Introduction to probability and statistics

Main topic: randomness

inherent: casino, radioactive decay, coins, dices

uncertainties: models, unknown quantities

errors: measurement devices, etc.

philosophical component: determinism, quantum physics, etc.

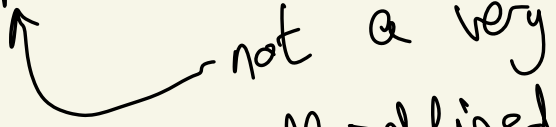
Mathematical framework allows us to circumvent these questions.

Literature:

Chan: Introduction to Probability for Data Science

Downey: Think Stats: Probability and Statistics for programmers.

Part 1: "Simple" random variables



Part 2: "Complicated" random variables

Throughout: discuss a few applications in data science and stats.

Part: "Simple" random variables

Examples: 1) Flipping a coin:

outcome: heads or tails.

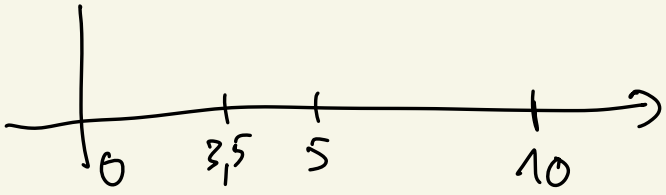
fair/unfair ?

2) Roll a die

outcomes: 1, 2, 3, 4, 5, 6

3) Choose a point between 0 and 10,

a) uniformly at random



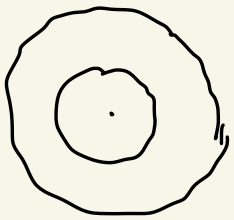
outcomes: infinitely many.

b) ..., but only take integer values

outcomes: $0, 1, 2, \ldots, 10$

c) ..., "so that larger values are more likely than smaller values".

4) Darts



random point in a disk

5) a) Temperature tomorrow in Berlin

b) rain / no rain.

Terminology: All of these are called "experiments"

↳ procedure with a random outcome that can be recorded.

Random variables and probability distributions.

---

1) **Random variables**: a quantity/object whose value/realisation depends on chance ("is random")

Examples: 2) Rolling a die:

random variable:

$X = 1, 2, 3, 4, 5$ or 6

(more mathematical notation:

$X \in \{1, 2, 3, 4, 5, 6\}$.

[As an aside:
the notion of "depends on chance" can be made very precise:

$$X : \Omega \longrightarrow \{1, 2, 3, 4, 5, 6\}.$$

(Ω: sample space)

would be a random variable]

[Chan, chapter 2].

1) Coin flip: $X =$ "heads" or "tails",

$$X \in \{\text{heads}, \text{tails}\}.$$

3a) "uniformly at random between 0 and 10":

$$X \in [0, 10].$$

Notice: X can take infinitely

many values.

Discrete vs. continuous random variables

Discrete: Examples 1, 2, 3b, 5b:

The random variables takes only finitely many[*1] values

Continues: Examples 3a, c:

The random variables takes (uncountably) infinitely many values.

Discrete

Continuous

*1: or countably infinite

Remark: Usually (in this course), random variables take numerical values, but the framework is very general: (Almost) anything can be random!

E.g. random words, sentences, graphs:

Think: chemical reactions.

2). Probability distributions:

roughly: assigns to every possible outcome a probability,

→ that is, a number between 0 and 1 (equivalently between 0% and 100%).

Start with discrete random variables.

Coin example:

$p(\text{"heads"}) = \frac{1}{2}$ (50%),

$p(\text{"tails"}) = \frac{1}{2}$ (50%)

OR (biased coin):

$p(\text{"heads"}) = 1/4$ (25%)

$p(\text{"tails"}) = 3/4$ (75%)

Die:

$p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = 1/6$

Observations (about these examples):

(can be turned into axioms)

- $0 \leq p(x) \leq 1$ for every outcome x.

  "Probabilities are numbers between 0 and 1 (included)".

- $\sum_i p(x_i) = 1$.

  If we have outcomes x_1, x_2, x_3, then this means $p(x_1) + p(x_2) + p(x_3) = 1$

  "The total probability is 1 (or 100%)".

Remarks: If all outcomes are equally likely (fair/unbiased coin or dice),

then $p(x) = \frac{1}{\text{total number of outcomes.}}$

(This related to the concept of symmetry.)

