American Community Survey Analysis

Frank Neugebauer

October 3, 2019

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.1

# Import the data  
comm\_sur <- read.csv("acs-14-1yr-s0201.csv", header=TRUE)

The data set is documented using different built-in functions.

str(comm\_sur)

## 'data.frame': 136 obs. of 8 variables:  
## $ Id : Factor w/ 136 levels "0500000US01073",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ Id2 : int 1073 4013 4019 6001 6013 6019 6029 6037 6059 6065 ...  
## $ Geography : Factor w/ 136 levels "Alameda County, California",..: 56 70 98 1 20 43 62 68 92 106 ...  
## $ PopGroupID : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ POPGROUP.display.label: Factor w/ 1 level "Total population": 1 1 1 1 1 1 1 1 1 1 ...  
## $ RacesReported : int 660793 4087191 1004516 1610921 1111339 965974 874589 10116705 3145515 2329271 ...  
## $ HSDegree : num 89.1 86.8 88 86.9 88.8 73.6 74.5 77.5 84.6 80.6 ...  
## $ BachDegree : num 30.5 30.2 30.8 42.8 39.7 19.7 15.4 30.3 38 20.7 ...

nrow(comm\_sur)

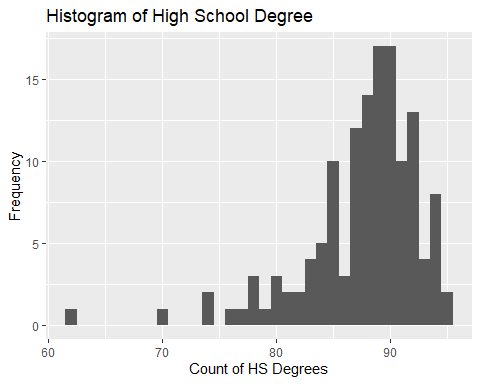
## [1] 136

ncol(comm\_sur)

## [1] 8

Use of ggplot to create a histogram.

ggplot(comm\_sur, aes(x = HSDegree)) + geom\_histogram(binwidth=1) +  
 ggtitle("Histogram of High School Degree") +  
 labs(x = "Count of HS Degrees", y = "Frequency")



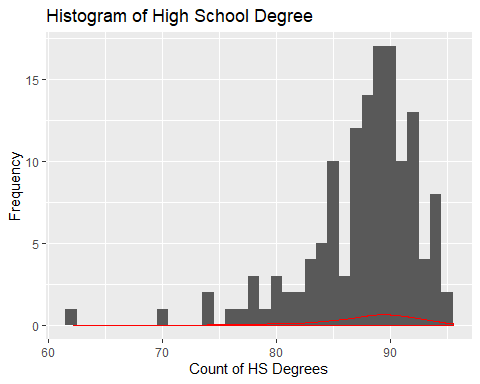
Histogram interpretation:

* The histogram does not appear to be unimodel. While there is one dominant hump at 90, there’s also other smaller ones at 85 and ~94.
* The histogram is clearly skewed right, with counts between 85 and 95 dominating the graph.
* This histogram is not bell-shaped; there are clear sub-peaks at 85 and ~94 and valleys at ~86 and ~93. (There are also smaller peaks and valleys from ~76 to ~85.)
* Looking at the historgram as a whole (and without any other calculations), I’d say it is not a normal distribution; the distribution goes from about 65 - 100 with the majority of degrees between 85 and 100, which implies the distribution is skewed towards the higher values.

Continuing on…

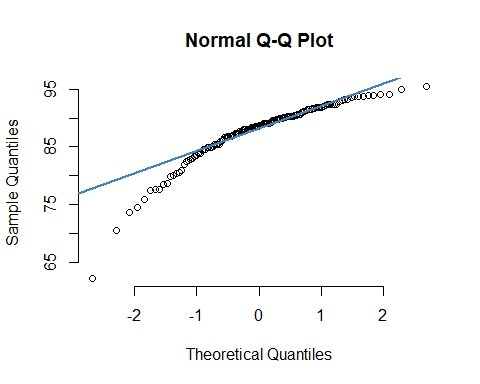
Adding a normal curve to the histogram for comparision.

