Stat 330: Homework #9

Due on April 9, 2014 at 3:00pm

Mr. Lanker, Section A

Josh Davis

Problem 1

Use this data to answer the questions below:

Wolverines:	27	38	21	37	23	28	23	24	34	26	20	21
Cornhuskers:	38	27	56	49	29	35	69	45	77	27	54	42

Part A

Draw a back to back stem and leaf display of the two point distributions. Put Wolverines on the left. Also include a legend.

Solution

For reading the Wolverines data, the middle column indicates the tens digit and the left column indicates the ones digit. For example, there were three games in which the Wolverines scored 30 or more points, 34, 37, and 38 points.

The same holds for the Cornhuskers, except in the opposite direction. So the Cornhuskers scored more than 70 only once when they scored 77.

Wolverines		Cornhuskers
011334678	2	779
478	3	58
	4	259
	5	46
	6	9
	7	7

Part B

Construct a frequency table and histogram for the Cornhuskers' point distribution. Your table should have the following column headings: points, frequency, relative frequency. Choose a 10-point interval – large enough so the data are grouped together but yields enough intervals to adequately show features of the distribution.

Solution

Here is the frequency table:

```
##    Freq Cumul relative
## 27    2     2  0.16667
## 29    1     3  0.08333
## 35    1     4  0.08333
## 38    1     5  0.08333
## 42    1     6  0.08333
## 45    1     7  0.08333
## 49    1     8  0.08333
## 54    1     9  0.08333
## 56    1    10  0.08333
## 69    1    11  0.08333
## 77    1    12  0.08333
```

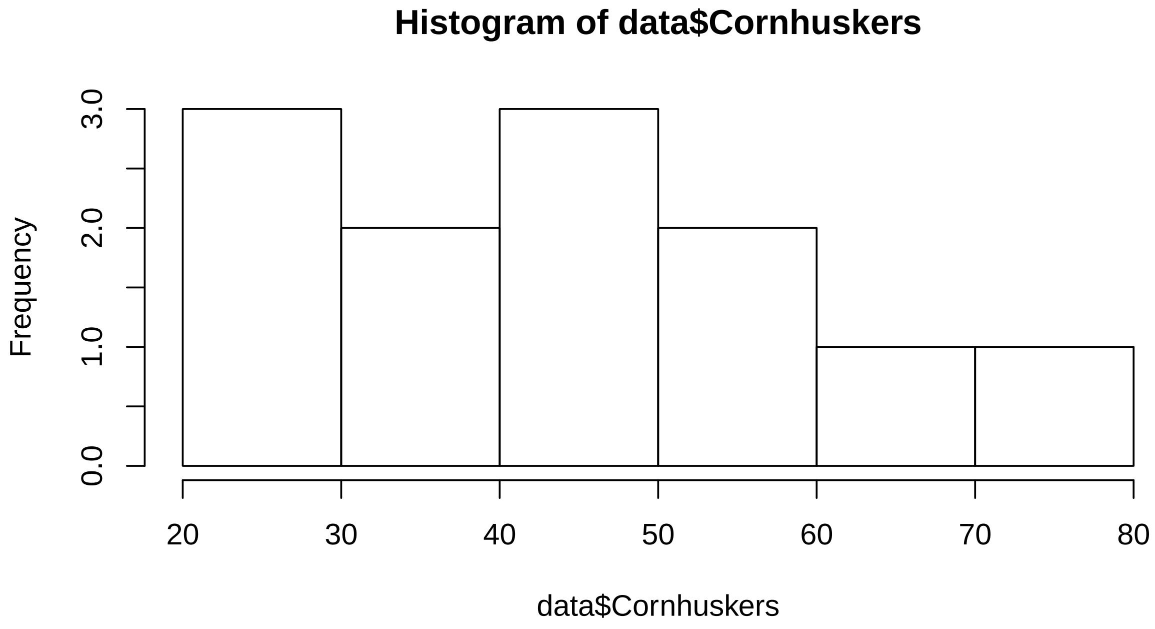

Figure 1: Histogram of Cornhuskers

Part C

From this histogram, describe the Cornhuskers' point distribution. Is it unimodal or bimodal? Is it symmetric or skewed? If skewed, in what direction?

Solution

It sort of is bimodal. It is skewed a bit to the left.

Part D

For both the Cornhuskers and the Wolverines, calculate the first quartile, the second quartile (median), the third quartile, the range, and the interquartile range (IQR).

Solution

Stat	Wolverines	Cornhuskers
1st Quartile	22	32
2nd Quartile	25	43.5
3rd Quartile	31	55
4th Quartile	38	77
Range	20 - 38	27 - 77
IQR $= Q_3 - Q_1$	9	23

Part E

Construct side-by-side boxplots comparing the data. Make the axis range from 20 to 80 points in increments of 10. Based on the boxplots, are there any *unusual* observations for either school?

