Northeastern University

INFO 6105
Data Sci Eng Tools & Mthds
Lecture 3 Statistics and Data Science

Probabilities and Bayesian Statistics

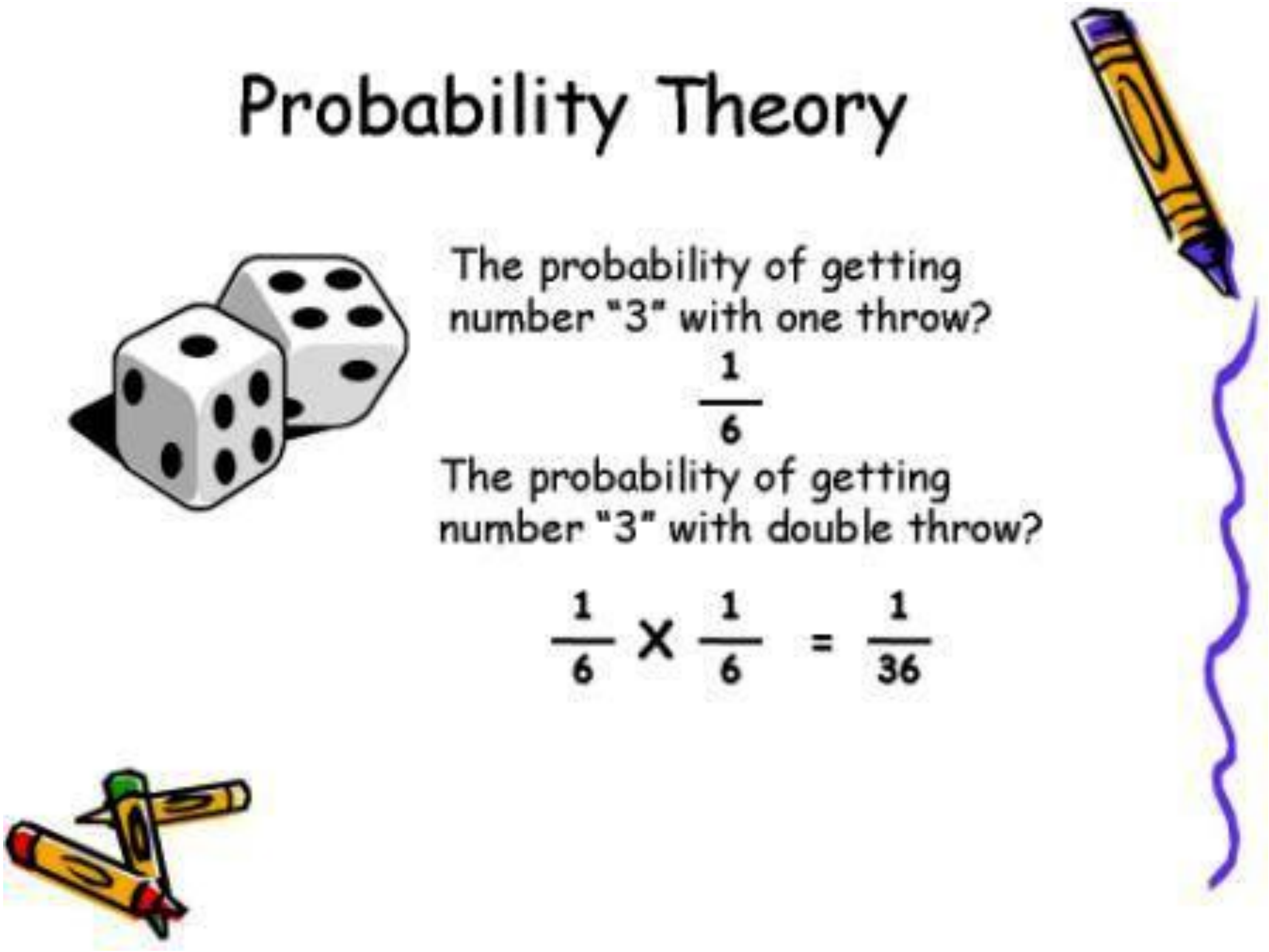
23 January 2019

New material

- Bayes, Occam, and Shannon
- Optimizing Pandas for speed

Part 1

PROBABILITY THEORY



Probability Theory

The probability of getting number "3" with one throw?

$$\frac{1}{6}$$

The probability of getting number "3" with double throw?

$$\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

Starting point of probability theory

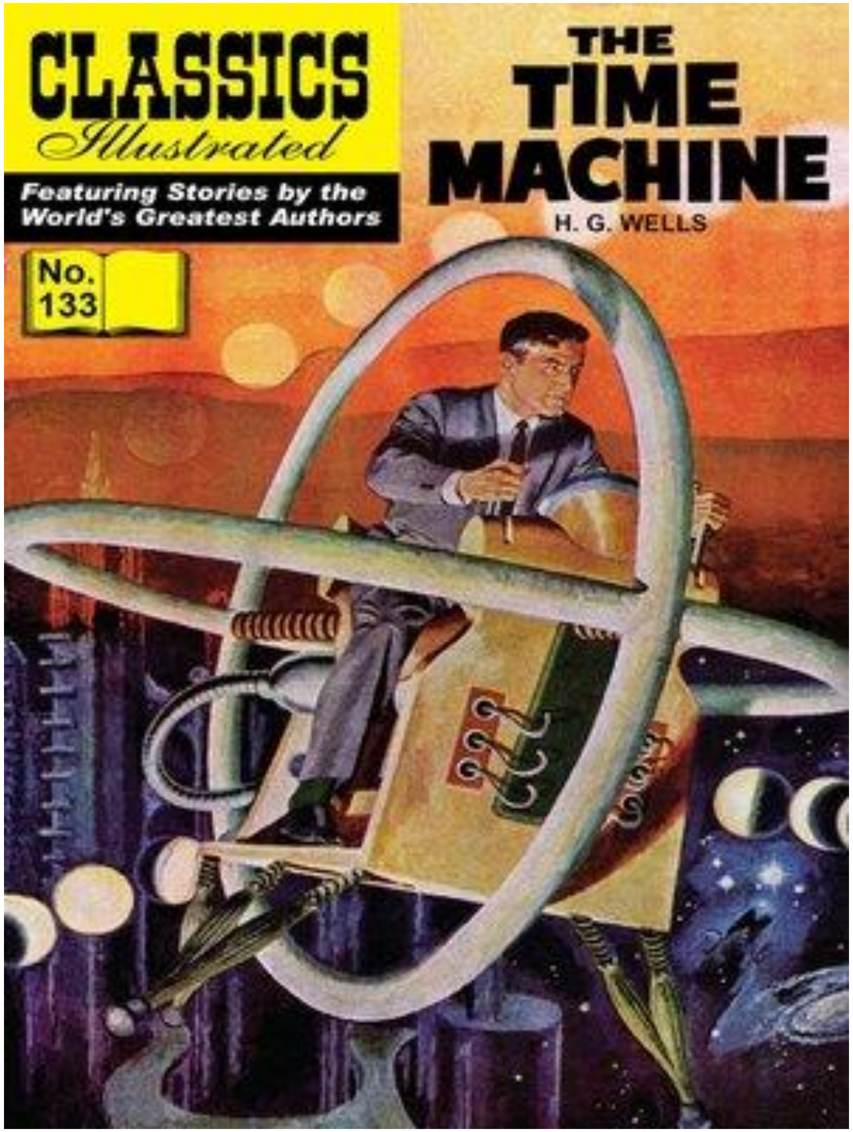
- **Given the probabilities of *two* events, the probability of *both* at the same time is:**

  $$P(a, b) = P(a) + P(b) – P(a \cap b)$$

- **Given the probabilities of two events, the probability of one event *after* the other is:**

  $$P(b|a) = P(a) * P(b)$$

State Machines

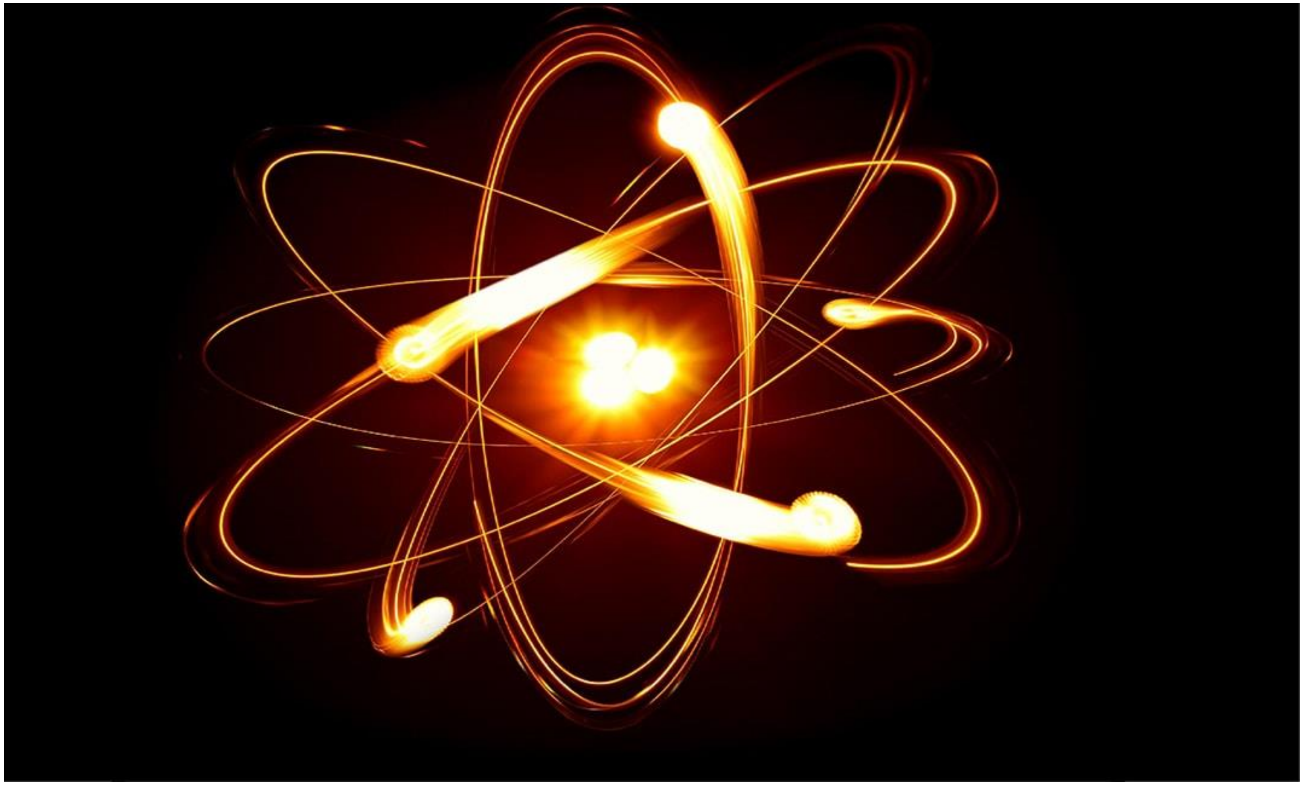


State Machines

- *Every* computer program is a Finite State Machine, because
  - Even if you convert all matter in the entire universe into a computer, it will *still* only have finite memory, thus a finite amount of states, and a finite amount of transitions between those states
- State Machines are *models*
  - Just like Lambda Calculus, Turing Machines, Random-Access Machines, Actor Systems, Object Systems, and so on
- Some problems lend themselves to being modeled by a state machine, some don’t
  - A traffic light?
  - Business Processes were often modeled as State Machines, even before computers existed
  - An elevator?
  - Netflix?

Probabilistic State Machines

- A state vector is a superposition of possible states, with only one of these states being an actual realization
- Example: Quantum Mechanics



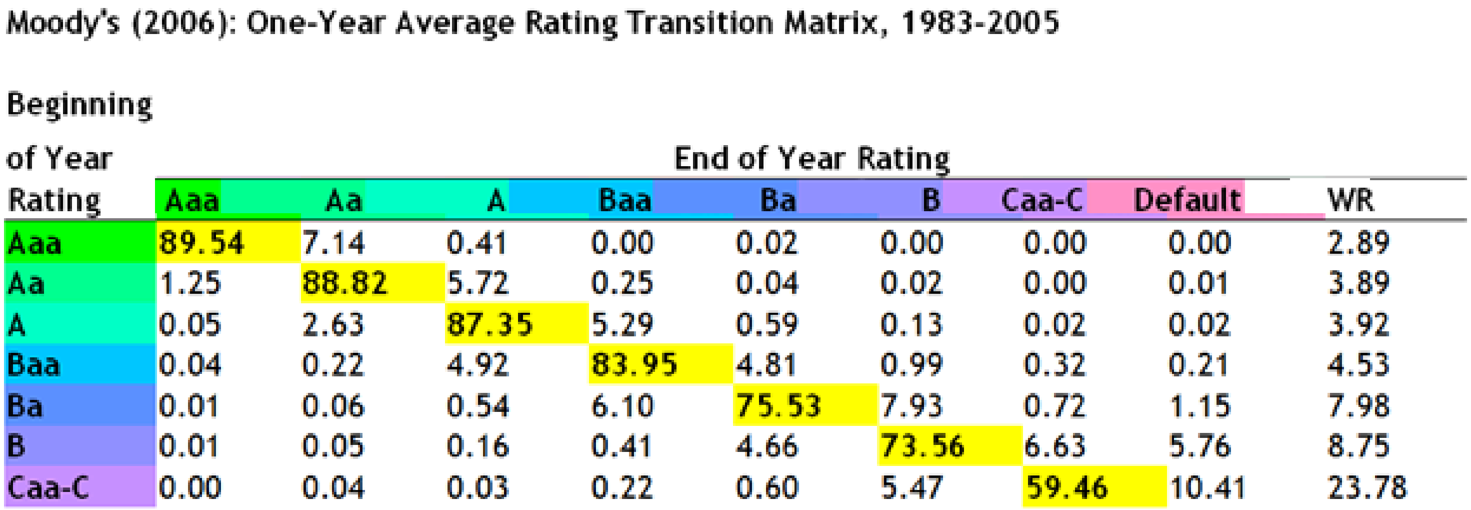
State Transition Matrix for probabilistic State Machines

- Each row represents a state
- Each column represents a state
- In row i, in each cell are the probabilities of moving from the state i, to the other states represented as column j
- Thus the rows of a Markov transition matrix each add to one
  - All state transition probabilities add to 1
- Sometimes such a matrix is denoted something like Q(x' | x) which can be understood this way
  - Q is a matrix, x is the existing state, x' is a possible future state, and for any x and x' in the model, the probability of going to x' given that the existing state is x, are in Q

Markov?

