BIOS 6611 2018
Homework 1 Answer Key

- Exercise 1
  - Part a:
    The three distributions can be generated using the following R code:

```
n <- 10000
norm <- rnorm(n, mean = 125, sd = 8)
pois <- rpois(n, lambda = 1.5)
binom <- rbinom(n, size = 5, prob = 0.15)
```

  - Part b:
    In order to verify the generated numbers have approximately the correct mean and standard deviation, the mean and standard deviation for each distribution must be either derived or looked up. The actual mean and sd for each sample is found using mean() and sd(). The code shows results for a seed of 123.

Commented [KAM1]: 15 points

5 points per distribution

	Theoretical Mean	Theoretical SD	Actual Mean	Actual SD
Normal	μ = 125	σ = 8	124.981	7.989
Poisson	λ = mean of Poisson = 1.5	$\sqrt{\lambda}$ = sd of Poisson = ~1.22	1.502	1.221
Binomial	n*p = mean of binomial = 0.75	$\sqrt{n * p * (1 - p)}$ = sd of binomial = 0.7984	0.737	0.803

○ Part c:

Commented [KAM2]: 15 points for code and histogram plots

5 points per distribution

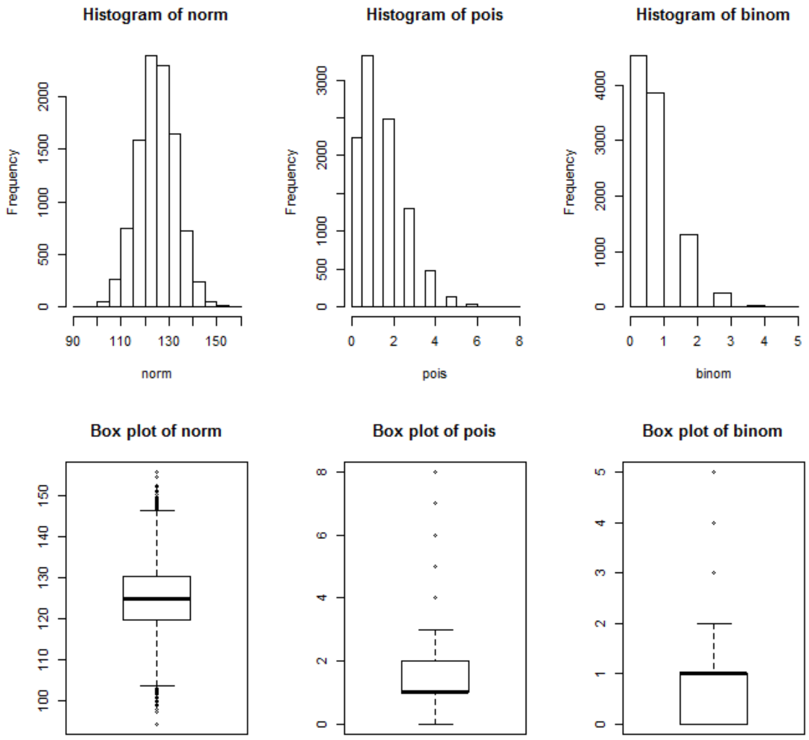
The following code can be used to create the box plots and histograms:

```
hist(norm)
hist(pois)
hist(binom)

boxplot(norm, main= 'Box plot of norm')
boxplot(pois, main= 'Box plot of pois')
boxplot(binom, main= 'Box plot of binom')
```

Note, you can plot all six plots on one figure if you run the following code first:

