SOLUTIONS BIOS6611 Homework 2: Discrete Distributions, Expected Value, and Properties of Estimators

Solutions

Exercise 1: Discrete Distributions: Binomial and Poisson

Your classmate is backpacking in Patagonia. While there, she discovers that 2.5% of the people she meets in the region are affected by pulmonary sarcoidosis. Wondering whether this sample prevalence is unusually high, she begins calculating the probability of her sample prevalence using a binomial distribution. However, this quickly becomes too computationally intensive. She wonders whether she could use the Poisson approximation instead to reduce the computational burden. Needing help, she writes you for advice.

1a. Calculate the probability that 2.5% of Patagonians have the disease, assuming a sample size of 120 and population prevalence of 1%. Use both the exact binomial probability and the Poisson approximation of it. Compare the two.

```r
# Set up
n = 120  # sample size
p = 0.01  # probability
k = 0.025*n  # number of patagonians with the disease
lambda = n*p  # Use approximation for lambda

# probability of patagonians with the disease
prob_bin = choose(n,k)*p^k*(1-p)^(n-k)  # exact binomial probability
prob_pois = (lambda^k*exp(-lambda))/factorial(k) # Poisson approximation to b
inomial probability

# Compare
prob_pois-prob_bin

## [1] 9.23039e-05
```

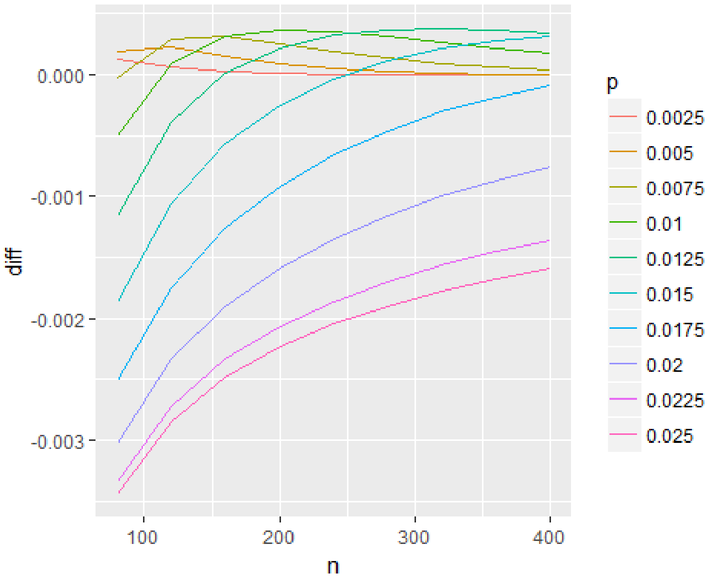
Solution: The probability that 2.5% of Patagonians in the sample have pulmonary sarcoidosis (assuming a sample size of 120 and population prevalence of 1% is 8.665% using the exact binomial, or 8.674% using the Poisson approximation of the binomial. Thus, the Poisson approximation of the binomial overestimates the exact probability by 0.009% (they're basically the same!), under these parameters.

1b. Allow your sample size to vary between 80 and 400 (by an increment of 40), while the population prevalence varies between 0.25% and 2.5% (by an increment of 0.25%).[1] The prevalence in your sample is still 2.5%. Calculate the difference between the exact binomial probability and the Poisson approximation of the binomial, under all combinations of parameters. Plot the results.[2]

```
# Libraries needed
library(ggplot2)

# Set up
n=seq(80,400,by=40) # varying sample size
p=seq(0.0025,.025,by=.0025)  # varying probability
np<-expand.grid(n=n,p=p)  # find every combination
np$k<-.025*np$n   # find the number in your sample with sarcoidosis
np$lambda <- np$n*np$p  # find lambda using the approximation equation

# Calculate the probabilities
np$prob_bin<-choose(np$n,np$k)*np$p^np$k*(1-np$p)^(np$n-np$k)  # exact binomi
al probability
np$prob_pois = (np$lambda^np$k)*exp(-np$lambda)/factorial(np$k)  # Poisson ap
proximation to binomial probability
np$diff <- np$prob_pois-np$prob_bin

# Plot the results
np$p<-factor(np$p)
ggplot(data=np,aes(x=n,y=diff,group=p,color=p))+geom_line()
```

[1] Note: There are 90 different combinations here

[2] Hint:
n=seq(80,400,by=40)
p=seq(0.0025,.025,by=.0025)
np<-expand.grid(n=n,p=p)



1c. At what sample size and prevalence would you recommend that your friend use the Poisson approximation to the binomial? How does this compare to the general recommendation given by Rosner?