PHOTO: BRYAN ALLEN/CORBIS

Example: Moody’s Credit rating transitions



Moody's (2006): One-Year Average Rating Transition Matrix, 1983-2005

Beginning of Year Rating	End of Year Rating								
	Aaa	Aa	A	Baa	Ba	B	Caa-C	Default	WR
Aaa	89.54	7.14	0.41	0.00	0.02	0.00	0.00	0.00	2.89
Aa	1.25	88.82	5.72	0.25	0.04	0.02	0.00	0.01	3.89
A	0.05	2.63	87.35	5.29	0.59	0.13	0.02	0.02	3.92
Baa	0.04	0.22	4.92	83.95	4.81	0.99	0.32	0.21	4.53
Ba	0.01	0.06	0.54	6.10	75.53	7.93	0.72	1.15	7.98
B	0.01	0.05	0.16	0.41	4.66	73.56	6.63	5.76	8.75
Caa-C	0.00	0.04	0.03	0.22	0.60	5.47	59.46	10.41	23.78

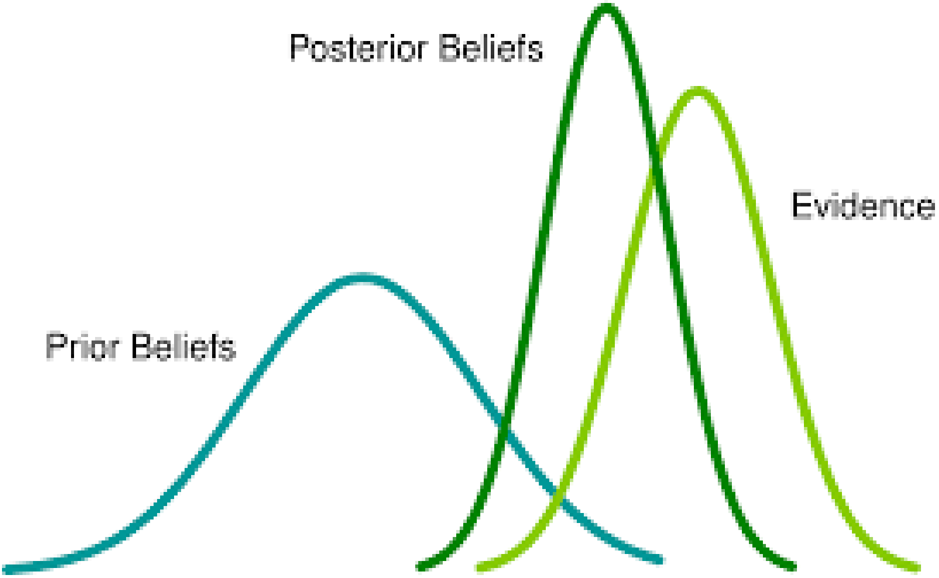
Two turns probabilities

- **Given these one-turn transition probabilities, can we compute the probability for each state to transition into any other state in *two* turns?**
- **Since the transitions are independent, the probability of *a* transitioning into *b* and then *b* into *c* is just the product $P_{c\leftarrow b}$ $P_{b\leftarrow a}$**
- **The probability of *a* transitioning into *c* via *any* intermediate state *b* is just the sum of those products over all possibilities *b:***

$$P^{(2)}_{c\leftarrow a} = P_{c\leftarrow 1}P_{1\leftarrow a} + P_{c\leftarrow 2}P_{2\leftarrow a} + \cdots = \sum_b P_{c\leftarrow b}P_{b\leftarrow a}$$

Matrix multiplication

- $\Sigma_k\ a_{ik}b_{kj}$ is *exactly* the formula for the multiplication of matrices [a_{ij}] and [b_{ij}]
  - That is why we need to lean *linear algebra, too*
- But let’s start with Probability theory..



Part 2

DATA SCIENCE & BAYESIAN STATISTICS

Statistics

- Statistics, and in particular *Bayesian statistics*, remains tough going to many
  - Amazed by the incredible power of machine learning, a lot of people have abandoned statistics and focus has narrowed down to exploring machine learning
  - Machine Learning is an n-dimensional surface printer and Machine prediction is an automated geometric surface extrusion approach
- Machine Learning (geometric approach) excels in Big Data problems, but many problems are not Big Data problems!
  - *Driving a car when it snows, at dusk, when the sun right behind the traffic light*, do we have Big Data for that?
- In 1770s, Thomas Bayes introduced ‘Bayes Theorem’
  - Centuries later, the importance of ‘Bayesian Statistics’ hasn’t faded
  - In fact, Bayesian ML is the most popular kind of ML today

Philosophy of Bayesian Inference

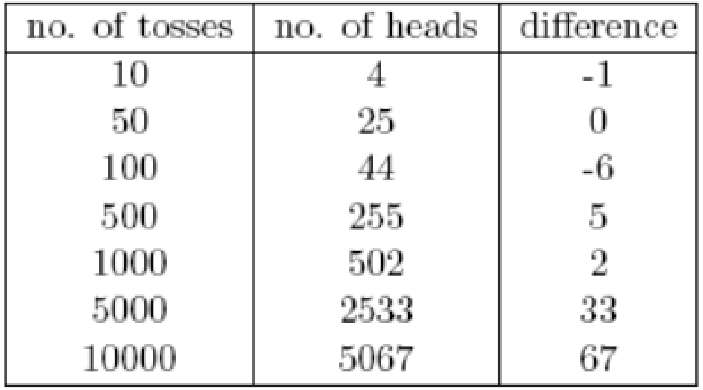
- You are a skilled programmer, but bugs still slip into your code. After a particularly difficult implementation of an algorithm, you decide to test your code on a trivial example. It passes. You test the code on a harder problem. It passes once again. And it passes the next, *even more difficult*, test too! You are starting to believe that there may be no bugs in this code...
- **If you think this way, then congratulations, you already are thinking *Bayesian*!**
  - **Bayesian inference is simply *updating your beliefs* after considering *new evidence***
  - **A Bayesian can rarely be certain about a result, but he or she can be very confident**
- We can never be 100% sure that our code is bug-free unless we test it *on every possible problem*
  - Instead, we can test it on a *large* number of problems, and if it succeeds we can feel more *confident* about our code, but still not certain
  - Bayesian inference works identically: We update our beliefs about an outcome

Frequentist statistics

- ..is the more *classical, non-Bayesian* version of statistics
- Assumes that probability is the long-run frequency of events
  - For example, the *probability of plane accidents* under a frequentist philosophy is interpreted as *long-term frequency of plane accidents*
  - *The probability of bugs in your code is the number of buggy SLOCs over total number of SLOCs you write* <u>*in your lifetime*</u>
- This makes logical sense for many probabilities of events, but becomes more difficult to understand when events have no long-term frequency of occurrences
  - We often assign probabilities to outcomes of presidential elections, but the election itself only happens once!
  - Frequentists get around this by invoking *alternative realities* and saying across all these realities, the frequency of occurrences defines the probability. Like Star Trek parallel universes. Yuck!
- Frequentist methods are still useful in some areas
  - Tools such as least squares linear regression, LASSO regression, and expectation-maximization algorithms are all powerful and fast
  - Bayesian methods complement these techniques by solving problems that these approaches cannot

Flaws in frequentist statistics

- Dependence of the result of an experiment on the number of times the experiment is repeated



no. of tosses	no. of heads	difference
10	4	-1
50	25	0
100	44	-6
500	255	5
1000	502	2
5000	2533	33
10000	5067	67

- **p-values measured against a sample (fixed size) statistic with some stopping intention changes with change in intention and sample size**
  - If two persons work on the same data and have different stopping intention, they may get two different p- values and t-scores for the same data, which is undesirable
- Confidence Intervals (C.I) are not probability distributions therefore they do not provide the most probable value for a parameter and the most probable values

Bayesian statistics

- Bayesians interpret a probability as measure of *belief*, or confidence, of an event occurring
- Simply, a probability is a summary of an opinion
  - An individual who assigns a belief of 0 to an event has no confidence that the event will occur
  - Conversely, assigning a belief of 1 implies that the individual is absolutely certain of an event occurring
  - Beliefs between 0 and 1 allow for weightings of other outcome
- This definition agrees with the probability of a plane accident example, for having observed the frequency of plane accidents, an individual's belief should be equal to that frequency, excluding any outside information
- If frequentist and Bayesian inference were programming functions, with inputs being statistical problems, then the two would be different in what they return to the user
  - Frequentist inference function would return a number, representing an estimate (typically a summary statistic like the sample average)
  - The Bayesian function would return *probability functions*

The `prior`

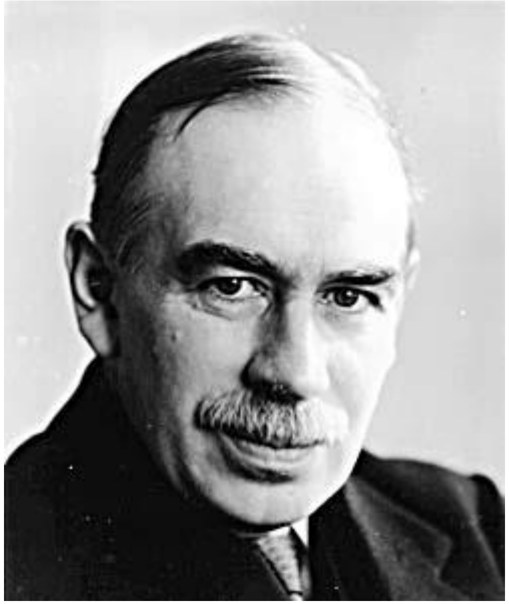
- **In our debugging problem above, calling the frequentist function with the argument "My code passed *all* XX tests; is my code bug-free?" would return a *YES***
- **Asking our Bayesian function "*Often my code has bugs*. My code passed all XX tests; is my code bug-free?" would return something very different. The function might return:**
  - **YES, with probability 0.8; NO, with probability 0.2**
- **This is very different from the answer the frequentist function returned**
  - **Notice that the Bayesian function accepted an additional argument: "*Often my code has bugs*“**
  - **This parameter is the *prior***
  - **By including the prior parameter, we are telling the Bayesian function to include our belief about the situation**

More and more evidence

- **As we gather an infinite amount of evidence, say as $N \to \infty$, our Bayesian results (often) align with frequentist results**
- Hence for large N, statistical inference is more or less objective
- On the other hand, for small N, inference is much more unstable: frequentist estimates have more variance and larger confidence intervals (thus less confidence)
- **This is where Bayesian analysis excels**
  - By introducing a prior, and returning probabilities (instead of a scalar estimate), we preserve the uncertainty that reflects the instability of statistical inference of a small N dataset
- Paradoxically, big data's predictive analytic problems are actually solved by *relatively simple algorithms*
  - Big data's prediction difficulty does not lie in the algorithm used, but instead on the computational difficulties of storage and execution on big data
  - The much more difficult analytic problems involve *medium data* and, especially troublesome, *really small data*

John Maynard Keynes & the `posterior`

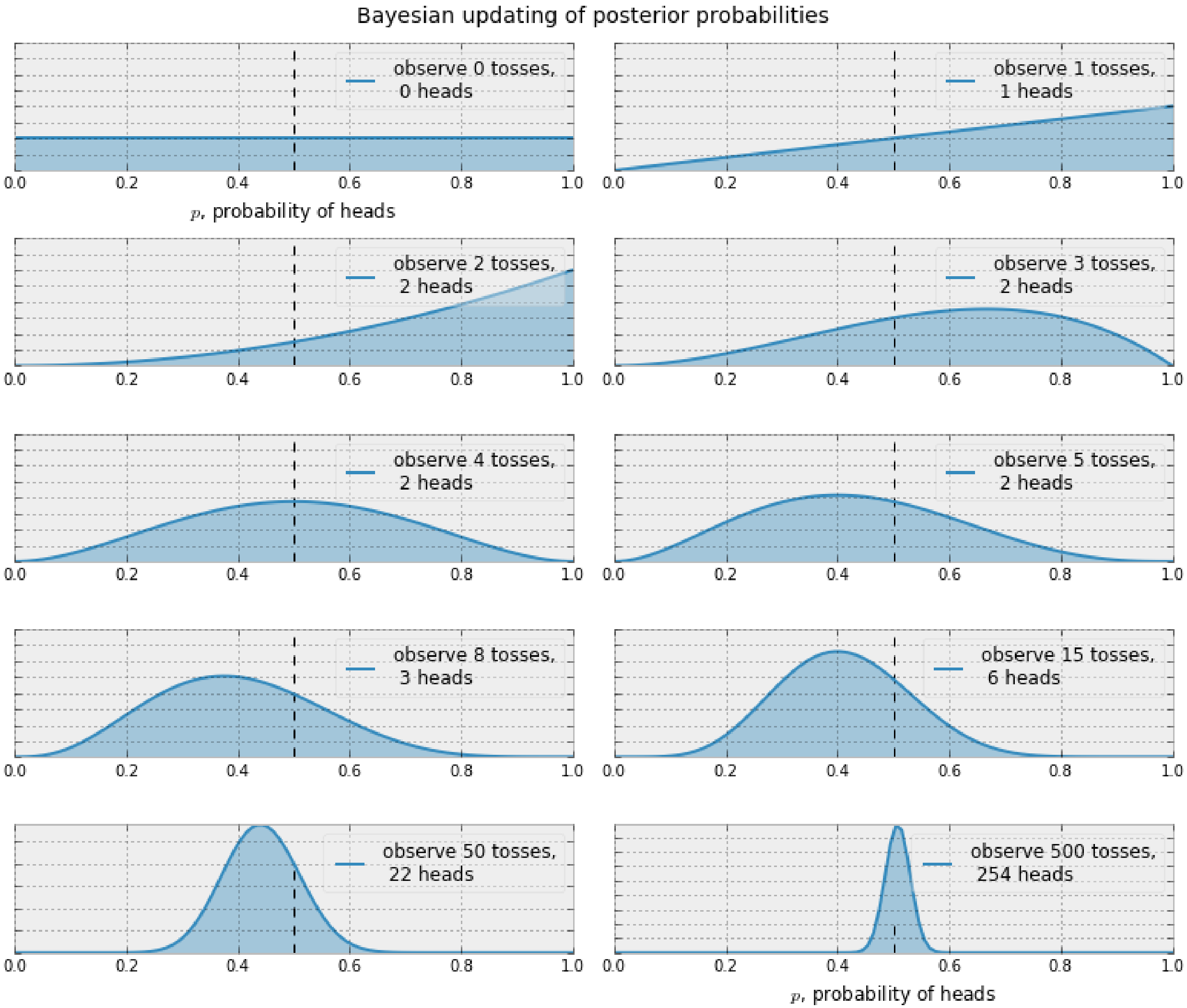
- John Maynard Keynes, a great economist and thinker, said "*When the facts change, I change my mind. What do you do, sir*?"
- This quote reflects the way a Bayesian updates his or her beliefs after seeing evidence
- Even — especially — if the evidence is counter to what was initially believed, the evidence cannot be ignored
- We denote our updated belief as P(A|X) , interpreted as the probability of A given the evidence X
- We call the updated belief the *posterior probability* so as to contrast it with the *prior probability*



Posterior probabilities

- Suppose you are unsure about the probability of heads in a coin flip (spoiler alert: it's 50%)
- You believe there is some true underlying ratio, call it pp, but have no prior opinion on what pp might be.
- We begin to flip a coin, and record the observations: either H or T .This is our observed data
- An interesting question to ask is how our inference changes as we observe more and more data?
  - More specifically, what do our *posterior probabilities* look like when we have little data, versus when we have lots of data
- We plot a sequence of updating posterior probabilities as we observe increasing amounts of data (coin flips)

Posterior probabilities of coin flips



- Posterior probabilities are represented by the curves, and our uncertainty is proportional to the width of the curve
- As we start to observe data our posterior probabilities start to shift and move around
- Eventually, as we observe more and more data (coin-flips), our probabilities will tighten closer and closer around the true value of p=0.5

Part 3: Lab

UNDERSTANDING BAYES

Slides by Brandon Rohrer

What does “Bayesian inference” even mean?

