STAT 511A Homework 10

Kathleen Wendt

11/27/2019

Load packages

```r
library(epitools)
library(metafor)
```

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 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.

Create data tables

Study 1 data

```r
tb_study1 <- matrix(c(619, 10, 2537, 8), nrow = 2, byrow = TRUE)
colnames(tb_study1) <- c("tbneg", "tbpos")
rownames(tb_study1) <- c("ctrl", "trt")
tb_study1
```

```
##      tbneg tbpos
## ctrl   619    10
## trt   2537     8
```

Study 2 data

```r
tb_study2 <- matrix(c(87892, 499, 87886, 505), nrow = 2, byrow = TRUE)
colnames(tb_study2) <- c("tbneg", "tbpos")
rownames(tb_study2) <- c("ctrl", "trt")
tb_study2
```

```
##      tbneg tbpos
## ctrl 87892   499
## trt  87886   505
```

Study 3 data

```
tb_study3 <- matrix(c(7232, 45, 7470, 29), nrow = 2, byrow = TRUE)
colnames(tb_study3) <- c("tbneg", "tbpos")
rownames(tb_study3) <- c("ctrl", "trt")
tb_study3
```

```
##      tbneg tbpos
## ctrl  7232    45
## trt   7470    29
```

Part 1A

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

Study 1 odds ratio

```
tb1_odds <- epitools::oddsratio(tb_study1, method = "wald")
```

```
## Warning in chisq.test(xx, correct = correction): Chi-squared approximation
## may be incorrect
```

```
tb1_odds$measure
```

```
##                          NA
## odds ratio with 95% C.I.  estimate     lower     upper
##                     ctrl 1.0000000        NA        NA
##                     trt  0.1951912 0.0767186 0.4966148
```

Study 2 odds ratio

```
tb2_odds <- epitools::oddsratio(tb_study2, method = "wald")
tb2_odds$measure
```

```
##                          NA
## odds ratio with 95% C.I. estimate     lower    upper
##                     ctrl 1.000000        NA       NA
##                     trt  1.012093 0.8940029 1.145782
```

Study 3 odds ratio

```
tb3_odds <- epitools::oddsratio(tb_study3, method = "wald")
tb3_odds$measure
```

```
##                          NA
## odds ratio with 95% C.I.  estimate     lower     upper
##                     ctrl 1.0000000        NA        NA
##                     trt  0.6239119 0.3907892 0.9961027
```

Part 1B

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 we combine information across studies? (4 pts)

Create data array

```
tb <- array(c(619, 2537, 10, 8,
              87892, 87886, 499, 505,
              7232, 7470, 45, 29),
            dim = c(2, 2, 3),
            dimnames = list(group = c("ctrl", "trt"),
                            response = c("tbneg", "tbpos"),
                            study = c("1", "2", "3")))
tb
```

```
## , , study = 1
##
##       response
## group  tbneg tbpos
##   ctrl   619    10
##   trt   2537     8
##
## , , study = 2
##
##       response
## group  tbneg tbpos
##   ctrl 87892   499
##   trt  87886   505
##
## , , study = 3
##
##       response
## group  tbneg tbpos
##   ctrl  7232    45
##   trt   7470    29
```

```
cmh <- metafor::rma.mh(ai = tb[1, 1,],
                       bi = tb[1, 2,],
                       ci = tb[2, 1,],
                       di = tb[2, 2,])
cmh$BDp
```

```
## [1] 0.0001456754
```

The Breslow-Day Test for equality of odds ratios of three tuberculosis studies yielded a p-value $< .001$. There is evidence to suggest a difference between odds ratios by study; it is not appropriate to 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.

Create data list

```
# observations
obs <- c(229, 211, 93, 35, 8)
y <- seq(from = 0, to = 4, by = 1)
bombs <- cbind(y, obs)
bombs
```

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

Part 2A

Find the sample mean (μ) bomb hits per grid.

```
# mean
muhat <- sum(obs*y)/sum(obs)
muhat
```

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

$\hat{\mu} = 0.9270833$ bomb hits per grid.