Recall: experiment, random variable, probability distribution

(synonyms: prob. law, prob. measure).

Interpretation:

i) Frequentist: If the experiment is repeated many times (say N times),

then the relative frequency of an outcome x will approach $p(x)$, as $N \to \infty$.

Simulation shows the following points:

- (Pseydo-) random number generator in Python
- relationships to histograms:

Theory vs. practice

probability distribution: theoretical object (describing the "essence" of a phenomenon).

histograms (approaches the prob. distr. as $N \to \infty$)

exhibits random fluctuation (observed data, realisation, sample).

ii) Bayesian (afte Thomas Bayes, 1701 - 1761)

probability as degree of belief.

"The chance of rain tomorrow is 30%."

"The chance that my mom had a coffee on the day of her wedding is 95%".

Next: What can we do with random variables / prob. distributions?

RV	1	2	3	4	5	6	
prob. distr	1/6	1/6	1/6	1/6	1/6	1/6	unbiased die
	0.6	0.1	0.1	0.05	0.05	0.1	dodgy die

Lumping together outcomes:

"even numbers" = $\{2, 4, 6\}$

event

Other events: "elementary events":

$\{1\}$, $\{2\}$, $\{3\}$.

"smaller than 4" = $\{1, 2, 3\}$.

certain event / whole sample space = $\{1, 2, 3, 4, 5, 6\}$.

Sum rule of probility.

("How do we calculate the probability of events?").

p("outcome is even") = $p(\{2, 4, 6\})$

$= p(2) + p(4) + p(6)$

$= 1/6 + 1/6 + 1/6 = 1/2 = 50\%$

(for the fair die.)

Dodgy die:

P_{DD} ("outcome is even")

$= p_{DD}(2) + p_{DD}(4) + p_{DD}(6)$

$= 0.1 + 0.05 + 0.1$

$= 0.25 = 25\%$

Warning. Suppose you conduct a study on healthy life style,

you record whether people exercise regularly, or whether they eat a balanced diet

event 1, ER

event 2, BD

$P(\text{"exercise regularly"}) = 0.6 = 60\%$

$P(\text{"eat a balanced diet"}) = 0.7 = 70\%$

$P(\text{"exercises regularly" or "eats a balanced diet"})$

$\neq 60\% + 70\% = 130\%$ ↯

The sum rule can only be used if the outcomes/events that are added are

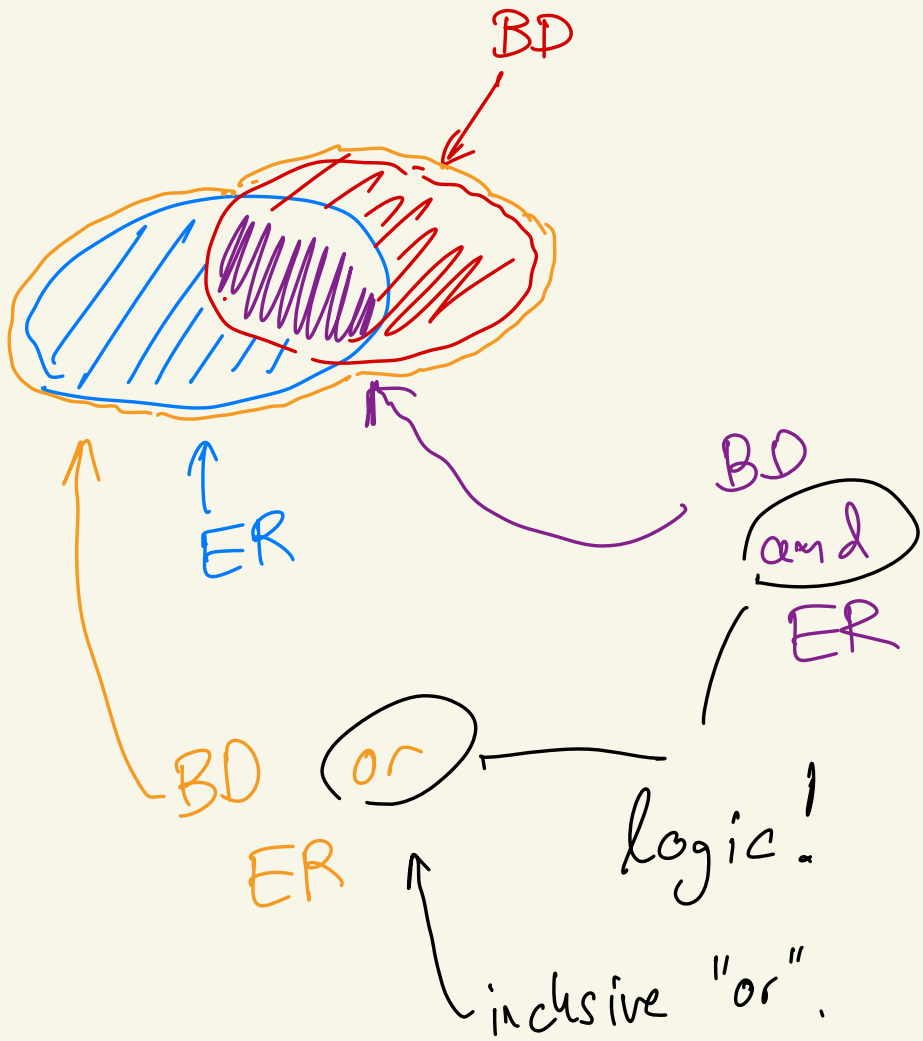
mutually exclusive!

