Lab Project STMATH390

This is a 3 part lab. At the end of each part, there are questions to answer. For each lab (you can turn it in as one file), submit a copy of your r code, your output, and the answers to the question sets at the end of each part.

Part I

Empirical rule

The Empirical Rule or the 68-95-99.7 Rule states that for if a frequency distribution of a set of sample data is symmetric (and approximately normally distributed) then:

- Approximately 68% of the data falls within 1 standard deviation of the mean i.e. within $x \pm s$.
- Approximately 95% of the data falls within 2 standard deviations of the mean i.e. within $x \pm 2s$.
- Approximately 99.7% of the data falls within 3 standard deviations of the mean i.e. within $x \pm 3s$.

We are going to use the Dataset "faithful" which is a built in data set for R and R-studio

Let's test the data set "the length of time of eruptions of the Old Faithful Geyser in Yellowstone" to see if it satisfies the Empirical Rule. In R or R-studio, load the data:

```
> data(faithful)
> help(faithful)      %% Shows info on the data
> attach(faithful)      %% Loads the 'names' in faithful into R
> names(faithful)
```

Let's look at the eruption times now in variable eruptions.

```
> length(eruptions)        %% How much data
> hist(eruptions,20)
```

We want to see what percentage of the data is within one, two and three standard deviations. Let's compute the mean and the standard deviation and save the numbers:

```
> mean(eruptions)
> sd(eruptions)
```

Let us save the mean and standard deviation in the variables emean and textttesd respectively.

```
> emean=mean(eruptions)
> esd=sd(eruptions)
```

To find the number of observations within ±1 standard deviation of the mean we use the R command sum:

> sum(erupstions>emean-esd & eruptions < emean+esd)

We can get the percentage by dividing by the number of observations which is given by length function:

> sum(erupstions>emean-esd & eruptions < emean+esd)/length(eruptions)

The answer is 55.1%. This means that the data set eruptions DOES NOT satisfy the Empirical rule, which in turn means that the data set is NOT symmetric and thus NOT normally distributed. This can also be seen from the histogram.

Questions

1. Include your histograms for the three data sets along with the answers to the following questions in your submission
2. For the above data set, what percentage of data falls in the range of $x \pm 1.25s$?
3. For the above data set, what percentage of data falls in the range of $x \pm 2s$, $x \pm 3s$?
4. How does it compare with the Empirical Rule? With Chebyshev's theorem (we talked about this in class)?
5. Find percentages of data which falls in the range $x \pm s$, $x \pm 1.25s$, $x \pm 2s$ and $x \pm 3s$ for the data set called "AirPassengers" and "LakeHuron". NOTE: After the data(datasetname) command, type attach(datasetname) to see the variables in the data set.
6. Compare your results with Empirical Rule and Chebyshev's theorem if it isn't symmetric.

Part II

Normal Distributions

A large part of statistical analysis is based on the properties of normally distributed random variables. There's a famous LAW that states that data coming from a large number of independent experiments will produce a Normal Distribution and as such, a lot of statistical models assume 'Normality'. We have seen the Empirical Rule for normal distributions before.

Recall: the Empirical Rule or the 68-95-99.7 Rule states that for if a frequency distribution of a set of sample data is normally distributed then

- Approximately 68% of the data falls within 1 standard deviation of the mean i.e. within $x \pm s$.
- Approximately 95% of the data falls within 2 standard deviations of the mean i.e. within $x \pm 2s$.
- Approximately 99.7% of the data falls within 3 standard deviations of the mean i.e. within $x \pm 3s$.

How Normal is a Distribution?

In this project, we will take what we know about normal distributions and compare the theoretical

distribution to some real data. We can find the cumulative area on the left of a z score for a standard normal distribution in R using the command:

> pnorm(z, mean=0, sd=1) > pnorm(1, mean=0, sd=1)
[1] 0.8413447

Let's load a data set containing information on the air quality in New York.

>data(airquality)
>attach(airquality) ## to allow us to access the named variables by name

The data set contains measurements of
>names(airquality)

[1] “Ozone” “Solar R” “Wind” “Temp” “Month” “Day”

Ozone is the primary ingredient in ‘smog’; the more ozone in the air, the worse the air quality. Too much ozone is dangerous to one’s health and poses real hazards for the elderly and those with lung ailments. A ‘bad air’ day is one with high ozone concentrations. Let’s concentrate on the Wind measurements. Since we want to compare these measurements to the Standard Normal Distribution, the first thing we’d like to do is normalize the data by converting to z-scores. In R, define a new measurement:

> zwind = (Wind -mean(Wind))/sd(Wind)
> hist(zwind, Prob=T)

Now we can compare the distribution of zwind to the Standard Normal Distribution. First, we know that 68% of Normally Distributed data lies between ±1 standard deviation of the mean. To see if this is true for the Wind data, we need to compute the percentage of data in zwind that lies between ±1 sd.

InR:

> sum(zwind > -1 & zwind < 1) ## Number of data points in 1 sd of mean 100
> sum(zwind > -1 & zwind < 1)/length(zwind) ## Percentage of data in 1 sd

[1] 0.6535948

Pretty close. Using the pnorm command we find that 86.64 % of Normally Distributed data lies within ±1.5 standard deviations of the mean. Check this for the Wind data:

> sum(zwind > -1.5 & zwind < 1.5)/length(zwind)
[1] 0.869281

Again, the Normal Distribution prediction is quite close.

