STAA 552: HW 1

Matthew Stoebe

Contents

See Canvas Calendar for due date.
52 points total, 4 pts per problem unless otherwise noted.
Content for Q1-Q7 is from section 01.
Content for Q8-Q14 is from section 02.
Add or delete code chunks as needed.
For full credit, your numeric answers should be clearly labeled, outside of the R output.

Q1 - Q3

For this group of questions, identify each variable as **nominal, ordinal or numeric**.

Q1 (2 pts)

Anxiety rating (none, mild, moderate, severe)

Response

Ordinal

Q2 (2 pts)

Favorite grocery store (Safeway, King Soopers, Whole Foods, other)

Response Nominal *****

Q3 (2 pts)

Annual income ($)

Response Numeric *****

Poisson Distribution (Q4 - Q6)

Note: These questions are "self-checking" because the theoretical and simulated distributions should yield similar results.

Q4

Simulate 5000 independent replicates of the random variable $Y \sim \text{Poisson}(\mu)$ with $\mu = 7.5$. Use set.seed(5821) for reproducibility. Calculate the sample mean and sample variance.

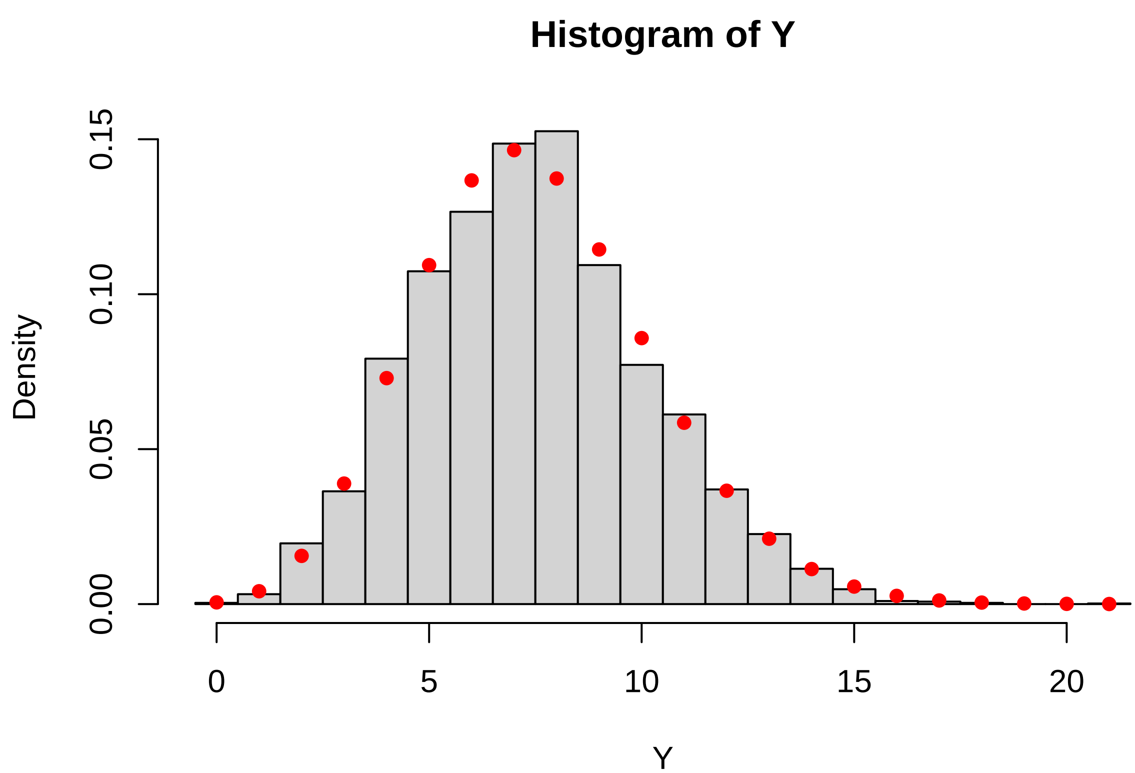
Response

```
## sample Mean: 7.4662
```

```
## Sample Variance: 7.377533
```