ggplot(comm\_sur, aes(x = HSDegree)) + geom\_histogram(binwidth=1) +   
 ggtitle("Histogram of High School Degree") +  
 labs(x = "Count of HS Degrees", y = "Frequency") +  
 geom\_density(aes(y=0.045\*..count..), colour="red")



Created a Probability Plot of the HSDegree variable.

qqnorm(comm\_sur$HSDegree, pch = 1, frame = FALSE)  
qqline(comm\_sur$HSDegree, col = "steelblue", lwd = 2)



* The solid line is a reference line of the probability plot.
* While visualizaton makes it looks visually ‘close’ to normal, the fact remains that The pattern of data is non-linear. This says to me that the distribution is (still) not normal, providing more statistical evidence to the visual conclusion earlier.
* The values are far more closely distributed between ~-.5 and 1 and on a scale of -2 to 2, this implies a skew to the right.

Using stat.desc() function from pastecs.

library(pastecs)

## Warning: package 'pastecs' was built under R version 3.6.1

stat.desc(comm\_sur$HSDegree, basic = TRUE, desc = TRUE, norm = TRUE)

## nbr.val nbr.null nbr.na min max   
## 1.360000e+02 0.000000e+00 0.000000e+00 6.220000e+01 9.550000e+01   
## range sum median mean SE.mean   
## 3.330000e+01 1.191800e+04 8.870000e+01 8.763235e+01 4.388598e-01   
## CI.mean.0.95 var std.dev coef.var skewness   
## 8.679296e-01 2.619332e+01 5.117941e+00 5.840241e-02 -1.674767e+00   
## skew.2SE kurtosis kurt.2SE normtest.W normtest.p   
## -4.030254e+00 4.352856e+00 5.273885e+00 8.773635e-01 3.193634e-09

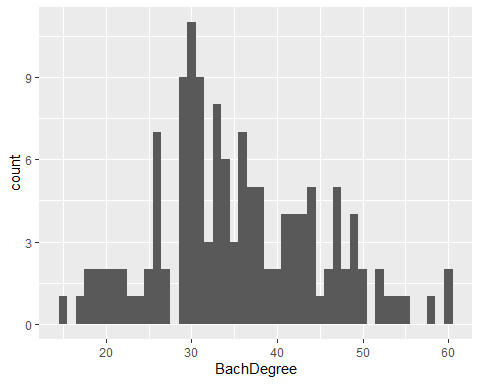
Tesults produced for skew, kurtosis, and z-scores.

* Skew - the value -1.67 says that the distribution has a long tail in the negative direction. It also says that median is larger than the mean, which is true (88.70 to 87.63). The significant number of values on the tail pulls the mean down. More to the points made previously, the mass of values is on the right.
* Kurtosis - this measure desribes the measure of both tails (unlike skew). Stated another way, it’s a measure of tailedness. (Sources: <https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics> and <https://en.wikipedia.org/wiki/Kurtosis>). The value 4.35 shows that HSDegree is leptokurtic (>0) with too many scores in the tails (in this case on the left tail).
* z-scores - generally, according to the Discovering Statistics Using R “text”, z-score is “the value of an observation expressed in standard deviation units.” However, the z-score is not computed using stats.desc(). The value can be calculated, however, using z = (X - μ) / σ where z is the z-score, X is the value of the element, μ is the population mean, and σ is the standard deviation. In this case, we have no X, so I’m unsure what is expected.

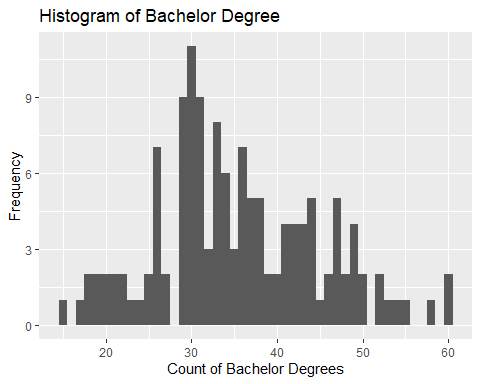
Repeating the process for the BachDegree variable.

