Lec3

1.3.3 Population Mean

The average of all values in the population is called pop. mean and is denoted by μ. When there are N values in the population, then $\mu = \frac{\sum \text{values in whole pop}}{N}$

One of our first tasks in statistical inference will be to present methods based on the sample mean for drawing conclusions about a pop. mean

For example, Ex. 1.14 $\bar{x} = 16.36$

guess for μ

1.3.4 Population Median

in whole pop.

Analogous to $\widetilde{x}$ as the middle value in the sample, the **population median** is middle value denoted by $\widetilde{\mu}$. As with $\bar{x}$ and μ, we can think of using sample median, to make an inference about pop. median.

Summary:

sample mean + median vs. pop. mean + median

1.3.5 Other Measures of Location

Maximum

Maximal value

Minimum

Minimal value

An **outlier** is an atypical data pt (too large or too small)

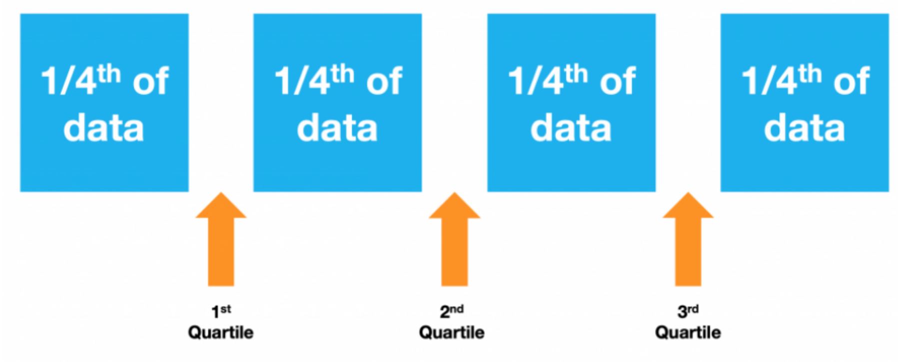
Sometimes max and/or min are outliers in the data set.

Quartiles and Percentiles

Quartiles divide data set into 4 equal parts

$\underbrace{}_{25\%}$ Q_1 $\underbrace{}_{25\%}$ Q_2 = Median $\underbrace{}_{25\%}$ Q_3 $\underbrace{}_{25\%}$

with the observations above the third quartile constituting the upper quarter of the data set, the second quartile being identical to the Median, and the first quartile separating the lower quarter from the upper three-quarters.



If the quantiles divide the data into 100 groups, then they're called percentiles

1.3.6 The effect of skewness on the mean and median

$\tilde{x} \approx \bar{x} \rightarrow$ symmetric

$\tilde{x} < \bar{x} \rightarrow$ right skewed

$\tilde{x} > \bar{x} \rightarrow$ left skewed

Think about the graph in extreme cases.

Suppose we have 11 data in a set of students grades, in which ten of them are 60s and one of them is 100.

60 60 60 60 ... 60 $\underline{\underline{100}}$

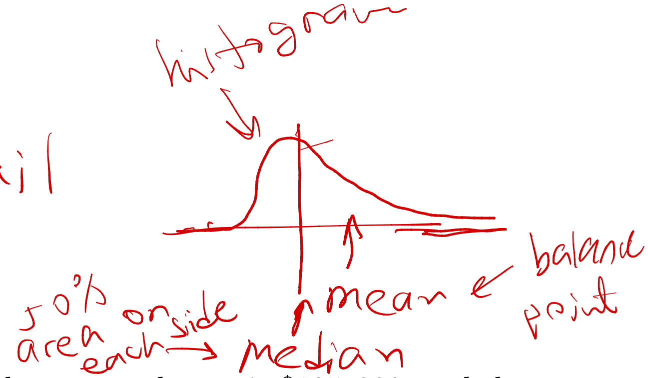
$\tilde{x} = 60 \quad \bar{x} > 60$

On the other hand, suppose we have 11 data in a set of students grades, in which ten of them are 100s and one of them is 60.

60 100 100 100 ... 100

$\tilde{x} = 100 \quad \bar{x} < 100$

Note. mean follows the tail



Example 4. Suppose we have 10 people in a room, the mean salary $\bar{x}$ is \$105,000, and the median salary $\widetilde{x}$ is \$65,000. What can we say about the distribution?

