S1201 - Introduction

Wenda Zhou

May 22, 2017

What is statistics?

Merriam-Webster

“A branch of mathematics dealing with the collection, analysis, interpretation and presentation of masses of numerical data”.

What is statistics?

What is statistics?

Prediction

Infer future or unknown observations.

E.g. given information about person A, how much do they make?

What is statistics?

Prediction

Infer future or unknown observations.
E.g. given information about person A, how much do they make?

Estimation

Infer latent or unobservable parameters.
E.g. given information about person A, how much more would they make if they had a university diploma?

What is statistics?

Prediction

Infer future or unknown observations.
E.g. given information about person A, how much do they make?

Estimation

Infer latent or unobservable parameters.
E.g. given information about person A, how much more would they make if they had a university diploma?

Testing

Truth under uncertainty.
E.g. given a survey of 100 random people among which all those having a university diplom earn more, how confident are we that it is not due to chance?

Achieving the goals of statistics

Achieving the goals of statistics

Mathematics

- conceptualize and understand data
- theoretical guarantees

Achieving the goals of statistics

Mathematics

- conceptualize and understand data
- theoretical guarantees

Software

- implement algorithms to analyze data
- facilitate the collection of data

Achieving the goals of statistics

Mathematics

- conceptualize and understand data
- theoretical guarantees

Software

- implement algorithms to analyze data
- facilitate the collection of data

Domain knowledge

- provide questions and problems
- improve models using expert knowledge

Data

Wide variety of data today:

- Surveys and experiments
- Financial and economic
- Text and other media
- Sensor data

Data

Wide variety of data today:

- Surveys and experiments
- Financial and economic
- Text and other media
- Sensor data

Unified mathematical framework to work with data.

Rectangular data

Table: First 10 row of SOCR MLB player data set

Name	Team	Position	Height (in)	Weight (lbs)	Age (yr)
Adam Donachie	BAL	Catcher	74	180	22.99
Paul Bako	BAL	Catcher	74	215	34.69
Ramon Hernandez	BAL	Catcher	72	210	30.78
Kevin Millar	BAL	First Baseman	72	210	35.43
Chris Gomez	BAL	First Baseman	73	188	35.71
Brian Roberts	BAL	Second Baseman	69	176	29.39
Miguel Tejada	BAL	Shortstop	69	209	30.77
Melvin Mora	BAL	Third Baseman	71	200	35.07
Aubrey Huff	BAL	Third Baseman	76	231	30.19
Adam Stern	BAL	Outfielder	71	180	27.05

Rows = observations Columns = variables

What is a data type?

Each variable measures a specific attribute.
The type is a inherent property of the attribute describing the possible values the attribute can take and the semantic of those values.

What is a data type?

Most common data types fall along a dichotomy of numerical vs categorical data.

What is a data type?

Most common data types fall along a dichotomy of numerical vs categorical data.

Numerical data

- Represents a quantity
- Takes a range of numerical values
- Continuous (real-valued) or discrete (integer-valued)
- Has the semantics of a quantity

What is a data type?

Most common data types fall along a dichotomy of numerical vs categorical data.

Numerical data

- Represents a quantity
- Takes a range of numerical values
- Continuous (real-valued) or discrete (integer-valued)
- Has the semantics of a quantity

Categorical data

- Represents discrete categories
- If no relation between categories: nominal
- If categories are ordered: ordinal

Examples of numerical and categorical data

Weight in kg Numerical: can add, multiply, average. Always ≥ 0.

Player position Categorical and nominal.

Air humidity Numerical. Always between 0 and 1.

Number of customers per day Numerical. Integer quantity.

Qualitative weight (Under, Normal, Over, Obese) Categorical, ordinal.

Examples of numerical and categorical data

Zip Code Categorical, despite being a number.

Time of day (e.g. seconds since midnight) Numerical. Be careful when averaging!

Colour (by name) Categorical.

Colour (by wavelength) Numerical.

Examples of numerical and categorical data

Likert scales

- Likert scales appear in survey and similar designs.
- Strongly disagree, disagree, neither agree nor disagree, agree, strongly agree
- Usually 3, 5, or 7 points
- Categorical ordinal – but usually treated as numerical in practice

Examples of numerical and categorical data

Discretized data

Can turn numerical variable into categorical variable by discretizing.

BMI and Obesity

BMI is a numerical continuous measure.
WHO guidelines on obesity:

BMI	Classification
$\mathrm{BMI} < 18.5$	Underweight
$18.5 \leq \mathrm{BMI} \leq 24.9$	Normal weight
$25.0 \leq \mathrm{BMI} \leq 29.9$	Overweight
$\mathrm{BMI} > 30.0$	Obese

Why descriptive statistics?

Datasets can be complex.
Obtain simple and widely applicable summaries of the data.

Measures of centrality

Answer the question: where are the values? Are they large? small?

Measures of centrality

Answer the question: where are the values? Are they large? small?

mean or average

$$\bar{x} = \frac{1}{n} \sum_i x_i \tag{1}$$

Measures of centrality

Answer the question: where are the values? Are they large? small?

mean or average

$$\bar{x} = \frac{1}{n} \sum_i x_i \tag{1}$$

median

Value such that 50% of observations are smaller.

Measures of dispersion

Answer the question: how spread out are the values?

Measures of dispersion

Answer the question: how spread out are the values?

Variance and standard deviation

$$\sigma^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 \tag{2}$$

Variance $= \sigma^2$, standard deviation $= \sigma$.

Measures of dispersion

Answer the question: how spread out are the values?

Variance and standard deviation

$$\sigma^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 \tag{2}$$

Variance $= \sigma^2$, standard deviation $= \sigma$.

Interquartile range (IQR)

Different between third and first quartile.
Rarely used outside of box plots.

Always ≥ 0.

Measures of association

Answer the question: how related are two sets of values?

Measures of association

Answer the question: how related are two sets of values?

Covariance

$$\sigma_{xy} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) \tag{3}$$

Measures of association

Answer the question: how related are two sets of values?

Covariance

$$\sigma_{xy} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) \tag{3}$$

Correlation

$$\mathrm{corr}_{\mathrm{xy}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{4}$$

Measures of association

Answer the question: how related are two sets of values?

Covariance

$$\sigma_{xy} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) \tag{3}$$

Correlation

$$\mathrm{corr}_{\mathrm{xy}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{4}$$

Correlation and covariance have the same sign.
Correlation is between -1 and 1.

Caveats of descriptive statistics

Descriptive statistics can sometimes be misleading.