BIOS 6611 2019
Homework 1 Answer Key

Exercise 1

1a. Reproducibly simulate a sample of 10,000 from each of the following distributions:

- *Normal($\mu = 125, \sigma = 8$)*
- *Poisson($\lambda = 1.5$)*
- *Binomial($n = 5, p = 0.15$)*

Solution: 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)
```

1b. Determine the theoretical mean and standard deviation for each distribution and verify that the generated numbers have approximately the correct mean and standard deviation. Note, you can derive or look-up and cite your source for the theoretical mean and standard deviation.

Solution: 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.

	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	$\sqrt{\mathrm{n} * \mathrm{p} * (1 - \mathrm{p})}$ = sd of binomial = 0.7984	0.737	0.803

1c. Create a histogram and boxplot depicting each of the mock samples.

Commented [KAM1]: 15 points for code and histogram plots --5 points per distribution

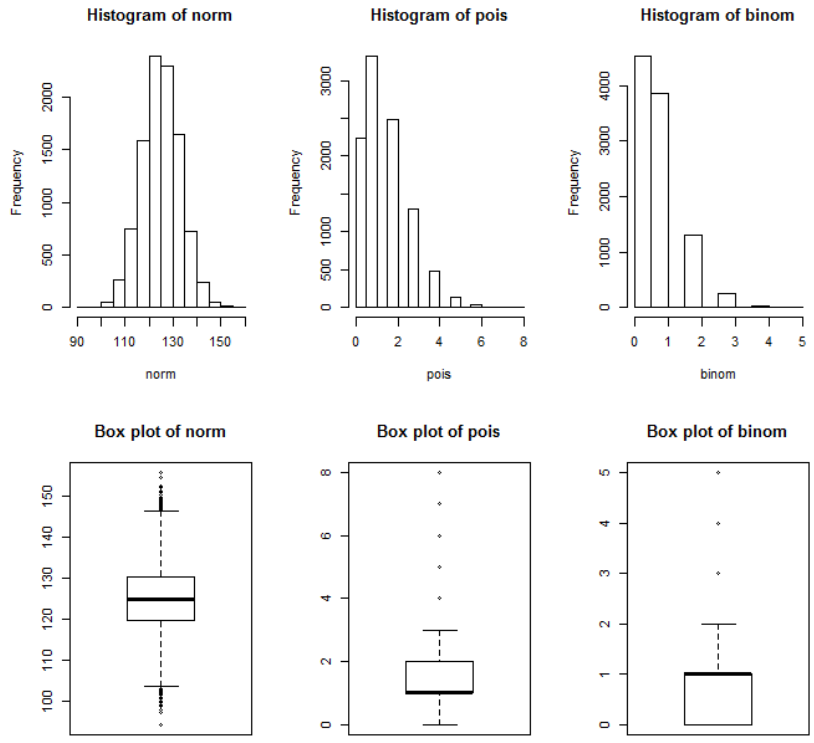
Solution: The following code can be used to create the box plots and histograms:

```
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.



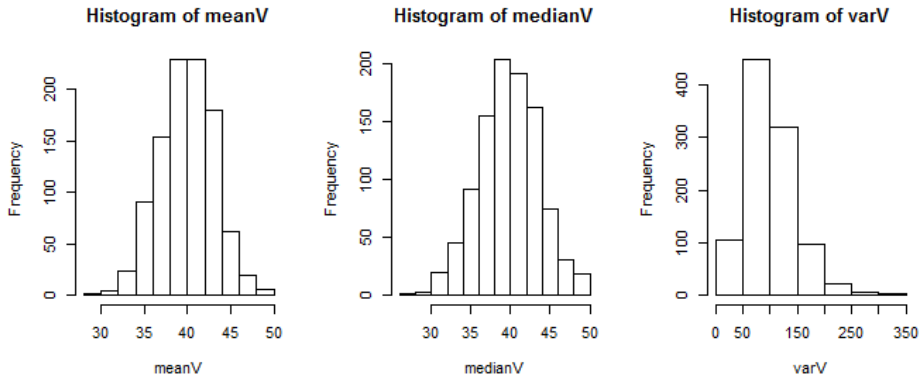
Exercise 2

2a. For a population that is normally distributed with mean 40 and standard deviation 10, generate histograms showing the sampling distribution of the mean, median, and variance. Use 1,000 simulation iterations and a sample size of n = 10.

