Math 11: Discussion 2

Recap of discussion 2:

1. Data visualization tools, their pros and cons: histograms, boxplots, scatter-plots, ...;
2. Quantifying the correlation between data points, within variables and among multiple variables;
3. Interpretation of linear correlation;

Contents

2.1 Upcoming assignments

Assignments	Chapters	Deadlines
Homework	Ch. 3	Wed. 04/08
Quiz	Ch. 3	Thu. 04/09
Homework	Ch. 4, 6	Fri. 04/10
Quiz	Ch. 4, 6	Sat. 04/11
Homework	Ch. 7	Mon. 04/13
Quiz	Ch. 7	Tue. 04/14
Lab 2		Fri. 04/10

Note: Assignments are spread out across the week.

Chapters:

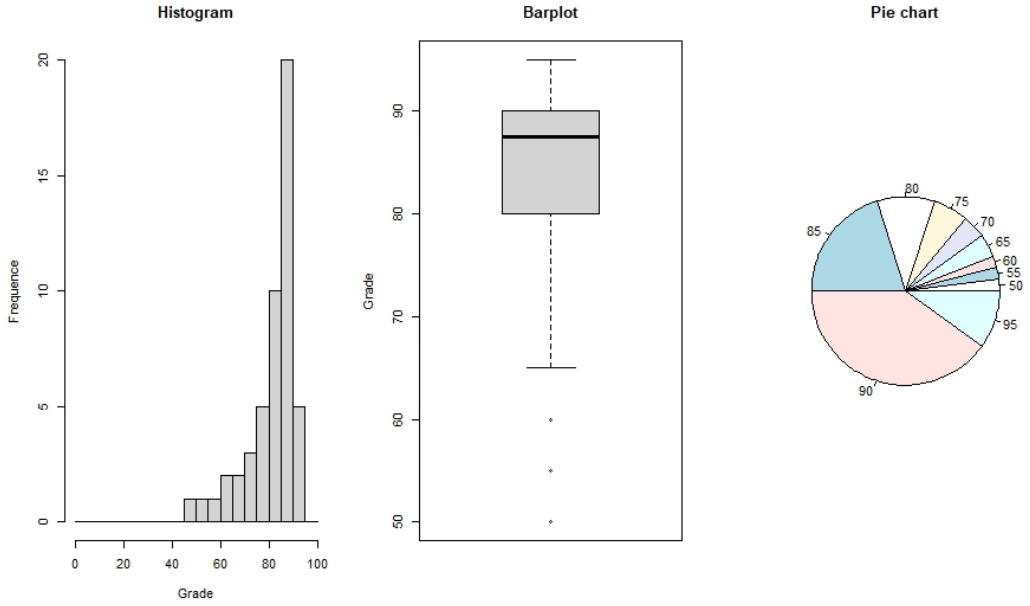
3. Displaying and Describing Categorical Data
4. Displaying and Summarizing Quantitative Data
6. Standard Deviation as a Ruler and the Normal Model
7. Scatterplots, Association, and Correlation

Key concepts (not exhaustive):

1. *contigency table*
2. *histogram, scatter-plot, boxplot*
3. *correlation, z-score*

2.2 Describing Data

Suppose we have a number of data points, it is natural to describe the data in some graphical ways. Some are histograms and barplots. For example:



2.2.1 Histogram

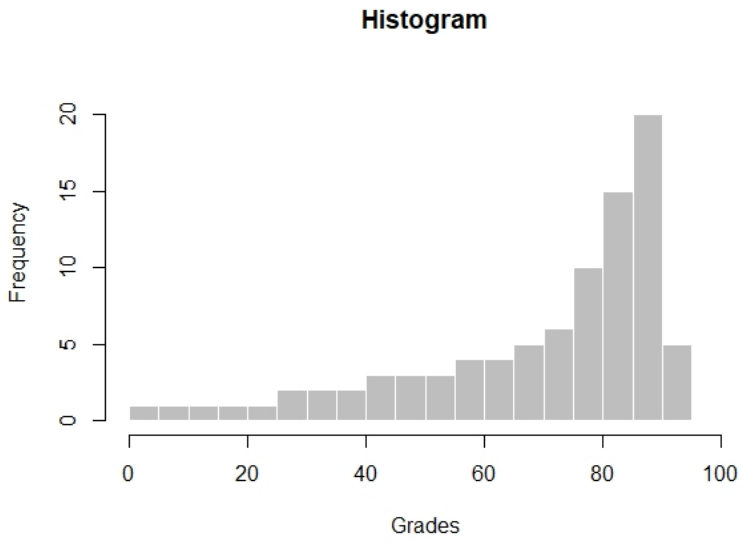
The most popular are histograms, which directly show us the frequency of the data. In describing a histogram, we have a number of notions:

