample standard deviation. Also construct a histogram and qqplot of the data and include them in your assignment. (4 pts)

NOTE: Because the data comes from the Poisson distribution, you should find that the mean and the sample variance (s^2) are close. However, you should also find from the histogram and qqplot that the data looks approximately normal.

```r
library(ggplot2)
poisson_data <- read.csv("../Data/PoissonData.csv")

# mean
poisson_mean <- mean(poisson_data$Y)

# standard deviation
poisson_sd <- sd(poisson_data$Y)

# histogram
ggplot(poisson_data, aes(Y)) +
  geom_histogram(fill = "pink", bins = 7) +
  ggtitle("Poisson Distribution")
```




```r
# qqplot
qqnorm(poisson_data$Y, main = "QQplot of Poisson Distribution"); qqline(poisson_data$Y)
```



The mean is 48.38.

The standard deviations is 7.342607.

Based on the histogram and qqplot, the data looks pretty normal.

B. Give a standard t-based 95% confidence interval for μ.

```
t.test(poisson_data$Y)
```

```
##
##  One Sample t-test
##
## data:  poisson_data$Y
## t = 46.591, df = 49, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  46.29325 50.46675
## sample estimates:
## mean of x
##     48.38
```

CI: (46.29, 50.47)

C. Following the example on CH10 Slide 106 (Death by Mule Kick CI), construct a 95% confidence interval for μ based on the normal approximation to the Poisson distribution. (4 pts) In order to do this, you will start by constructing a CI on the total number of events, then divide by the number of units.

NOTE: The CIs from parts B and C should be similar.

$$(y \pm z_{0.025} * \sqrt{y})/n$$

```
y = sum(poisson_data$Y)
s = sd(poisson_data$Y)

CI_low <- (y - 1.96*sqrt(y))/50
CI_high <- (y + 1.96*sqrt(y))/50
```

CI: (46.4520134, 50.3079866)