Math 1180:Lab5

Kyle Gaffney

February 5th, 2014

1 Mean, Median, Expected Value, and Mode

First let us generate a data set for the following probability distribution:

$$P(x = 1) = 0.4$$
$$P(x = 2) = 0.3$$
$$P(x = 3) = 0.2$$
$$P(x = 4) = 0.1$$

```
Probs=c(0.4,0.3,0.2,0.1)
N=100
Data=sample(1:4,N,prob=Probs,replace=T)
```

Now we have our data set, Data. Let us see what it looks like:

```
hist(Data,freq=F,breaks=c(0,1,2,3,4))
```

1.1 Median

The median value is right in the middle of the probabilities. The best way to determine the median is to plot the cdf and see where it crosses the line y=0.5 (the half way point).

First let us collect the CDF of our data set.

```
accumulation<-function(t){
N=length(Data)
length(Data[Data<t])/N}
tlist2=seq(1,4.1,length=25)
approxcdf=seq(1,25)
for(i in seq(1,25)){approxcdf[i]<-accumulation(tlist2
[i])}
```

Now let us plot it.

```
plot(tlist2,approxcdf)
lines(tlist2,rep(0.5,length(tlist2)))
```

1.2 Mean

The mean value is the average value of the data set. The best way to determine the mean of a data set is to sum all the entries of the set and divide by the size of the set.

```
Datamean=sum(Data)/length(Data)
```

1.3 Expected value

Expected value is the sum over all events of the probability of an event happening times the value of that event. This is a feature of the Discrete Random Variable.

$$ExpectedValue = \sum_{1}^{4} P_i v_i = (0.4) * 1 + (0.3) * 2 + (0.2) * 3 + (0.1) * 4 = 0.4 + 0.6 + 0.6 + 0.4 = 2$$

1.4 Mode

Mode is a statistic of the most commonly occurring value in a data set. In the sense of a Discrete Random Variable the mode is the value where the maximum of the PDF occurs

```
DataMode=max(table(Data))
```

To understand why DataMode is the mode explore what the table(Data) command did without using the max command.

2 Quartiles,Range,Mean Absolute Deviation, Variance, and Standard Deviation

These are statistics that give information about how wide or varied the values of the data set are.

2.1 Quartiles

Quartiles are similar to the median (in fact the median is the 50 percent quartile). The best way to find the quartiles is to find when the lines y=0.25, y=0.5, and y=0.75 cross the approximate CDF of the data set. Then:

```
plot(tlist2,approxcdf,type='l')
lines(tlist2,rep(0.25,length(tlist2)))
lines(tlist2,rep(0.5,length(tlist2)))
lines(tlist2,rep(0.75,length(tlist2)))
```

the minimum value and the maximum value are easily found with the min,max commands:

```
min(Data)
max(Data)
```

2.2 Range

Range is the difference between the minimum and maximum values of the data set.

```
Datarange=max(Data)-min(Data)
```

2.3 Mean Absolute Deviation

The mean absolute deviation is a way of measuring how far the data values are away from the mean on average. For discrete random variables like our data set we have that:

$$MAD = \sum_{i=1}^{n} |x_i - \bar{X}| p_i = |1-2| * 0.4 + |2-2| * 0.3 + |3-2| * 0.2 + |4-2| * 0.1 = 0.4 + 0 + 0.2 + 0.2 = 0.8$$

2.4 Variance

Variance is defined as follows:

$$\sum_{i=1}^{n} |x_i - \bar{X}|^2 p_i = |1-2|^2 * 0.4 + |2-2|^2 * 0.3 + |3-2|^2 * 0.2 + |4-2|^2 * 0.1 = 0.4 + 0 + 0.2 + 0.4 = 1$$

2.5 Standard Deviation

The standard deviation is the square root of the variance. We can also calculate the standard deviation of our data set and see how it compares:

```
sd(Data)
```

3 Assignment for the Week

Generate a Data set of size N=100000 from the following Discrete Probability Density Function:

$$P(x = i) = \frac{1}{(i + 3)}$$

for i=1..6

and P(x=7)=11/2520

Find the following values for this data set for the Probability Density function as well as for the data set that you generated when applicable. How do the compare?

1. Mean
2. Median
3. Mode (The Mode of a Discrete Random Variable is the largest value of the PDF)
4. Expected Value
5. Range
6. Quartiles
7. Mean Absolute Deviation
8. Variance
9. Standard Deviation

What would happen if we made N larger? What about if N became smaller?