Inference = Educated guessing

Thomas Bayes = Nonconformist Presbyterian minister in London back when the United States were still The Colonies

He also wrote an essay on probability. His friend Richard Price edited and published it after he died

Bayesian inference = Guessing in the style of Bayes

What does “Bayesian inference” even mean?

Inference = Educated guessing

Thomas Bayes = A nonconformist Presbyterian minister in London back when the United States were still The Colonies.

He also wrote an essay on probability. His friend Richard Price edited and published it after he died.

Bayesian inference = Guessing in the style of Bayes

Dilemma at the movies

This person dropped their ticket in the hallway.

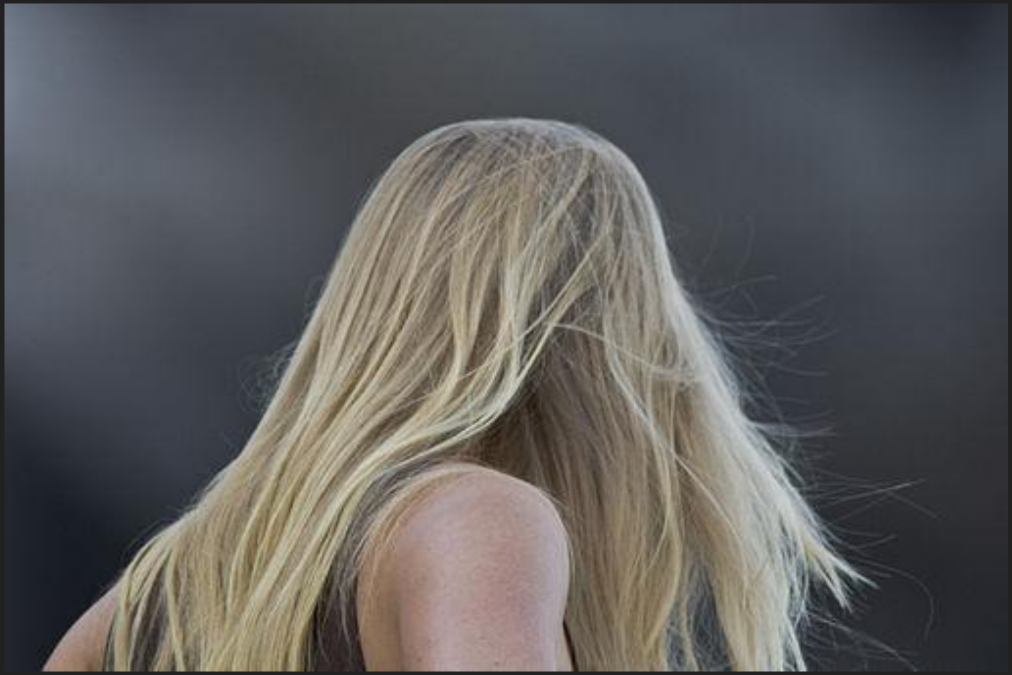
Do you call out

“Excuse me, ma’am!”

or

“Excuse me, sir!”

You have to make a guess.



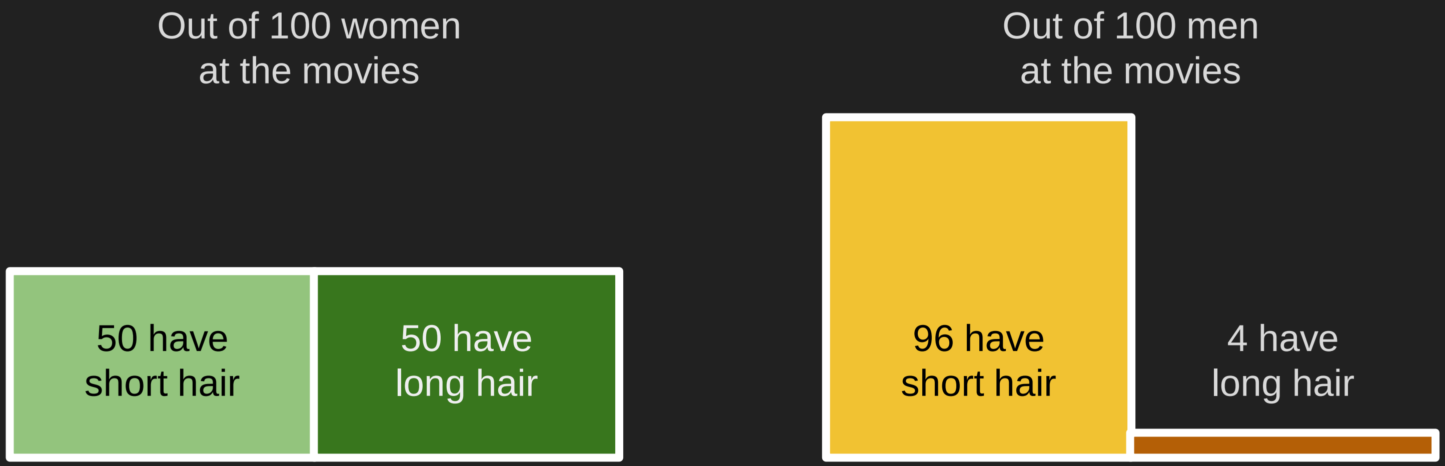
Dilemma at the movies

What if they’re standing in line for the men’s restroom?

Bayesian inference is a way to capture common sense.

It helps you use what you know to make better guesses.

Put numbers to our dilemma



Out of 100 women
at the movies

50 have
short hair

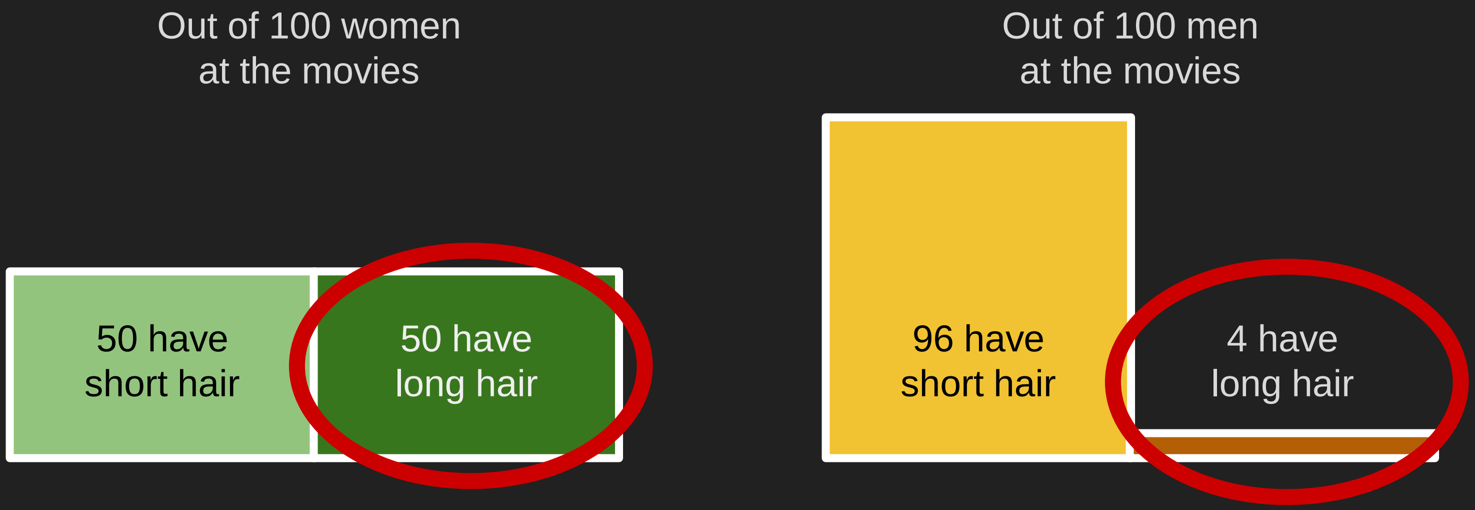
50 have
long hair

Out of 100 men
at the movies

96 have
short hair

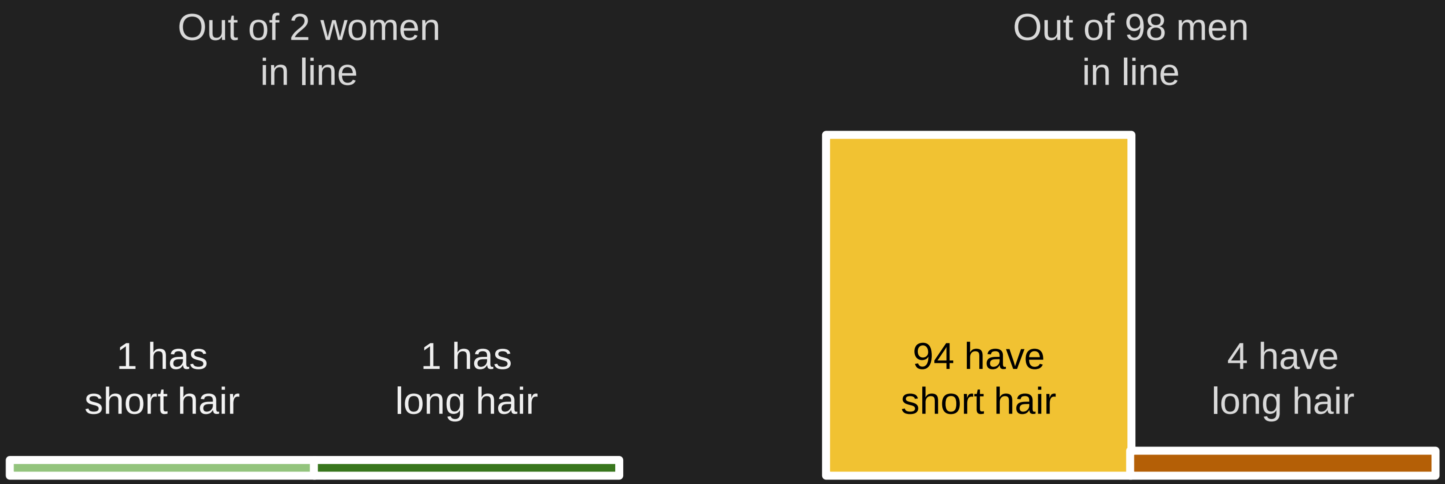
4 have
long hair

Put numbers to our dilemma



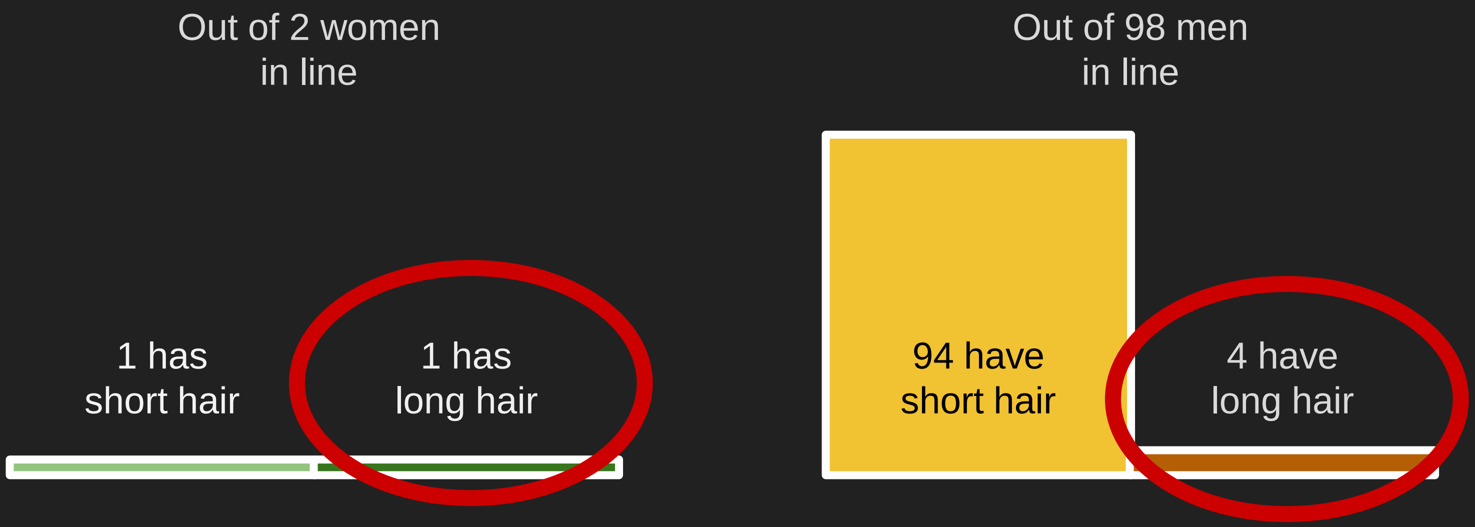
About 12 times more women have long hair than men.

Put numbers to our dilemma



But there are 98 men and 2 women in line for the men’s restroom.

