# STA13: Elementary Statistics

## Lecture 4
## Book Section 2.5

Dmitriy Izyumin

January 17 2018

# Review

Last time we covered three measures of center:

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

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

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

# Spread

A measure of spread is a quantitative measure that describes the variability or dispersion of the data along the horizontal axis.

Measures of center tell us a value around which the observations fall.

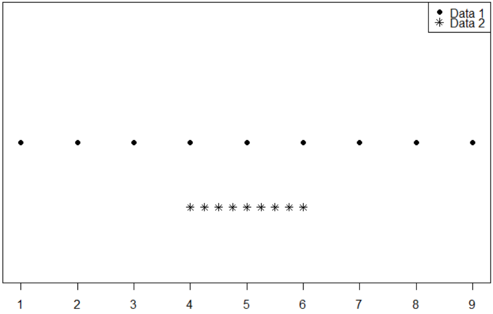
Measures of spread tell us how widely the observations are dispersed around that center.

# Spread

Data 1: $1, 2, 3, 4, 5, 6, 7, 8, 9$

Data 2: $4, 4.25, 4.5, 4.75, 5, 5.25, 5.5, 5.75, 6$

Both are centered around 5, but the first data set has greater spread.



# Range

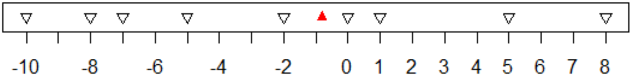
The range of a data set is found by subtracting the minimum value from the maximum: range=max-min

- Pro: very easy to compute
- Con: only depends on two observations in the set
- Con: strongly affected by outliers

# Summarizing Spread

11 observations: -10,-8,-7,-5,-2,0,1,1,5,8,8

Sample mean: $\bar{x} = -0.8182$




How can we quantify the spread?

# Deviations



11 observations: -10,-8,-7,-5,-2,0,1,1,5,8,8
Sample mean: $\bar{x} = -0.8182$

The deviations are the individual differences between each observation and the sample mean.

| | |
|---|---|
| -10-(-0.8182)=-9.182 | 1-(-0.8182)=1.818 |
| -8-(-0.8182)=-7.182 | 1-(-0.8182)=1.818 |
| -7-(-0.8182)=-6.182 | 5-(-0.8182)=5.818 |
| -5-(-0.8182)=-4.182 | 8-(-0.8182)=8.818 |
| -2-(-0.8182)=-1.182 | 8-(-0.8182)=8.818 |
| 0-(-0.8182)=0.818 | |

There are as many deviations as there are observations. How can we summarize them?

# Summarizing Deviations



| | |
|---|---|
| Observations: | $x_1, x_2, \cdots, x_n$ |
| Sample mean: | $\bar{x}$ |
| Deviations: | $(x_i - \bar{x})$ for $i = 1, 2, \cdots, n$ |

Let's see what happens when we add up the deviations:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i) - \sum_{i=1}^{n}(\bar{x}) = n\bar{x} - n\bar{x} = 0$$

The deviations always add up to 0. Why does that happen?

This means we can't take a simple average.
What else can we do?

# Summarizing Deviations

Some observations are greater than the mean, some are less.

Some deviations are positive, some are negative.

If we add them, they cancel out and the sum is always 0.

We want to add the deviations without them cancelling out.

We should focus on the *size* of the deviations, not their *sign*.

One way is to use absolute value: $|x_i - \bar{x}|$

It turns out that squaring works best: $(x_i - \bar{x})^2$

# Variance

| | |
|---|---|
| Observations: | $x_1, x_2, \cdots, x_n$ |
| Sample mean: | $\bar{x}$ |
| Deviations: | $(x_i - \bar{x})$ for $i = 1, 2, \cdots, n$ |

## Sample Variance

- Denoted by $s^2$
- An "average" of the squared deviations:

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

- Measured in squared units
- Non-negative: $s^2 \geq 0$ (when is $s^2$ exactly 0?)

# Standard Deviation

The variance is measured in squared units.

If the variable is monthly income (in dollars), the variance would be measured in dollars$^2$. What does this mean?

It's useful to have a measure of spread measured in the same units as the variable.

## Sample Standard Deviation

- Denoted by $s$
- Square root of the sample variance: $s = \sqrt{s^2}$
- Measured in original units
- Non-negative: $s \geq 0$ (when is $s$ exactly 0?)
- On average, an observation is $s$ units away from $\bar{x}$

# Formulas

Computing $s^2$ from its definition:

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

A less time-consuming approach:

$$s^2 = \frac{1}{n-1}\left[\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2\right]$$

Regardless of the way you compute the variance $s^2$, the standard deviation $s$ is just the square root:

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

# Computing Variance and SD

Long way:

1. Compute the sample mean $\bar{x}$
2. Compute the deviations $(x_i - \bar{x})$
3. Compute the squared deviations $(x_i - \bar{x})^2$
4. Use formula $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
5. Take the positive square root for SD

Quicker way:

1. Compute the sample mean $\bar{x}$
2. Compute the sum of squared values $\sum_{i=1}^{n} x_i^2$
3. Use formula $s^2 = \frac{1}{n-1} \left[ \left( \sum_{i=1}^{n} x_i^2 \right) - n\bar{x}^2 \right]$
4. Take the positive square root for SD

# Why Divide by $n-1$?

We compute the sample variance (or SD) to get a sense of the population variance (or SD).

Dividing by $n-1$ makes $s^2$ a more accurate estimate of $\sigma$.

This is especially true when $n$ is small.

We will not go into the mathematical details of this.

# Linear Transformations

Obs. $x_1, x_2, \cdots, x_n$
Sample mean $\bar{x}$
Sample variance $s^2$

Multiply each observation by $c$. $(x_i \to cx_i)$

| | |
|---|---|
| The new sample mean | $c\bar{x}$ |
| The new variance | $c^2 s^2$ |
| The new standard deviation | $\lvert c \rvert s$ |

Add $d$ to each observation. $(x_i \to x_i + d)$

| | |
|---|---|
| The new sample mean | $\bar{x} + d$ |
| The new variance | $s^2$ |
| The new standard deviation | $s$ |

# Linear Transformations

Obs. $x_1, x_2, \cdots, x_n$
Sample mean $\bar{x}$
Sample variance $s^2$

Multiply by $c$ and add $d$. $(x_i \to cx_i + d)$

| | |
|---|---|
| The new sample mean | $c\bar{x} + d$ |
| The new variance | $c^2 s^2$ |
| The new standard deviation | $\lvert c \rvert s$ |

Exercise: what would happen to the median and range?

# Example

We have 10 temperature measurements in $^\circ\mathrm{F}$.
The mean value is $67\,^\circ\mathrm{F}$, and the standard deviation is $2\,^\circ\mathrm{F}$.

What would happen to the mean and SD if the values are converted to $^\circ\mathrm{C}$?

If $x$ is measured in $^\circ\mathrm{F}$, it can be converted to $^\circ\mathrm{C}$ as follows:

$$x \to \frac{5}{9}(x - 32)$$

$$x \to \frac{5}{9}x - \frac{160}{9}$$

$$x \to 0.556x - 17.778$$

New mean: $0.556(67) - 17.78 = 19.472$

New SD: $0.556(2) = 1.112$