Anthony Cunningham

STAT: 3200

1/27/2017

**Homework 1**

1). A. > mydata = rnorm(100, 2, 5)

> mydata[1:10]

[1] 4.19218692 0.05897917 0.15505759 7.21795588 3.07124710 4.81172880

[7] 4.60032086 -7.52794243 -2.17382118 1.63853242

> mean(mydata)

[1] 3.33458

> var(mydata)

[1] 23.80816

> hist(mydata)

> boxplot(mydata)

> qqnorm(mydata)

> qqline(mydata)



B. Yes, I would say that the data does follow an approximate normal distribution. The histogram is relatively symmetric, while the box plot doesn’t lean too far to any part of the Interquartile Range; it is also relatively symmetric. The box plot also shows a few outliers that substantially deviate from the mean, perhaps making the data set less normal than if we were to ignore those observations. The QQ plot shows that most observations fall near the normal reference line and that there are only a couple of observations that deviate from a normal distribution.

2). A. > head(Prestige)

education income women prestige census type

gov.administrators 13.11 12351 11.16 68.8 1113 prof

general.managers 12.26 25879 4.02 69.1 1130 prof

accountants 12.77 9271 15.70 63.4 1171 prof

purchasing.officers 11.42 8865 9.11 56.8 1175 prof

chemists 14.62 8403 11.68 73.5 2111 prof

physicists 15.64 11030 5.13 77.6 2113 prof



B. 1. The data doesn’t appear to be very symmetrical. In fact, it appears skewed to the right. The data appears to be unimodal, with the apex topping right near $5,000. It seems that any observation greater than $15,000 appears to be extreme; certainly anything greater than $20,000 would be considered extreme in this dataset.

2. There appears to be some deviation from normality in this dataset. Those extreme values in the tail of the graph could be problematic for our assumption of normality when analyzing this dataset.

C. > summary(Prestige$income)

Min. 1st Qu. Median Mean 3rd Qu. Max.

611 4106 5930 6798 8187 25880

1. The mean ($6798) is larger than the median ($5930) due to the data being skewed to the right.

2. Yes, the numerical summary and histograms match very well. The majority of observations occur right around $5,000 according to the numerical summary, which is where the box on the boxplot hovers around.

D. > boxplot(Prestige$income)



1. The boxplot shares the same distribution as the histograms.

2. However, what the boxplot does show (and the histograms do not) are the quantifiable outliers in the dataset. High outliers appear to begin at about $15,000, about where I estimated the extreme value threshold would be.

E. > qqnorm(Prestige$income)

> qqline(Prestige$income)



1. The distribution of the dataset doesn’t appear to be very normal. The outliers near the top-right of the QQ plot contribute to making the dataset seem less normal. If one were to ignore those points, most of the other observations would be on or near the normal reference line, signaling normality.

2. The upward bow shows that the data is skewed to the right, as the values of corresponding percentiles are significantly above the normal reference line. This can also be verified by looking at a histogram and a boxplot, where there is a long tail where income increases.

3). A. The last problem showed a distribution that was positively skewed. In this situation, decreasing the power would transform the distribution of the data and make it more normally-distributed.

B. > income2 = (Prestige$income)^2

> par(mfrow=c(2,1))

> hist(income2)

> qqnorm(income2)

> qqline(income2)



\* The square transformation made the data even more skewed to the right and farther depart from normality. All of those outlier values became even greater outliers compared to the bulk of the data since they were large outliers, outliers greater than the bulk of the original dataset. Ascending the ladder of powers made matters worse, further deviating from normality.

> logincome = log(Prestige$income)

> par(mfrow=c(2,1))

> hist(logincome)

> qqnorm(logincome)

> qqline(logincome)



\* The log transformation appears to have transformed the distribution of the dataset more towards normality as compared to the distribution of the original dataset. While the distribution does appear slightly skewed to the left, it doesn’t appear to be as blatant as the original dataset. I believe that the log transformation (descending the ladder of powers) succeeded in transforming the distribution of the data to be more normally-distributed.

> sqrtincome = sqrt(Prestige$income)

> par(mfrow=c(2,1))

> hist(sqrtincome)

> qqnorm(sqrtincome)

> qqline(sqrtincome)



\* The square root transformation appears to have succeeded in making the distribution of the original dataset more normal. There is still a slight skew to the right, but I don’t think it will have as much an effect when analyzing the data as the original dataset.

C. I think that the log transformation and the square root transformation succeeded in making the distribution of the original dataset appear more normal. I think that one would be fine with choosing either transformation. If I had to choose one, I would probably go with the log transformation since its tail appears to be a bit smaller than the tail on the square root transformation.