Put numbers to our dilemma

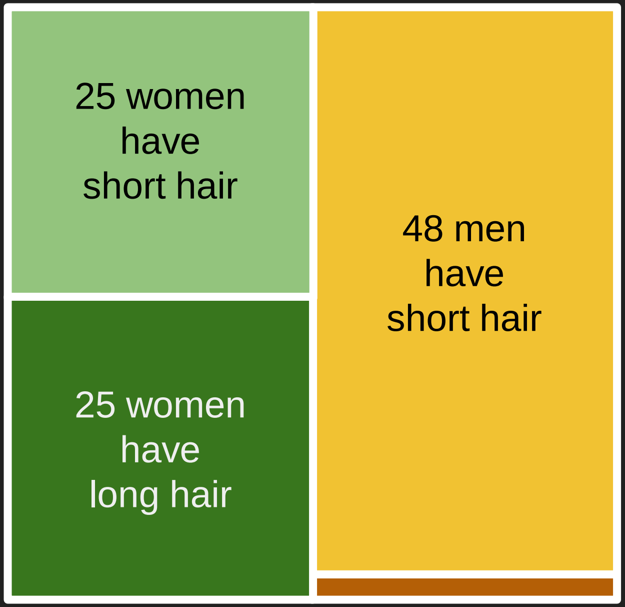


In the line, 4 times more men have long hair than women.

Out of 100 people
at the movies

50 are women

50 are men



25 women have short hair

25 women have long hair

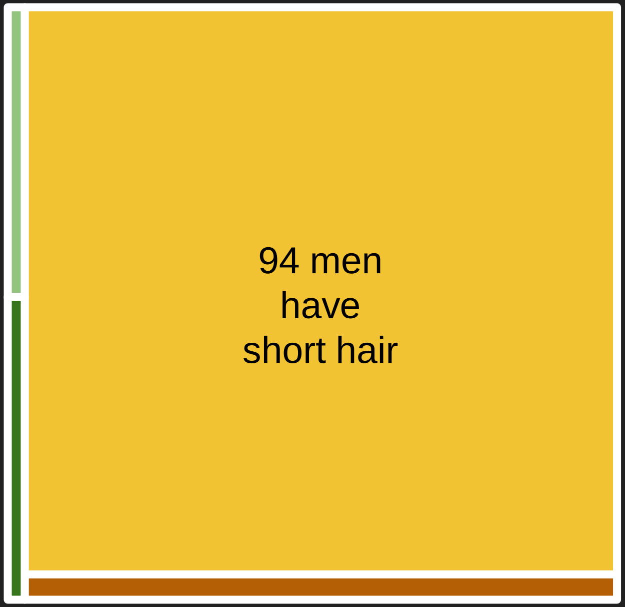
48 men have short hair

2 men have long hair

Out of 100 people
In line for the
men’s restroom

2 are
women

98 are men



One
woman
has
short
hair

One
woman
has
long
hair

94 men
have
short hair

4 men have long hair

Translate to math

P(something) = # something / # everything

P(woman) = Probability that a person is a woman

= # women / # people

= 50 / 100 = **.5**

P(man) = Probability that a person is a man

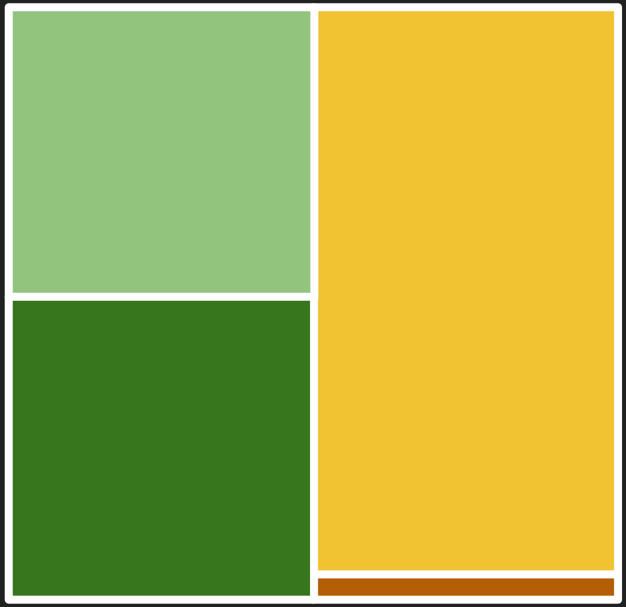
= # men / # people

= 50 / 100 = **.5**

Out of 100 people
at the movies

50 are women

50 are men



Translate to math

P(something) = # something / # everything

P(woman) = Probability that a person is a woman

= # women / # people

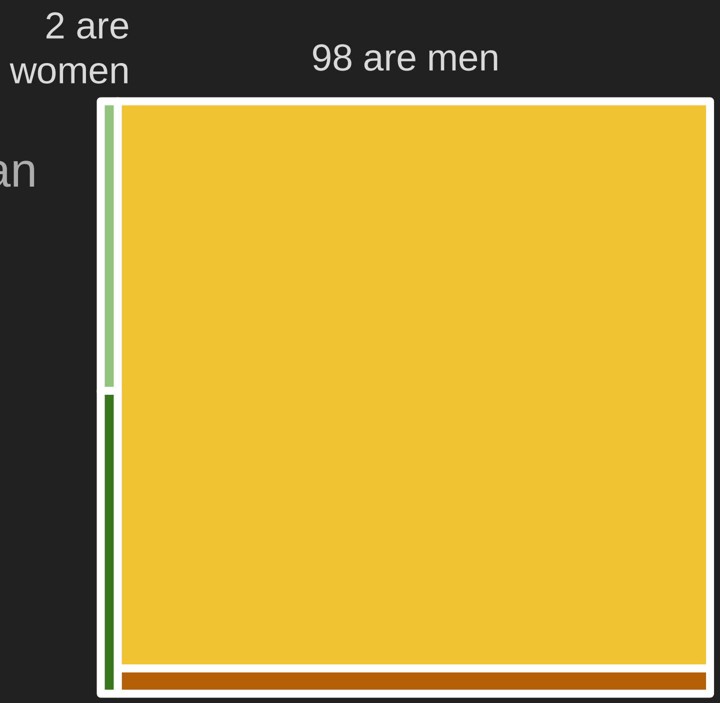
= 2 / 100 = **.02**

P(man) = Probability that a person is a man

= # men / # people

= 98 / 100 = **.98**

Out of 100 people
In line for the
men's restroom



Conditional probabilities

P(long hair | woman)

If I know that a person is a woman, what is the probability that person has long hair?

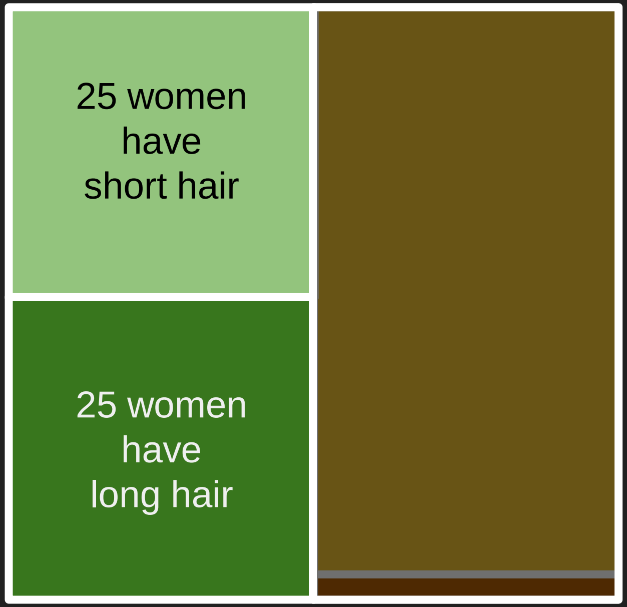
P(long hair | woman)

= # women with long hair / # women

= 25 / 50 = **.5**

Out of 100 people
at the movies

50 are women



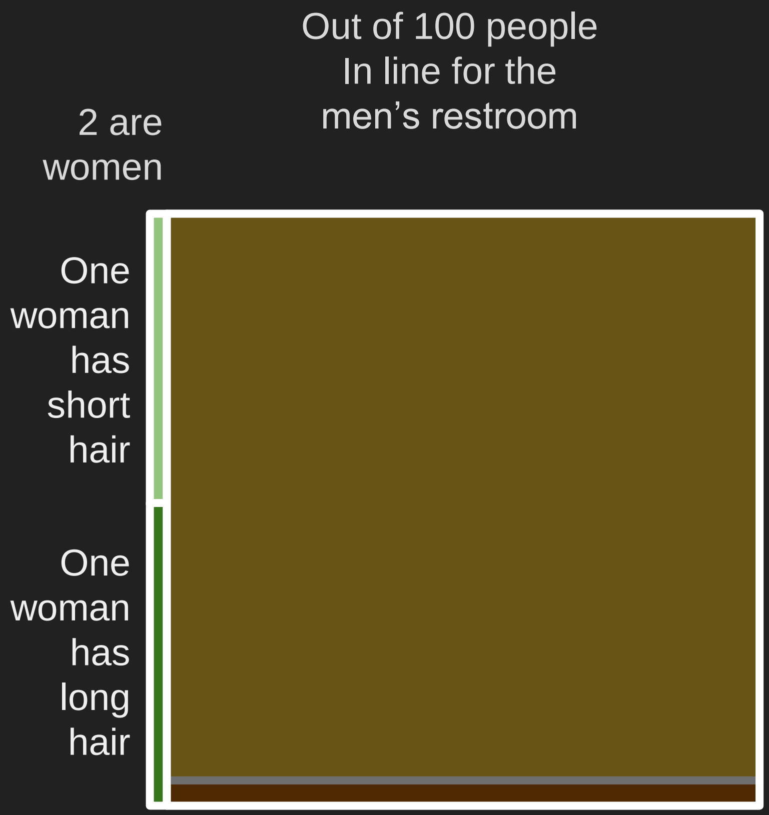
Conditional probabilities

This doesn’t change when we consider people in line.

P(long hair | woman)

= # women with long hair / # women

= 1 / 2 = **.5**



Conditional probabilities

If I know that a person is a man, what is the probability that person has long hair?

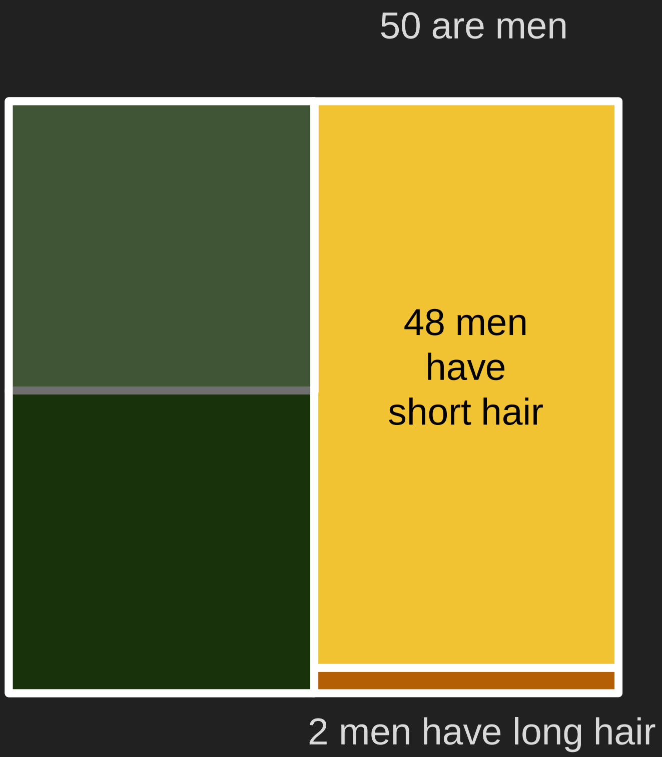
P(long hair | man)

= # men with long hair / # men

= 2 / 50 = **.04**

Whether in line or not.

Out of 100 people
at the movies



Conditional probabilities

P(A | B) is the probability of A, given B.

“If I know B is the case, what is the probability that A is also the case?”

P(A | B) is not the same as P(B | A).

P(cute | puppy) is not the same as P(puppy | cute)

If I know the thing I’m holding is a puppy, what is the probability that it is cute?

If I know the the thing I’m holding is cute, what is the probability that it is a puppy?

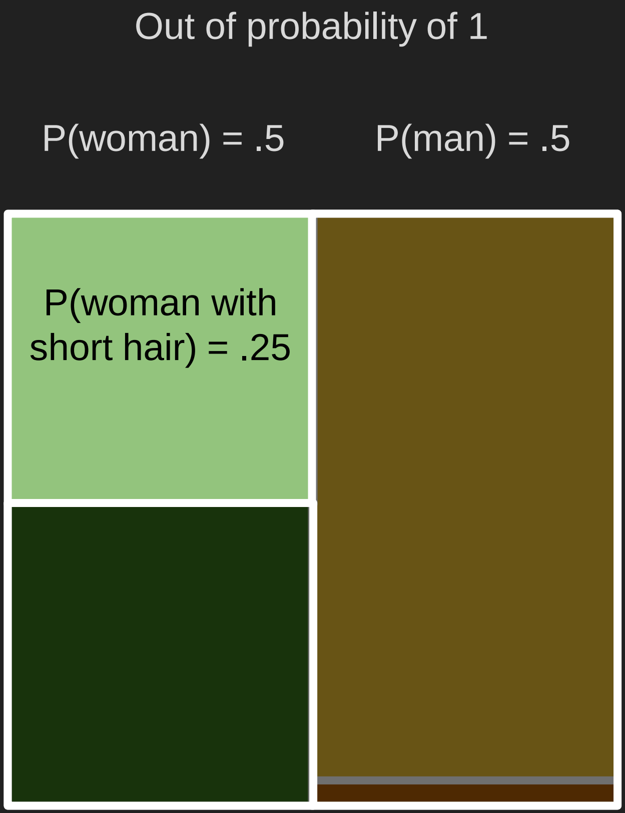
Joint probabilities

What is the probability that a person is both a woman and has short hair?