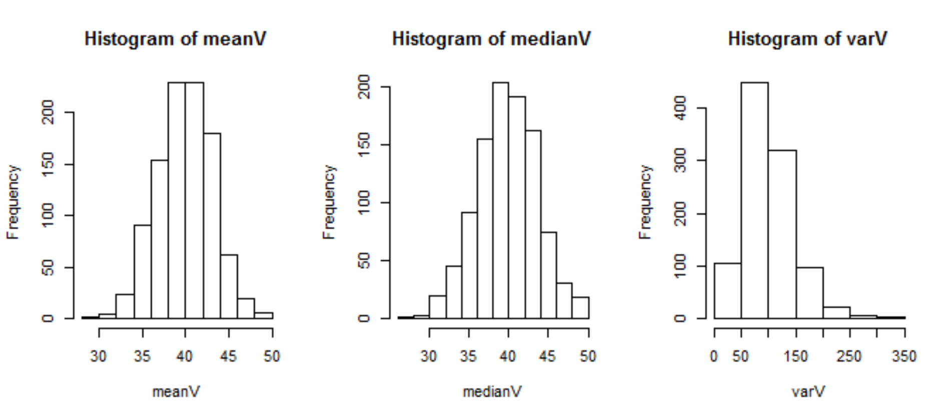
```
par( mfrow=c(2,3) )
```

This tells R to create a figure with 2 rows and 3 columns to insert plots/figures to.



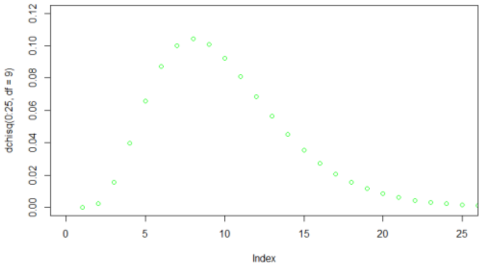
- Exercise 2
  - Part a: the following code generates the desired output:

```
nsim <- 1000
n <- 10
meanV <- rep(NA, 1000)
medianV <- rep(NA, 1000)
varV <- rep(NA, 1000)

for(i in 1:nsim){
  random <- rnorm(n, mean = 40, sd = 10)
  meanV[i] <- mean(random)
  medianV[i] <- median(random)
  varV[i] <- var(random)
}

hist(meanV)
hist(medianV)
hist(varV)
```

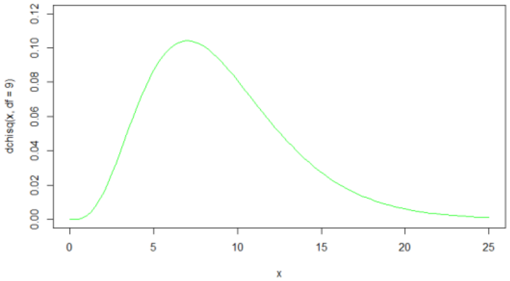


  - Part b:
    The sample mean and sample median should be distributed normally.
  - Part c:
    The following code (or a variation of it) plots the variance and plots a chi-squared curve, so you can see the two distributions are the same

```
varVshift <- varV * (9/10^2)
hist(varVshift)
plot(dchisq(0:25, df = 9), col="green", xlim = c(0,25),
      ylim = c(0,0.12))
```



```
# or use curve() function
curve(dchisq(x, df = 9), col="green", xlim = c(0,25),
      ylim = c(0,0.12))
```



- Exercise 3

  ○ Part a and b: The same code could be used four times, or a for loop in a for loop, the code for which is below:

```
nsim <- 500
sizeVec <- c(10,20,30,40,50)
meanMatrix <- matrix(NA, nrow = 500, ncol = 5)

for(j in 1:5){
  for(i in 1:nsim){
    binomData <- rbinom(sizeVec[j], size = 1, prob = 0.15)
    meanMatrix[i,j] <- mean(binomData)
  }
}
```

  ○ Part c: The following code can be used, or the mean() and sd() functions for each:

