Assigment 2.1

Roni Kaakaty

6/13/2020

2014 American Community Survey

1. What are the elements in your data (including the categories and data types)? Categories: Data Type: Id Character Id2 Integer Geography Character PopGroupID Integer Total Population Character Races Reported Integer HS Degree Numerical Bach Degree Numerical

library(ggplot2)  
census\_df <- read.csv("data/acs-14-1yr-s0201.csv")  
x\_data <- census\_df

1. Output from the following functions: str(), nrow(), and ncol().

str(x\_data)

## 'data.frame': 136 obs. of 8 variables:  
## $ Id : chr "0500000US01073" "0500000US04013" "0500000US04019" "0500000US06001" ...  
## $ Id2 : int 1073 4013 4019 6001 6013 6019 6029 6037 6059 6065 ...  
## $ Geography : chr "Jefferson County, Alabama" "Maricopa County, Arizona" "Pima County, Arizona" "Alameda County, California" ...  
## $ PopGroupID : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ POPGROUP.display.label: chr "Total population" "Total population" "Total population" "Total population" ...  
## $ 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(x\_data)

## [1] 136

ncol(x\_data)

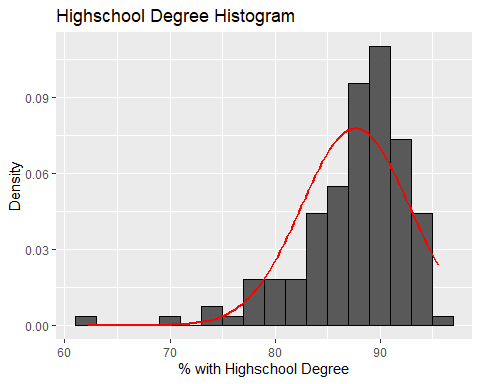
## [1] 8

1. Create a Histrogram of the HSDegree variable:

summary (x\_data$HSDegree)

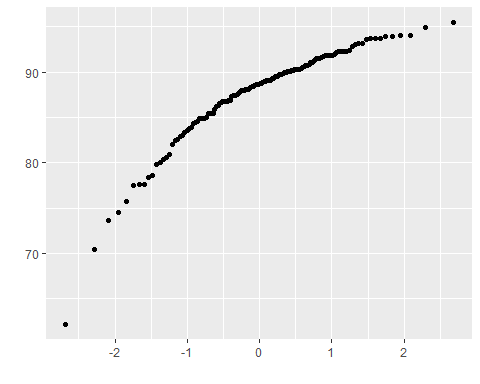
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 62.20 85.50 88.70 87.63 90.75 95.50

hs\_mean <- mean(x\_data$HSDegree)  
hs\_std <- sqrt(var(x\_data$HSDegree))  
hist.HS <- ggplot(x\_data, aes(HSDegree)) + geom\_histogram(binwidth = 2, colour = "black", aes(y = ..density..)) + ggtitle("Highschool Degree Histogram") + xlab("% with Highschool Degree") + ylab("Density")  
hist.HS + stat\_function(fun = dnorm, args = list(mean = hs\_mean, sd = hs\_std), colour = "red", size = 1)



1. Answer the following questions based on the histogram above:
2. Is the distribution unimodal? Yes, there is a peak followed by a descension.
3. Is it approximately symmetrical? No.
4. Is it bellshaped? Yes.
5. Is it approximately normal? No.
6. If not normal, is it skewed? If so, what direction? Yes, skewed to the left.
7. Normal curve included.
8. I believe a normal curve can be used for this data as the curve aligns with a majority of the data.
9. Create a Probability plot of the HSDegree variable.

qqplot.HSDegree <- qplot(sample = x\_data$HSDegree)  
qqplot.HSDegree



1. Answer the following questions based on the Probability plot:
2. No, the distribution isn’t normal because normal distrubtion would result in a linear line.
3. Yes, skewed to the left (negatively skewed), since most of the clustering is at the high end of the scale.
4. Utilize stat.desc():

install.packages('pastecs', repos='http://cran.us.r-project.org')

## Installing package into 'C:/Users/Roni Kaakaty/Documents/R/win-library/4.0'  
## (as 'lib' is unspecified)

## package 'pastecs' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\Roni Kaakaty\AppData\Local\Temp\Rtmpuahbkr\downloaded\_packages

library(pastecs)  
stat.desc(x\_data$HSDegree, basic = FALSE, norm = TRUE)

## median mean SE.mean CI.mean.0.95 var   
## 8.870000e+01 8.763235e+01 4.388598e-01 8.679296e-01 2.619332e+01   
## std.dev coef.var skewness skew.2SE kurtosis   
## 5.117941e+00 5.840241e-02 -1.674767e+00 -4.030254e+00 4.352856e+00   
## kurt.2SE normtest.W normtest.p   
## 5.273885e+00 8.773635e-01 3.193634e-09

1. Explain skew, kurtosis and z-score.In addition, what would a change in sample size do?

A negative skew score that’s given above indicates that the mean of the data values is less than the median, which results in the data distribution being left-skewed. A kurtosis greater than 3 (in this instance 4.353) indicates a heavier tail which typically means more outliers.A z-score indicates how many standard deviations from the mean your data is. Large sample sizes will produce small standard errors so it’s important to increase the absolute value criteria in large samples.In samples greater than 200, it’s important to look at the shape of the distribution, rather than calculate their significance.