Ans : right skewed

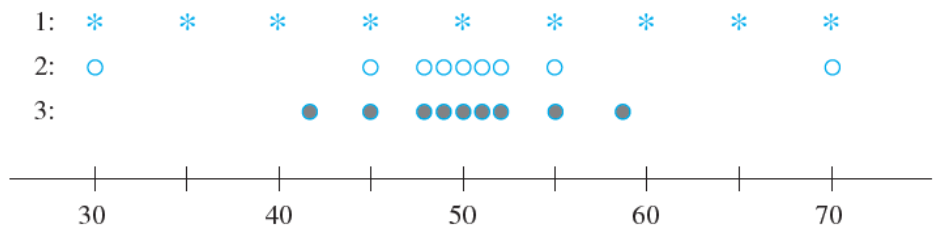
Theres likely an outlier

What is the more appropriate measure of center when there are extreme valus in the data set and why?

Ans : Median

1.4 Measures of Variability

Figure below shows dotplots of three samples with the same mean and median, but the extent of spread about the center is different for all three samples. The first sample has the largest amount of variability, the third has the Smallest amount, and the second is intermediate



Samples with identical measures of center but different amounts of variability

Measures of Variability for Sample data

The simplest measure of variability in a sample is the range, which is span of data set

- max − min

The value of the range for sample 1 in Figure above is much larger than sample 3, reflecting more variability in the first sample than in the third. A defect of the range, though, is sensitive to outliers.

Samples 1 and 2 in Figure above have the same range, yet when the observations between the two extremes are taken into account, there is much less variability in the second sample than in the first.

Our primary measures of variability involve values − mean

$x_1 - \bar{x}$ $x_2 - \bar{x}$ $x_3 - \bar{x}$ ...

A deviation will be **positive** (+) if $x_i > \bar{x}$ and **negative** (−) if $x_i < \bar{x}$

If all the deviations are small, then all x_i's are close to $\bar{x}$ and there is little variability. Alternatively, if some of the deviations are large, then some x_i's far from $\bar{x}$, suggesting a larger variability

A simple way to combine the deviations into a single quantity is to average out.

Unfortunately, this is a bad idea:

$$\text{sum of deviations} = (x_1-\bar{x})+(x_2-\bar{x})+\cdots+(x_n-\bar{x})$$

so that the average deviation is ______. $= \underbrace{x_1+\cdots+x_n}-n\bar{x} = n\bar{x}-n\bar{x} = 0$

To prevent negative and positive values from counteracting one another when they are combined, we consider instead squared deviations

$(x_1-\bar{x})^2 \quad (x_2-\bar{x})^2 \quad \cdots$ think Euclidean distance

To get the averaged square deviation, for several reasons we will not cover here, we divide the sum of squared deviations by $n-1$ instead of n.

Definition 1. The sample **variance**, denoted by s^2, is given by

$$s^2 = \frac{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}$$

The sample **standard deviation(SD)**, denoted by s, is the $\sqrt{\ }$ of the variance:

$$s = \sqrt{s^2} = \sqrt{\frac{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}}$$

The standard deviation is our preferred measure of variability, because it has the same unit as data.

A shortcut formula for s,

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

This is because we can write s^2 as

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

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

$-2n\bar{x}^2$

$$n\left(\frac{\sum x_i}{n}\right)^2 = \frac{(\sum x_i)^2}{n}$$

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

Example 5. Suppose we have a data set $x_1, x_2, \cdots, x_5$ where $\sum_{i=1}^5 x_i = 10.9$ and $\sum_{i=1}^5 x_i^2 = 29.97$. What is the SD?

Ans :

$$\sqrt{\frac{29.97 - \frac{(10.9)^2}{5}}{4}} = 1.246$$

$$\bar{x} = \frac{10.9}{5}$$

Check equals the s you get from the original formula!

Properties of the mean and SD:

Let $x_1, x_2, \cdots, x_n$ be a sample and c be any nonzero constant.

1. If $y_1 = x_1 + c,\ y_2 = x_2 + c, \cdots, y_n = x_n + c$,

   then mean $\bar{y} = \bar{x} + c$

   the sample variance $s_y^2 = s_x^2$

   the SD $s_y = s_x$

2. If $y_1 = cx_1,\ y_2 = cx_2, \cdots, y_n = cx_n$,

   then mean $\bar{y} = c\bar{x}$

   the sample variance $s_y^2 = c^2 s_x^2$

   the SD $s_y = |c| s_x$

$$\frac{(x_1+c)+\cdots+(x_n+c)}{n} = \frac{x_1+\cdots+x_n+nc}{n} = \bar{x}+c$$

$$\left((x_i+\cancel{c})-(\bar{x}+\cancel{c})\right)^2$$

$$\frac{cx_1+\cdots+cx_n}{n} = c\bar{x}$$

$$\sqrt{c^2} = |c|$$

$$\left(\sqrt{(-5)^2} = 5 \text{ not } -5\right)$$

$$(cx_i - c\bar{x})^2 = c^2(x_i - \bar{x})^2 =$$