STA13: Elementary Statistics

Lecture 3

Book Sections 2.3-2.4

Dmitriy Izyumin

January 12 2018

Index Notation

Sometimes it's easier to talk about a general set of observations rather than specific numbers.

We'll denote the n observations as $x_1, x_2, \cdots, x_n$.

This is index notation, and the subscript i is called the index.

Here x_1 deonotes the first observation, x_2 denotes the second observation, and so on.

Another way to express the set of n observations is "x_i for $i = 1, 2, \cdots, n$"

Sigma Notation

Once we have observations $x_1, x_2, \cdots, x_n$, we may want to add them all up.

There's shorthand notation for that: $\sum_{i=1}^{n} x_i$

- The "$i = 1$" under Σ denotes that we're summing over the index i, starting at the first observation.
- The "n" on top of Σ indicates that we stop summing at the n^{th} observation.
- In general we don't have to start at 1 or end at n.
- To the right of Σ are indexed objects that will be summed. These can get more complicated than just x_i. For example, we may want to add x_i^2 or $\sqrt{(x_i - 5)}$.

What does $\sum_{i=3}^{7}(x_i - 10)^2$ mean?

Sigma Notation

- We will use this notation throughout the course
- Section 2.3 of the book
- If you are unfamiliar with the notation, then come to office hours, and do practice problems from section 2.3.

Center

When dealing with quantitative variables, graphical summaries are good for a quick assessment, numerical summaries are needed for further analysis.

A measure of center

- locates the "center" of the distribution on the x-axis
- tells us "around which value" the observations fall
- is measured in the same units as the variable

Mean

One measure of center is the mean. It is also called the arithmetic mean or average.

The mean of n observations is their sum divided by n.

Data: 1, 3, 4, 4, 7

Mean: $\frac{1+3+4+4+7}{5} = 3.8$

Median

The median of n observations is the value that falls in the middle when the observations are arranged from smallest to largest.

Once the data are ordered, the position of the median is given by $\frac{n+1}{2}$.

Median

Whether n is odd or even affects the way we calculate the median.

n is odd

$5, 6, 8, 10, 11, 11, 11, 13, 15$

median: 11

position: $\frac{n+1}{2} = \frac{9+1}{2} = 5$

n is even

$5, 6, 8, 8, 10, 11, 11, 11, 13, 15$

median: $\frac{10+11}{2} = 10.5$

position: $\frac{n+1}{2} = \frac{10+1}{2} = 5.5$

Mode

The mode of n observations is the value that occurs most frequently.

The mode may not be unique:
1, 3, 3, 3, 6, 6, 8, 8, 8, 11, 12

When dealing with a histogram, the class with the highest frequency is called the modal class, and the midpoint of that class is called the mode.

Effect of Outliers

We say a measure is robust if it is not strongly affected by outliers.

The median is robust.

The mean is not robust.

The median is often used as a measure of center if outliers are present.

Effect of Outliers

Ages of 10 children on a playground:

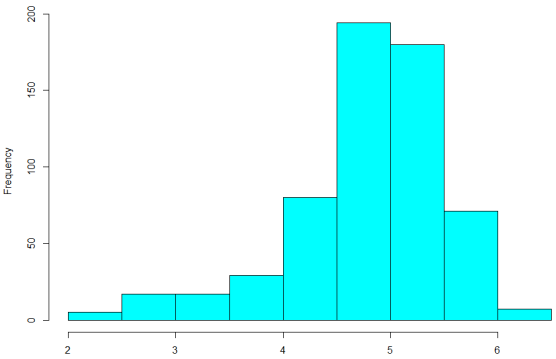
Data: 2, 2 ,3 ,4, 4, 5, 5, 5, 6, 7
Mean: 4.3
Median: 4.5

Ages of 9 children and one grandma:

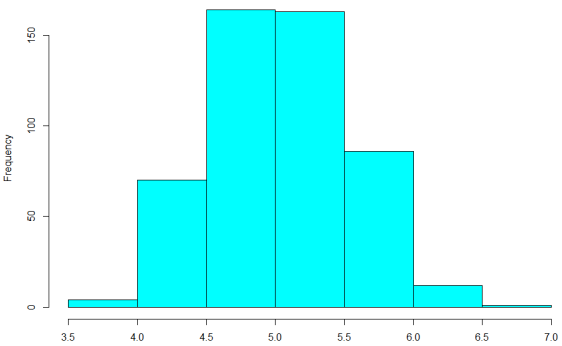
Data: 2, 2, 3, 4, 4, 5, 5, 5, 6, 79
Mean: 11.5
Median: 4.5

When one value was replaced by an outlier, the mean changed a lot, but the median remained the same. Why?

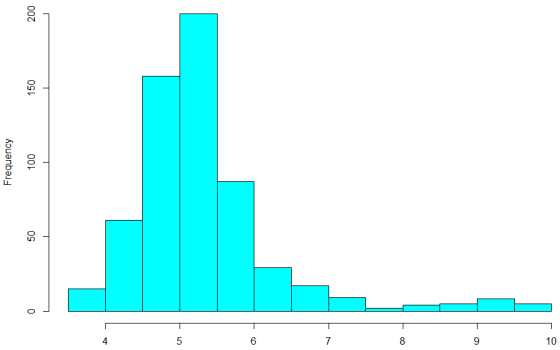
Effect of Skewness



Skewed left:
mean $<$ median



Symmetric:
mean $\approx$ median



Skewed right:
mean $>$ median

Robustness

We say a measure is robust if it is not strongly affected by outliers.

The median is robust.

The mean is not robust.

The median is often used as a measure of center if outliers are present.