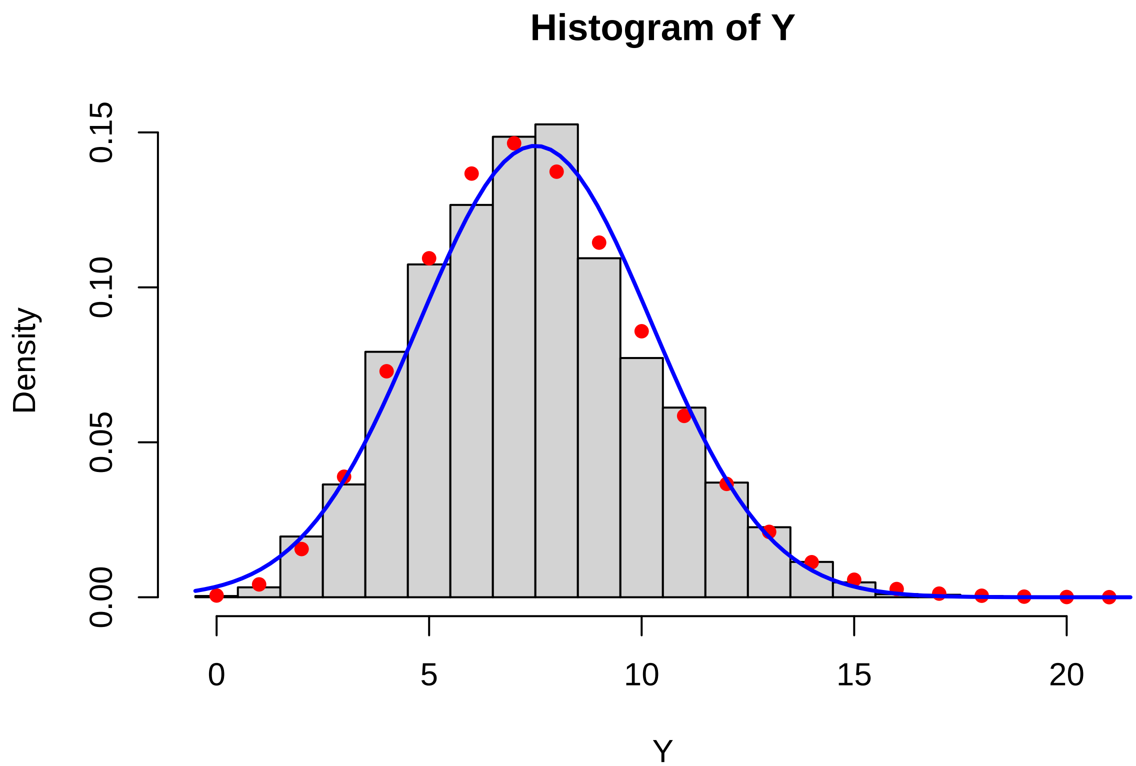
Q5

Plot the empirical probability mass function (ex: density histogram) of your simulated values. Overlay the true probability mass function for Poisson(7.5) in red.
Note: Some example R code has been provided to get you started. But other approaches are allowed.



Q6

Plot the empirical probability mass function (ex: density histogram) of your simulated values. Overlay the normal approximation with mean = 7.5 and variance = 7.5.
Notes: Since $\mu < 10$ we don't expect a great fit. Some example R code has been provided to get you started. But other approaches are allowed.



Diabetes (Q7 - Q10)

Suppose it is known that 12% of US adults have diabetes. Hence, π = 0.12. Consider a random sample of n = 160 US adults. Let Y represent the number of people with diabetes (in a given sample). Let $\hat{\pi} = (Y/n)$ represent the sample proportion of people with diabetes.

Note: Q8-Q10 are "self-checking" because the theoretical and simulated distributions should yield similar results.

Q7 (6 pts)

Specify the (exact) distribution for Y. Give E(Y) and Var(Y).

Response

This is a Binomial Distribution with p = .12 (probability of selecting a person with diabetes) and n = 160

```
## Expected Value: 19.2
```

```
## Expected Variance: 16.896
```

Q8 (6 pts)

Specify the (approximate) distribution for $\hat{\pi}$. Give $E(\hat{\pi})$ and $Var(\hat{\pi})$.

Response

The Binominal Distribution can be approximated by the Normal distribution when N is large and p is small.

```
## Expected Value: 0.12
```

```
## Expected Variance: 0.00066
```

Q9

Simulate 5000 independent replicates from the distribution specified in Q7. Use set.seed(4966) for reproducibility. Then calculate the corresponding $\hat{\pi}$ values. Calculate the sample mean and sample variance of the $\hat{\pi}$ values.

