DATA8

Summer 2022

Lecture 23

Correlation

Meme Monday

Announcements

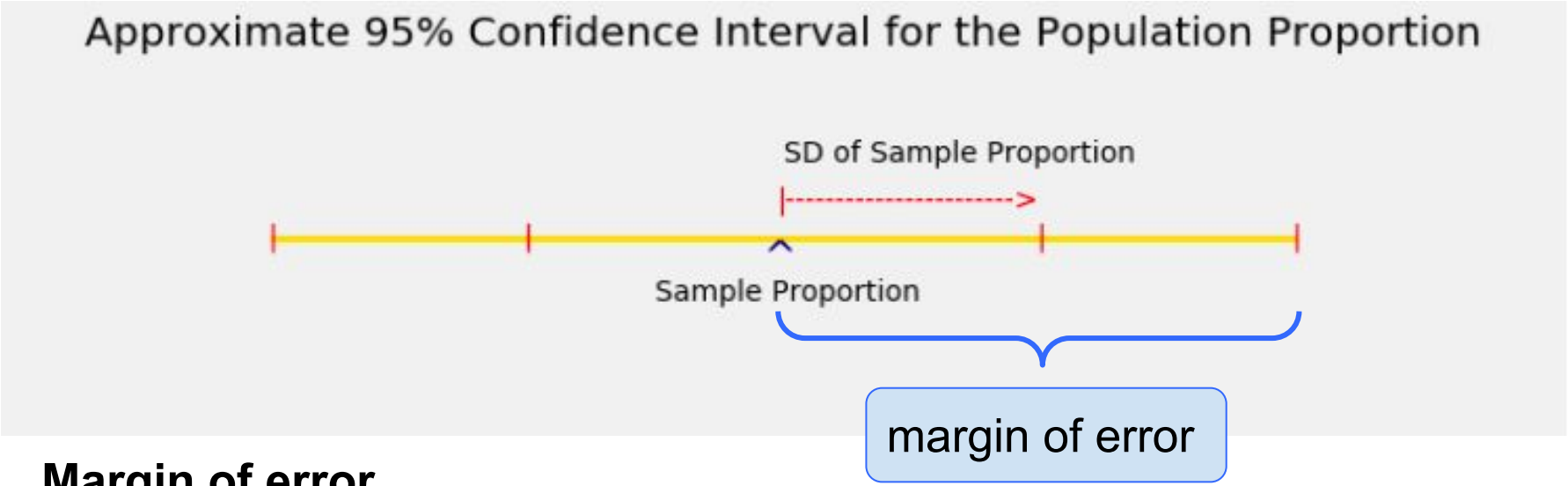
- Assignments
  - HW 8 is due Tue 7/26 (EC 7/25)
  - Today's lab: Lab 8
  - Lab 9 released later today
- Grade reports released on Gradescope.
- Last module of the course is fast-paced. Make sure to keep up with the course.

Weekly Goals

- Today
  - Intuitive approach to prediction
  - A measure of linear association
- Tuesday
  - Predicting one numerical variable based on another
  - Linear Regression
- Wednesday/Thursday
  - The “best” linear predictor
  - The method of least squares
  - Evaluating lines of best fit

Review: Margin of Error

Margin of Error in Polls



Margin of error

- Distance from the center to an end
- Half the width of the interval
- 2 * SD of sample proportion

95% CI for Population Proportion

- Based on a large random sample
- Total width = 4 * (SD of 0/1 population) / $\sqrt{\text{sample size}}$
- "Margin of error"
  = distance from the center to the end
  = 2 * (SD of 0/1 population) / $\sqrt{\text{sample size}}$
- The SD of a 0/1 population is at most 0.5

Prediction

Guessing the Future

- One branch of machine learning *(supervised machine learning)* uses data to making predictions

This week

Predicting numbers

Predicting labels

Guessing the Future

- Based on incomplete information

- One way of making predictions:
  - To predict an outcome for an individual,
  - find others who are like that individual
  - and whose outcomes you know.
  - Use those outcomes as the basis of your prediction.

(Demo)

Remember **Association**?

Two Numerical Variables

- Trend
  - Positive association
  - Negative association
- Pattern
  - Any discernible “shape” in the scatter
  - Linear
  - Non-linear

Visualize, then quantify

(Demo)

Quantifying linear associations

Correlation Coefficient

The Correlation Coefficient *r*

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

(Demo)

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

Care in Interpretation

Watch Out For ...

- False conclusions of causation
- Nonlinearity
- Outliers
- Ecological Correlations

(Demo)

Chocolate and Nobel Prizes

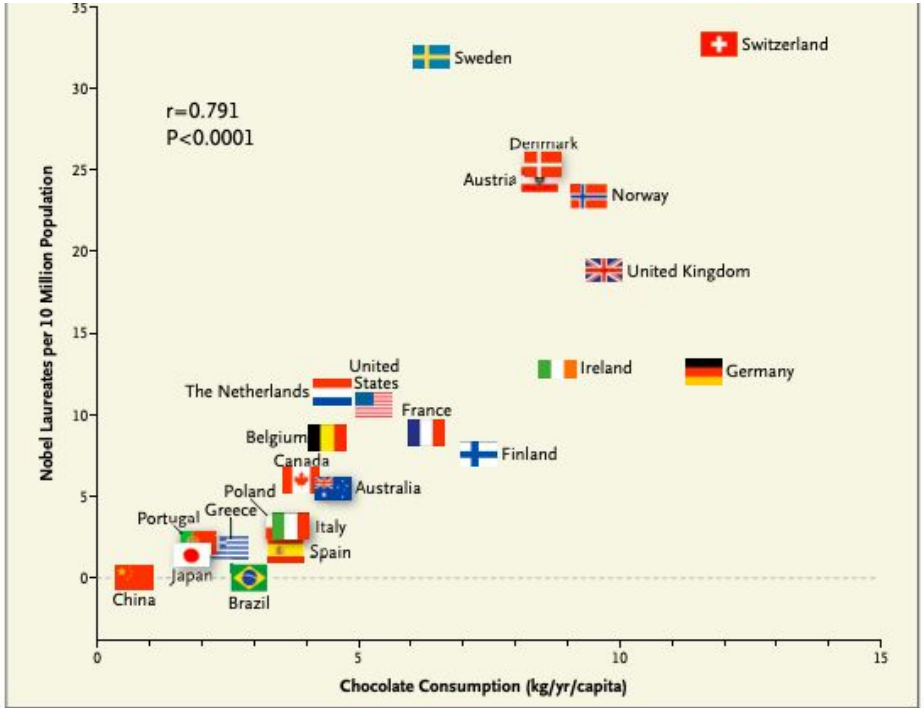


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Why do we care about **r** in light of prediction?