# Semana 11

Wednesday, October 13, 2021 8:18 PM

## Center and spread

### Goals

- Quantify natural concepts like "center" and "variability"
- Examine bell shaped distributions
- Understand why many of the empirical distributions that we have generated are bell shaped

### The average and the histogram

The average (or mean)

- Need not be a value in the collection
- Need not be an integer even if the data are integers
- Somewhere between min and max, but not necessarily halfway in between
- Same units as the data
- Smoothing operator: collect all the contributions in one big pot, then split evenly

### Relation to the histogram

- The average of a list depends only on the proportions in which the distinct values appear, not on the number of entries in the list.
- The average is the center of gravity of the histogram.
- It is the point on the horizontal axis where the histogram balances.

### The average and the median

- **Average:** Balance point of the histogram
- **Median:** Halfway point of the data; half the area of histogram is on either side of the median.
- If the distribution is symmetric about a value, then that value is both the average and the median.
- If the histogram is skewed, then the mean is pulled away from the median in the direction of the tail.

### Standard deviation

Defining variability

- **Plan A:** "biggest value - smallest value"
    - Doesn't tell us much about the shape of the distribution
- **Plan B:**
    - Measure variability around the mean
    - Need to figure out a way to quantify this

The standard deviation (SD) measures roughly how far the data are from their average

- SD = root mean square of deviation from average

(Steps: 5 4 3 2 1)

- The SD has same units as the data.

### Why use the SD?

There are two main reasons.

- **The first reason:** No matter what the shape of the distribution, the bulk of the data in the range "average ± a few SDs"
- **The second reason:** Relation with bell shaped curves

### Chebyshev's Bounds

| Range | Proportion |
|---|---|
| average ± 2 SDs | at least 1 - 1/4 (75%) |
| average ± 3 SDs | at least 1 - 1/9 (88.888...%) |
| average ± 4 SDs | at least 1 - 1/16 (93.75%) |
| average ± 5 SDs | at least 1 - 1/25 (96%) |

No matter what the distribution looks like

### How big are most of the values?

No matter what the shape of the distribution, the bulk of the data in the range "average ± a few SDs"

**Chebyshev's Inequality**

No matter what the shape of the distribution, the proportion of the data in the range "average ± z SDs" is at least 1-1/z^ 2

# Normal curve

## Goals

- Describe what is meant by "bell shaped curved"
- Explain how bell shaped curves arise in inference

## Standard units

- The standard units measures "how many SDs above average?"
- Z = (value-average)/SD
    - Negative z: value below average
    - Positive z: value above average
    - Z= 0: value equal to average
- When values are in standard units: average = 0, SD = 1
- Chebyshev: At least 96% of the values of z are between -5 and 5

## The SD and the histogram

- Usually, its not easy to estimate the SD by looking at a histogram
- But if the histogram has a bell shape, then you can.

## The SD and Bell-Shaped curves

If a histogram is bell-shaped, then

- The average is at the center
- The SD is the distance between the average and the points of inflection on either side

## How big are most values?

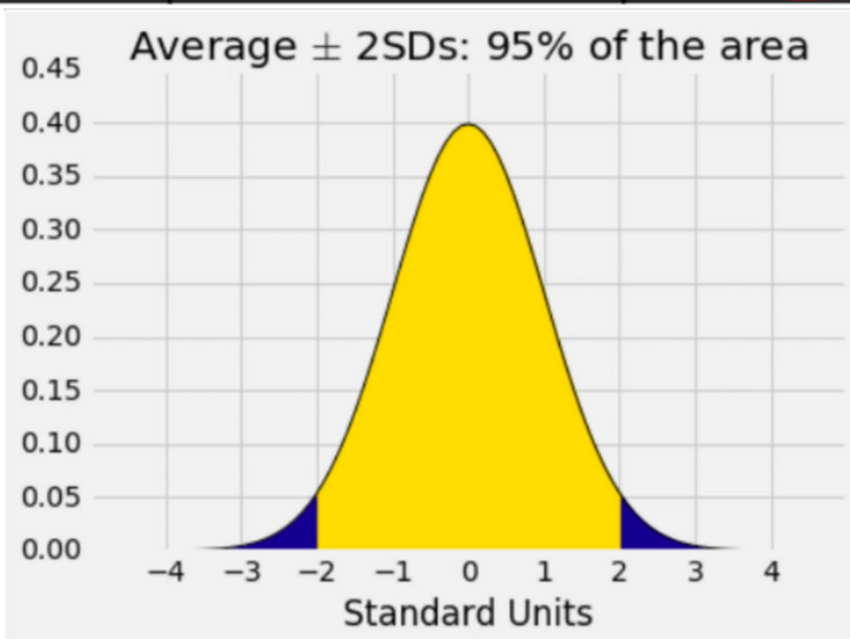
No matter what the shape of the distribution, the bulk of the data are in range "average ± a few SDs". If a histogram is bell-shaped, then

- Almost all of the data are in the range "average ± 3 SDs"

## Bounds and Normal Approximations

| Percent in Range | All Distributions | Normal Distribution |
|---|---|---|

| Percent in Range | All Distributions | Normal Distribution |
|---|---|---|
| average ± 1 SD | at least 0% | about 68% |
| average ± 2 SDs | at least 75% | about 95% |
| average ± 3 SDs | at least 88.888...% | about 99.73% |




## Second reason for using the SD

If the sample is

- Large, and
- Drawn at random with replacement,

Then, regardless of the distribution of the population, the probability distribution of the sample sum (or of the sample average) is roughly normal

## Sample averages

- Often we only have a sample; we don't know much about the population which it was drawn.
- The central limit theorem says that the probability distribution of the average of a large random sample is roughly normal, regardless of the distribution of the population.
- This allows us to make inferences based on averages of large random samples.

EXTRA:
What is the mean and standard deviation of a list converted into standard units? Mean is 0, SD is 1

# Correlation

## Prediction

- To predict the value of a variable,
  - Identify attributes that are associated with that variable and that you can measure.
  - Describe the relation between the attributes and the variable you want to predict
  - Use the relation to make your prediction

## Visualization
Two numerical variables

- Trend: Just some general upward or downward movement.
  - Positive association
  - Negative association
- Patern
  - Any discernible "shape" in the scatter
  - Linear
  - Non-linear

**Motto: Visualize, then quantify**

## The correlation coefficient *r*

- Measures linear association
- Based on standard units
- $-1 \leq r \leq 1$
  - R = 1: scatter is perfect straight line sloping up
  - R = -1: scatter is perfect straight line sloping down
- R = 0: no linear association; uncorrelated

## Definition of *r*
Correlation Coefficient *(R)* =

| | | | | |
|---|---|---|---|---|
| Average of | Product of | X in standard units | and | Y in standard units |

Measures how clustered the scatter is around a straight line

## Operations that leave *r* Unchanged
The correlation coefficient is not affected by:

- Changing the units of measurement of the data
  - Because r is based on standard units
- Which variable is plotted on the horizontal axis and which on the vertical
  - Because the product of standard units is the same either way

## Casual conclusion
Be careful …

- Correlation measures linear association
- Association doesn't imply causation
- Just because two variables are correlated, that doesn't mean one causes the other

## Nonlinearity and Outliers
Both of these can affect correlation

- Draw a scatter plot before you decide to compute r

## Ecological Correlation

- Correlations based on groups or aggregated data
- These can be misleading
  - For example, they can be artificially high