Commented [KAM1]: 25 points, answers can vary from example solution

Solution: [This is a slightly subjective answer. In line with suggestions by Hesterberg about potential need for high accuracy in some settings we may choose larger sample sizes, but our problem's setting may not need really high accuracy relative to other contexts.] Assuming that the Poisson approximation to the binomial is "good enough" if it falls within a difference of 0.001, I would recommend that she use the Poisson approximation to the binomial at sample sizes ≥ 100, and a population prevalence of ≤ 1%. Rosner recommends using the Poisson approximation to the binomial at sample sizes of ≥ 100 and population prevalences of ≤ 1%. Thus, my recommendation is in line with that from Rosner.

Exercise 2: Expected Value and Variance

Sally just started the Master's program in biostatistics at the University of Colorado Anschutz Medical Campus. She is originally from Iowa. While she loves Iowa and all of its cornfields, she desires to establish residency in Colorado, so that she will qualify for in state tuition the following year. Sally goes to the Division of Motor Vehicles with all of her forms. Before she enters the building, Sally wonders how long she should expect to wait in line before being helped. Assume the service times follow an exponential distribution with a rate of 3 people helped per hour.

2a. How long should Sally expect to wait in line? Use the definition of expected value and calculus to answer this problem.[3]

Solution:

$$
\begin{aligned}
E[X] \quad &= \int_0^\infty x\,\lambda e^{-\lambda x} dx \\
&= (x\lambda)\frac{-e^{-\lambda x}}{\lambda}|_0^\infty - \int_0^\infty \frac{-e^{-\lambda x}}{\lambda}\lambda dx \\
&\qquad \text{(Integration by parts with } u = x\lambda,\ dv = e^{-\lambda x}dx \Longrightarrow du = \lambda dx,\ v = \frac{-e^{-\lambda x}}{\lambda}\text{)} \\
&= \frac{-x}{e^{\lambda x}}|_0^\infty - [\frac{e^{-\lambda x}}{\lambda}|_0^\infty] \\
&= [0 - 0] - [0 - \frac{1}{\lambda}] \\
&= \frac{1}{\lambda} = \frac{1}{3} \approx 0.33
\end{aligned}
$$

Sally should expect to wait 0.33 hours, or 20 minutes. (Note: You can also plug in "integrate x lambda exp(-lambda x) from 0 to infinity" to wolframalpha.com and it will return $\frac{1}{\lambda}$ to check your answer.)

2b. What is the variation around this estimate?[4]

Commented [KAM2]: 25 points, if they derived 2a by hand but didn't do this part still give full credit since question wasn't explicit

Solution: Using the hint, we first need to calculate $E(X^2)$:

[3] Hint: Integration by parts necessary. You can also use some math software or a website like wolframalpha.com to calculuate the integration and check you answers.

[4] Hint: $Var(X) = E[X^2] - E[X]^2$

$$E[X^2] = \int_0^\infty x^2\,\lambda e^{-\lambda x}dx$$

$$= (x^2\lambda)\frac{-e^{-\lambda x}}{\lambda}|_0^\infty - \int_0^\infty \frac{-e^{-\lambda x}}{\lambda}2x\lambda dx$$

$$\text{(Integration by parts with } u = x^2\lambda,\ dv = e^{-\lambda x}dx \Longrightarrow du = 2x\lambda dx,\ v = \frac{-e^{-\lambda x}}{\lambda}\text{)}$$

$$= [0-0] - \int_0^\infty \frac{-e^{-\lambda x}}{\lambda}2x\lambda dx \quad \text{(Factor out } -\frac{2}{\lambda}\text{)}$$

$$= \frac{2}{\lambda}\int_0^\infty e^{-\lambda x}\,xdx$$

$$\text{(Note that we now have the integral from 2a: } E[X] = \int_0^\infty x\,\lambda e^{-\lambda x}dx = \frac{1}{\lambda}\text{)}$$

$$= \frac{2}{\lambda}\times\frac{1}{\lambda}$$

$$= \frac{2}{\lambda^2}$$

Now, based on our hint and our value of $\lambda = 3$:

$$Var(X) = E[X^2] - E[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} = \frac{1}{3^2} \approx 0.11$$

The variation around our estimate of 0.33 hours is 0.11 hours^2.

