**Lab 3: Normal Distribution**

In this lab, we will learn how to use R instead of Table A. Specifically, we will

* Find area under the curve to the left of the z-score
* Find the critical z-value for a given area under the Standard Normal curve
* Compare theoretical and empirical distributions

You will submit:

* a **Word/PDF** of your properly formatted answers to the lab questions, including ***relevant*** output where appropriate;
* a copy of your nicely commented *do-file* that can be used to reproduce all of the analysis that you conducted.

**Part 1:**

For Part 1 of this lab, we will learn to use R as a “calculator” for normal distribution quantities.



1. The BMI of American adults follows a Normal Distribution with mean 26.5 and standard deviation 4.2. Use R to:

1. Find the probability that a randomly selected adult is underweight (i.e. has a BMI less than 18.5).
2. Find the probability that a randomly selected adult is overweight or obese (i.e. has a BMI greater than 25.0)
3. Find the proportion of American adults who have a healthy weight (i.e. a BMI between 18.5 and 25.0)

2. Use R to:

1. Find the critical z-score such that 30% of the area under the normal curve is to the left of this value.
2. Find the BMI level such that 30% of American adults have a BMI less than that number.

**Part 2:**

For Part 2 of this lab, we will return to the *tango* dataset from Lab 1.

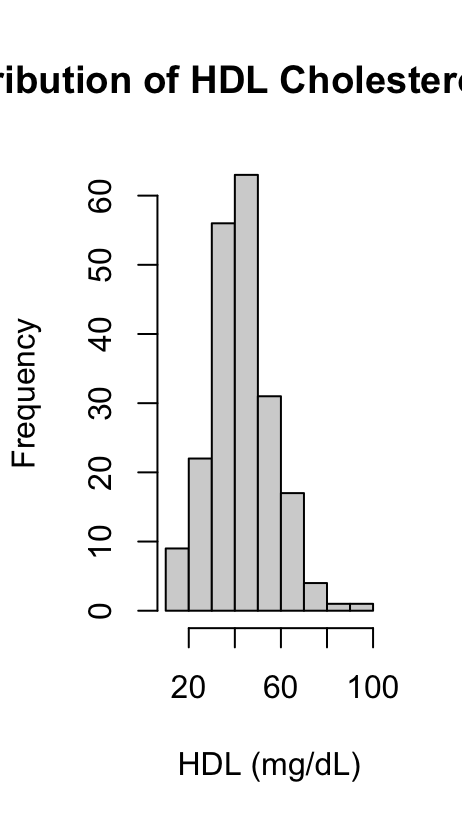
**Preparation:**

1. Create a new folder on your computer called *Lab 3*.
2. Download the dataset *tango.csv* from Canvas into the Lab 3 folder (or copy/paste the *tango.csv* dataset from your Lab 1 folder to your Lab 3 folder).

1. Read the dataset *tango.csv* into R, naming it **tg**.

tg<-read.csv("/Users/zhaozhan/Downloads/Tufts\ University/principles\ of\ Biostatistics/week3/IN\_class/tango-5.csv")

2. Conduct an exploratory data analysis of the variable **High Density Lipoprotein (HDL) Cholesterol** (*hdl)*. This should include both numerical and graphical summaries. Write a very short paragraph describing what this variable is (you may use Google to learn the basics of HDL and the units it is measured in) and the distribution of this variable in this dataset.



Answer: *The High-Density Lipoprotein cholesterol,commonly referred to as “good cholesterol,”helps remove excess cholesterol from the bloodstream and is typically measured in milligrams per deciliter. In this dataset, the variable hdl is approximately symmetric and follows a roughly bell-shaped distribution.The mean HDL Level is 43.51 mg/dL and the median is 43.99 mg/dL, with a standard deviation of 13.34 mg/dL and an interquartile range iqr of 17mg/dL. These values indicate moderate variability around the center. A histogram provides a clear view of the shape ,center ,and spread,and it shows that the distribution is close to normal .Two outliers were identified at 81 and 92 mg/dL,and there is also one missing value in the dataset.*

*Hint:* For numerical summaries, if the variable is fairly symmetric, then the mean and standard deviation are appropriate measures of center and spread. If the distribution is skewed, then median and IQR are more appropriate measures of center and spread.

1. Look for outliers, and if there are any, look into them.
2. For a quantitative variable, graphical summaries include histograms and boxplots. You don’t always have to include both – depending on the situation you may choose to include one or the other, or both. Think about when one is more useful than another.
3. A *nice* graphic includes an appropriate title and properly labeled x- and y-axes (including correct measurement units where appropriate).
4. When describing the distribution of the variable, make sure you use the appropriate measurement units.

3. (Re)draw a histogram of *hdl*, and superimpose the normal curve on it.