Commented [KAM2]: 15 points for code and histogram plots --5 points for each measure (mean, median, variance)

Solution: 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)
```



2b. Based on theory, what is the distribution of the sample mean and sample median in this case (e.g., uniform, exponential, gamma, normal, etc. distributed)?

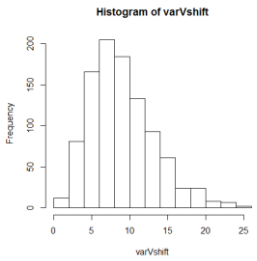
Solution: The sample mean and sample median should be distributed normally.

2c. When the underlying population is normally distributed, the sampling distribution of $(n-1)s^2/\sigma^2$ *is chi-squared with n - 1 degrees of freedom (n - 1 = 9 in this case).*

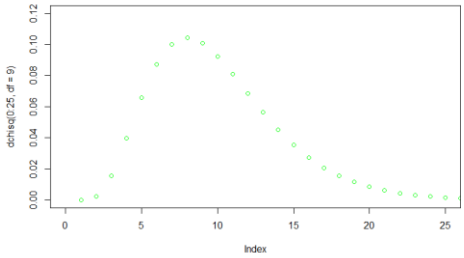
Use the base R `plot()` *and* `dchisq()` *function to plot the theoretical sampling distribution of the variance (i.e., instead of a histogram, either predict the density and present either density curve or set of points which could be connected together to form an approximate curve). Compare it to your result in 2a (hint: for ease of comparison, you might want to multiply your sample variance vector from 2a by a factor of* $9/10^2$*).*

Solution: 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:

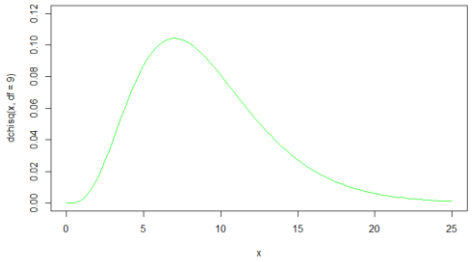
```
varVshift <- varV * (9/10^2) # multiply varV, o.w. x-axis range is diff.
hist(varVshift)
```




```
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

Drs. Bob and Billy are friends with a shared interest in heights. They learned that the average height for the population of adult male patients at the University of Colorado Hospital (UCH) is 70 inches. Dr. Bob hypothesizes that the median height of adult male patients that arrive at the hospital the next day will be close to the population average. Assume the heights of adult male patients seen at the UCH follow a normal distribution with mean=70 inches and variance=15 inches^2.

*3a. Assume that 100 patients are seen the next day. If Dr. Bob calculates the median height for the adult male patients seen on that day, what is the **bias** of his median estimate wrt the population mean? (Hint: Simulate a normal distribution of size n=100)*

Solution: The following code can simulate the 100 patients and be used to calculate the bias.

```
set.seed(0203) # Set seed for reproducibility

sim_norm <- rnorm(n=100,mean=70,sd=sqrt(15)) # Simulate a normal dist.

med <- median(sim_norm)  # Calculate the median

bias <- med-70  # Calculate the bias (Estimate - population mean)

bias

## [1] 0.2409348
```

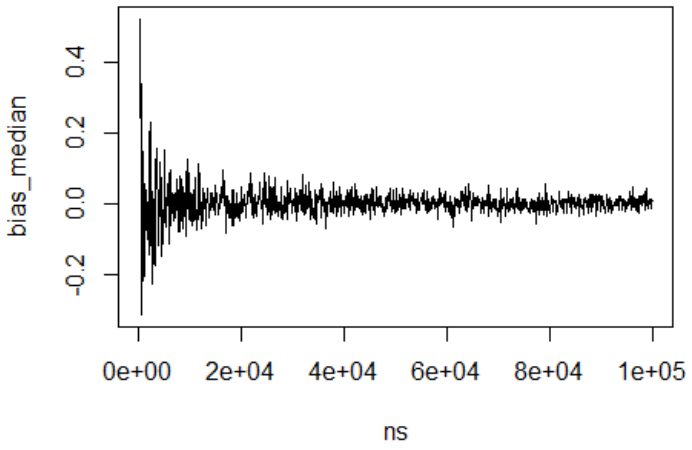
The median height from the sample is 70.24. Thus, the median is biased 0.24 units wrt to the population mean of 70.