Part 2B

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 the corresponding Poisson Probabilities
prob <- dpois(y, muhat)
prob
```

```
## [1] 0.39570617 0.36685260 0.17005146 0.05255063 0.01217970
```

```
length(prob)
```

```
## [1] 5
```

```
sum(prob)
```

```
## [1] 0.9973406
```

```
# "fix" the final entry so that the probabilities sum to 1
prob[5] <- 1 - sum(prob[1:4])
prob
```

```
## [1] 0.39570617 0.36685260 0.17005146 0.05255063 0.01483914
```

```r
length(prob)
```

```
## [1] 5
```

```r
sum(prob)
```

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

```r
# calculate expected values and contributions to chi-square test statistic
exp <- prob*576
x2 <- (obs-exp)^2/exp
cbind(y, obs, prob, exp, x2)
```

```
##      y obs       prob        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
```

```r
# run GOF test
gof_chi <- sum(x2)
gof_chi
```

```
## [1] 1.030062
```

```r
gof_df <- 5-2
```

```r
gof_p <- 1 - pchisq(gof_chi, gof_df)
gof_p
```

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

We fail to reject the null hypothesis that data are from a Poisson distribution, $p = 0.7939783 > \alpha = 0.05$.

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).

```r
poisson <- readr::read_csv("PoissonData.csv")
```

```
## Parsed with column specification:
## cols(
##   Y = col_double()
## )
```

Part 3A

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)

Sample mean

```r
mu <- mean(poisson$Y)
mu
```

```
## [1] 48.38
```

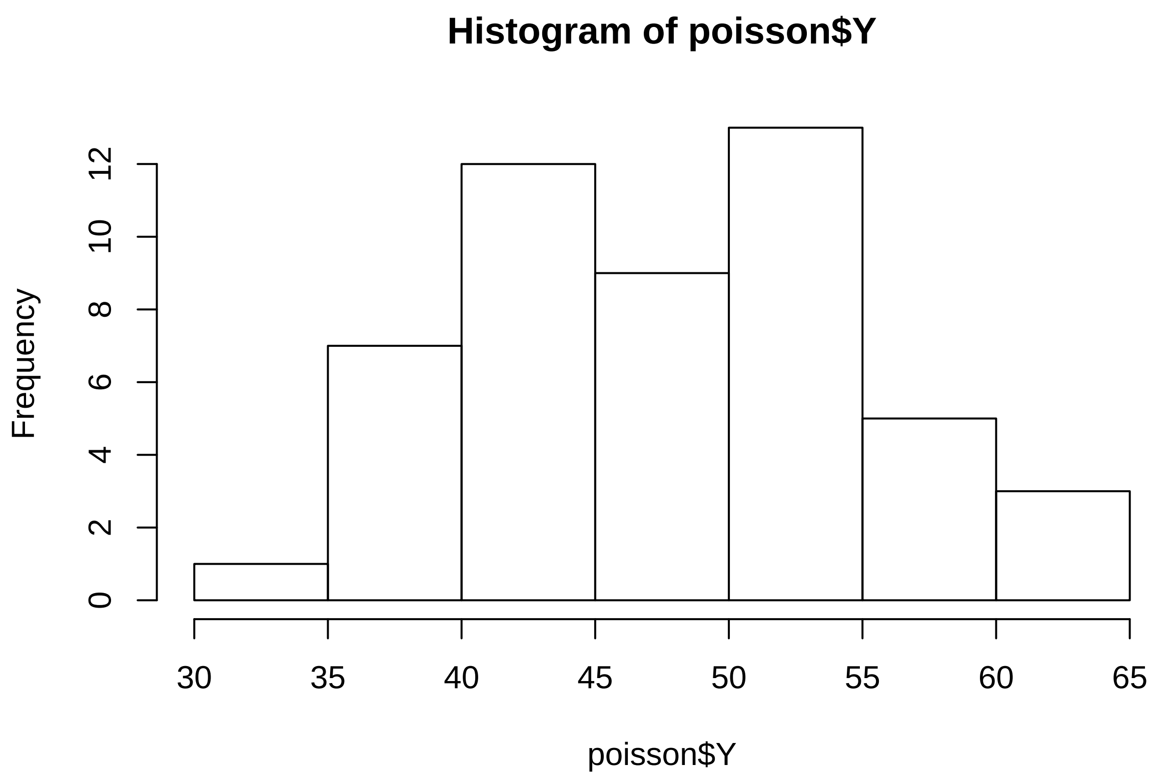
Sample standard deviation

