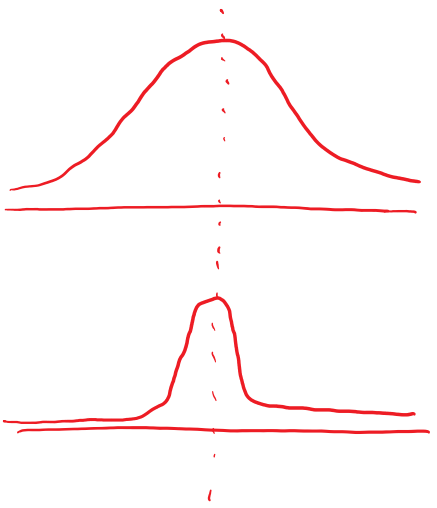
Section 2.2: Measures of Spread/ Variability

Thursday, October 24, 2019 11:12 AM



two distributions with the same mean

measures of spread - indication of how "wide" or how "spread out" a data set is

When do you want a small spread?

- when trying for uniformity

example: manufacturing identical objects

When do you want a large spread?

- when you are trying to make distinctions

high quality vs low quality

rankings

2019/10/28

range - difference between the maximum and minimum values

example: 3, 7, 15, 42, 54

range is 51
↑
single number

good part: easy to calculate

bad part: almost completely useless

→ heavily influenced by outliers

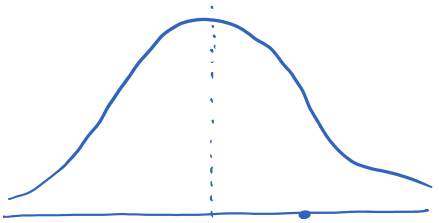
→ only depends on the values of two data points out of the entire set

---

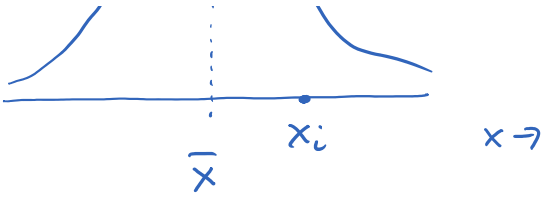
the annoying measures to calculate

- variance
- standard deviation ← commonly used

a little bit of background on how you calculate these:



sample data



consider some point x_i in this data set

- how far from the mean is x_i?

$$(x_i - \bar{x})$$

if we add up all of these as is the positive values will cancel the negative values and we'll end up with zero

but if we square $(x_i - \bar{x})$ and take the sum

$$\Sigma (x_i - \bar{x})^2$$

then this is a measure of how far away from the mean the data points are

population variance:

$$\sigma^2 = \frac{\Sigma (x_i - \mu)^2}{N}$$

Greek letter sigma (lower case)

where μ = population mean
N = population size

population standard deviation:

$$\sigma = \sqrt{\sigma^2}$$

sample variance:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$\bar{x}$ = sample mean
n = sample size

Sample standard deviation

$$s = \sqrt{s^2}$$

---

things I would like you to know:

- the standard deviation (std dev) is a measure of how "wide" or "scattered" or "spread out" a distribution / data set is
- it measures how far, on average, each data point is from the mean

---

examples: for the following pairs of data sets, which one has the higher standard dev? or are they the same?

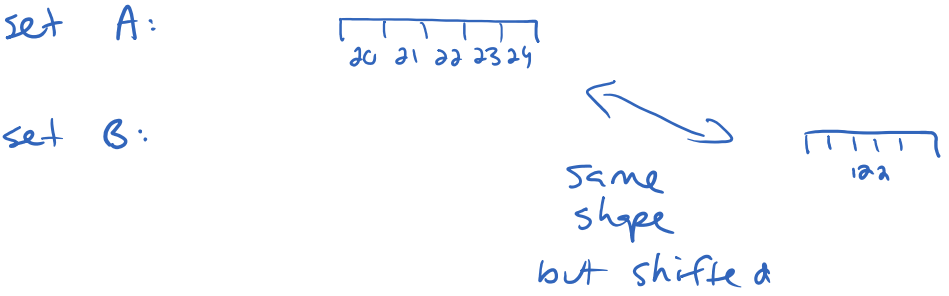
Set A: 5, 10, 15, 20, 25

Set B: 5, 5, 15, 25, 25

answer: B because the data points shaded in blue are farther from the middle

set A: 20, 21, 22, 23, 24

set B: 120, 121, 122, 123, 124



A and B have same standard deviation

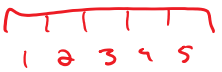
2019/10/29

set A: 1, 2, 3, 4, 5

set B: 2, 4, 6, 8, 10

answer:

set A:



set B:

(B) has higher std dev

(in fact, it's exactly twice as high)

example: The Gizmo store has to raise its prices because of a platinum shortage. Every device in the store has a different price.

a) If every device has its price increased by $5, what happens to the mean, median, range, and standard deviation of the prices of the devices? Be as specific as you can!

mean : increases by $5

median : "

range : same

std dev : "

b) Only one device needs to have its price changed. The most expensive device has its price increased by $25. What happens to the mean, median, and range?

mean - increases (by how much will depend on how many devices there are - if you want to be specific, it increases by $\frac{25}{n}$, where n is the number of devices)

median - stays the same*

(* unless there are two or fewer devices)

range - increases by $25

c) All devices increase in price by 10%. What happens to the mean, median, range, and std dev?

all increase by 10%