SECTION 2

Chapter 6, 7, 8 and 9

Introduction

Yu (Zoey) Zhu

Email: yuzhu201@ucsc.edu

Sections: Friday, 5 - 6 PM

No Office Hour and Attend Lectures

Section Materials: https://github.com/YuZoeyZhu/STAT05-TANotes

CHAPTER 6

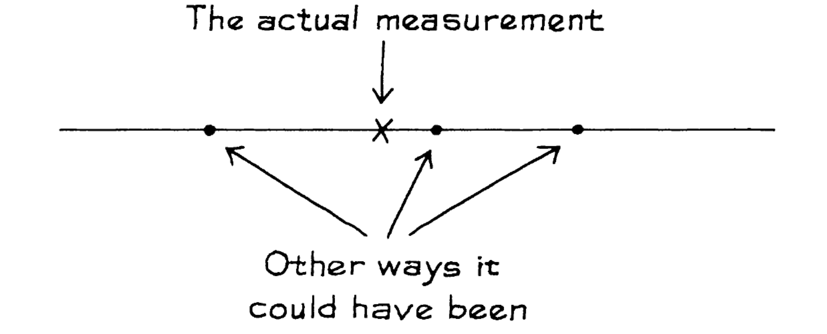
Measurement Error

- Chance Error
- Bias / System Error
- Outliers

Chance Error

If the same thing is measured several times, the same result would be obtained each time. But in practice, there are differences because of chance error.

Chance error: error changes from measurement to measurement



> The SD of a series of repeated measurements estimates the likely size of the chance error in a single measurement.

Whenever a measurement is taken, and no matter how carefully it is made any measurement is subject to chance error.

individual measurement = exact value + chance error

Bias / System Error

Example: a butcher weighs a steak with his thumb on the scale

- Bias affects all measurements the same way, pushing them in the same direction, either too high or too low.
- Chance errors change from measurement to measurement, sometimes up and sometimes down.

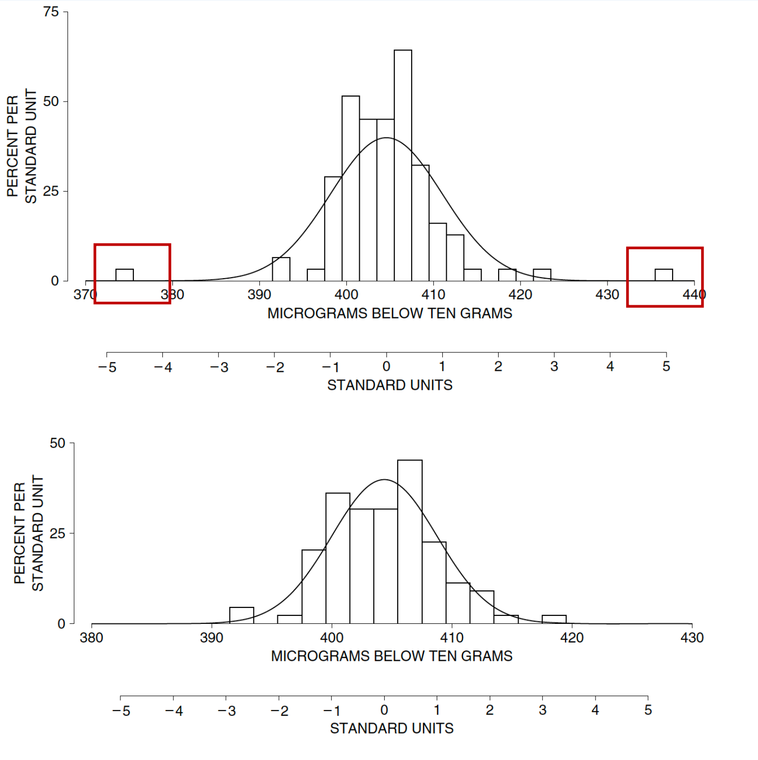
So the modification of the previous equation:

individual measurement = exact value + bias + chance error

Outliers

Measurements which are far out of the range of the other measurement.