Histogram shape	Preferred statistical measures
symmetric	mean & standard deviation/variance
skewed	median & IQR

Example 1. *Let's describe this histogram:*
Questions:

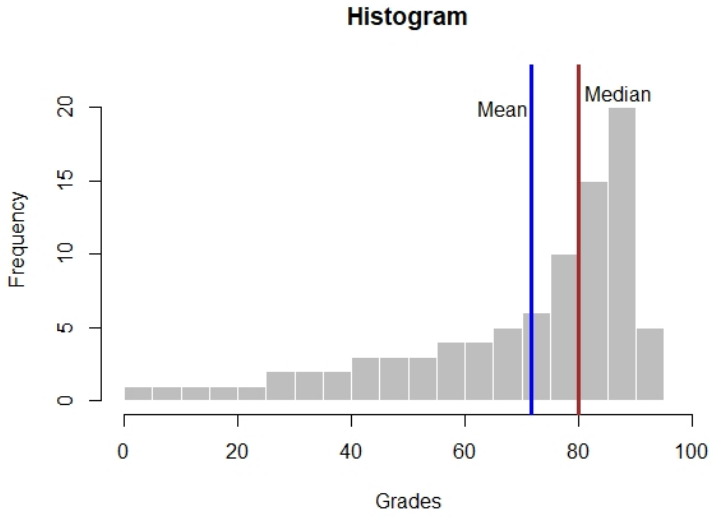
1. *Is it unimodal, bimodal or what else?*
2. *Is it skewed?*
3. *What can we tell about the mean and median?*
4. *Between* (mean, *standard deviation*) *and* (median, IQR)*, which one is more meaningful?*

Math 11: Discussion 2



Solution:

1. *unimodal* since there is only 1 peak
2. *left-skewed* since the peak is on the far right
3. mean $<$ median, think about the effects from the outliers.
4. (median, IQR) is more meaningful, because of the presence of outliers.



In particular, we have some observations regarding a skewed histogram:

left skewed	mean < median
right skewed	mean > median

Additionally, report any outliers and modes, if appropriate.

2.2.2 Correlation, z-score

1. Correlation tells us if there is a linear relationship between variables (think of the graph of a straight line).
2. z-score tells us how far away a particular data point is from the mean.

Example 2. *Consider the data points which describe the number of hours spent on study and the grades obtained:*

Hours	8	8	8	9	10	10	11	12	12	14
Grades	60	65	65	80	80	90	90	95	95	100

Questions:

1. *Calculate the correlation.*
2. *Suppose a student studies* 13 *hours a week, calculate the z-score.*

Solution:

1. We first recall the definition of *correlation*:

$$corr[x, y] = r = \frac{1}{n-1}\sum_{i=1}^{n}\Big(\frac{x_i - \bar{x}}{s_x}\Big)\Big(\frac{y_i - \bar{y}}{s_y}\Big)$$

   Note we first need $\bar{H}, s_H, \bar{G}, s_G$, where H is hours, and G is grades:

$$\bar{H} = 10.2; \quad s_H = 2.04; \qquad \bar{G} = 82; \quad s_G = 14.38;$$

   which gives the correlation r

$$r \;=\; \frac{1}{9}\sum_{i=1}^{10}\Big(\frac{H_i - 10.2}{2.04}\Big)\Big(\frac{G_i - 82}{14.38}\Big) \;=\; .93$$