→ Getting two different numbers on a die is mutually exclusive.

→ In contrast, one person can both exercise regularly and eat a balanced diet

event 1

event 2.



probability ⟷ area / volume associated to set, substance distributed in some space.

$$p(ER \cup BD)$$

or, write as $\vee$ a logical symbol,

set-theoretic symbol.

$$= p(ER) + p(BD) - p(ER \cap BD)$$

logical and, intersection / overlap.

Mean and variance.

Think back to the dice:

RV	1	2	3	4	5	6	X
prob. distr	1/6	1/6	1/6	1/6	1/6	1/6	unbiased die
	0.6	0.1	0.1	0.05	0.05	0.1	dodgy die

The mean / expectation / expected value

is a weighted average over possible outcomes:

$$\mu = \underset{\text{mb}}{\mathbb{E}}[X] = \sum_i x_i \, p(x_i)$$

notation — definition

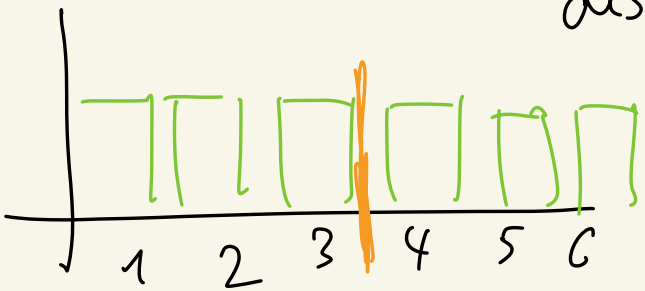
$$= x_1 \cdot p(x_1) + x_2 \cdot p(x_2) + \dots$$

$$= 1 \cdot \tfrac{1}{6} + 2 \cdot \tfrac{1}{6} + 3 \cdot \tfrac{1}{6} + 4 \cdot \tfrac{1}{6} + 5 \cdot \tfrac{1}{6} + 6 \cdot \tfrac{1}{6} = 3{,}5$$

In words: The average outcome is 3.5.

Different ways of thinking about it:

- Geometrically: mean characterises the center of the distribution:



unbiased die

Dodgy dice.

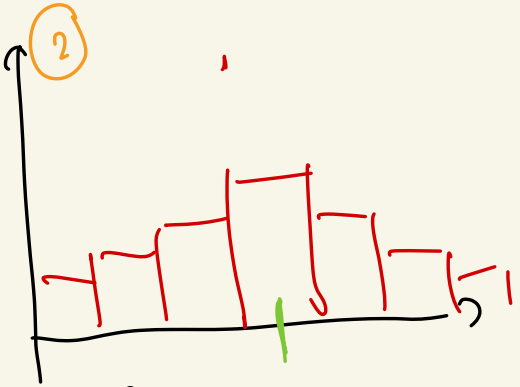
- Frequentist interpretation:

Law of large numbers.

If we repeat the same experiment a large number of times, then

the average over the outcomes will approach the expected value (the mean).