2c. Reproducibly simulate an Exponential(3) distribution of size 100,000. Calculate the mean and variance of your simulated distribution. How similar are these values to your answers above?

Commented [KAM3]: 15 points

```
set.seed(0202)
sim_exp<-rexp(100000,3)
mean(sim_exp)

## [1] 0.3344283

var(sim_exp)

## [1] 0.111996
```

Solution: The mean of the simulated Exponential(3) distribution is 0.334, which is nearly identical with our 1/3 theoretical value. Additionally, the variance of the Exponential(3) distribution is 0.112, which is, again, very similar to our 1/9 theoretical value.

2d. Now suppose that Sally has been at the DMV for 10 minutes and has not been helped. Assume Sally is still just as oblivious about the number of people ahead of her as when she got there. How long should Sally expect to wait now?[5]

Solution: Sally should expect to wait the same amount of time in line (20 minutes), even after waiting for ten minutes. That is, the expected amount of waiting time does not change even after some time has passed, when the waiting time is distributed exponentially.

Exercise 3: Properties of Estimators: Bias, Consistency, and Efficiency

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)

```
set.seed(0203) # Set seed for reproducibility
sim_norm <- rnorm(n=100,mean=70,sd=sqrt(15)) # Simulate a normal distribution
med <- median(sim_norm)  # Calculate the median
bias <- med-70  # Calculate the bias (Estimate - population mean)
bias

## [1] 0.2409348
```

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

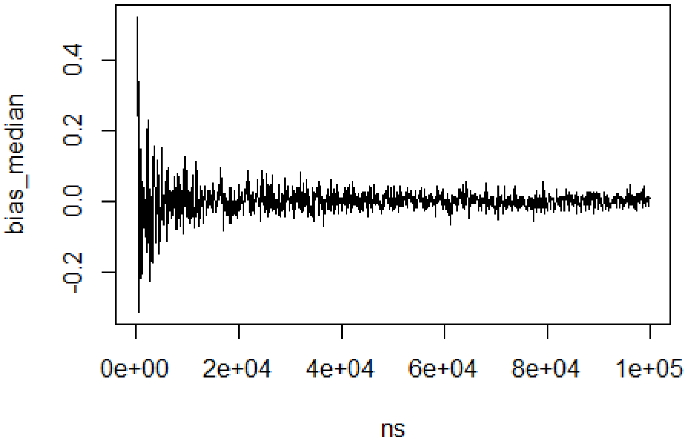
3b. Although improbable, now assume the number of patients seen in a day increases. Increase the sample size from 100 to 100,000, by 100 person increments. Calculate the bias for the varying sample sizes. Plot the results. What does this say about the **consistency** of the median estimate wrt the population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

Commented [KAM4]: 15 points

```
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 distributions for each
n
```

[5] Hint: Memoryless property of the exponential function.

```
  medians<-median(sim) # Calculate the median of the samples
  bias<-medians-70 # Calculate the bias of the samples
  return(bias)
})

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



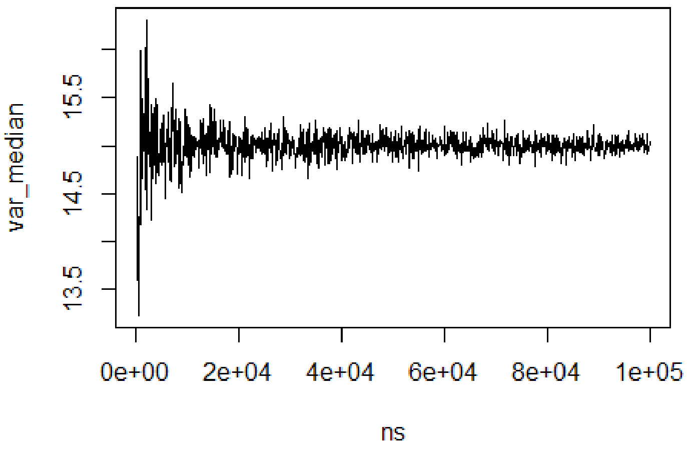
Solution: When the sample size increases to 100,000, the bias of the median appears to vary about 0. Thus, this simulation exercise suggests that the median becomes unbiased as the sample size increases infinitely, and the median estimator could be considered **consistent** wrt to the population mean.

3c. How does the variance of the data wrt the median estimator change as the sample size increases?[6]

```
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)
```

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

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

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



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

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 [KAM5]: 20 points

```
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)
})
```

```r
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
```

Solution: 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 [KAM6]: 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 [KAM7]: 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

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