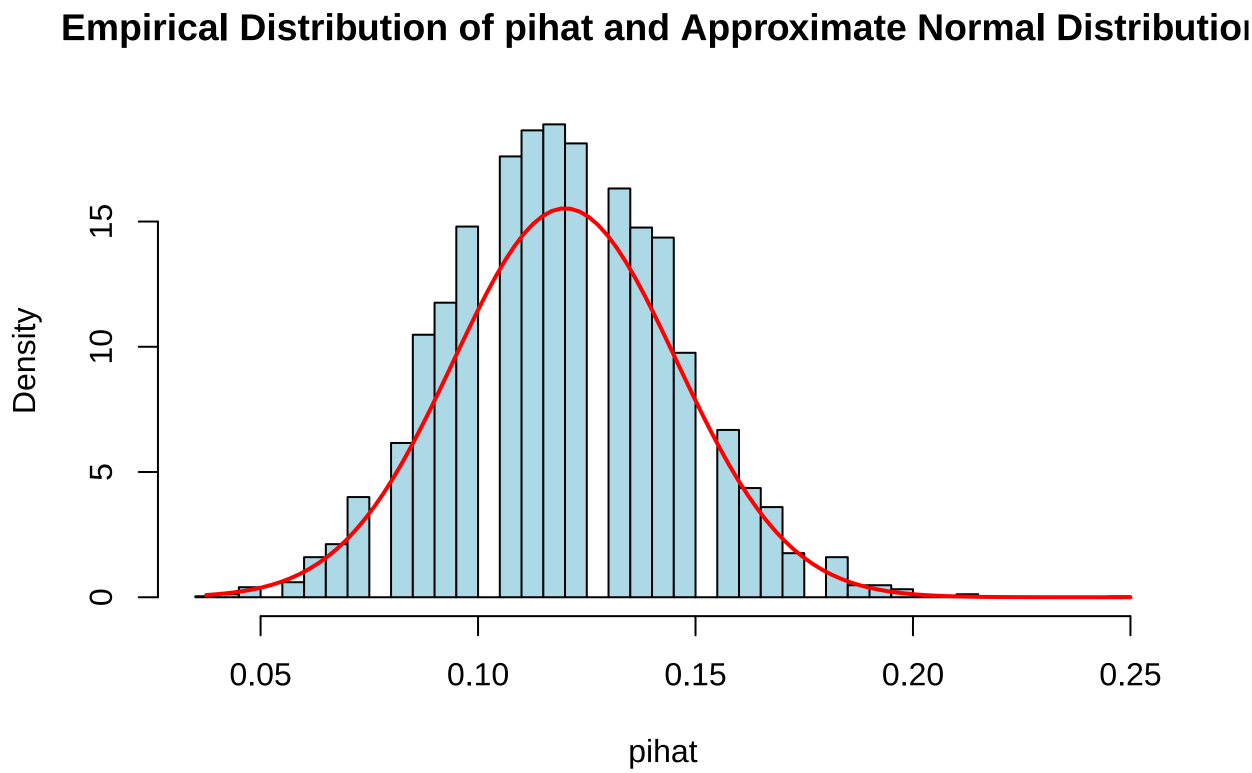
Response

```
## sample Mean: 0.1201425
```

```
## Sample Variance: 0.0006620058
```

Q10

Plot the empirical distribution (ex: density histogram) of your observed $\hat{\pi}$ values from the previous question. Overlay the approximate distribution from Q8.



Male Births (Q11 - Q14)

Let π be the true proportion of babies that are male in some large population. Suppose we have a random sample of n = 1574 birth records this population. Of these, y = 803 are male. We will use this data to test $H_0 : \pi = 0.5$ vs $H_A : \pi \neq 0.5$.

Note: Q12-Q14 are "self-checking" because all three methods will give very similar results.

Q11 (2 pts)

Calculate $\hat{\pi}$.

Response

```
## observed rate pihat: 0.5101652
```

Q12

Run the test of interest using a **Wald** test. Report the χ^2 test statistic and corresponding p-value. Do this "by hand" (using R as a calculator) and echo your R code.

Response

```
#Q12

pi0 <- .5

z <- (pihat - pi0) / sqrt(pihat * (1 - pihat)/n)

x <- z**2

p <- pchisq(x, df=1, lower.tail = FALSE)

cat("wald test chi squared statistic is:", x, "\n")
```

```
## wald test chi squared statistic is: 0.6508408
```

```
cat("with a p value of", p, "\n")
```

```
## with a p value of 0.4198122
```

Q13

Run the test of interest using a **Score** test. Report the χ^2 test statistic and corresponding p-value. Do this "by hand" (using R as a calculator) and echo your R code.