2. We first recall the definition of z-score of a data point x_i:

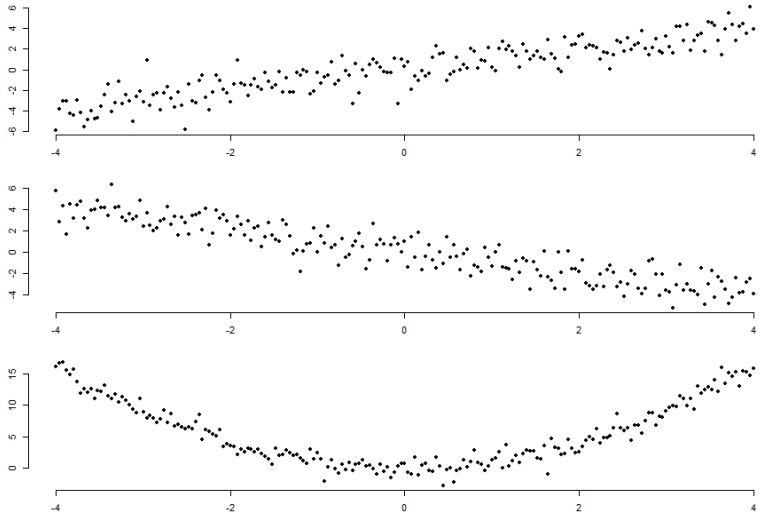
$$z(x_i) = \frac{x_i - \bar{x}}{x_y}$$

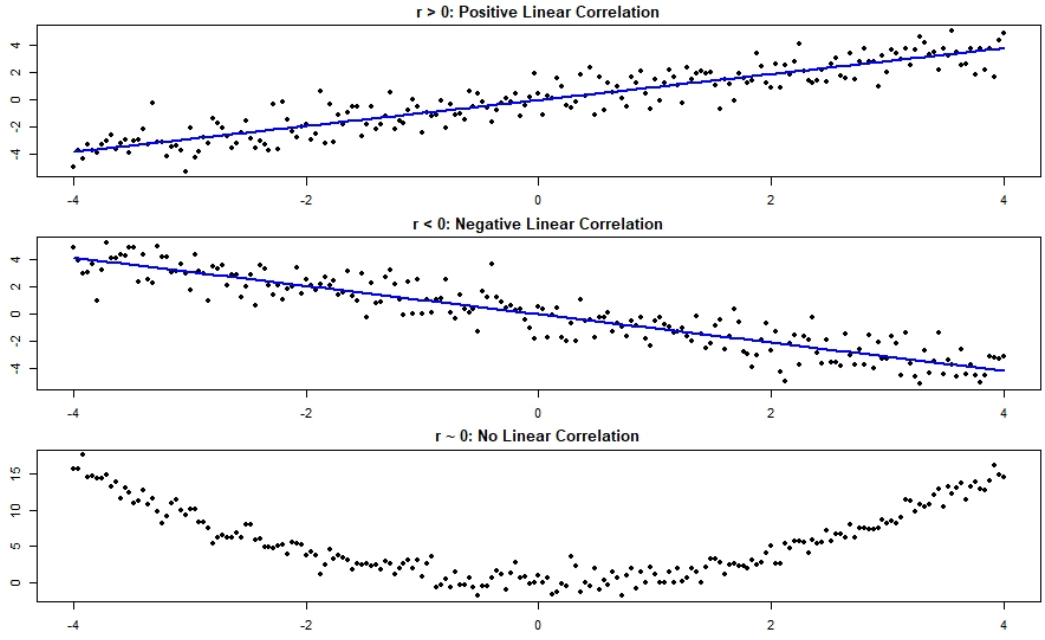
   Let $H_i = 13$, this gives:

$$z(13) \;=\; \frac{13 - 10.2}{2.04} \;=\; 1.38$$

2.2.3 Linear correlation

Example 3. *Let's look at these examples and determine if there is any correlation, if yes, what kind.*





Remark.

- *Correlation does not mean causation*: just because there is a linear (or any, in principle) correlation does not necessarily mean 1 variable causes the other.
- Just because there is no *linear* correlation does not necessarily mean the two variables are uncorrelated, they can be correlated in some other (more complicated) ways (graph C above).