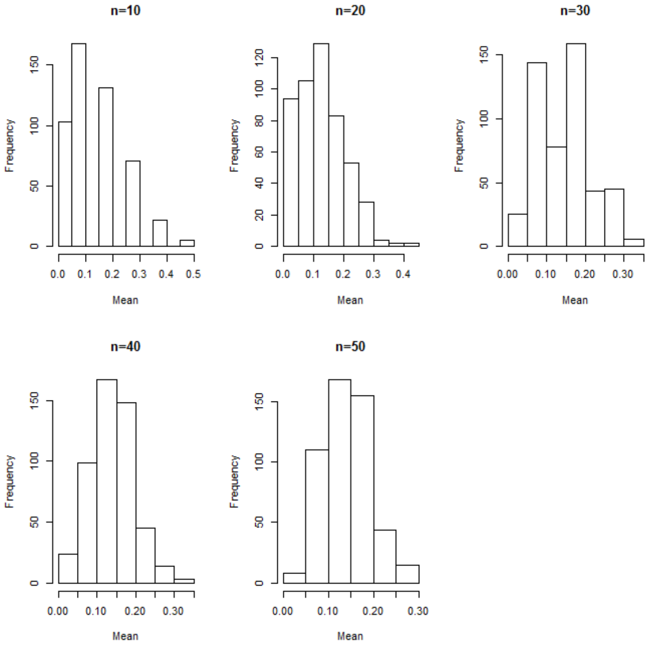
```
apply(meanMatrix,2,mean)
apply(meanMatrix,2,sd)
```

Commented [KAM3]: 15 points for code

Any variation on this that works is fine!

- Part d: The following code can be used, or the hist() function:

```
sapply(1:5, function(x) hist(meanMatrix[,x], xlab='Mean',
      main=paste0('n=',sizeVec[x]) ) )
```



- Part e: Around a sample size of 40, the distributions begin to look more normal. However, the “correct” answer will depend on what the simulated mean values look like for each seed/person.

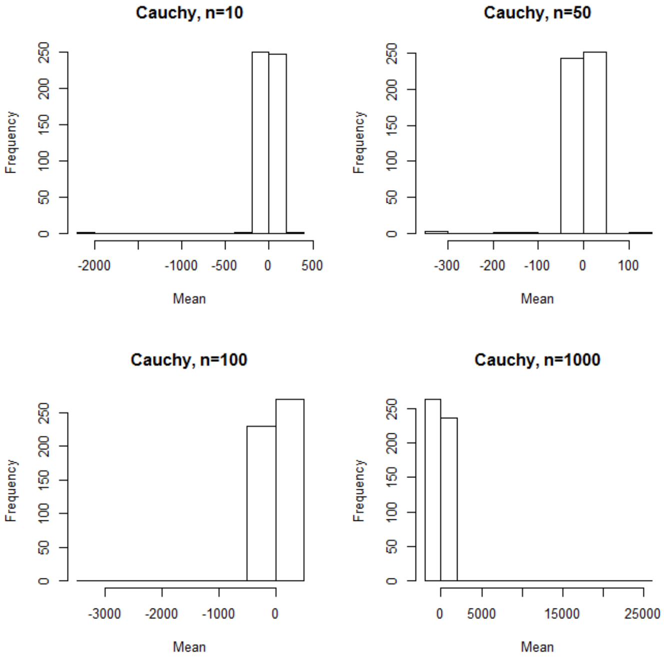
Commented [KAM4]: 15 points for answer

- Exercise 4: The following code can be used to produce the random data from the Cauchy distribution and plot it:

Commented [KAM5]: Extra credit

+ 3 points for completely correct code and answer

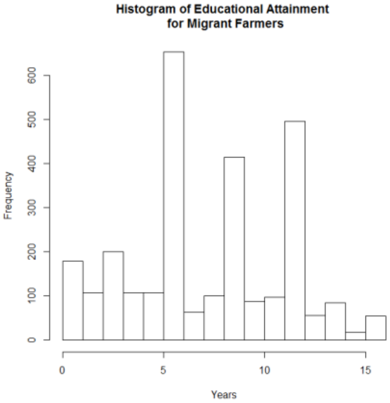
```
nsim <- 500
sizeVec <- c(10,50,100,1000)
meanMatrix <- matrix(NA, nrow = 500, ncol = 4)

for(j in 1:4){
  for(i in 1:nsim){
    cauchyData <- rcauchy(sizeVec[j])
    meanMatrix[i,j] <- mean(cauchyData)
  }
}

sapply(1:4, function(x) hist(meanMatrix[,x], xlab='Mean',
      main=paste0('Cauchy, n=',sizeVec[x]) ) )
```