P(woman with short hair)

= P(woman) * P(short hair | woman)

= .5 * .5 = **.25**



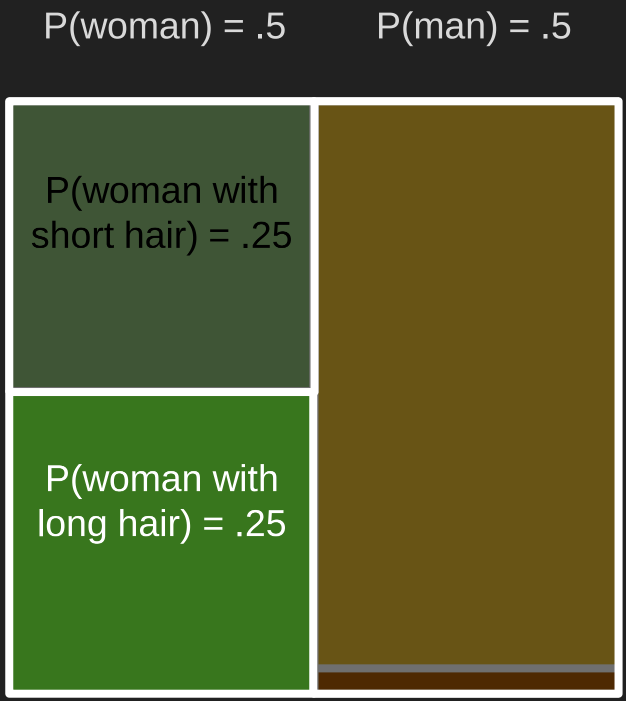
Joint probabilities

P(woman with long hair)

= P(woman) * P(long hair | woman)

= .5 * .5 = **.25**

Out of probability of 1



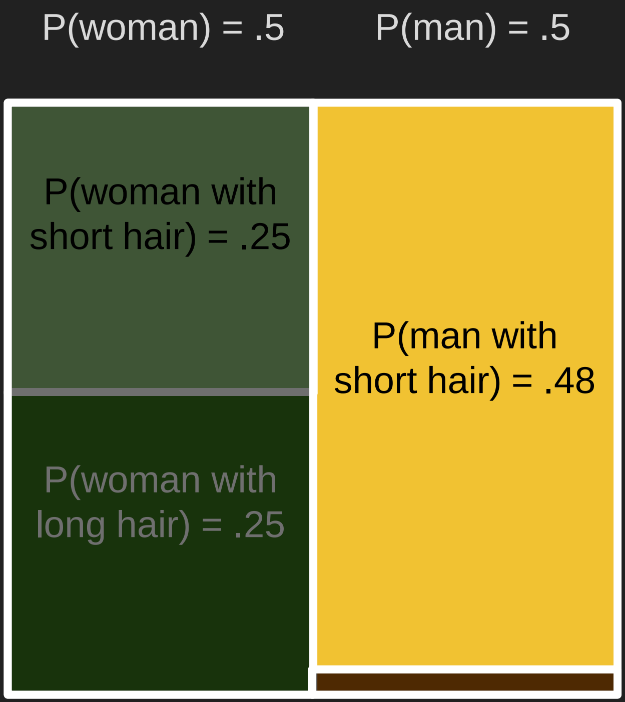
Joint probabilities

P(man with short hair)

= P(man) * P(short hair | man)

= .5 * .96 = **.48**

Out of probability of 1

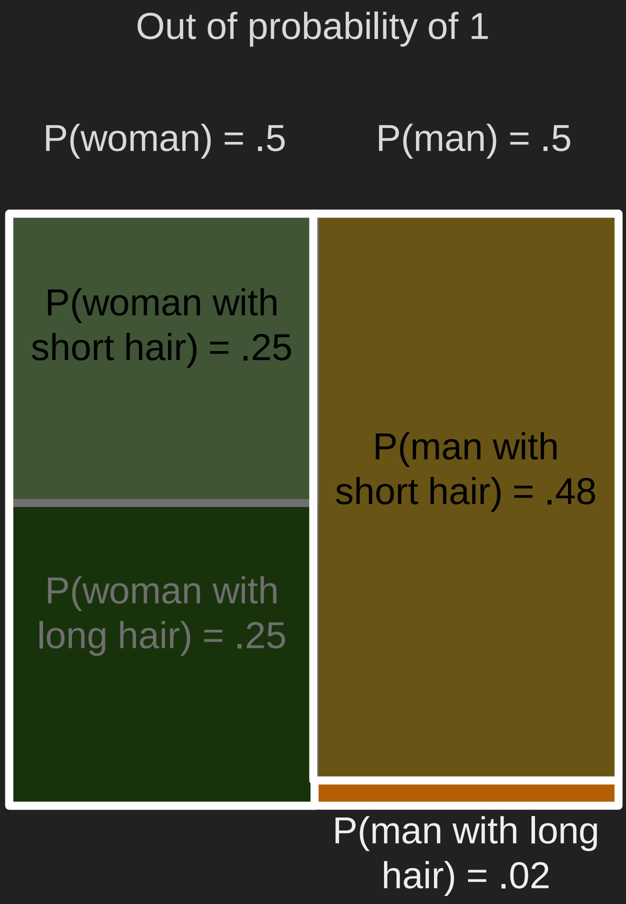

P(woman) = .5
P(man) = .5
P(woman with short hair) = .25
P(woman with long hair) = .25
P(man with short hair) = .48

Joint probabilities

P(man with long hair)

= P(man) * P(long hair | man)

= .5 * .04 = **.02**


Out of probability of 1
P(woman) = .5
P(man) = .5
P(woman with short hair) = .25
P(woman with long hair) = .25
P(man with short hair) = .48
P(man with long hair) = .02

Joint probabilities

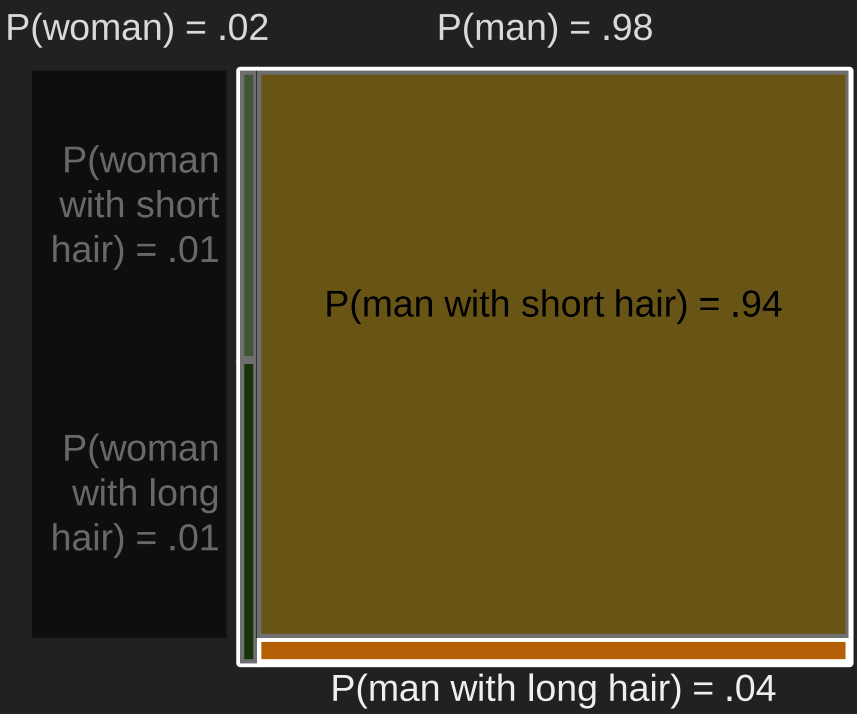
If P(man) = .98 and P(woman) = .02, then the answers change.

P(man with long hair)

= P(man) * P(long hair | man)

= .98 * .04 = **.04**

Out of probability of 1


P(woman) = .02
P(man) = .98
P(woman with short hair) = .01
P(woman with long hair) = .01
P(man with short hair) = .94
P(man with long hair) = .04

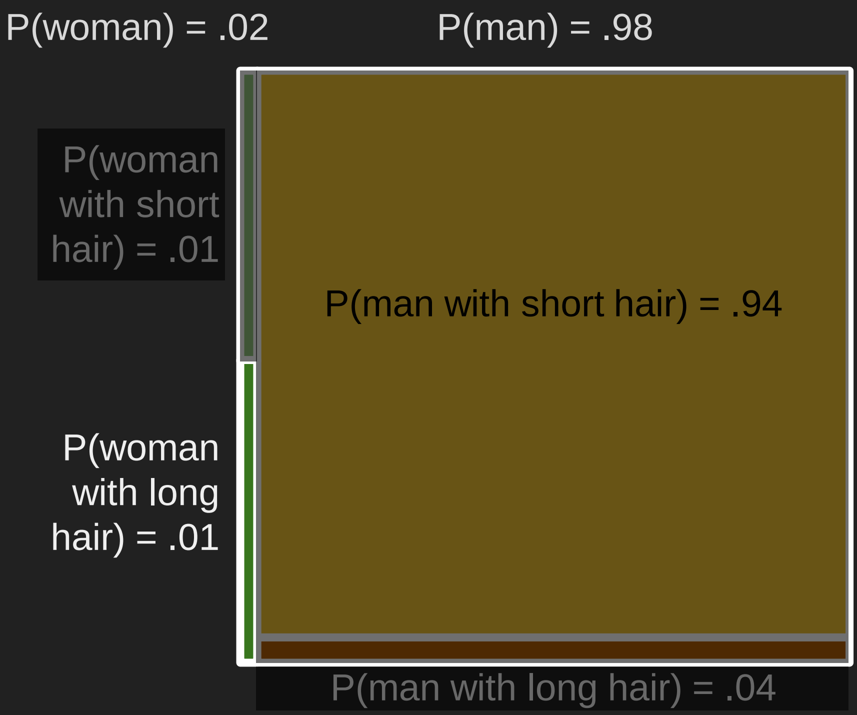
Joint probabilities

P(woman with long hair)

= P(woman) * P(long hair | woman)

= .02 * .5 = **.01**

Out of probability of 1


P(woman) = .02
P(man) = .98
P(woman with short hair) = .01
P(man with short hair) = .94
P(woman with long hair) = .01
P(man with long hair) = .04

Joint probabilities

P(A and B) is the probability that both A and B are the case.

Also written P(A, B) or P(A ∩ B)

P(A and B) is the same as P(B and A)

The probability that I am having a jelly donut with my milk is the same as the probability that I am having milk with my jelly donut.

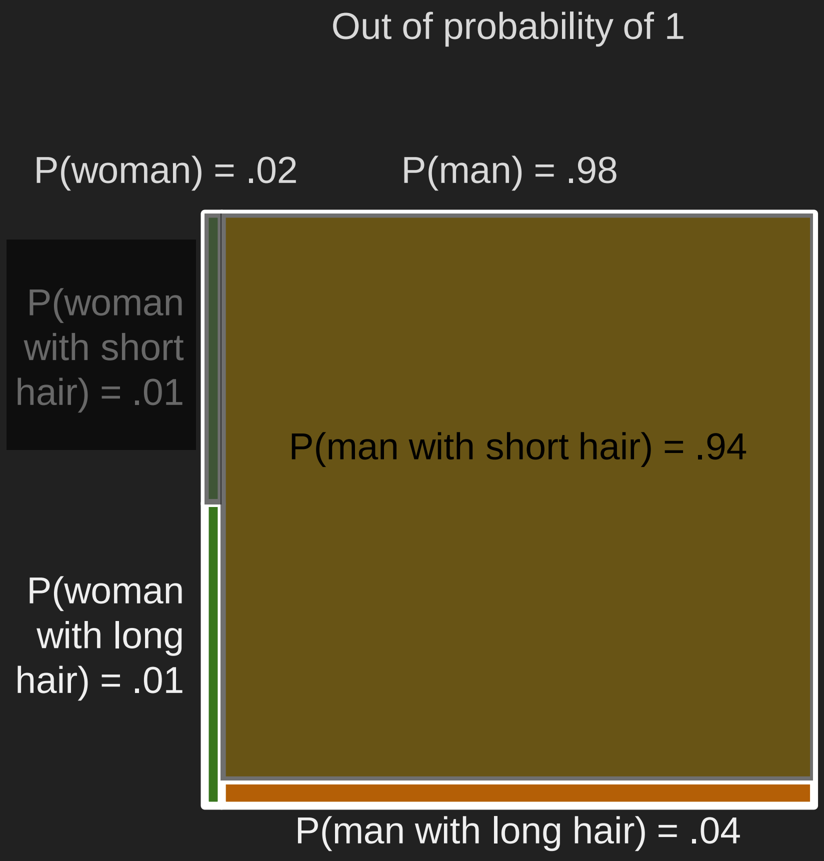
P(donut and milk) = P(milk and donut)

Marginal probabilities

P(long hair) = P(woman with long hair) +

P(man with long hair)

= .01 + .04 = **.05**

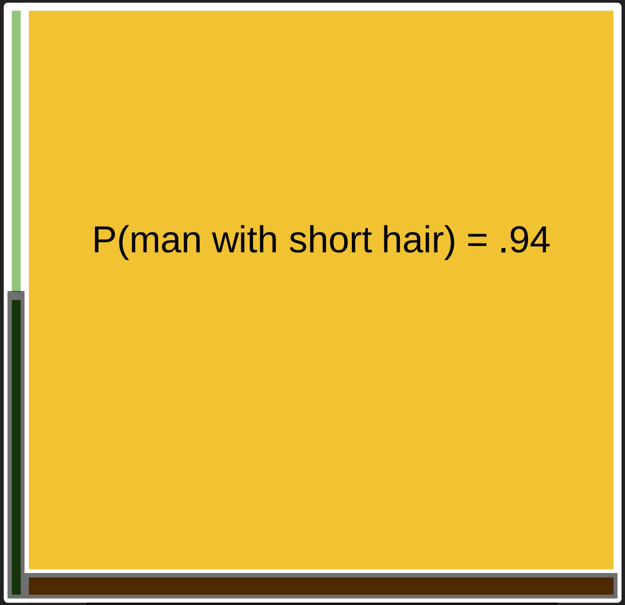

Out of probability of 1
P(woman) = .02
P(man) = .98
P(woman with short hair) = .01
P(man with short hair) = .94
P(woman with long hair) = .01
P(man with long hair) = .04

Marginal probabilities

P(short hair) = P(woman with short hair) + P(man with short hair)

= .01 + .94 = **.95**

Out of probability of 1



P(woman) = .02

P(man) = .98

P(woman with short hair) = .01

P(man with short hair) = .94

P(woman with long hair) = .01

P(man with long hair) = .04

What we really care about

We know the person has long hair.
Are they a man or a woman?

P(man | long hair)

We don’t know this answer yet.