*3b. Although improbable, now assume the number of patients seen in a day increases. Increase the sample size from 100 to 100,000, by 100 person increments. Calculate the bias for the varying sample sizes. Plot the results. What does this say about the **consistency** of the median estimate wrt the population mean?*

Commented [KAM3]: 15 points

Solution:

```
set.seed(0203) # Set seed for reproducibility

# Increasing sample size
ns <- seq(100,100000,by=100)

bias_median<-sapply(ns, function(x){

  sim<-rnorm(n=x,mean=70,sd=sqrt(15)) #Simulate normal dist. for each n

  medians<-median(sim) # Calculate the median of the samples

  bias<-medians-70 # Calculate the bias of the samples

  return(bias)
}) # Note, ")" matches to the left parenthesis for sapply

# Plot the results
plot(x=ns,y=bias_median,type='l')
```

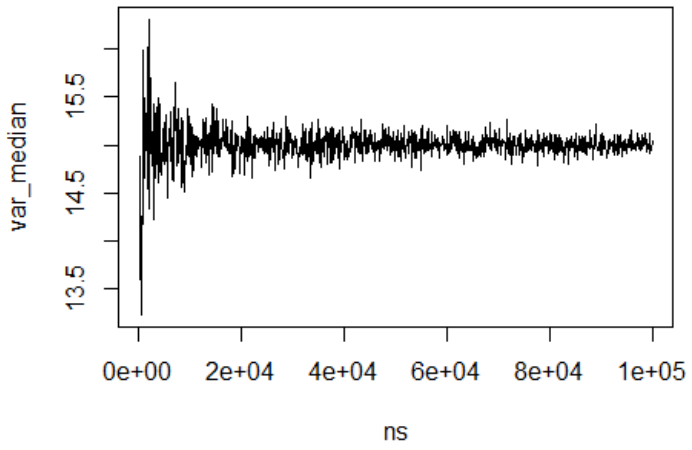


When the sample size increases to 100,000, the bias of the median appears to vary about 0. Thus, this simulation exercise suggests that the median becomes unbiased as the sample size increases infinitely, and the median estimator could be considered **consistent** with respect to the population mean.

3c. How does the variance of the data wrt the median estimator change as the sample size increases?[1]

Solution:

```
set.seed(0203) # Set seed the same as before

# Increasing sample size
ns<-seq(100,100000,by=100)

var_median<-sapply(ns,function(x){

  sim<-rnorm(n=x,mean=70,sd=sqrt(15))

  med<-median(sim)

  var_median<-sum((sim-med)^2)/length(sim)

  return(var_median)
})

# Plot the results
plot(x=ns,y=var_median,type='l')
```



As the sample size increases, the variance about the median varies less and centers about the population variance.

[1] $Var(X) = \sum_{\forall i}(x_i - median)^2/n$, where $\forall i$ means "for all" values of i

*3d. Dr. Billy bets Dr. Bob that the sample mean is more **efficient** (i.e. less variable about the population mean) than the sample median. To compare the relative efficiency of estimators, simulate 10,000 normal distributions with sample size n=1000, population mean=70 inches, and variance=15 inches^2. Calculate the median and mean for each simulation. Then compare the variance of the set of sample medians to the variance of the set of sample means. Using the results of your simulation, which estimator is more efficient?*

Commented [KAM4]: 15 points

Solution:

```
set.seed(0203) # Set seed the same as before

# Increasing sample size
ns<-rep(1000,100000)

medians<-sapply(ns,function(x){
  sim<-rnorm(n=x,mean=70,sd=sqrt(15))
  med<-median(sim)
  return(med)
})

set.seed(0203) # Set seed the same as before

# Increasing sample size
ns<-rep(1000,100000)

means<-sapply(ns,function(x){
  sim<-rnorm(n=x,mean=70,sd=sqrt(15))
  m<-mean(sim)
  return(m)
})

var(medians)
## [1] 0.02358404
var(means)
## [1] 0.01498784
var(means)/var(medians) # Relative efficiency
## [1] 0.6355079
```

The variance of the sample medians is 0.0236, and the variance of the sample means is 0.0150. Thus, the relative efficiency of the sample median wrt to the sample mean is 0.6355. Thus, the sample mean is more efficient (i.e. less variable about the population mean) than the sample median.

