Introduction to Data Science Name: LUCÍA CARRERA

HW5 – DESCRIBING DISTRIBUTIONS AND RELATIONSHIPS

1. **Finding standard deviation:** For the following questions, please either type your work, using Draw for part b, OR. write your work and answers by hand and scan with a phone app (e.g., adobe scan or genius scan). In either case, upload to Blackboard.

   a. For the small data set below, find the standard deviation by hand, as we did in class, using the table below, and showing all your work.

i	X	Deviation	Dev^2
1	-6	-5	25
2	-3	-2	4
3	-2	-1	1
4	0	+1	1
5	1	+2	4
6	4	+5	25

1) Calculate the mean

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{(-6)+(-3)+(-2)+0+1+4}{6} = \frac{-6}{6} = -1$$

2) Calculate the deviations $dev_i = x_i - \bar{x}$

$dev_1 = -6-(-1) = -5$, $dev_2 = -3-(-1) = -2$, $dev_3 = -2-(-1) = -1$

$dev_4 = 0-(-1) = 1$, $dev_5 = 1-(-1) = 2$, $dev_6 = 4-(-1) = 5$

3) Square the deviations

$dev_1^2 = (-5)^2 = 25$, $dev_2^2 = (-2)^2 = 4$, $dev_3^2 = (-1)^2 = 1$

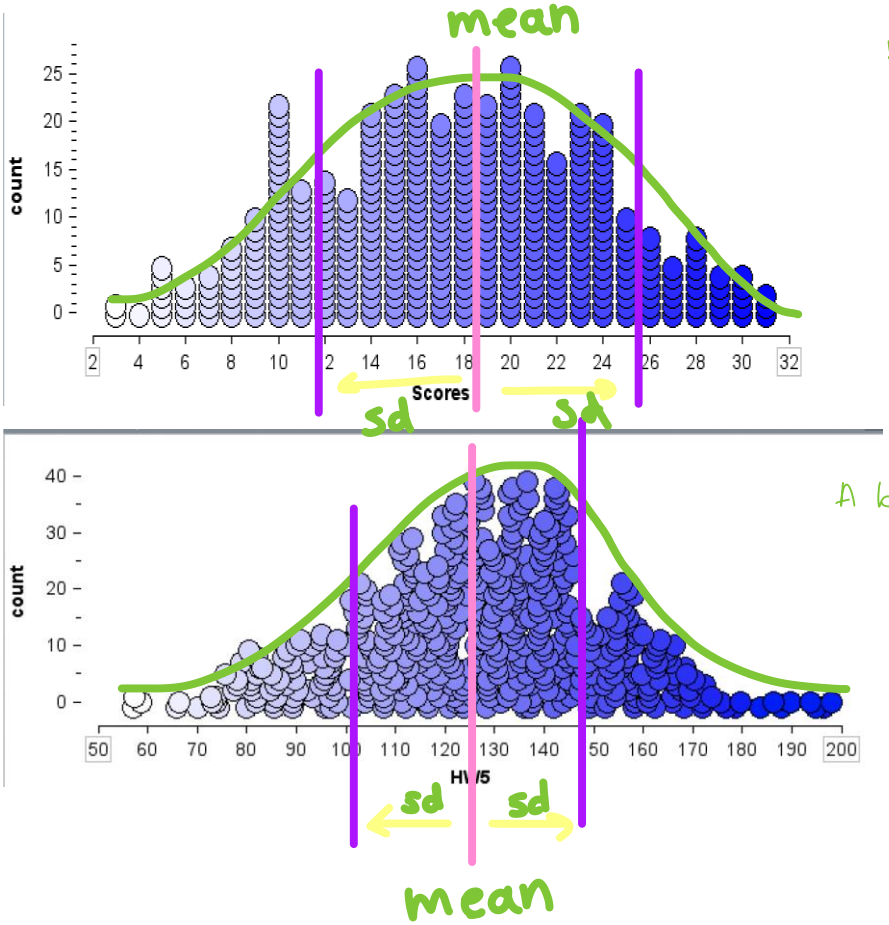
$dev_4^2 = (1)^2 = 1$, $dev_5^2 = (2)^2 = 4$, $dev_6^2 = (5)^2 = 25$

4) Find the standard deviation

$$sd = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{(n-1)}} = \sqrt{\frac{\sum_{i=1}^{n} dev_i^2}{(n-1)}} = \sqrt{\frac{25+4+1+1+4+25}{(6-1)}} = \sqrt{\frac{60}{5}} = \sqrt{12} = 3.464$$

sd = 3.464

   b. For each of the histograms below, estimate the mean and standard deviation, as we did in class, showing your work on the plot:



tiny bit left skewed?

A bit left skewed (more than previous)

Introduction to Data Science Name: LUCÍA CARRERA

2. **Measurement data.** Put the R code for all of the following tasks into an R script and upload it, in addition to the word document with answers to 1.

a. Download MeasureF21.csv, from the 'DATA' link on Blackboard, to your class folder, and read it into R as the data frame, **m**, using your working directory. Do include the 'set-up' code -- read.csv and loading packages -- in your homework program. Be sure to include the 'data cleaning' that we did in class.

b. Using ggplot, create a scatterplot of Armspan by Height, adding a regression line to the plot. Do remove the confidence band on the regression line. Repeat with RFoot by Height, then with Rhand by Height.

c. Create the linear model object for Armspan by Height, as we did in class, and do a summary of the object. Repeat with RFoot by Height, then Rhand by Height.

d. Note the slope of each regression line, and summarize each slope, as we did in class: print("For each….")

e. Using dplyr, find the correlation coefficients for each relationship, using as much data as you can (i.e., don't remove values that are missing for variables other than the two under consideration).

f. Summarize your findings, comparing the results for the three relationships. What was the general phenomenon responsible for the relationships? Which relationship is strongest? weakest? Why would that be, in terms of what you know about human bodies?