Solution

Problem 1 continued on next page. . .

See Figure 2. I'd say the only unusual observation is the time that the Cornhuskers scored 77 points in a single game. That's 11 touchdowns and pretty crazy. It is unusualy but (obviously) it can happen rarely.

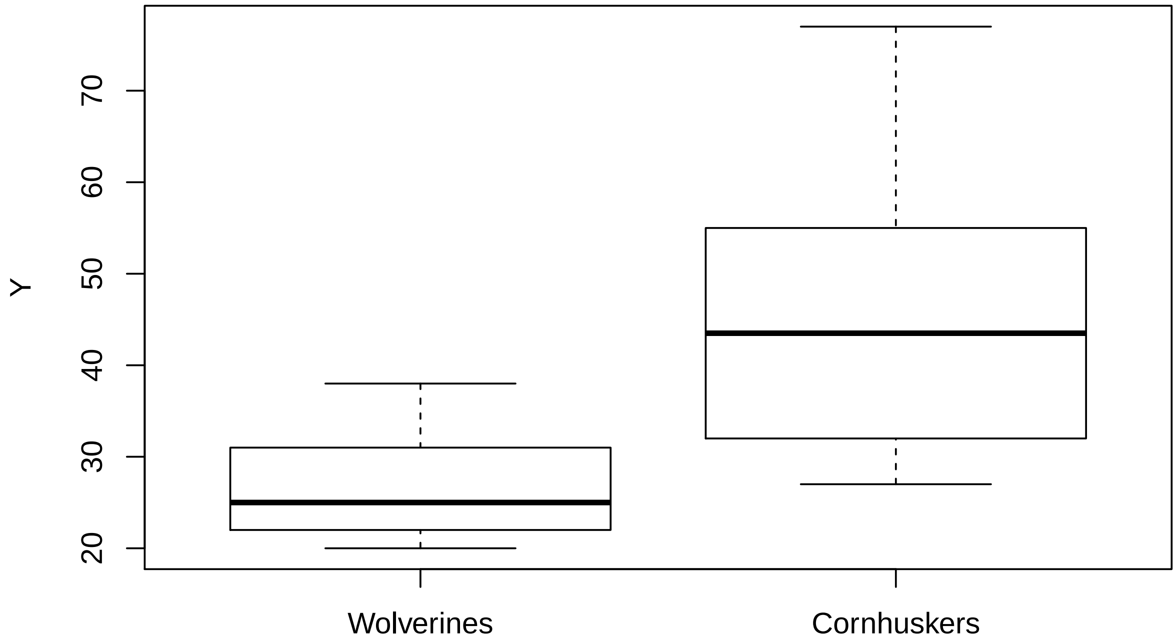

Figure 2: Wolverines vs Cornhuskers

Part F

Calculate the sample mean, sample variance, and sample deviation for each school.

Solution

We'll denote the statistic as having a subscript w for Wolverines and a subscript c for the Cornhuskers.

The sample mean is given to be $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. This gives us:

$$\bar{X}_w = 26.8333, \quad \bar{X}_c = 45.6667$$

The sample variance is given to be $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. This gives us:

$$s_w^2 = 39.4242, \quad s_c^2 = 261.3333$$

The sample deviation is then $s = \sqrt{s^2}$. This gives us:

$$s_w = 6.2789, \quad s_c = 16.1658$$

Problem 2

Below are the ages of all 44 US presidents at inauguration:

42, 43, 46, 46, 47, 47, 48, 49, 49, 50, 51, 51, 51, 51, 51, 52, 52, 54, 54, 54, 54, 54, 55, 55, 55, 55, 56, 56, 56, 57, 57, 57, 57, 58, 60, 61, 61, 61, 62, 64, 64, 65, 68, 69

Part A

Construct a stem and leaf plot of this data (use interval size of 5).

Solution

Where the first number is the range 0 to 4 and the second range is 5 to 9. Thus the first 4 has ages of 42 and 43.

4	23
4	6677899
5	0111112244444
5	555566677778
6	0111244
6	589

Part B

Do you think that these data come from a normal distribution? Justify your answer.

Solution

By tipping the stem plot on its side, this gives us the histogram. It looks very symmetrical and very unimodal. Thus the histogram ends up looking quite a bit like the Normal distribution.

Part C

Given your answer for (2b), which is more appropriate to use to describe this data set, the mean and standard deviation or the median and IQR? Explain. *Hint:* See top part of pg. 217 and the beginning of 8.2.6 of Baron.

Solution

We know that the median is far more robust than the mean when it comes to outliers. If our data had a lot of them, we might want to look at using the median to accurately capture the majority of the data.

However, our data doesn't appear to have any significant potential outliers. Therefore I'd say ysing the mean and IQR would be the best way to describe our data.