Questions

1. As above, find the percentage of Wind data which is less than the following 4 values of z: −0.75, −1.25, 1.85, 2.85 Compare the percentages in the attached table for the standard normal distribution (theoretical) to the percentages of Wind data found using R (observed).

z	−0.75	−1.25	1.85	2.85
Observed Value				
Theoretical Value				

2. Compute the following Probabilities using (a) the Wind Data and (b) the Normal Distribution. (Note: you will need to convert these to their z–scores).
   (a) Prob(Wind > 10 mph)
   (b) Prob(Wind > 15 mph)
   (c) Prob(Wind > 20 mph)
   (d) Prob(Wind < 5 mph)

3. Take a look at histograms for Ozone, Wind and Temp. From the histograms, which could best be described by a normal distribution? Give numbers to support your conclusion. Which of the three is 'least' normal (i.e. compare numbers from with Empirical rule)?

Part III

Normal Approximation to the binomial

In this project we will compare the binomial distribution, its approximation using the normal distribution and the approximation using the continuity correction. We will also learn commands to find probability for the binomial distribution.

Binomial Distributions using R

For the Binomial Distribution let p be the probability of a success and q =1 - p be the probability of failure. The probability of exactly k successes in n trials is given by $P(k) = \binom{n}{k} p^k q^{n-k}$. The mean μ = np and the standard deviation $\sigma = \sqrt{n \cdot p \cdot q}$

The R command

dbinom(k,size=n,prob=p)

gives the probability P (k). For example probability of getting 4 heads when 7 coins are tossed is:

```
> dbinom(4,size=7,prob=0.5)
0.2734375 > 0.5^4*0.5^3*choose(7,4)    #check answer using formula
0.2734375
```

The R command

pbinom(k,size=n,prob=p)

gives the probability for the binomial distribution for at most k successes. This can also be done by summing P (k) for k from 0 to n. For example the probability of getting at most 4 heads when 7 coins are tossed is:

```
> pbinom(4,size=7,prob=0.5)
```

0.7734375 > sum(dbinom(0:4,size=7,prob=0.5)) #check answer by adding
0.7734375

Example 1 The probability of getting between 3 and 6 heads when 7 coins are tossed is given by:

> pbinom(6,size=7,prob=0.5)-pbinom(2,size=7,prob=0.5) #Note the 2 instead of 3
0.765625 > sum(dbinom(3:6,size=7,prob=0.5)) #check answer by adding
0.765625

Example 2 Use the pbinom command to find the probability of getting between 5 and 15 heads when 25 coins are tossed. (answer = 0.8847833)

The command

pnorm(x,mean=0,sd=1)

gives the probability for that the z-value is less than x i.e. the cumulative area on the left of a x for a standard normal distribution. The area which pnorm computes is shown here. For example, probability of getting a number less than 1 in the standard normal distribution is:

> pnorm(1,mean=0,sd=1)
[1] 0.8413447

The command

pnorm(x,mean=m,sd=s)

gives the probability for selecting a number less than x from a normal distribution with mean m and standard deviation s.

Example 3 The probability of getting a number between 1 and 4 in the a normal distribution with mean 2 and standard distribution 0.7 is given by:

> pnorm(4,mean=2,sd=.7)-pnorm(1,mean=2,sd=0.7)
[1] 0.9978535

Example 4 Use the pnorm command to find the probability of getting a number between 5 and 15 heads for a normal distribution with mean 8 and standard deviation 4. (answer =0.7333135).

Approximating the Binomial distribution

Now we are ready to approximate the binomial distribution using the normal curve and using the continuity correction.

Example 5 Suppose 35% of all households in Carville have three cars, what is the probability that a random sample of 80 households in Carville will contain at least 30 households that have three cars.

Solution : For this problem n = 80 and p = 35% = 0.35, q =0.65. The mean μ = n × p =

80 × 0.35 = 28 and the standard deviation $\sigma = \sqrt{n \cdot p \cdot q}$ = $\sqrt{80 \cdot .25 \cdot .65}$ 4.26.

Using binomial distribution:

```
> pbinom(80,size=80,prob=0.35)-pbinom(29,size=80,prob=0.35)
[1] 0.3588295
```

Using the normal distribution:

```
> 1-pnorm(30,mean=28,sd=4.26)
[1] 0.319362
```

Using continuity correction:

```
> 1-pnorm(29.5,mean=28,sd=4.26)
[1] 0.3623769
```

You can see that the answer using continuity correction is much closer to the actual value !

Questions

About two out of every three gas purchases at Cheap Gas station are paid for by credit cards. 480 customers buying gas at this station are randomly selected. Find the following probabilities using the binomial distribution, normal approximation and using the continuity correction.

1. Find n,p, q, the mean and the standard deviation.
2. Find the probability that greater than 300 customers will pay for their purchases using credit card.
3. Find the probability that between 220 to 320 customers will pay for their purchases using credit card.
4. Generate a random number using the command

   >floor(rnorm(1, mean=200, sd=50))
5. Write this number down. Lets call it N. (This number will be different for each student.)
6. Find the probability that at most N (from #4) customers will pay for their purchases using credit card.