How far out does it have to be to be an outlier? A rule of thumb is 3 SD's. But it still depends.



What can we do about outliers?

- Ignore them
- Drop those as bad data points and make the left data be closer to the normal curve
- Analyze separately and report them as outliers

CHAPTER 7

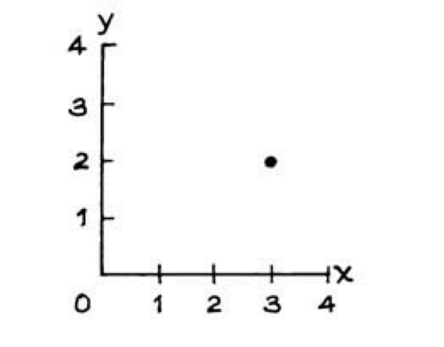
Plotting Points and Lines

- Read and Plot Points
- Slope and Intercept
- Algebraic Equation for a Line

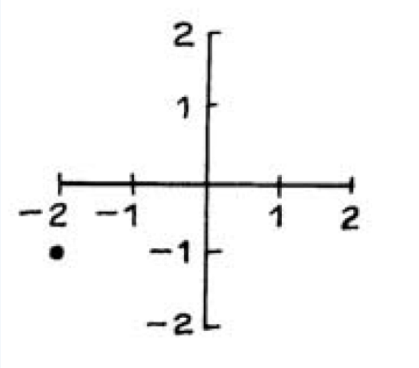
Read and Plot Points

<u>Read points:</u> in x- and y-coordinates, (x, y)

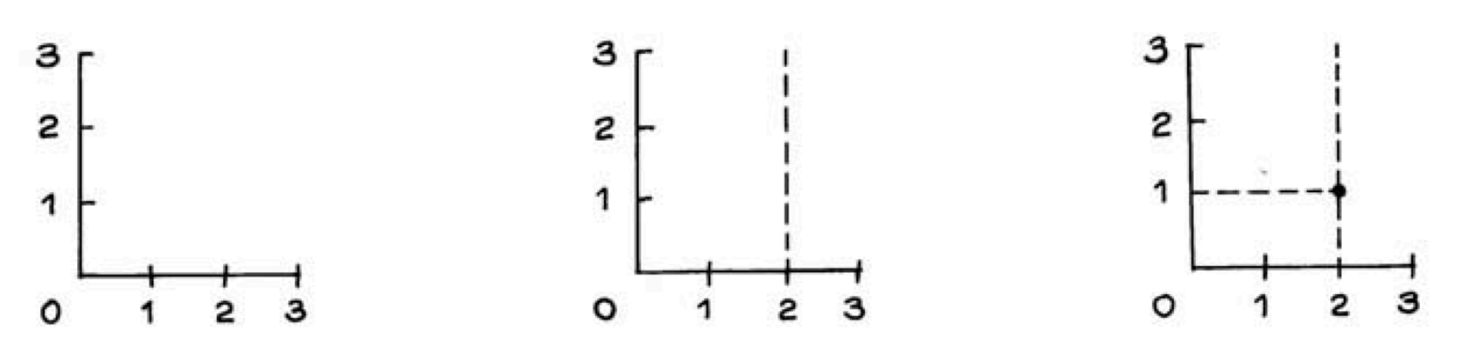
x = 3, y = 2, (3, 2)



x = -2, y = -1, (-2, -1)



<u>Plot points:</u> in x- and y-coordinates, for data point (x, y), locate x on the x-axis and y on the y-axis



(2, 1)