3e. Extra Credit: What is the Cramer-Rao Lower Bound, and why does it relate to this exercise?

Commented [KAM5]: 3 extra credit points possible

Solution:

The Cramer-Rao Lower Bound is the smallest possible variance of an unbiased estimator. It is used to compare estimators. The estimator that reaches this lower bound is known as the **best unbiased estimator**. Thus, it relates to this exercise, because we are trying to compare whether the sample mean or sample median is a better estimator of the population mean. From our results, we concluded that both the mean and the median estimators are consistent; however, the sample mean is more efficient that the sample median, so the sample mean is a better estimator for the population mean, when the distribution is symmetric.

Commented [KAM6]: Note: This part of the answer is sufficient for full extra points (or something along these lines).

The Cramer-Rao Inequality Theorem is as follows: Let $X_1, \dots, X_n$ be a sample with pdf $f(x|\theta)$, and let $W(X) = W(X_1, \dots, X_n)$ be any estimator satisfying

$$\frac{d}{d\theta} E_\theta W(X) = \int_X \frac{\delta}{\delta\theta} [W(x) f(x|\theta)] dx$$

and

$$Var_\theta W(X) < \infty$$

Then

$$Var_\theta W(X) \leq \frac{(\frac{d}{d\theta} E_\theta W(X))^2}{E_\theta((\frac{\delta}{\delta\theta} log\ f(X|\theta))^2)}$$

Exercise 4

Study design sprints. See R code file for examples.

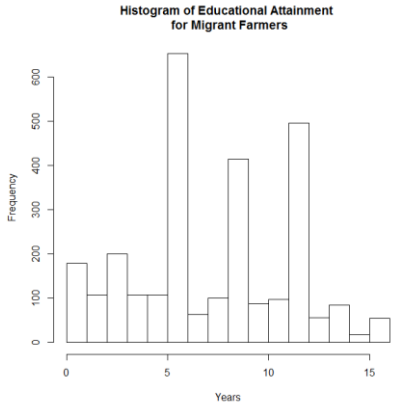
Exercise 5

5a. Read NAWS2014.csv from the Canvas site into R with the name NAWS. This is survey data publicly available from the Department of Labor.

5b. Plot a histogram of the A09 column, which asks how many years of school migrant farmers have completed (the same population referenced earlier in R Lab 1). Label the title and axes appropriately.

Solution: 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")
```



5c. As mentioned earlier in this document, it is commonly reported that the average educational attainment among migrant farmers is 8 years. In your opinion, does reporting just this average tell the whole story? Why or why not? (Feel free to speculate on why you think the histogram has its unique shape.)

Solution: 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.

Commented [KAM7]: 15 points

Part a) 5 for reading in data

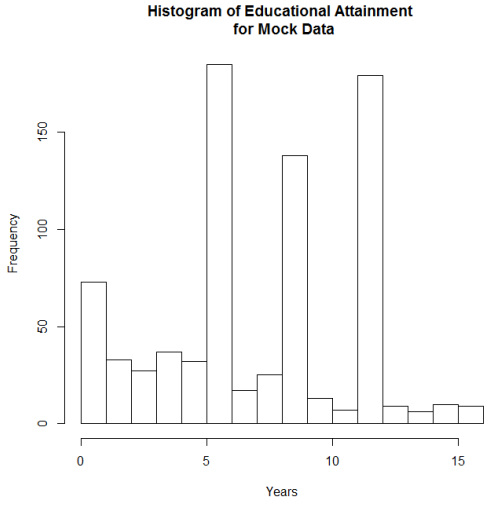
Part b) 10 for histogram with title and similar labels

5d-5f. See R Lab 1 for questions/steps.

Solutions: See R code for example code

5g. Finally, plot a histogram of `mockdata$educ_years`*. Does your mock data set look similar to the data it is based off of?*

Solution: Yes, the histogram looks similar to that from before.



Problem 2: Goodman (1999) p-value fallacy reading

Goodman (1999) paper on the p-value fallacy, touching on:

a. Inductive vs. deductive reasoning

b. What is the "p-value fallacy"

c. Confidence intervals vs. p-values

Solution: There are many possible write-ups of the Goodman paper, with one example provided 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.

Commented [KAM8]: 25 points,

7 per item plus 4 for overall framing

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