Response

```
#Q13
pi0 <- .5

z <- (pihat - pi0) / sqrt(pi0 * (1 - pi0)/n)

x <- z**2

p <- pchisq(x, df=1, lower.tail = FALSE)

cat("score test chi squared statistic is:", x, "\n")
```

```
## score test chi squared statistic is: 0.6505718
```

```r
cat("with a p value of", p, "\n")
```

```
## with a p value of 0.4199083
```

Q14

Run the test of interest using a **likelihood ratio** test. (See notes or CDA 1.4.1 for details and formula for test statistic.) Report the χ^2 test statistic and corresponding p-value. Do this "by hand" (using R as a calculator) and echo your R code.

Response

```r
#Q14

L0 <- y * log(pi0) + (n - y) * log(1 - pi0)

L1 <- y * log(pihat) + (n - y) * log(1 - pihat)


likelyhood_ratio = -2*(L0 - L1)


p <- pchisq(likelyhood_ratio, df = 1, lower.tail = FALSE)


cat("likelyhood_ratio test chi squared statistic is:", likelyhood_ratio, "\n")
```

```
## likelyhood_ratio test chi squared statistic is: 0.6506166
```

```r
cat("with a p value of", p, "\n")
```

```
## with a p value of 0.4198923
```

Appendix

```r
#Retain this code chunk!!!
library(knitr)

knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
#Q4
set.seed(5821)

n = 5000

Y = rpois(n, lambda = 7.5)

sample_mean <- mean(Y)
sample_variance = var(Y)

cat("sample Mean:", sample_mean, "\n")
cat("Sample Variance:", sample_variance, "\n")


#Q5
x = seq(0, max(Y))
hist(Y, freq = FALSE, breaks = seq(-0.5, max(Y)+0.5, 1))
#This choice of breaks is suggested to center each bar at an integer value since the Poisson distributi
points(x, dpois(x, 7.5), col = "red", pch = 16)
#Q6
x = seq(0, max(Y))
hist(Y, freq = FALSE, breaks = seq(-0.5, max(Y)+0.5, 1))
#This choice of breaks is suggested to center each bar at an integer value since the Poisson distributi
points(x, dpois(x, 7.5), col = "red", pch = 16)
#This choice of breaks is suggested to center each bar at an integer value since the Poisson distributi
curve(dnorm(x, mean = 7.5, sd = sqrt(7.5)), col="blue", lwd=2, add = TRUE)

#Q7
n = 160
p = .12

E_Y <- n*p
Var_y <- n*p*(1-p)

cat("Expected Value:", E_Y, "\n")
cat("Expected Variance:", Var_y, "\n")


#Q8
n = 160
p = .12

E_pihat <- p
Var_pihat <- p*(1-p)/n

cat("Expected Value:", E_pihat, "\n")
cat("Expected Variance:", Var_pihat, "\n")
```

```r
#Q9
set.seed(4966)

i = 5000


Y = rbinom(i, n,p)

sample_mean <- mean(Y)
sample_variance = var(Y)

pihat = Y/n

pihat_mean <- sample_mean / n
pihat_variance <- sample_variance/n**2

cat("sample Mean:", pihat_mean, "\n")
cat("Sample Variance:", pihat_variance, "\n")

#Q10

hist(pihat, freq = FALSE, breaks = 50,
     main = "Empirical Distribution of pihat and Approximate Normal Distribution",
     xlab = "pihat", col = "lightblue", border = "black")

x_vals <- seq(min(pihat), max(pihat), length = 100)
y_vals <- dnorm(x_vals, mean = E_pihat, sd = sqrt(Var_pihat))
lines(x_vals, y_vals, col = "red", lwd = 2)


#Q11
n <- 1574
y <- 803
pi <- .5

pihat <- y/n

cat("observed rate pihat:", pihat, "\n")

#Q12

pi0 <- .5

z <- (pihat - pi0) / sqrt(pihat * (1 - pihat)/n)

x <- z**2

p <- pchisq(x, df=1, lower.tail = FALSE)

cat("wald test chi squared statistic is:", x, "\n")
cat("with a p value of", p, "\n")

#Q13
```

```r
pi0 <- .5

z <- (pihat - pi0) / sqrt(pi0 * (1 - pi0)/n)

x <- z**2

p <- pchisq(x, df=1, lower.tail = FALSE)

cat("score test chi squared statistic is:", x, "\n")
cat("with a p value of", p, "\n")
#Q14

L0 <- y * log(pi0) + (n - y) * log(1 - pi0)

L1 <- y * log(pihat) + (n - y) * log(1 - pihat)


likelyhood_ratio = -2*(L0 - L1)


p <- pchisq(likelyhood_ratio, df = 1, lower.tail = FALSE)


cat("likelyhood_ratio test chi squared statistic is:", likelyhood_ratio, "\n")
cat("with a p value of", p, "\n")
```