**Solution**

1. **Load the file Lab Week 7 - NHANES subset.csv into Excel and R**

***A screenshot of a cell phone

Description automatically generated***

1. **Inspect the NHANES subset (n=1000) by computing descriptive statistics on SBP characterizing the cohort. Functions: summary, mean, median, sd, standard error, 25th and 75th percentile, IQR in Excel and R. Compare the values**

Summary

A close up of a logo

Description automatically generated

Mean

124.9

Median

122

SD

20.28431

25th Percentile

112

75th Percentile

136

IQR

24

In excel, we got –

|  |  |
| --- | --- |
| *SBP* | |
|  |  |
| Mean | 124.922 |
| Standard Error | 0.641446311 |
| Median | 122 |
| Mode | 116 |
| Standard Deviation | 20.28431338 |
| Sample Variance | 411.4533694 |
| Kurtosis | 3.007555303 |
| Skewness | 1.326154746 |
| Range | 148 |
| Minimum | 82 |
| Maximum | 230 |
| Sum | 124922 |
| Count | 1000 |
| 25th Percentile | 112 |
| 75th Percentile | 136 |
| IQR | 24 |

*In R and Excel, we got the same results.*

1. **Build histograms of SBP indicating mean and 1 SD (R). Try three different approaches (differing in the range of the bins)**

**A screenshot of a cell phone

Description automatically generated**

BREAKS= 60

A screenshot of a cell phone

Description automatically generated

BREAKS = 50

A screenshot of a cell phone

Description automatically generated BREAKS = 20

**4.Compare the standard deviation and the standard error of SBP. What do you notice?**

Standard Deviation = 20.284321

Standard Error = 0.6414463

SD tells us how the data is distributed around the mean.

The **standard error tells** us how accurate the mean of any given sample from that population is likely to be compared to the true population mean.

A close up of text on a white background

Description automatically generated

**5.Draw 3 random samples (n=100) and build three objects – how do standard deviation and standard error of SBP compare to each other and to the overall population. How do histograms of SBP compare?**

*Sample 1*

Standard Deviation

[1] 18.80378

Standard Error

[1] 1.880378

A screenshot of a cell phone

Description automatically generated

*Sample 2*

Standard Deviation

[1] 24.23183

Standard Error

[1] 2.423183

A screenshot of a cell phone

Description automatically generated

*Sample 3*

Standard Deviation

[1] 21.68903

Standard Error

[1] 2.168903

A screenshot of a cell phone

Description automatically generated

*The population figure is as follows –*

Standard Deviation = 20.284321

Standard Error = 0.6414463

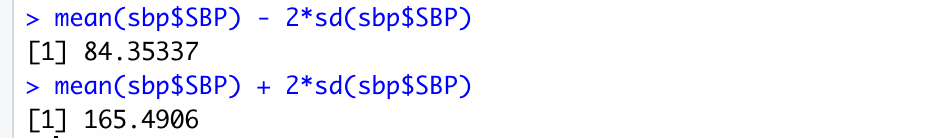
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SAMPLE 1 | SAMPLE 2 | SAMPLE 3 | POPULATION |
| SD | 18.80378 | 24.23183 | 21.68903 | 20.284321 |
| SE | 1.880378 | 2.423183 | 2.168903 | 0.6414463 |

Sample 1 has the SE closest to population. Sample 1 and 3 has SD close to population. The difference in the samples is due to sampling variation.

The classes of the histogram will be of lesser size in case of the samples of size 100.The histogram of sample 1 shows less skewness and resembles the histogram of population.

**6.What interval contains 95% of all SBP values in R and Excel?**

The following 2 limits contains 95% of all values in the dataset –



Lower Limit – 84.35337

Upper Limit – 165.4906

**CODE**

#Clear Enviornment

rm(list = ls())

pathname<-"C:/BME/Sem 2/Biostat/Class 7/"

#Load file

df<-read.csv(paste(pathname,"Lab Week 7 - NHANES subset.csv",sep=""))

summary(df)

mean(df$SBP)

median(df$SBP)

sd(df$SBP)

# New package for standard error formula

library(plotrix)

std.error(df$SBP)

quantile(df$SBP)

quantile(df$SBP)[2]

quantile(df$SBP)[4]

IQR(df$SBP)

pdf(paste(pathname,"Grade Lab.pdf",sep=""))

{

{

hist(df$SBP,breaks = 60)

abline(v=mean(df$SBP)+sd(df$SBP), lty=2,col="red")

abline(v=mean(df$SBP),lty=1,col="blue")

abline(v=mean(df$SBP)-sd(df$SBP), lty=2,col="red")

text(x=175,y=150,labels="mean+/-sd")

}

{

hist(df$SBP,breaks = 50)

abline(v=mean(df$SBP)+sd(df$SBP), lty=2,col="red")

abline(v=mean(df$SBP),lty=1,col="blue")

abline(v=mean(df$SBP)-sd(df$SBP), lty=2,col="red")

text(x=175,y=150,labels="mean+/-sd, SE")

}

{

hist(df$SBP,breaks = 20)

abline(v=mean(df$SBP)+sd(df$SBP), lty=2,col="red")

abline(v=mean(df$SBP),lty=1,col="blue")

abline(v=mean(df$SBP)-sd(df$SBP), lty=2,col="red")

text(x=175,y=150,labels="mean+/-sd")

}

}

dev.off()

# Standard deviation and standard error

sd(df$SBP)

std.error(df$SBP)

#Sample 1

set.seed(123)

index1 <- sample(1:nrow(df), 100)

Sample1 <- df[index1,]

View(Sample1)

nrow(Sample1)

sd(Sample1$SBP)

hist(Sample1$SBP)

library(plotrix)

std.error(Sample1$SBP)

#Sample 2

set.seed(124)

index2 <- sample(1:nrow(df), 100)

Sample2 <- df[index2,]

sd(Sample2$SBP)

hist(Sample2$SBP)

library(plotrix)

std.error(Sample2$SBP)

#Sample 3

set.seed(125)

index3 <- sample(1:nrow(df), 100)

Sample3 <- df[index3,]

sd(Sample3$SBP)

hist(Sample3$SBP)

library(plotrix)

std.error(Sample3$SBP)

mean(df$SBP) - 2\*sd(df$SBP)

mean(df$SBP) + 2\*sd(df$SBP)