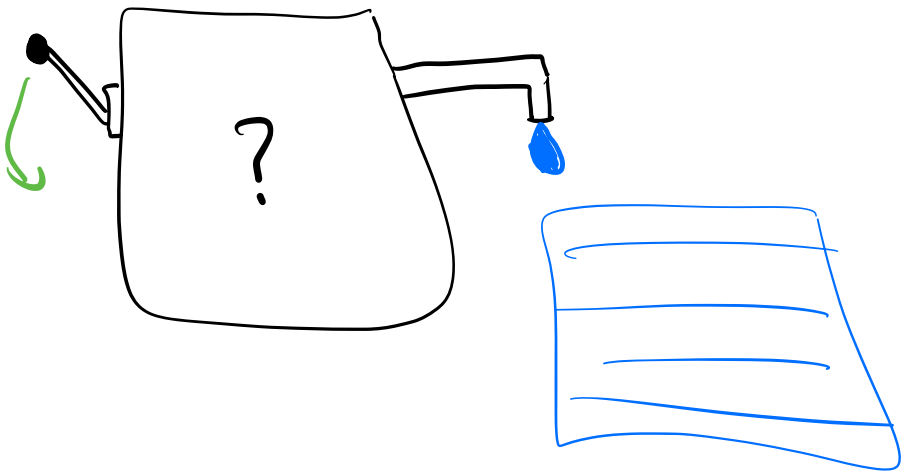
Probability and Statistics: A data scientist's perspective on the basics

Where does data come from?



Probability: how to model the ?

Statistics: Given data only, what can we learn about ?

Probability: basic terms

Coin flip

experiment - process that results in one of a set of outcomes

$\{H, T\}$

sample space - the set of possible outcomes

lands T

event - subset of sample space

the **probability** of an outcome s is written $P(s)$

and satisfies:

- $0 \leq P(s) \leq 1$ (between 0 and 1)
- $\sum_{s \in S} P(s) = 1$ (they all sum to 1)

$P(H) = 0.5$

$P(T) = 0.5$

A **random variable** is a function that maps an outcome to a number.

$V(T) = 0$

$V(H) = 1$

The **expected value** of a R.V. V is the sum of V's values, each weighted by their probability:

$$E(V) = \sum_{s \in S} P(s) V(s)$$

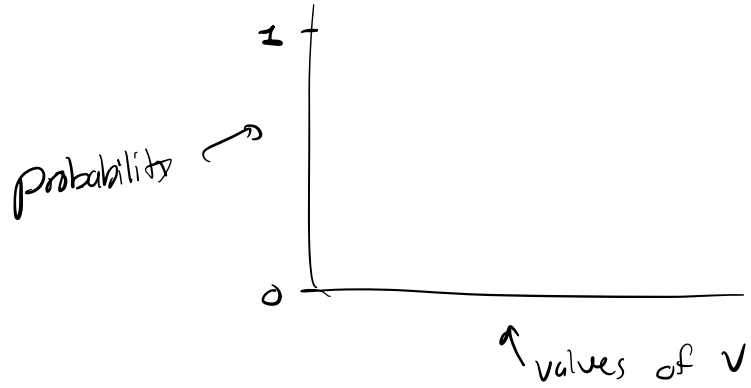
$E(V) = 0.5 \cdot 0 + 0.5 \cdot 1$

$= 0.5$

Exercises 1-3

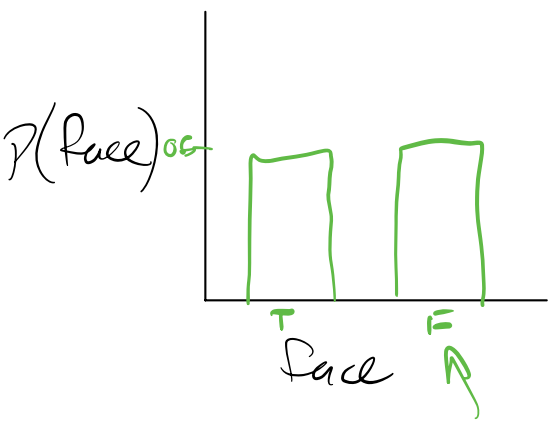
Probability Distributions:

A random variable V's probability density function (PDF) is a function mapping each value of V to its probability.

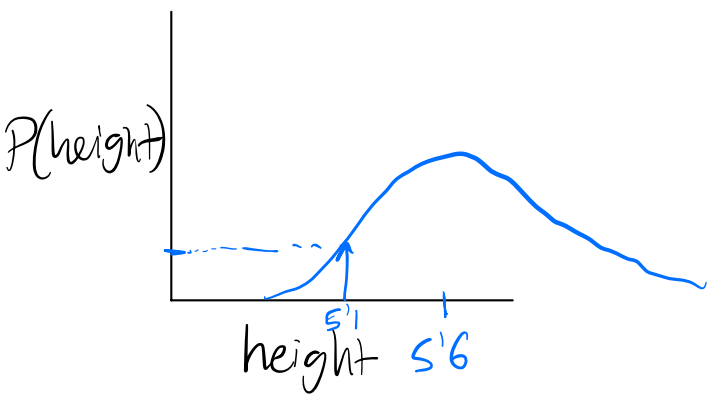


Examples:

Flip a fair coin



Grow a human to adulthood



Statistics

If all you have is data, what can you say about the generating process?

Histograms are essentially an empirical measurement of a PDF.

(see notebook)

Ex. 4

Summary Statistics

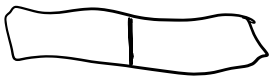
distill data to fewer numbers

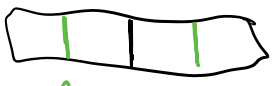
Central tendency measures

- mean (arithmetic) $\frac{1}{n}\sum_{x_i \in X} x_i$
- median middle value
- quantiles
- geometric mean $\left(\prod a_i\right)^{\frac{1}{n}}$ ↑ use for ratios

2 2 2 2 2 99

* ←





↑ 25th %ile ↑ 75th %ile

Variability measures

- standard deviation $\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sigma$
- variance σ^2

Ex. #5