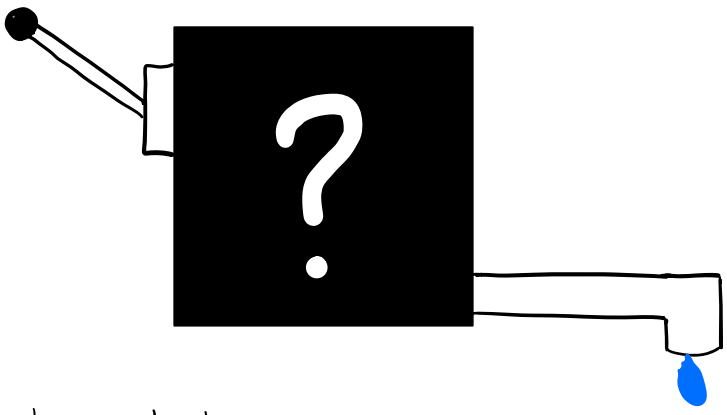
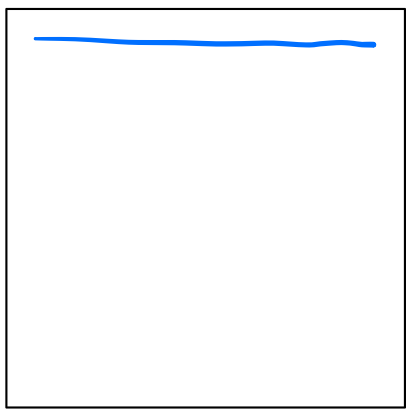
Probability and Statistics: A data scientist's take on the basics

Where does data come from?



Probability: modeling data-generating process

Statistics: "learning" the model from data



Probability: Basic Terms

Flip a fair coin	DS student takes an exam	**experiment** - a process that results in one of a set of possible **outcomes**
{H, T}	{A, B, C, D, F}	**sample space** - the set of possible outcomes
(1) {{H}, {T}, {H, T}}	{A} {A, B, C, D}	**event** - subset of the sample space

(2)

$P(H) = 0.5$

$P(T) = 0.5$

$P(A) = 0.6$
$P(B) = 0.3$
$P(C) = 0.05$
$P(D) = 0.03$
$P(F) = 0.02$

The probability of an outcome s is written $P(s)$ and satisfies:

- $0 \leq P(s) \leq 1$ (between 0 and 1)
- $\sum_{s \in S} P(s) = 1$ (total probability is 1)

(3)

$V(T) = 0$
$V(H) = 1$

$G(A) = 4$
$G(B) = 3$
$G(C) = 2$
$G(D) = 1$
$G(F) = 0$

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

(4)

$V(H) \cdot P(H)$
$1 \cdot 0.5$
+
$0 \cdot 0.5$
$V(T) \cdot P(T)$
$= 0.5$

3.43

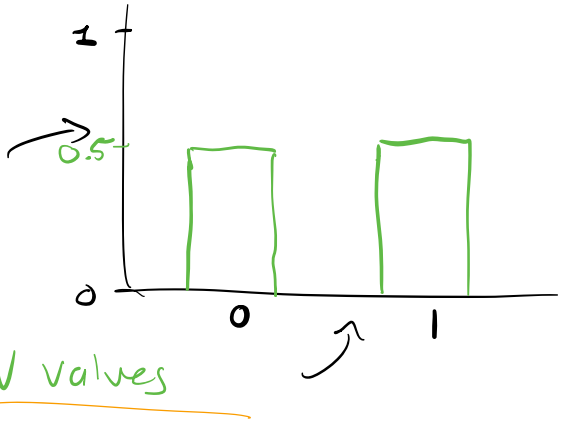
The expected value of a R.V. V is the sum of the values of each outcome, each weighted by its probability:

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

Extra Practice: Exercises 5-7

Probability Distributions:

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



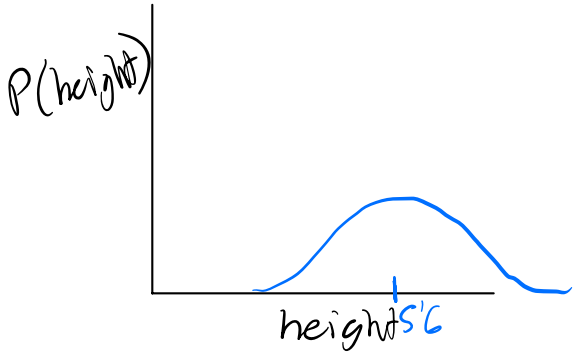
Examples:

Flip a fair coin

V = 1 if H, 0 if T

Grow a human to adulthood

H: height of human



Ex. 8

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)

Summary Statistics

distill data to fewer numbers

Central tendency measures

- (arithmetic) mean:
- median:
- geometric mean:
- quantiles:

Variability measures

- standard deviation
- variance

Ex. #9