Class 7: Selecting the right statistic to describe data

API-201
Quantitative Analysis and Empirical Methods I

Profs. Goel and Taylor
Harvard Kennedy School

Agenda

1. Central tendency
2. Spread
3. Correlations

Mean and median

A stylized example

In 2012 Harvard conducted a survey of the annual income of the entering class of 2002. They were interested in learning the *typical* salary 10 years after first entering.

Median salary was approximately $75,000
[The *middle* number. Sort everyone by their salary and pick the salary of the person who's halfway down the list.]

Mean salary was approximately $_____
[The *average* number. Add up everyone's salary and divide by the number of students.]

Outliers

Outliers are data points that are *extreme* in value relative to others. [What is *extreme* depends on context.]

What should we do with outliers that we encounter in a dataset?

Mean or median?

Mean

- Familiar to many people
- Building block of statistics [more later]
- But…sensitive to outliers and *skew* in the data

Median

- Robust to outliers and *skew* in the data

Mean or median?

It depends. Rule of thumb for characterizing the "**center**" of a distribution

- Use the **mean** when the distribution is symmetrical and lacks outliers.
- Use the **median** when the distribution is skewed or has outliers.

Try both, and if they're approximately the same, then your conclusions are robust to your choice!

Distributions matter

The mean and median give us useful information, but it's often important to look at the whole distribution.

London transit

In 2006, the main contractor in charge of London's subway system was compensated according to system-wide criteria.

- The contractor had to ensure that at least 85.75% of gates and ticket readers were operational on average.
- The contractor would receive a bonus if it maintained at least 95.75% availability.

Why might such a compensation scheme lead to problematic outcomes? [Discuss with your neighbor.]

Question

A student scores 86 in their Econ midterm and 76 in their Stats midterm. Both exams were out of a possible maximum of 100 points. The mean score for the class was 80 for Econ and 70 for Stats.

Relative to the rest of the class, in which exam did they do better?

A. Econ B. Stats C. The same D. Not enough information

Variance

A widely used measure of the spread of a variable is the *variance*.

The variance tells us — roughly — how far a variable is, on average, from its mean.

mean $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

Variance

A widely used measure of the spread of a variable is the *variance*.

The variance tells us — roughly — how far a variable is, on average, from its mean.

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

A large variance indicates that the data points are spread far from their mean, and a small variance indicates that they are clustered closely around their mean.

Variance and standard deviation

The variance involves *squared differences*, and so the units of variance are the square of the original units.
[If original units are **dollars**, variance is in **dollars squared**.]

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

It's often convenient to express spread in the original units, and so we also use the *standard deviation*, which is the square root of variance.

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

Correlation

Correlation is a measure of *association* between two variables.

Correlation takes values between -1 and 1.

- **Positive correlation** between X and Y means [roughly] that X and Y increase together.
- **Negative correlation** means [roughly] that X and Y move in opposite directions.
- The farther the correlation is from 0, the stronger the relationship

Correlation

- The correlation between kindergarten test score and number of children by age 27 is _____.
- The correlation between kindergarten test score and future earnings is _____.

Correlation is not causation

Increasing kindergarten test scores does not necessarily *cause* people to have fewer children or to earn higher salaries.

The correlation is just telling us that test scores are *associated* with number of future children and future earnings.

People who have higher kindergarten test scores also tend to have fewer children and earn higher salaries.

Sometimes prediction is enough!

Doctors aim to *predict* future illness so as to intervene early
[e.g., risk of heart disease]

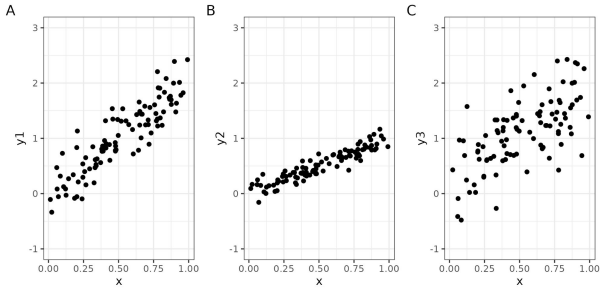
- Family history of heart disease is often a good predictor of your own health.
- If a relative gets heart disease, it does not generally *cause* you to develop heart disease.
- But people with relatives with heart disease have higher risk of developing heart disease due to shared genes and environments
- Family history is a useful **predictor** but **not a cause**

Correlation as prediction

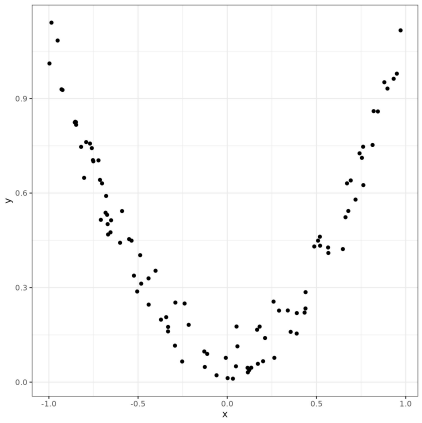
When X and Y are highly correlated [with values close to 1 or -1], then X is a good predictor of Y.

When X and Y have low correlation [close to 0], then it is often hard to predict Y from X.

Which plot shows the largest correlation?



Correlation measures *linear* relationships



Correlation in a nutshell

Key ideas

- Correlation is between -1 and 1
- Correlation measures **linear** relationships
- Correlation is **positive** when X and Y move in the same direction and is **negative** when X and Y move in opposite directions
- Correlation measures how well we can **predict** Y from X

Think about what's important

Central tendency?
To you want to reduce the influence of outliers?

Spread of distribution?
Are you concerned about a particular part of the distribution?

Relationship between two variables?
Correlation measures linear relationships

Something else entirely?

Key takeaways

Summary statistics can be very helpful at drawing conclusions from data. **But there's no perfect statistic**.

Each statistic is trying to **summarize** an entire distribution into a single number, and necessarily omits information

- Look at distributions
- Understand what each statistic tells you [and what it doesn't]