Variance and standard deviation.



more random than 1.

least random (or, not random at all)

(1) is spread out.

(2) is concentrated.

How can we describe this in mathematical terms?

Variance: $\operatorname{Var} X = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$

↑ definition

↓ in words: The variance is the expected (average) squared deviation from the mean.

$$= \sum_i p(x_i) \cdot (x_i - \mu)^2$$

↑ deviations from the mean.

weighting: How probable are these deviations?

← unbiased die

$$= \frac{1}{6}(1-3{,}5)^2 + \frac{1}{6}(2-3{,}5)^2 + \frac{1}{6}(3-3{,}5)^2 + \frac{1}{6}(4-3{,}5)^2 + \frac{1}{6}(5-3{,}5)^2 + \frac{1}{6}(6-3{,}5)^2$$

$$\approx 2{,}9167.$$

this is a measure for the variability/randomness inherent in throwing an unbiased die.

---

standard deviation:

$$\sigma(X) = \sqrt{\operatorname{Var} X}.$$

How to estimate mean and variance from data?

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{i=1}^{N}\left(\frac{1}{N}\right) x_i$$

data points

$\hat{\mu}$: mean from data

"hat" indicates that this is something that comes from data.

instead of $p(x_i)$, we use $\frac{1}{N}$, attaching equal weights/probability to the data points.

$$\hat{s}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \hat{\mu}\right)^2$$

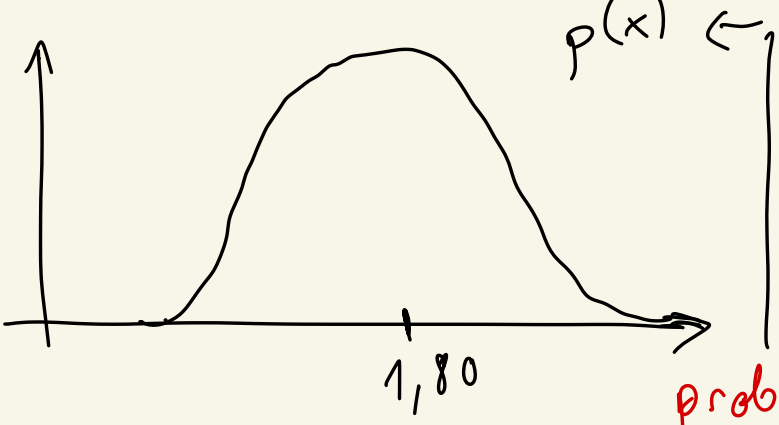
$\hat{s}^2$: variance from data

$\frac{1}{N}$ would also be ok.

Continuos random variables:

Random variables:

Probability distribution:



probability density function (pdf).

Quite similar to

$p(\text{"heads"}) = \frac{1}{2}$.

$p(\text{"tails"}) = 1/2$

$p(\text{"heads"}) + p(\text{"tails"}) = 1.$

Similar: • $p(x) \geq 0$ (probabilities are non-negative).

• higher values of $p(x)$ "indicate that the outcome x is more likely".

• $\int_{-\infty}^{\infty} p(x)\,dx = 1$,

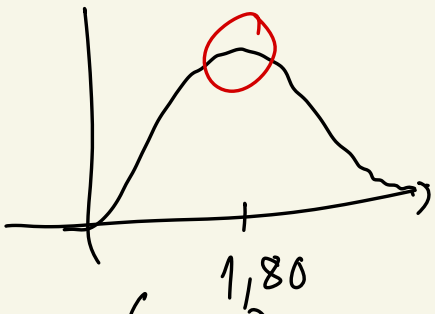
- mean and variance:

$$\mu = \int_{-\infty}^{\infty} x\,p(x)\,dx,$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2\,p(x)\,dx.$$

Warning! Q: What is the probability that a randomly chosen person's height is 1,80?

Temptation:



Read off. $p(1,80)$,
but it can happen that
$p(x) > 1$!