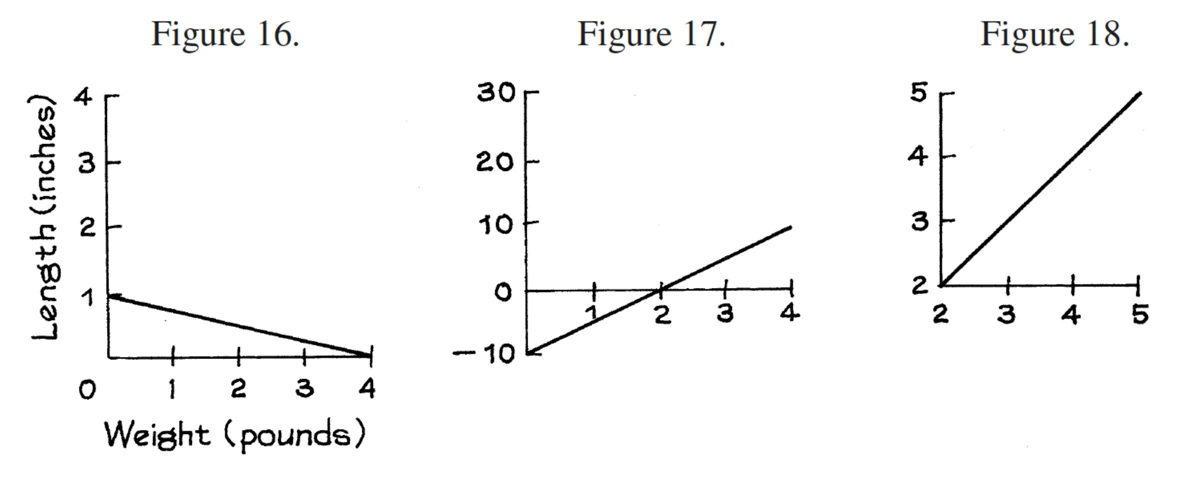
Slope and Intercept

Slope: the rate at which y increases with x, along the line. slope = rise/run

Intercept: the height when x = 0.

Slope Calculation Method:
Find two points in the line, A and B. Find the rise and run, then do the calculation

Exercises:



NOTE:

- For some lines with same slopes, they are parellel with each other
- With the positive slope, the line shows the increasing trend
- With the negative slope, the line shows the decreasing trend
- The larger the absolute value of slope, the steeper the line

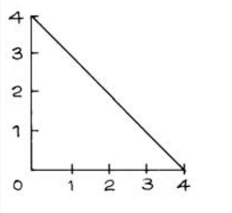
Algebraic Equation for a Line

Equation $y = mx + b$ is a straight line, with slope m and intercept b. **[Two points can define one line]**

Type I. Given an equation, plot the line.

Plot the line whose equation is $y = -\frac{1}{2}x + 4$.

Type II. Given a line, find out the equation.



Type III. Given a line/equation, find out if some points are on this line.

Eg: Given a line $y = x - 4$, if $(0.5, -5)$ is in this line

Type IV. Given three points, find out if they are on one line.

CHAPTER 8 & 9

Correlation

- Scatter Plot
- Correlation Coefficient
- SD Line
- Correlation Coefficient Computation

Scatter Plot

• A scatterplot or scatter diagram is a two-dimensional plot of data. The horizontal dimension is called x, and the vertical dimension is called y.

• Each point on a scatterplot or scatter diagram shows two values, an x value and a y value. Each point represents a single case. A single case could be a single person or object, but a single case could be a matched pair (e.g. father-son, twins, husband-wife)

• Scatter diagrams only show association, but association does not mean causation

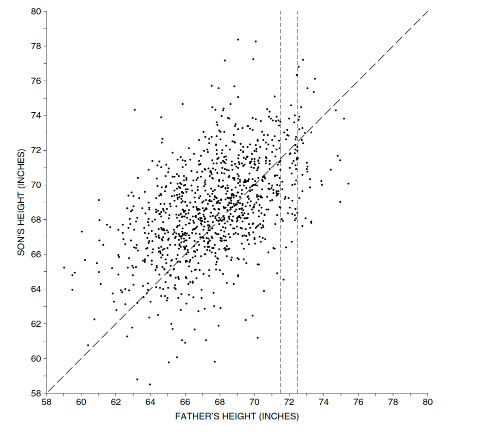
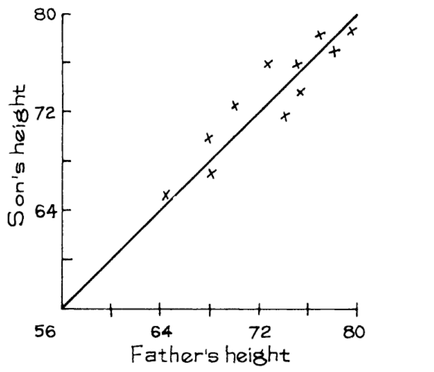