```r
s <- sqrt(mu)
s
```

```
## [1] 6.955573
```

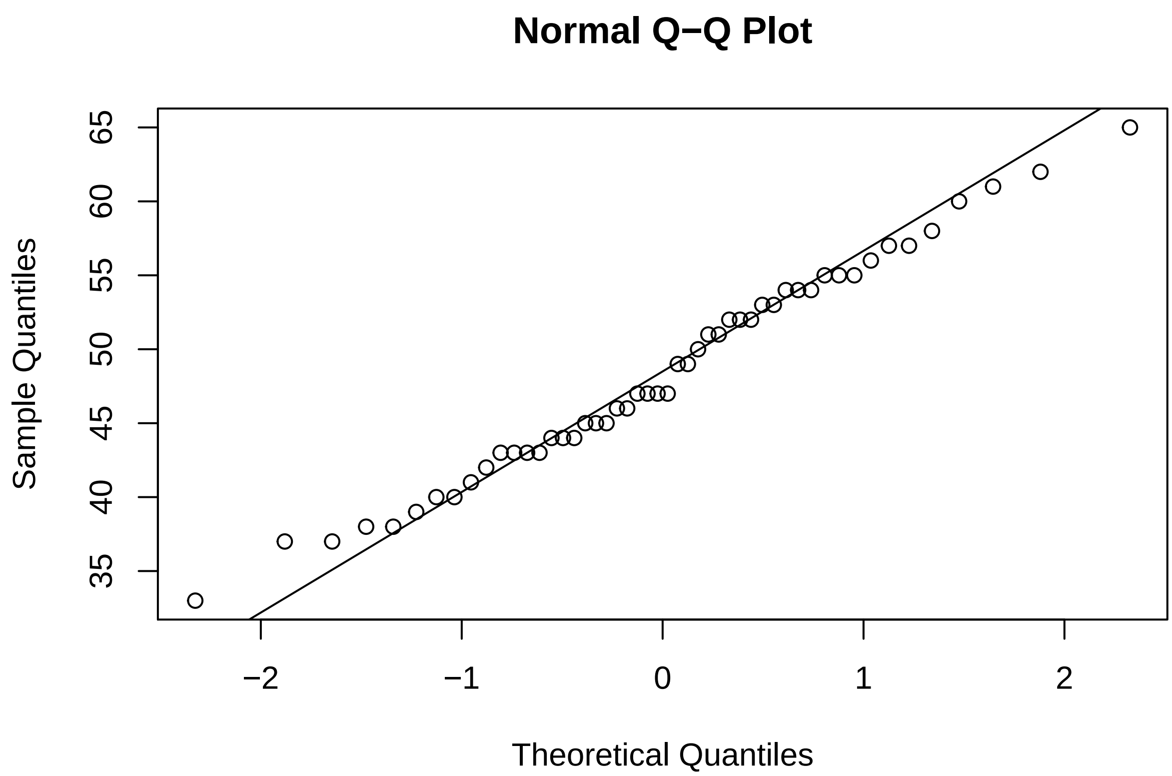
Histogram

```r
hist(poisson$Y)
```



Q-Q Plot

```r
qqnorm(poisson$Y)
qqline(poisson$Y)
```



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.

Part 3B

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

```
t <- 2.009
n <- nrow(poisson)
(mu + t*(s/(sqrt(n))))/n
```

```
## [1] 1.007124
```

```
(mu - t*(s/(sqrt(n))))/n
```

```
## [1] 0.9280763
```

Part 3C

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.

```
z <- 1.96
(mu + z*(sqrt(mu)))/n
```

```
## [1] 1.240258
```

```
(mu - z*(sqrt(mu)))/n
```

```
## [1] 0.6949415
```