Descriptive

Agenda

- Identify categorical, ordinal and metric variables
- Calculate some descriptive statistics with pandas
- Understand when to use which statistics
- Show how this looks in code

1. Introduction

- Mainly recap of most important statistical concepts
- For some of you, this might be very simple, but in this first week, we just want to lay down the basics

Why do we spend time on this?

Exploratory Data Analysis (EDA) is a crucial step in working with data, first thing that we do before doing anything related to machine learning and to help us understand characteristics and relationships in our data, we look at different statistics

- Also for later: helps preprocessing our data and feeding our model (outliers, which features to use, Feature engineering ideas,...)

They help us understand our data, they help us simplify, and make sense of piles of numbers that we can not intuitively grasp.

A little bit like plotting - simplify and summarise our data in different ways.

Descriptive Statistics:

- A set of methods and measures to **describe** a **dataset**. Ways of capturing properties of a dataset or sample.
- **summarization** of data and provides language to talk about it. Reducing the large dataset to a **smaller summary statistic** (e.g. mean, min, count,..): aggregation as data reduction.
- There is no uncertainty in descriptive statistics

Inferential statistics:

Inferential statistics aims to **infer** values for attributes of a population by observing a **subsample** of that population. There is no certainty in inferential statistics!

Example: what is the favourite food of people living in Berlin? we ask (preferably randomly) 1000 people what their favourite food is (this is our **sample**), the most popular food is pizza.

Then we infer, that in the whole **population** (People in Berlin) the favourite food is pizza.

2. Variable types

Before we dive into the most widely used measures, let's have a look at the type of variables that we encounter in datasets. Three you see in the notebook, these are the ones we encounter often, there are more.

2.1 Nominal = Categorical

names, labels, categories that are not connected, **no order.** E.g. **colors**, **languages**, etc. but you cannot perform arithmetic operations on them.

- can be analysed: mode (what is most occuring?)
- could be 2 or more

2.2 Ordinal

- variable with an **order** or relative rank. But no standardised intervals, difference between values **not measurable**
- Examples are penguin health-status 🐧: poor, medium, good, excellent

2.3 Metric

- Quantitative values
- Natural ordering, difference is measurable.
- Discrete
  - The values of a **discrete** variable are **countable** *e.g. number of eggs of a penguin*
- Continuous
  - The values of a **continuous** variable are **uncountable** *e.g. Flipper length and body mass*

3. Measures

3.1. Measures of central tendency

- What is a typical value of the variable? where is the middle?

Mean = Average (= Expectation)

Tells us something about the data as a whole, but nothing about individual datapoints

Good for measuring if samples are normally distributed.

- Arithmetic mean: (mu: μ)

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

- Weighted arithmetic mean

  Where w_i is the weight assigned to the i_th observation of x.
  - Example: a survey may gather enough responses from every age group to be considered statistically valid, but the 18-34 age group may have fewer respondents than all others relative to their share of the population. The survey

team may weight the results of the 18-34 age group so that their views are represented proportionately.

$$\frac{w_1 x_1 + w_2 x_2 + ... + w_n * x_n}{n}$$

- Geometric mean
  - product of all the observations, take n-th root.
  - also often used with rates of return.
  - Example: Consider a stock that grows by 10% in year one, declines by 20% in year two, and then grows by 30% in year three. The geometric mean of the growth rate is 4.6% annualy

$$\sqrt[n]{\prod_{i=1}^{n} x_i}$$

- Harmonic mean
  - Invert observations, then take the average
  - often used for speed, or any kind of rate (e.g. rate of acceleration of a train)

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

(The other means besides the arithmetic one might have to imported from `scipy`)

Median

The value that divides the sample into two groups of equal size. Middle number if we line up data from smallest to largest.

- The probability of observing a value larger than the median and the probability of observing a value smaller than the median in the sample are 50% each.

Example: At the University of North Carolina, geography students have the highest average starting salary (above 100000 USD)

Why ?

Michael Jordan was a student there and he is definetely an outliter that skews the mean. Here it would be better to use the median and the way to communicate this info would be median is at such and such number, 50% of people earn below it, 50% above.

Mode

The Mode of a variable is the value that occurs most often in the dataset.

If the distribution is symmetrical (esp. normal), the measures of central tendency are equal!

BREAK? ☕ ?

3.2 Measures of dispersion/spread 🥪

Range

- largest number minus smallest number. The larger the distance: the more spread out

Interquartile Range (IQR)

- Similar, but doesn't consider extreme values: looks at spread of middle 50% of the data.

Variance

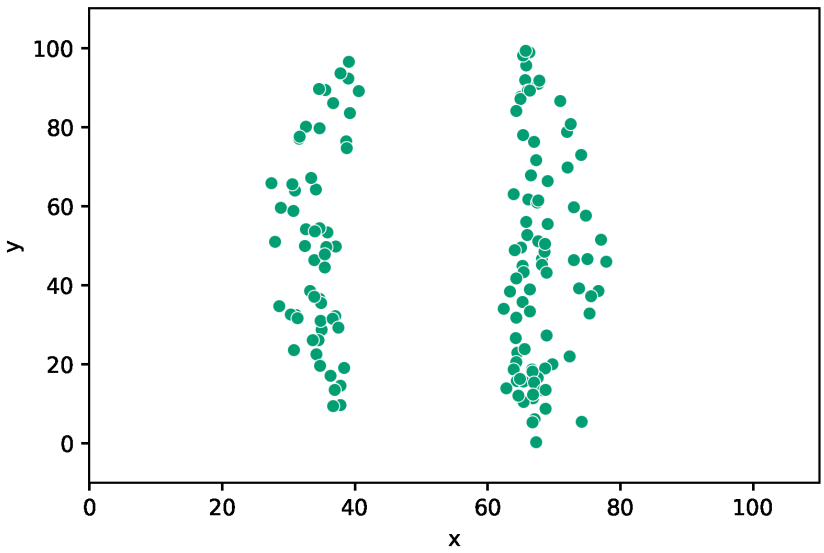
- includes all data, so gives a better idea. But still sensitive to outliers!
- sum of squared deviations from the mean divided by number of samples (average squared deviation)→ unit returned is not the same as our data!