Figure 3. Son's height close to father's height.

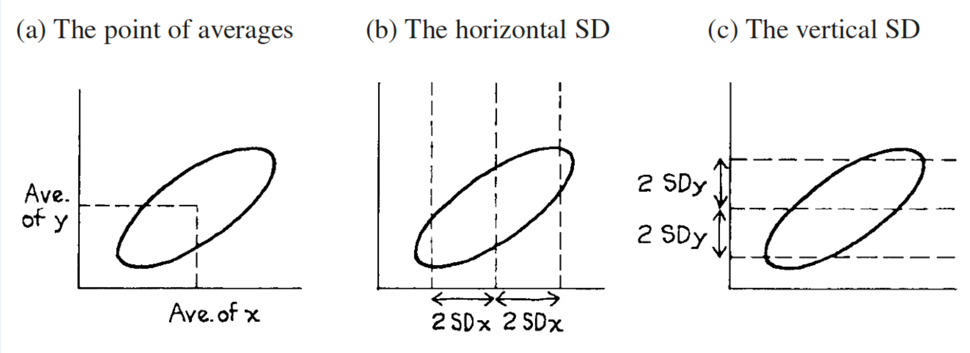


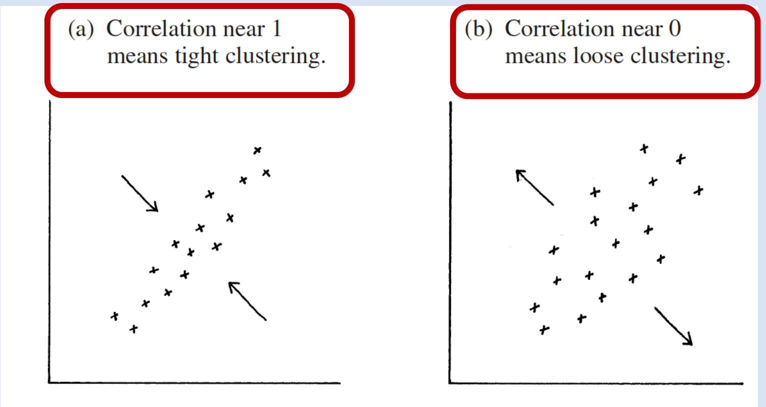
A ***positive association***
between the heights of fathers and sons

The swarm of points slopes upward to the right, the y-coordinates of the points tending to increase with their x-coordinates.

Correlation Coefficient

The Correlation Coefficient, denoted r, measures how close the data are to a straight line, or in other words, it measures the strength of association.





The relationship between two variables can be summarized by:

- The <u>average</u> of the x-values, the <u>SD</u> of the x-values
- The <u>average</u> of the y-values, the <u>SD</u> of the y-values

- The <u>correlation coefficient r</u>

SD Line

The points in a scatter diagram generally seem to cluster around the SD line.