Syntax: **hist(tg$hdl, freq=FALSE)**

**curve(dnorm(x, mean(tg$hdl, na.rm=TRUE), sd(tg$hdl, na.rm=TRUE)), add=TRUE)**

hist\_Hdl2 <- hist(tg$hdl,

freq = FALSE,

main = "Distribution of HDL Cholesterol Levels",

xlab = "HDL (mg/dL)",

ylab = "Density")

curve(dnorm(x,

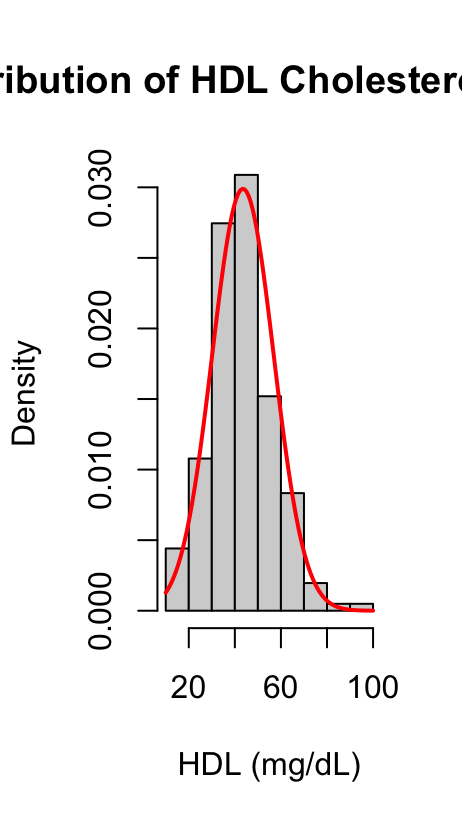
mean = mean(tg$hdl, na.rm = TRUE),

sd = sd(tg$hdl, na.rm = TRUE)),

add = TRUE,

col = "red",

lwd = 2)



*Notes:*

1. Make sure you modify the code above to make the graph *nice*.
2. What does adding **freq=FALSE** after the comma do? This tells R that on the y-axis, you want to see the “density” instead of the “frequency” (i.e., the number of events). (Try running the above syntax with **freq=TRUE** and see what the graph looks like).
3. What does adding the **curve** command with **add=TRUE** do? It “superimposes” the normal curve on the histogram. How? What this R command is doing is that it is using the variable *hdl*, calculating its *empirical* mean and standard deviation (perhaps like you already did in question 2 above), and then simply making a density curve for the normal distribution on top of the histogram.

4. Using the relevant numerical summaries from Question 2, if HDL is exactly normally distributed, what percentage of subjects would you expect to have HDLcholesterol between 40 and 60 mg/dL, inclusive?

P\_46<-pnorm(60,mean = mean(tg$hdl,na.rm = TRUE),sd=sd(tg$hdl,na.rm = TRUE))-pnorm(40,mean = mean(tg$hdl,na.rm = TRUE),sd=sd(tg$hdl,na.rm = TRUE))

P\_46

[1] 0.4955444

Answer: I expect the 49.55% to have HDLcholesterol between 40 and 60 mg/dL.

*Hint:* You can calculate this without looking at your histogram or dataset. Use Table A or the instructions in Part 1 of this lab to find the answer.

5. What percentage of subjects in our sample have HDL in this range?

Syntax: **sum(tg$hdl >= 40 & tg$hdl <= 60)**

Answer:

s46<-sum(tg$hdl>=40 & tg$hdl<=60,na.rm = TRUE)

nrow(tg)

P\_sa<-(s46/nrow(tg))\*100

49.26829

6. If HDL is exactly normally distributed, what percentage of subjects would you expect to have HDLcholesterol in the desirable range (i.e. greater than 60 mg/dL)?

pnorm(60,

mean = mean(tg$hdl, na.rm = TRUE),

sd = sd(tg$hdl, na.rm = TRUE))

P\_L60<-pnorm(60,

mean = mean(tg$hdl, na.rm = TRUE),

sd = sd(tg$hdl, na.rm = TRUE))

P\_G60<-(1-P\_L60)\*100

Answer:I expect to have 10.83338% HDLcholesterol in the desirable range.

7. Assume the population of interest has the same mean and standard deviation for HDL as this sample. What is the probability that the sample mean, for a sample of 30 participants, has HDL cholesterol in the desirable range (i.e. greater than 60 mg/dL)?

Answer:

St.er1<-(sd(tg$hdl,na.rm = TRUE)/sqrt(30))

z1<-(60-mean(tg$hdl,na.rm = TRUE))/St.er1

z2<-1-z1

Pg60<-pnorm(z2)

The probability is 4.04023e-07, not HDL cholesterol in the desirable range.

8. What percentage of subjects in our sample have HDL in the desirable range?

S\_desriable<-sum(tg$hdl>60,na.rm = TRUE)

> S\_tol<-nrow(tg)

> P\_desirable<-(S\_desriable/S\_tol)\*100

> P\_desirable

[1] 11.21951

Answer:There are 11.21951% subjects in our sample have HDL in the desirable range.