Histogram.

ggplot(comm\_sur, aes(x = BachDegree)) + geom\_histogram(binwidth=1)



ggplot(comm\_sur, aes(x = BachDegree)) + geom\_histogram(binwidth=1) +   
 ggtitle("Histogram of Bachelor Degree") +  
 labs(x = "Count of Bachelor Degrees", y = "Frequency")

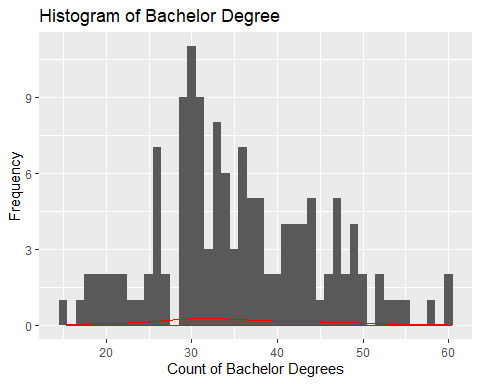


Analysis:

1. The histogram does not appear to be unimodel. While there is one dominant hump at 30, there’s also other smaller ones throughout the disribution, especially above 30.
2. The histogram is clearly skewed left, with counts below 40 dominating the graph.
3. There are clear sub-peaks throughout and the peak itself is (fairly far) left of center. It is not bell-shaped.
4. Looking at the historgram as a whole (and without any other calculations), I’d say it is not normal; the distribution is skewed towards the lower values.

Adding normal curve to the histogram:

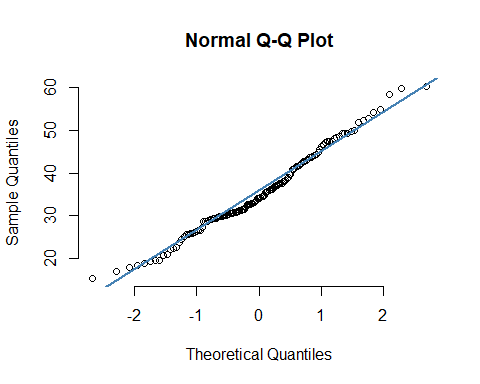
ggplot(comm\_sur, aes(x = BachDegree)) + geom\_histogram(binwidth=1) +   
 ggtitle("Histogram of Bachelor Degree") +  
 labs(x = "Count of Bachelor Degrees", y = "Frequency") +  
 geom\_density(aes(y=0.045\*..count..), colour="red")



A normal distribution cannot accurately be used as a model for this data because again, the data is so heavily skewed left and the data is also fairly random looking at all the peaks and valleys.

Probability Plot of the HSDegree variable:

qqnorm(comm\_sur$BachDegree, pch = 1, frame = FALSE)  
qqline(comm\_sur$BachDegree, col = "steelblue", lwd = 2)



The solid line is a reference line of the probability plot.

The Q-Q plot shows a pretty normal distribution, given how close the ideal line the values fall.

Adding stat.desc() function.

library(pastecs)  
stat.desc(comm\_sur$BachDegree, basic = TRUE, desc = TRUE, norm = TRUE)

## nbr.val nbr.null nbr.na min max   
## 136.00000000 0.00000000 0.00000000 15.40000000 60.30000000   
## range sum median mean SE.mean   
## 44.90000000 4822.70000000 34.10000000 35.46102941 0.81545273   
## CI.mean.0.95 var std.dev coef.var skewness   
## 1.61271456 90.43498856 9.50973126 0.26817415 0.32843046   
## skew.2SE kurtosis kurt.2SE normtest.W normtest.p   
## 0.79035382 -0.27742492 -0.33612576 0.98316075 0.09206162

Further analysis:

* Skew - the value .328 says that the distribution has a short tail in the positive direction. This appears to be a more normal distribution.
* Kurtosis - the value -0.28 shows that BachDegree is platykurtic (<0) with the values (marginally) flat vis-a-vis the tails.
* z-scores - generally, according to the Discovering Statistics Using R “text”, z-score is “the value of an observation expressed in standard deviation units.” However, the z-score is not computed using stats.desc(). The value can be calculated, however, using z = (X - μ) / σ where z is the z-score, X is the value of the element, μ is the population mean, and σ is the standard deviation. In this case, we have no X, so I’m unsure what is expected.