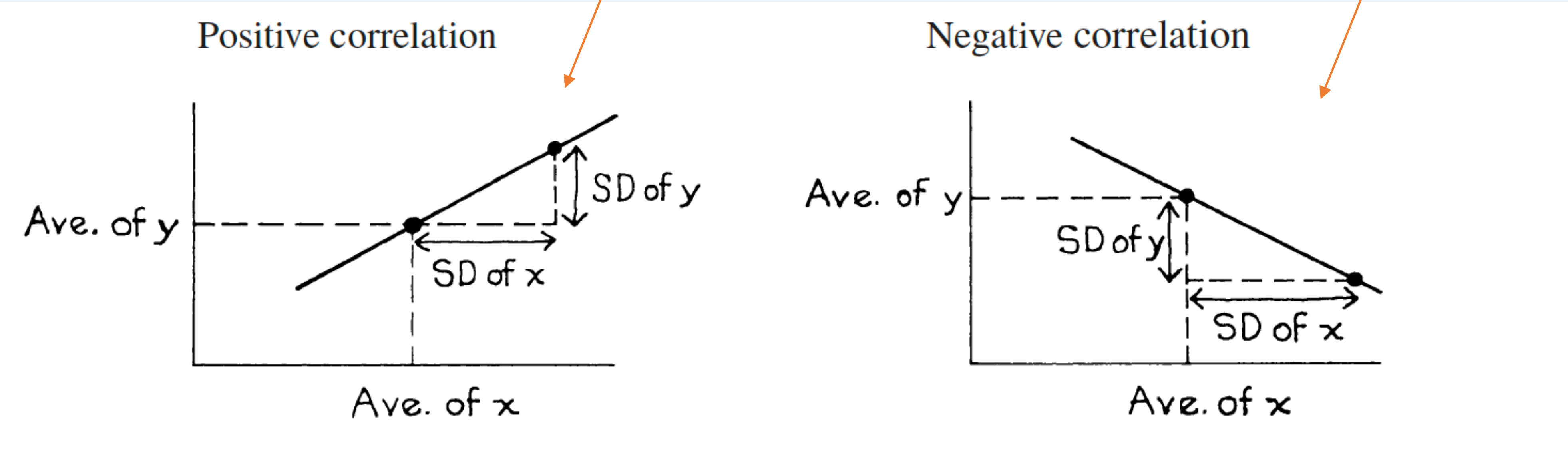
- Goes through the point of averages

Avg of y = Intercept + Slope × Avg of x

- Goes through all the points which are an equal number of SDs away from the average, for both variables

Slope: **(SD of y)/(SD of x)**

Slope: **-(SD of y)/(SD of x)**



Correlation Coefficient Computation

Method I:

$$r = \text{average of (x in standard units)} \times \text{(y in standard units)} \quad \text{where standard units} = \frac{\text{value} - \text{average}}{SD}$$

Method II:

$$r = \frac{cov(x, y)}{(SD\ of\ x) \times (SD\ of\ y)} \quad \text{where cov(x, y) = average of products xy} - \text{(average of x)} \times \text{(average of y)}$$

Table 1. Data.

x	y
1	5
3	9
4	7
5	1
7	13

Mean_x = 4, Mean_y = ; SD_x = 2, SD_y = 4

Method I:

Step 1. Convert the x-values to standard units

Step 2. Convert the y-values to standard units

Step 3. Work out the product for each (x, y) pair
(x in standard units) ×(y in standard units)

Step 4. Take the average of the products

Method II:

Step 1. Calculate the average of products xy, avg of x, avg of y

Step 2. Calculate covariance cov(x, y)

Step 3. Calculate SD of x, SD of y

Step 4. Divide covariance by the product of SD of x and SD of y

- **-1 <= r <= 1**
- The correlation r measures how close the data are to a line
- If r is close to 1 or -1, the data are close to a line
- If r is close to 0, the data are not close to a line
- r does **NOT** tell what percentage of the data fall on the line
- r = 0.80 does not indicate twice as much linearity as r = 0.40
- The correlation between x and y is the same as the correlation between y and x. r(x, y)=r(y, x)
- Invariant under addition: If some constant "a" is added to every one of the X or the Y values, the correlation is unchanged
- Invariant under multiplication: if all the x or the y values are multiplied by some positive constant "b", the correlation is unchanged. The correlation can change very dramatically if only ONE of the data points is changed