Thomas Bayes noticed something cool

P(man with long hair) = P(long hair) * P(man | long hair)

Thomas Bayes noticed something cool

P(man with long hair) = P(long hair) * P(man | long hair)

P(long hair and man) = P(man) * P(long hair | man)

Thomas Bayes noticed something cool

P(man with long hair) = P(long hair) * P(man | long hair)

P(long hair and man) = P(man) * P(long hair | man)

Because P(man and long hair) = P(long hair and man)

Thomas Bayes noticed something cool

P(man with long hair) = P(long hair) * P(man | long hair)

P(long hair and man) = P(man) * P(long hair | man)

Because P(man and long hair) = P(long hair and man)

P(long hair) * P(man | long hair) = P(man) * P(long hair | man)

Thomas Bayes noticed something cool

P(man with long hair) = P(long hair) * P(man | long hair)

P(long hair and man) = P(man) * P(long hair | man)

Because P(man and long hair) = P(long hair and man)

P(long hair) * P(man | long hair) = P(man) * P(long hair | man)

P(man | long hair) = P(man) * P(long hair | man) / P(long hair)

Thomas Bayes noticed something cool

P(man with long hair) = P(long hair) * P(man | long hair)

P(long hair and man) = P(man) * P(long hair | man)

Because P(man and long hair) = P(long hair and man)

P(long hair) * P(man | long hair) = P(man) * P(long hair | man)

P(man | long hair) = P(man) * P(long hair | man) / P(long hair)

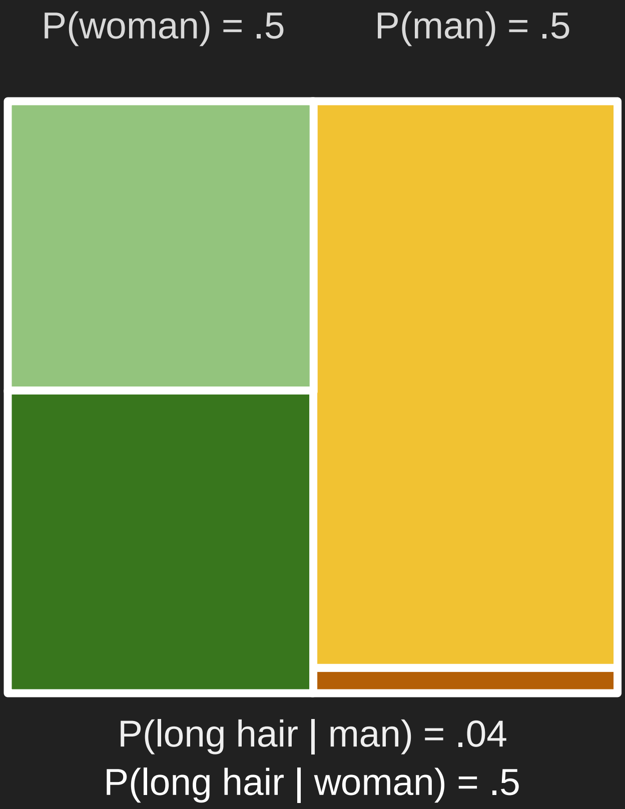
P(A | B) = P(B | A) * P(A) / P(B)

Bayes’ Theorem

$$P(A \mid B) = \frac{P(B \mid A)\ P(A)}{P(B)}$$

Back to the movie theater, this time with Bayes

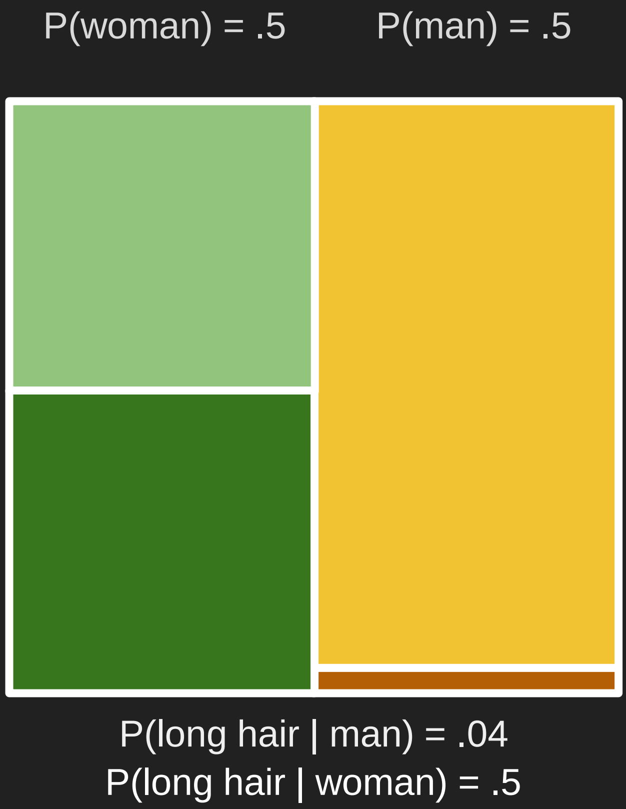
$$P(\text{man} \mid \text{long hair}) = \frac{P(\text{man}) * P(\text{long hair} \mid \text{man})}{P(\text{long hair})}$$



Back to the movie theater, this time with Bayes

$$P(\text{man} \mid \text{long hair}) = \frac{P(\text{man}) * P(\text{long hair} \mid \text{man})}{P(\text{long hair})}$$

$$= \frac{P(\text{man}) * P(\text{long hair} \mid \text{man})}{P(\text{woman with long hair}) + P(\text{man with long hair})}$$



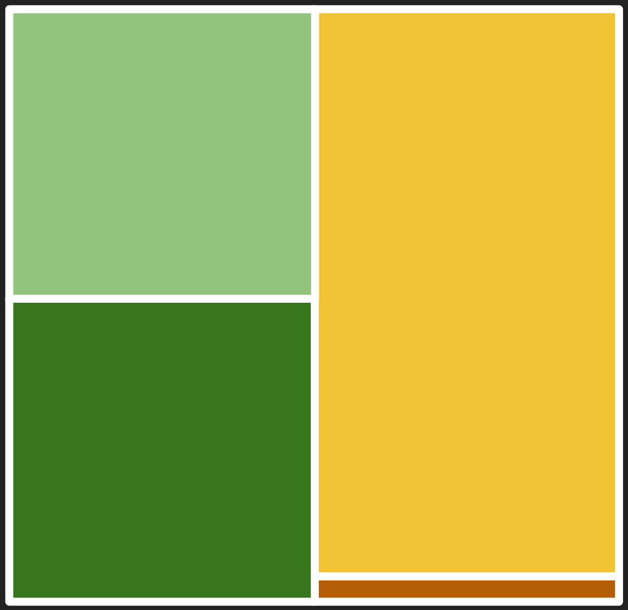
Back to the movie theater, this time with Bayes

$$P(\text{man} \mid \text{long hair}) = \frac{P(\text{man}) * P(\text{long hair} \mid \text{man})}{P(\text{long hair})}$$

$$= \frac{P(\text{man}) * P(\text{long hair} \mid \text{man})}{P(\text{woman with long hair}) + P(\text{man with long hair})}$$

$$P(\text{man} \mid \text{long hair}) = \frac{.5 * .04}{.25 + .02} = .02 / .27 = \boxed{.07}$$

P(woman) = .5 P(man) = .5



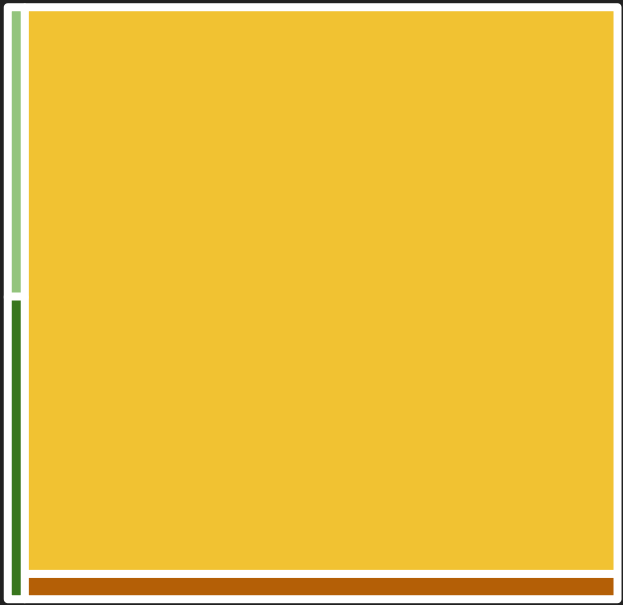
P(long hair | man) = .04
P(long hair | woman) = .5

Back to the movie theater, this time with Bayes

$$P(\text{man} \mid \text{long hair}) = \frac{P(\text{man}) * P(\text{long hair} \mid \text{man})}{P(\text{long hair})}$$

$$= \frac{P(\text{man}) * P(\text{long hair} \mid \text{man})}{P(\text{woman with long hair}) + P(\text{man with long hair})}$$

P(woman) = .02 P(man) = .98



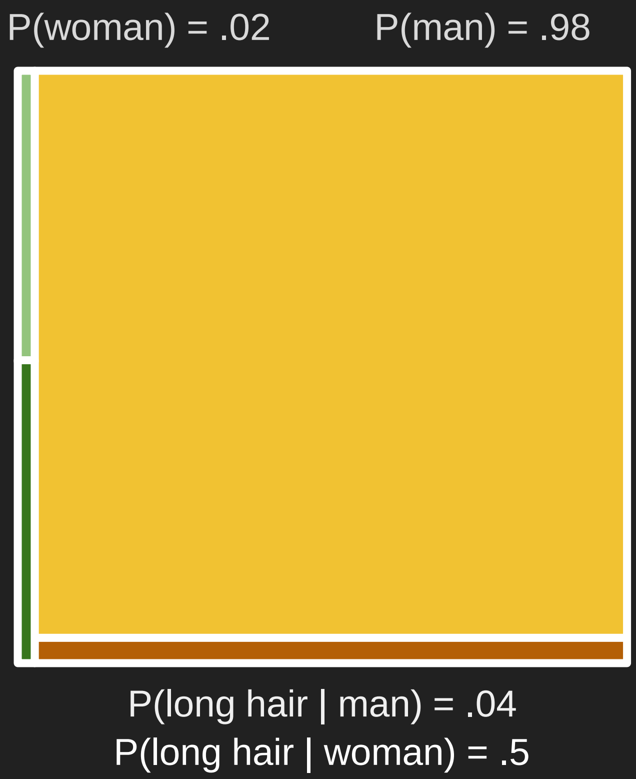
P(long hair | man) = .04
P(long hair | woman) = .5

Back to the movie theater, this time with Bayes

$$P(\text{man} \mid \text{long hair}) = \frac{P(\text{man}) * P(\text{long hair} \mid \text{man})}{P(\text{long hair})}$$

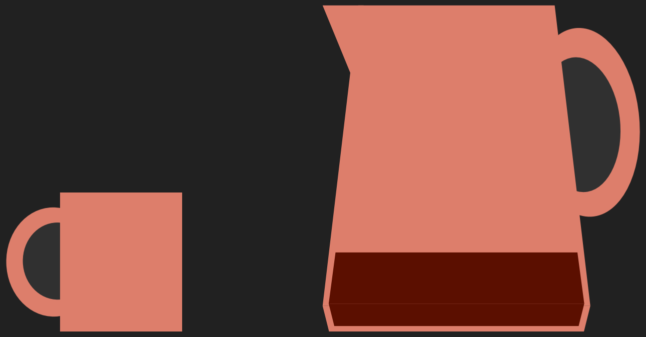
$$= \frac{P(\text{man}) * P(\text{long hair} \mid \text{man})}{P(\text{woman with long hair}) + P(\text{man with long hair})}$$

$$P(\text{man} \mid \text{long hair}) = \frac{.98 * .04}{.01 + .04} = .04 / .05 = \boxed{.80}$$



Probability distributions

Probability is like a pot with just one cup of coffee left in it.

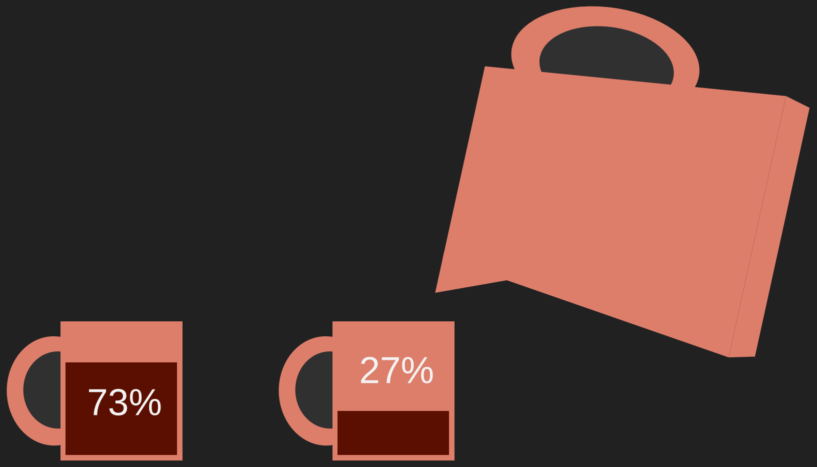


Probability distributions

If you only have one cup, you can fill it completely.

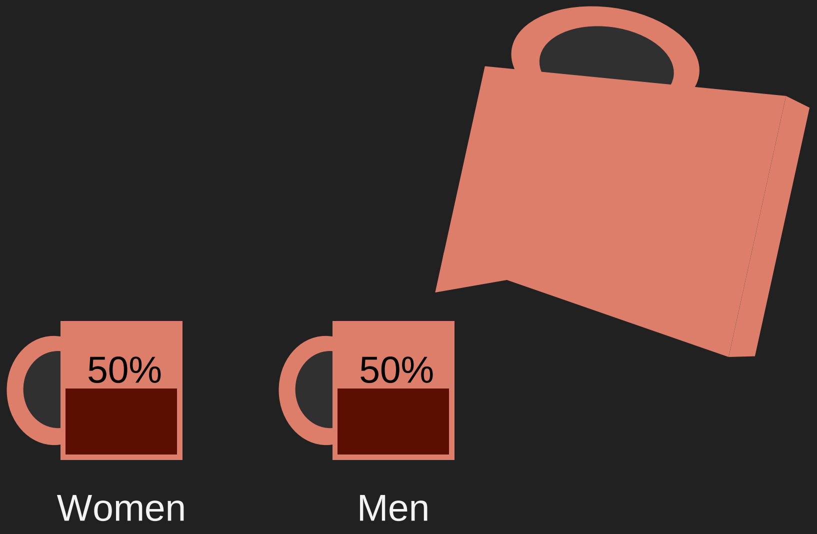
Probability distributions

If you have two cups, you have to decide how to share (distribute) it.



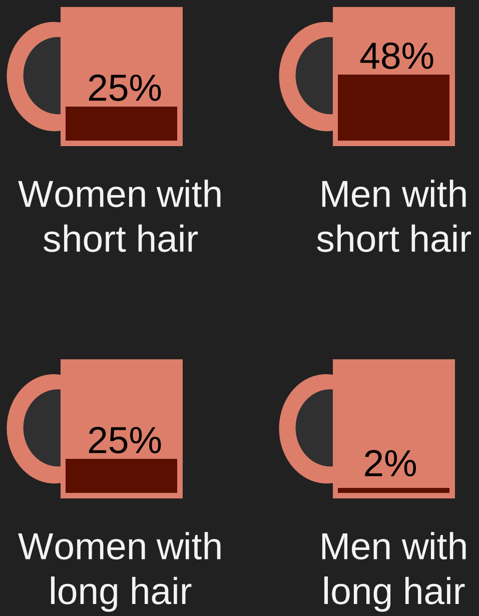
Probability distributions

Our people are distributed between two groups, women and men.