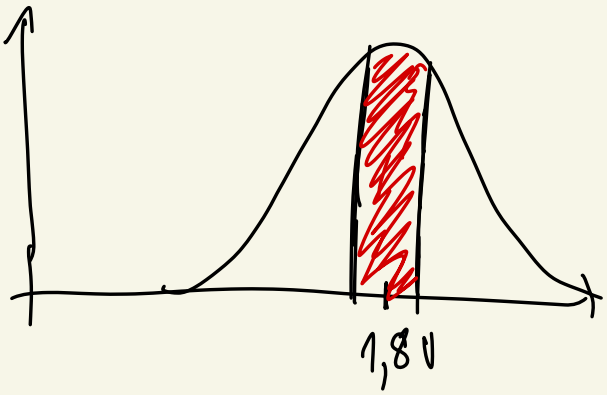
A: Literally, the answer is zero!

$1,80 = 1,8000000000\ldots$
different (e.g.) from $1,8000001$.

Better way of asking / interpreting the question:

What is the probability that a random person's height is approximately 1,80 (say, between 1,795 and 1,805 m).

↳ finite accuracy of measurement devices.



A: $p(1{,}795 \le X < 1{,}805)$

$$= \int_{1{,}795}^{1{,}805} p(x)\,dx.$$

might be difficult to compute.

The most important cont. distr. is normal / Gaussian distribution.

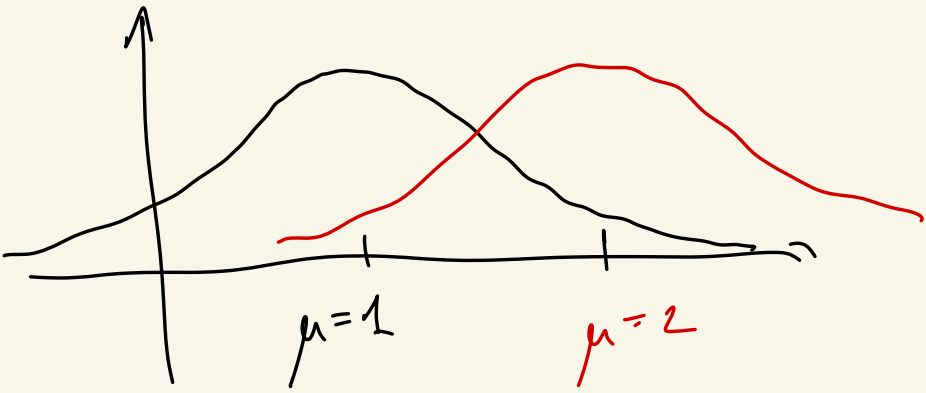
Formula:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

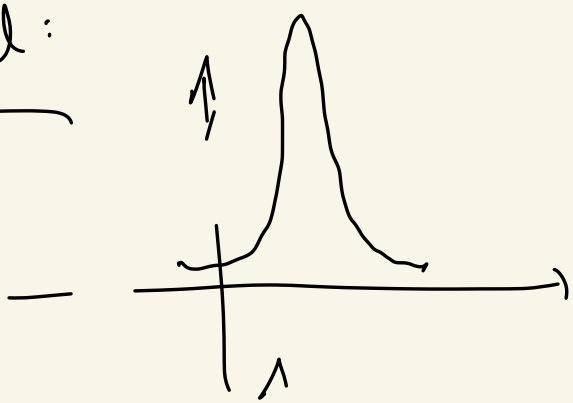
Picture:

- This is a family of distribution,

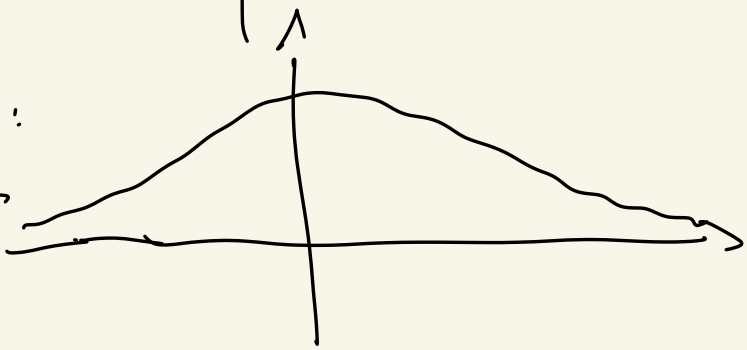
The role μ, the mean:



σ small:



σ large:



Why is it so important:

The Gaussian (normal) distribution shows up whenever a quantity can be thought of as arising from many small influences, which are

- independent from each other,
- none of them can be too important.