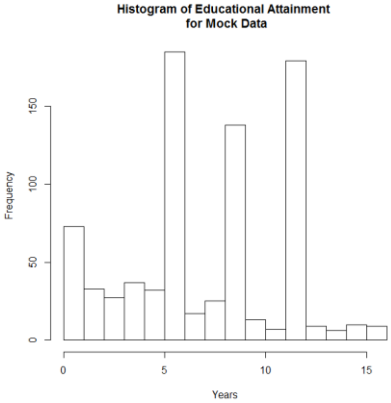
These histograms do not begin to look normal, even with large sample sizes. This is because the central limit theorem does not apply to the Cauchy distribution since it does not have finite variance.

- Exercise 5:
  - See R code file for examples.

- Exercise 6:
  - Part a: The following code will read in the data and make the histogram:
    ```
    naws <- read.csv("filepath/NAWS2014.csv", header=T)
    hist(naws$A09, main = "Histogram of Educational Attainment \n for
    migrant farmers", xlab = "Years")
    ```
  - Part b:



  - Part c: OPINION QUESTION. The mean doesn't tell the whole story. There are three peaks in the histogram, and by just reporting the mean, it isn't clear that the data has this pattern. These peaks correspond roughly to 6, 9 and 12 years of schooling, which would be like elementary, middle and high school grade achievements. Other summaries may be useful, such as the mode, interquartile range, or even providing the histogram.
  - Part d-f: See R code for example code
  - Part g: Yes, the histogram looks similar to that from before.



Commented [KAM6]: 15 points

Part a) 5 for reading in data

Part b) 10 for histogram with title and similar labels

2. Goodman (1999) paper on the p-value fallacy

a. Inductive vs. deductive reasoning

b. What is the “p-value fallacy”

c. Confidence intervals vs. p-values

Commented [KAM7]: 25 points,

7 per item plus 4 for overall framing

Ok to focus on specifics that are different from what’s below!

In the article by Goodman he gives a brief background on the origins of p-values and their use in hypothesis testing. The original intent of the p-value for Fisher was to assess the long-term frequency of unusual experimental outcomes, i.e., the discrepancy between the data observed and the null hypothesis. He advocated using background information in combination with a p-value to draw conclusions from any given experiment. Neyman and Pearson, on the other hand, thought about drawing conclusions based on a specific test that uses critical values or regions, based on minimizing the costs of making a wrong decision, to choose between one of two competing hypotheses. The two approaches have been melded together over time and the casual suggestion of a 5% level of significance has become ingrained in practice. P-values capture neither the long-term behavior of sampling nor the strength of evidence in support of one or the other hypothesis that Neyman and Pearson sought to capture through their approach. Thus, p-values serve neither of the intended interpretations. This is the p-value fallacy.

The practice of inductive reasoning in hypothesis testing is to determine the hypothesis with which the data are most consistent. Because there are many possible underlying hypotheses, e.g. many possible diagnoses that a patient could have based on a set of symptoms, this approach to proving what hypothesis is true is challenging. Deductive reasoning, in contrast, means that we use a hypothesis taken as the truth in order to predict an outcome, e.g. if disease is present then certain symptoms will be observed, a relatively simpler but not as useful inferential task.

A possible alternative to the use of p-values is the use of confidence intervals, which contain information not only on the magnitude of observed differences or effects but also on their variability. Because they are often used to categorically draw conclusions about hypotheses from data they suffer from the same limitations as p-values, including the fact that no external evidence is used to form a confidence. Fisher’s hope for how inference should be done is no better served by applying confidence intervals to draw conclusions than using p-values.