Probability distributions

We can distribute them more.



25%

Women with short hair

48%

Men with short hair

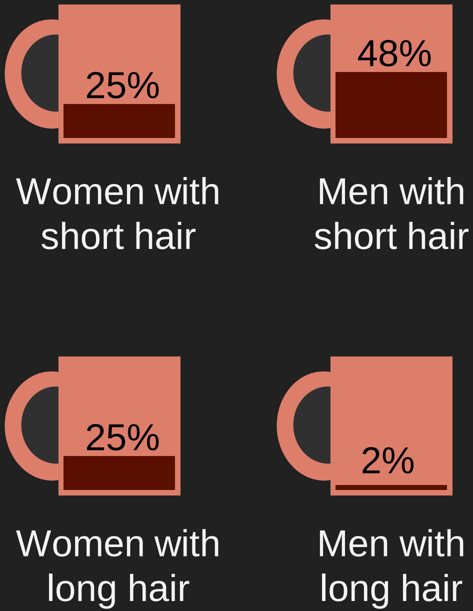
25%

Women with long hair

2%

Men with long hair

Probability distributions



25%

Women with short hair

48%

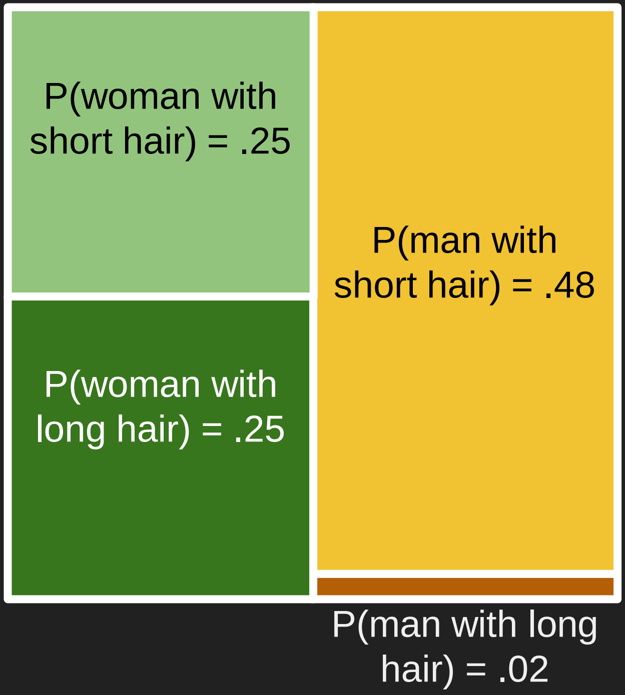
Men with short hair

25%

Women with long hair

2%

Men with long hair



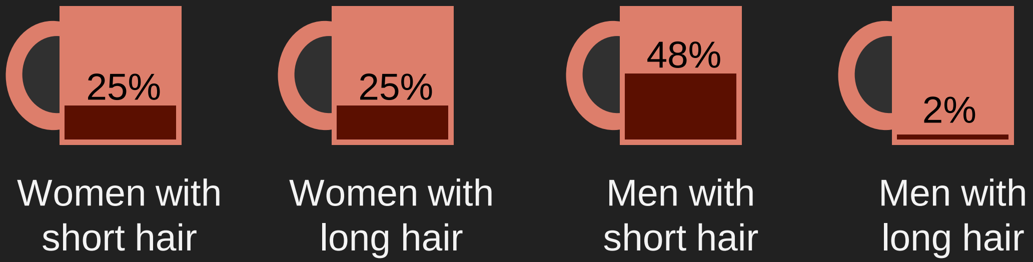
P(woman with short hair) = .25

P(man with short hair) = .48

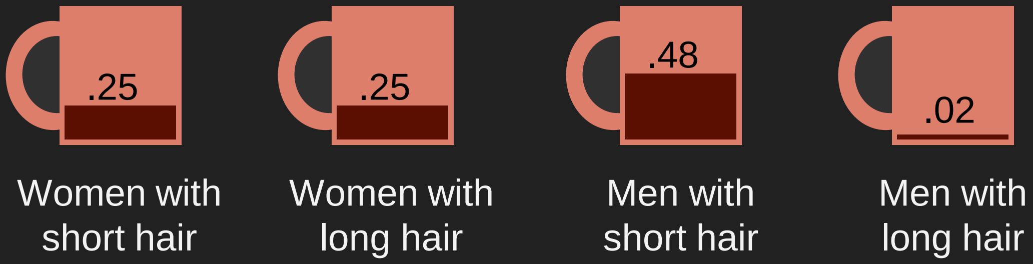
P(woman with long hair) = .25

P(man with long hair) = .02

Probability distributions

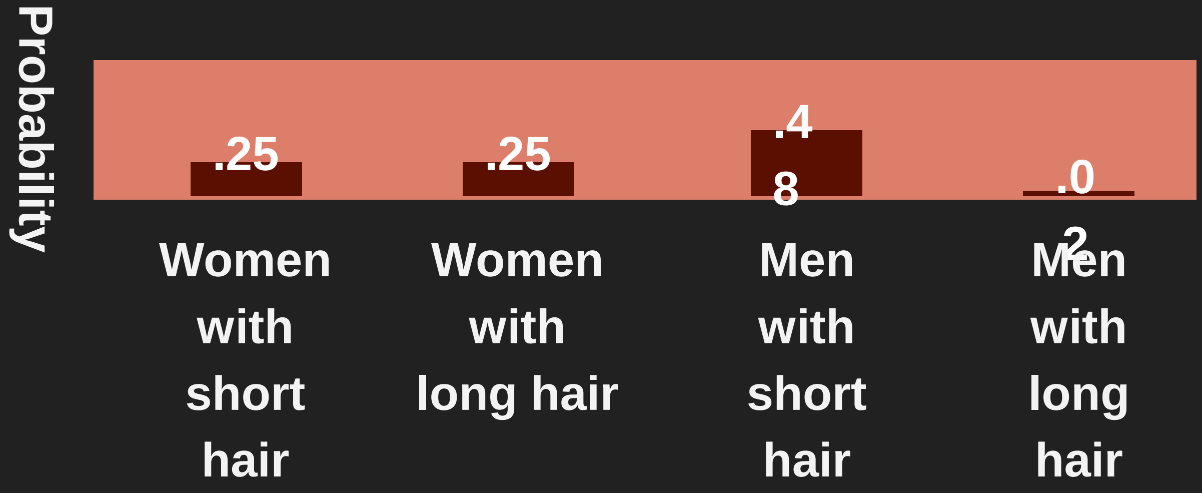


Probability distributions



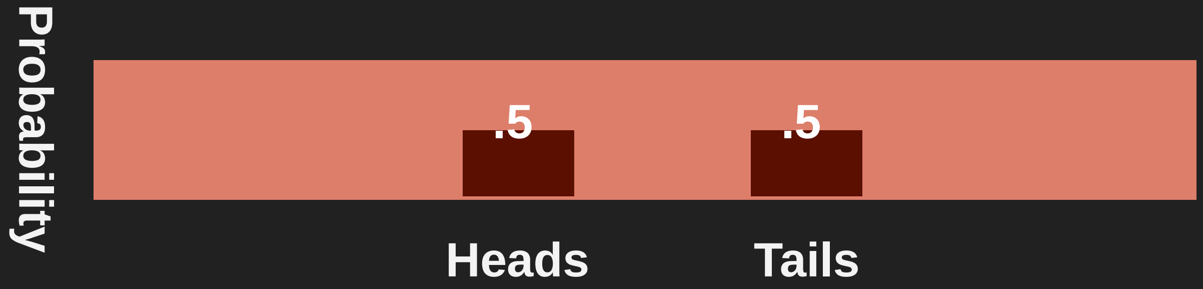
Probability distributions

It’s helpful to think of probabilities as beliefs



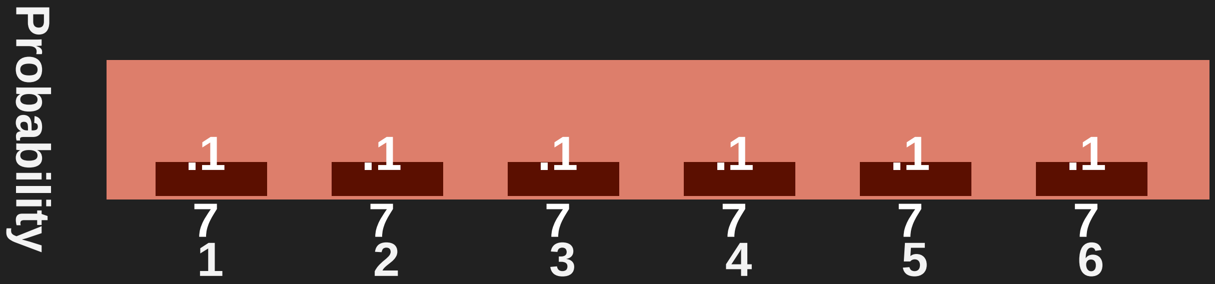
Probability distributions