Standard deviation

- squared root of the variance → returns our data-unit so we can interpret it.
- average we expect a datapoint to differ from the mean.

It’s always good to plot your data as some data with the same stats looks very different. so you might find something interesting there

Datasaurus Dozen. show gif

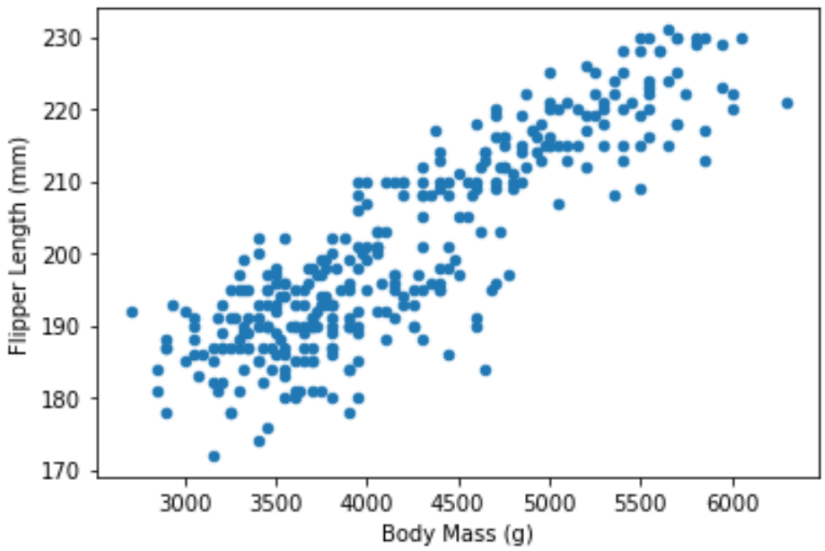


x mean: 54.27
y mean: 47.83
x standard deviation: 16.77
y standard deviation: 26.94
correlation: -0.07

4. Correlation and Causation

Scatter plots can tell us if there is a relationship between different variables in the data but we cannot **quantify** how strong the relationship. Correlation helps us answer that

Observe the plot: Is there a linear relationship?



- Yes. The bigger the penguin, the longer the flippers.
- And we can see, that it is a positive correlation:
  - **positive:** move in same direction (e.g. the more I run the more calories I burn).
  - **Negative:** move in opposite direction (e.g vaccination rates and rates of these illnesses)

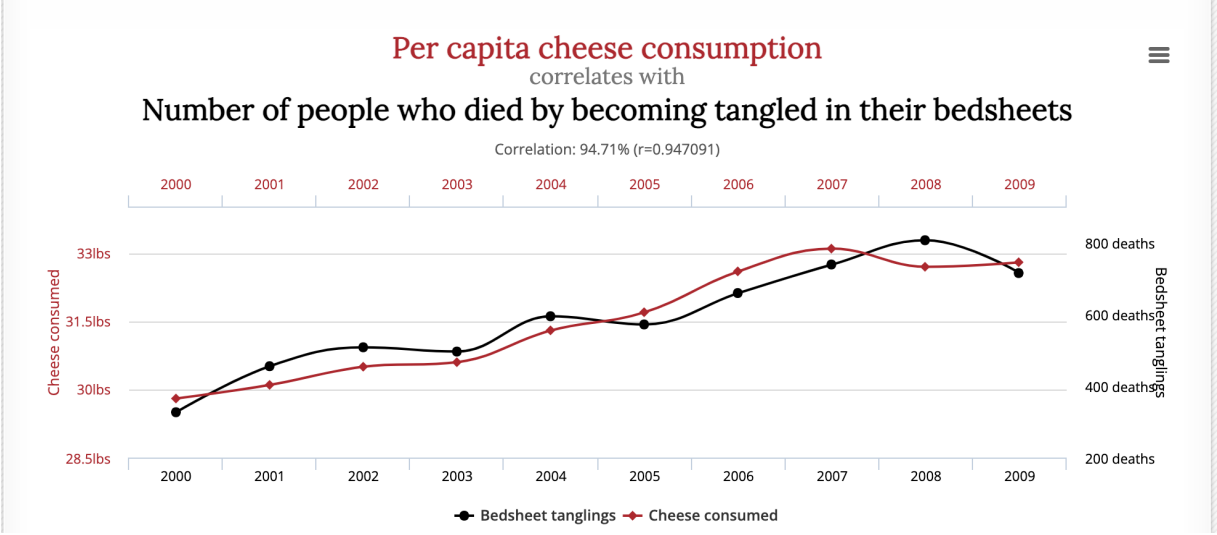
But how strong is it?

→ For this, we look at the **correlation coefficient (r)** (Pearsons, usually meant if not specificly stated because there are other types):

- describes the strength of the linear relationship between two variables
- -1 and 1 would be perfectly straight lines, 0 means no linear relationship.

Note: Correlation does not imply causation, always a rule of thumb.

Lets see an example:



More of those examples:

http://tylervigen.com/spurious-correlations

Exercises

- Open life expectancy data in pandas, and find an interesting statistic to share on Slack ✅