Problem 3

Problem 9.4 from Baron on pg. 301

A sample of 3 observations ($X_1 = 0.4, X_2 = 0.7, X_3 = 0.9$) is collected from a continuous distribution with density:

$$f(x) = \begin{cases} \theta x^{\theta-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Part A

Estimate θ using moment of methods.

Solution

Using the method of moments, we just let $\mu = \bar{X}$. This gives us:

$$\mu = \frac{1}{3}\sum_{i=1}^{3} X_i = 0.4 + 0.7 + 0.9 = \frac{2}{3}$$

Then we can use this for calculating our expected value of our density:

$$\int_0^1 X_i \cdot \theta X_i^{\theta-1} = \int_0^1 \theta X_i^{\theta} = \theta \int_0^1 X_i^{\theta} = \left[\frac{\theta X_i^{\theta+1}}{\theta+1}\right]_0^1 = \frac{\theta 1^{\theta+1} - \theta 0^{\theta+1}}{\theta+1} = \frac{\theta}{\theta+1} = m_1$$

By setting $m_1 = \mu_1$, we get:

$$m_1 = \frac{\theta}{\theta+1} = \frac{2}{3} = \mu$$

Which gives us $\theta = 2$.

Part B

Estimate θ using maximum likelihood.

Solution

Using the density, the maximum likelihood becomes:

$$\mathcal{L}(\theta \mid X) = \prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} \theta X_i^{\theta-1}$$

Using the sample that we were given, we have the values for it, this gives:

$$\mathcal{L}(\theta \mid X) = \theta(0.4)^{\theta-1} \cdot \theta(0.7)^{\theta-1} \cdot \theta(0.9)^{\theta-1} = \theta^3(0.252)^{\theta-1}$$

We can take the log of it to make it easier to work with, this gives us:

$$\log \mathcal{L}(\theta \mid X) = \log \theta^3(0.252)^{\theta-1} = 3\log\theta + (\theta-1)\log 0.252$$

We want the max or when the derivative is equal to 0, this gives:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}(\log \mathcal{L}(\theta \mid X)) \Rightarrow 0 &= \frac{3}{\theta} + \log 0.252 \\ \frac{3}{\theta} &= -\log 0.252 \\ \theta &= -\frac{3}{\log 0.252} \end{aligned}$$

This gives $\theta = 2.177$ as the estimator.

Problem 4

Problem 9.3, part c, from Baron on pg. 300

Estimate the parameter μ using maximum likelihood if a sample from Normal(μ, σ) distribution is observed, and we already know σ.

Solution

We know σ and have a sample $X = X_1, \dots, X_n$. We also know that the density for a normal distribution is given as:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

So if we determine our likelihood function, we want to maximize the probability given our sample and statistic to determine μ. Since we assume our sample is i.i.d., we get:

$$\mathcal{L}(\mu \mid \sigma, X_1, \dots, X_n) = \prod_{i=1}^{n} f(X_i)$$

we can simplify this by taking the log likelihood, and this becomes:

$$\begin{aligned} \log \mathcal{L}(\mu \mid \sigma, X_1, \dots, X_n) &= \sum_{i=1}^{n} \log f(X_i) \\ &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} + \left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \end{aligned}$$

We know we want to maximize the likelihood, so we can take the derivative and then determine where it equals 0. This gives:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}(\log \mathcal{L}(\mu \mid \sigma, X_1, \dots, X_n)) &= \sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}x}(\log f(X_i)) \\ &= \sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}x}(\log \frac{1}{\sqrt{2\pi}\sigma} + \left(-\frac{1}{2\sigma^2}(X_i-\mu)^2\right)) \\ &= \sum_{i=1}^{n} -\frac{1}{2\sigma^2}\left(2(X_i-\mu)\right) \\ &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) \end{aligned}$$

Settings this to zero gives:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}(\log \mathcal{L}(\mu \mid \sigma, X_1, \dots, X_n)) \Rightarrow 0 &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) \\ &= \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu \\ &= \sum_{i=1}^{n} X_i - n\mu \end{aligned}$$

Which means that $\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$ using maximum likelihood.

Problem 5

Problem 9.3, part a, from Baron on pg. 300

Estimate the parameters a and b using maximum likelihood if a sample from Uniform(a, b) distribution is observed.

Solution

We know that the pdf of the uniform distribution is:

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

Given a sample, $X = X_1, \ldots, X_n$, the likelihood is then:

$$\begin{aligned} \mathcal{L}(a, b \mid X) &= \prod_{i=1}^{n} f(X_i) \\ &= \prod_{i=1}^{n} \frac{1}{b-a} \\ &= \frac{1}{(b-a)^n} \end{aligned}$$

When thinking about this in terms of our sample, we know that we will have a range of values. Therefore it should be somewhat obvious that we can guarantee with probability 1 that if we choose the parameters such that $a = \min(X_1, \ldots, X_n)$ and $b = \max(X_1, \ldots, X_n)$, that all our samples will fall within the range a to b and the likelihood is maximized.