McKibben DSC520 Ex. 3.2

Makayla McKibben

2024-06-25

```
knitr::opts_chunk$set(echo = TRUE)
```

```
# Class: DSC520
# Assignment: Exercise 3.2
# Author: Makayla McKibben
# Date: 06.25.2024

# Installing the necessary packages
install.packages("ggplot2")
library(ggplot2)
install.packages("pastecs")
library(pastecs)

# Reading the data from the csv file
ACSData <- read.table(file = ’ACSData.csv’, header = TRUE, sep =",", stringsAsFactors = FALSE)

# Getting information about the data from the csv file
colnames(ACSData)
str(ACSData)
nrow(ACSData)
ncol(ACSData)

# Comments in lines 22-29 describe the headers from the data in the csv file
# Id-character I believe that Id is a unique identifier (like a name but anonymous)
# for the imported data
# Id2-integer I think Id2 is associated with the next column and denotes location information
# Geography-character The Geography header has to do with location information
# PopGroupID-integer is a group that all of these surveyed people fall into
# POPGROUP.display.label-character is a group that denotes the total population of the
# geography header’s location
# RacesReported-integer I think RacesReported is not actually having to do with race
# but is the total population of the county from the geography header
# HSDegree-numeric should be the percentage of how many respondents have a high school
# diploma in that county
# BachDegree-numeric is similarly the percentage of respondents that have a bachelor’s
# degree in that county

# This section of code takes some of the data from the csv file in order to line up the
# normal distribution on the plot
xmin <- min(data.frame(ACSData$HSDegree))
xmax <- max(data.frame(ACSData$HSDegree))
```

```
xrange <- xmax - xmin
change_increment <- 0.007353*xrange
x <- seq(xmin, xmax, by = change_increment)
norm_dis <- dnorm(x, mean = 88.8, sd = 6.8)
norm_dis <- norm_dis*10^2.42

# This is where the first bit of plotting actually happens
ACS_Histogram <- ggplot(ACSData, aes(HSDegree))
ACS_Histogram + geom_histogram(binwidth = 0.88, fill = "lightgrey", color = "black") +
  labs(title = "American Community Survey Data", x ="Percentage of County with a HS Diploma",
       y = "Frequency") +
  xlim(60, xmax) + geom_line(aes(x, norm_dis, color = "red"))

# Comments in lines 47-53 answer questions about the plotted data
# Yes, this is a unimodal distribution.
# It is not quite symmetrical.
# No, it is not evenly bell-shaped.
# No, it is not normal, it has a skew.
# It is negatively skewed.
# This data should not be modeled by a normal distribution due to the skewing of the data.
# There are other models that can account for skew and we should use one of those.

# Second bit of plotting here
ACS_Probability <- ggplot(ACSData, aes(HSDegree))
ACS_Probability + geom_density()+ xlim(70, 100)

# Comments in lines 60-65 answer questions about the plotted data
# The distribution is not normal.
# This is evidenced by the fact that as the ends of the lines around the curve
# approach the x-axis,
# there is a significant gap on the left hand side where the right approaches asymptotically.
# These addresses the next question as well. We can see the left hand side does
# not approach the x-axis
# in the same manner as the right, then there is a peak further to the right than
# left of the data indicating
# a negatively skewed distribution as there are more data points on the left than the right.

# Pulling statistical information about the data
stat.desc(ACSData$HSDegree, norm = TRUE)

# Comments in lines 71-81 answer questions about the statistical evaluation of the data
# Skew for a standard normal distribution is 0, we have a value of -1.67.
# This value matches with the observations of the graph that I made where I indicated
# it appeared to have a negative skew.
# Kurtosis for a standard normal distribution is 3.
# Our kurtosis is 4.35 which matches the heavier tail we see in the graph.
# For calculating z-scores we have a standard deviation of 5.12 and a mean of 87.6
# and if you calculate a z-score you will produce a value that tells you how
# many standard deviations you are away from the mean.
# If positive you are above the mean, if negative below. These z-scores are
# typically applied to normal distributions though, and should probably not be used for this data
# since it is not normal.
# If we were to increase the number of data points we would begin to come closer to
# a normal distribution.
# This makes sense if your population is not large enough you will not see the true distribution.
# Increasing the number of data points decreases the value of both skewness and kurtosis.
```