Flipping a fair coin



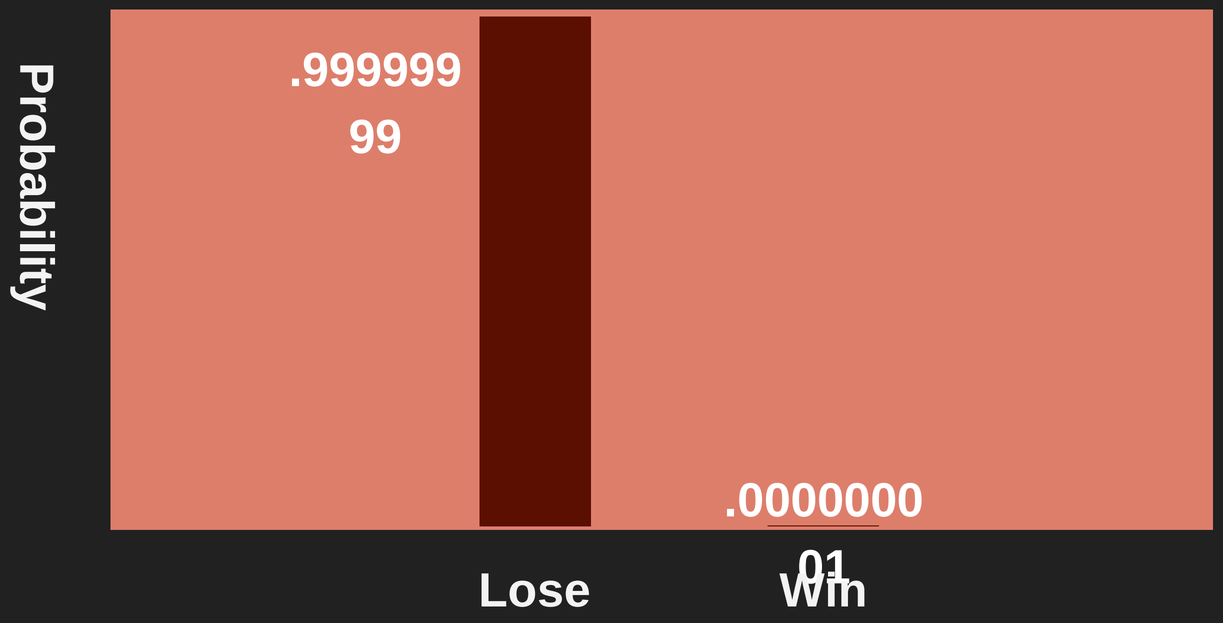
Probability distributions

Rolling a fair die



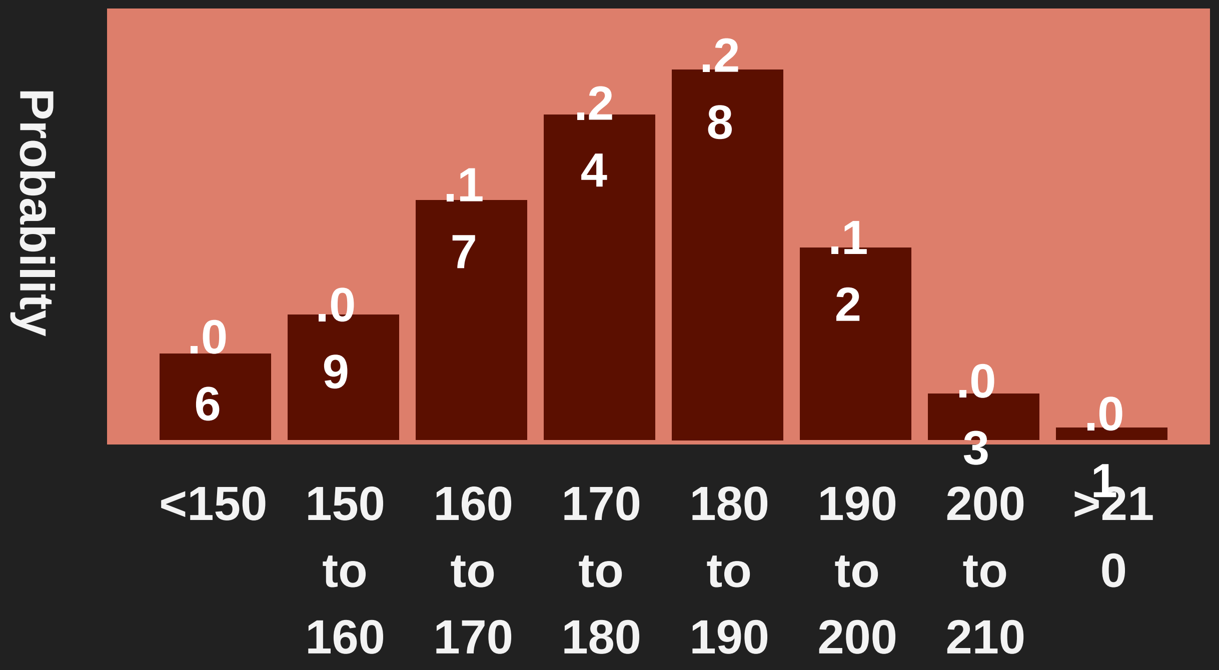
Probability distributions

Playing for the Powerball jackpot



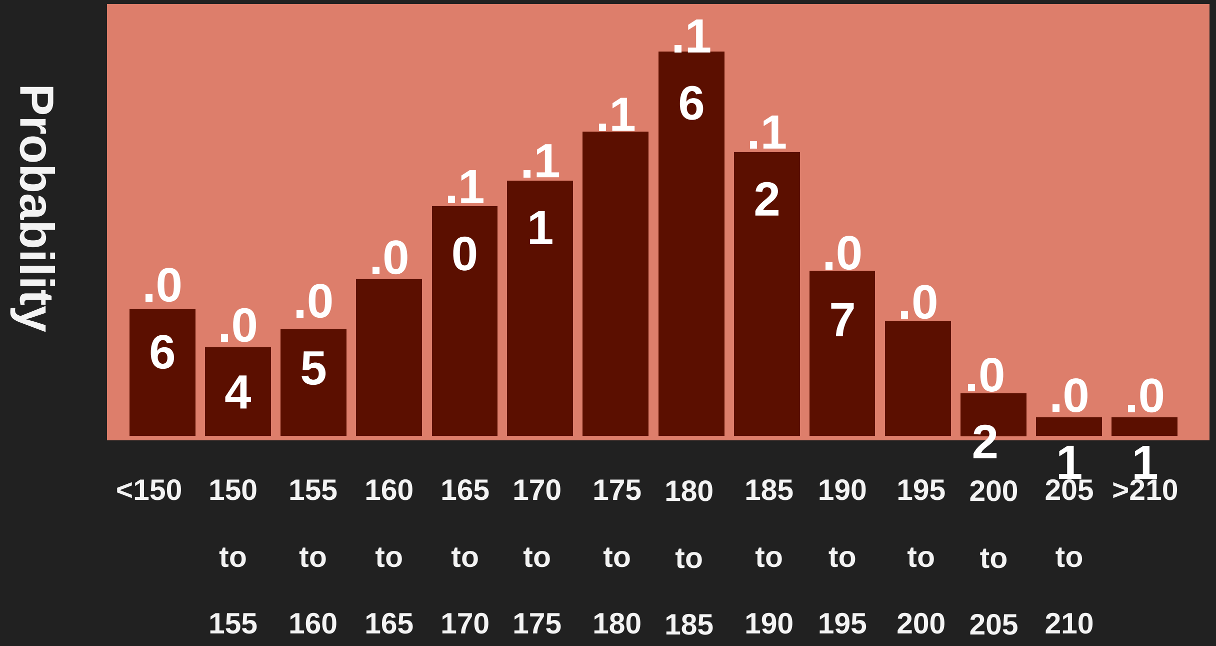
Probability distributions

Height of adults in cm



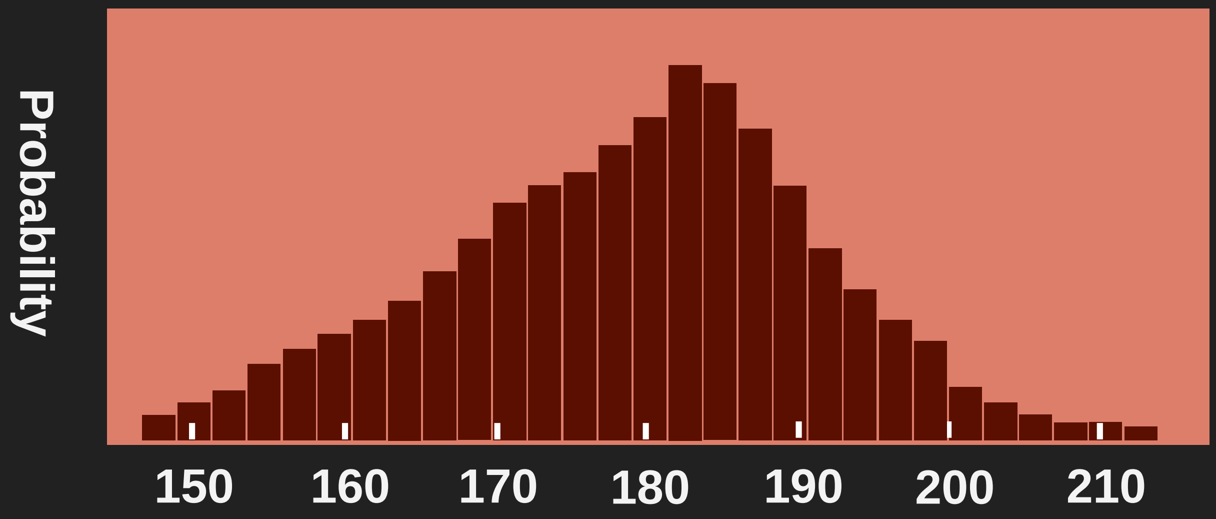
Probability distributions

Height of adults in cm



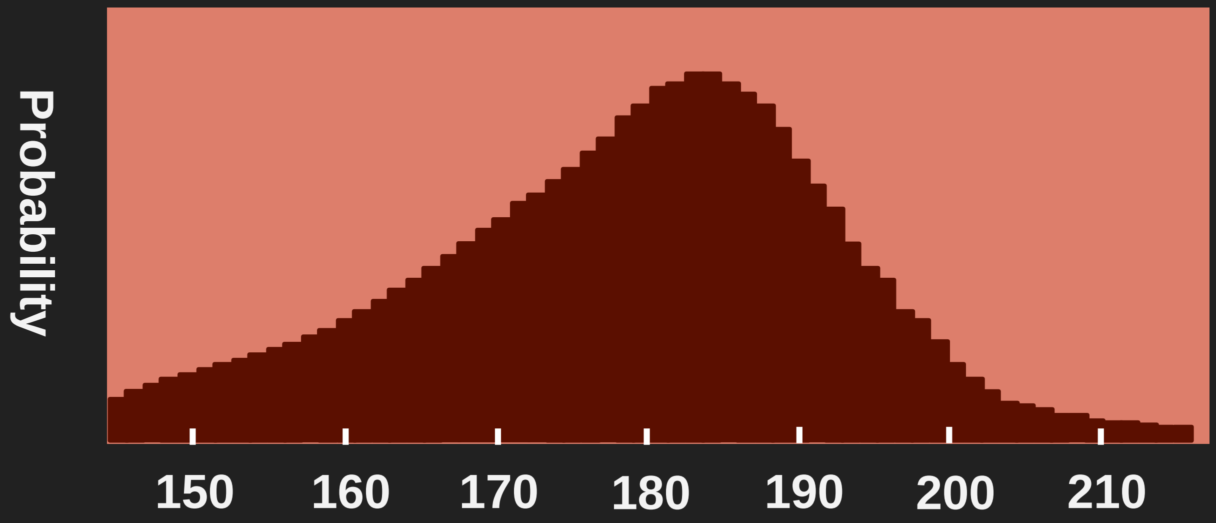
Probability distributions

Height of adults in cm



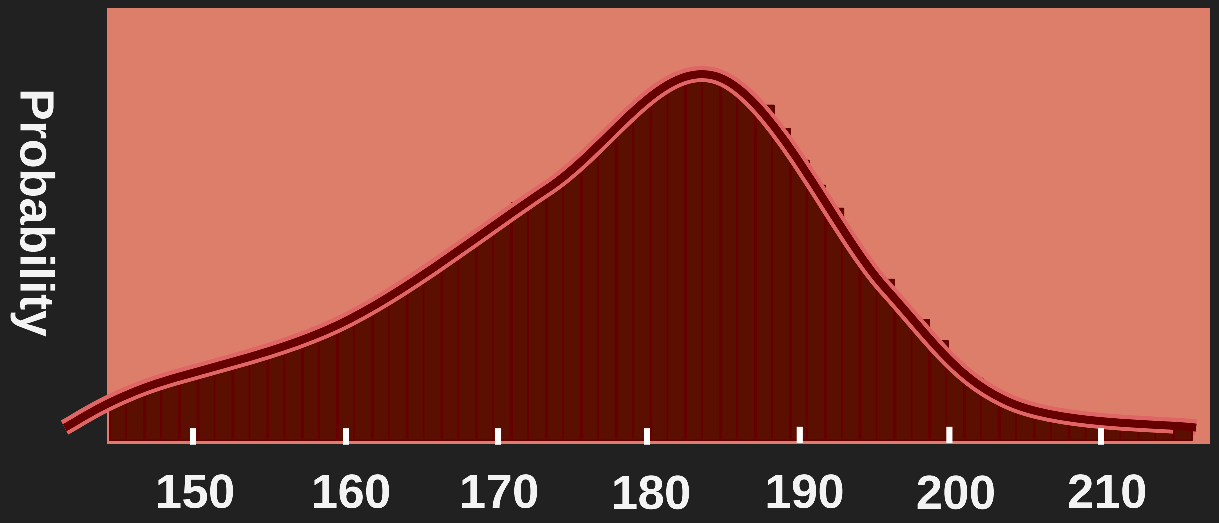
Probability distributions

Height of adults in cm



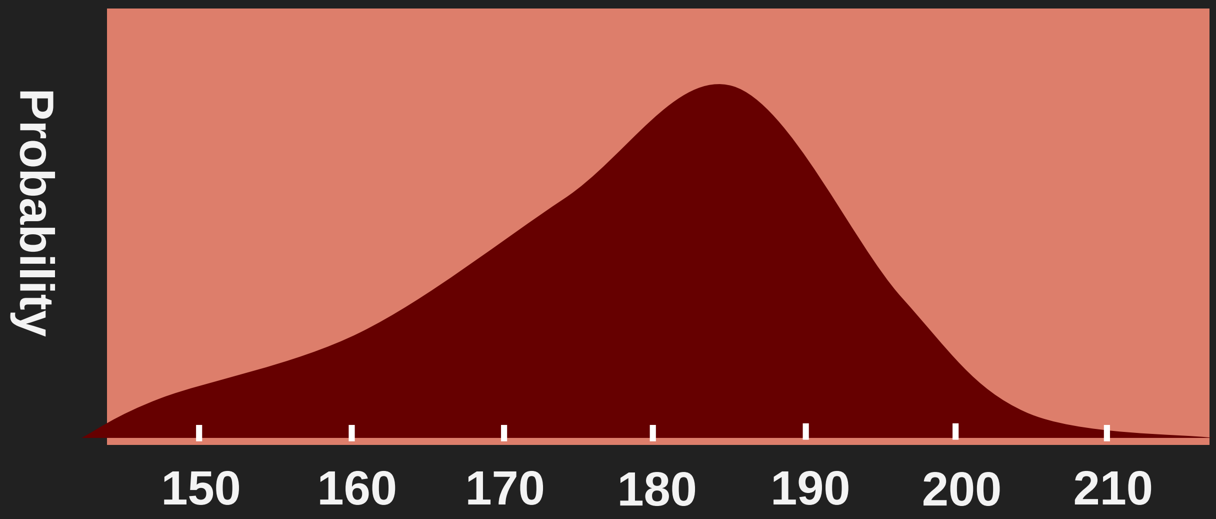
Probability distributions

Height of adults in cm



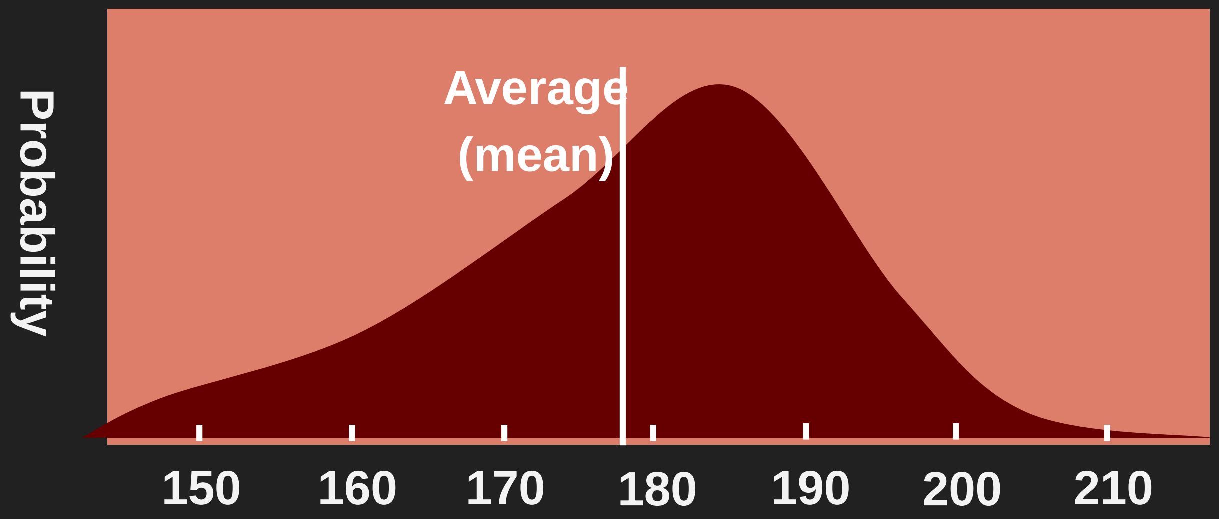
Probability distributions

Height of adults in cm



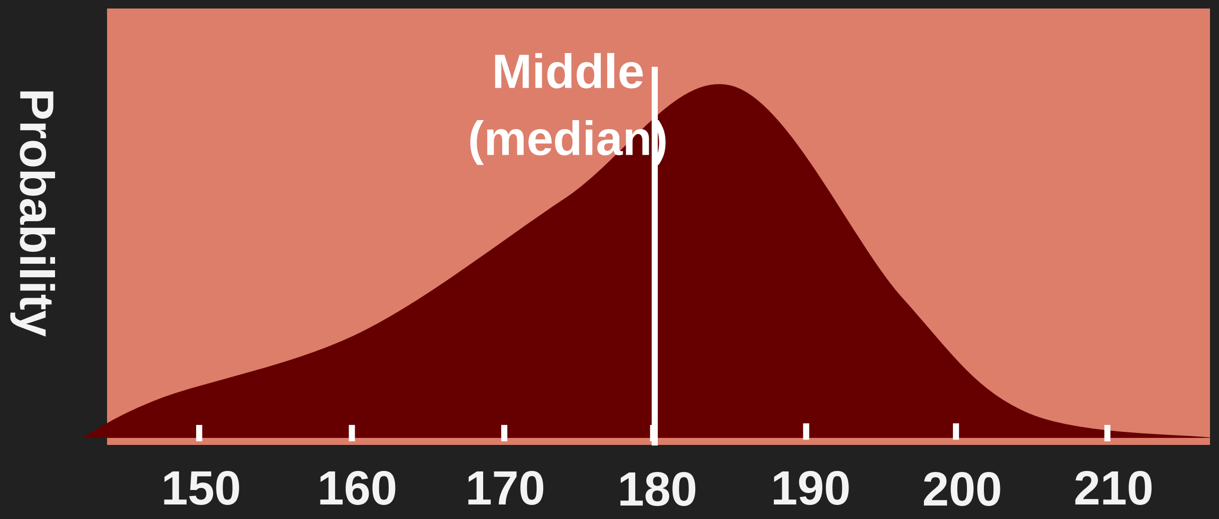
Probability distributions

Height of adults in cm



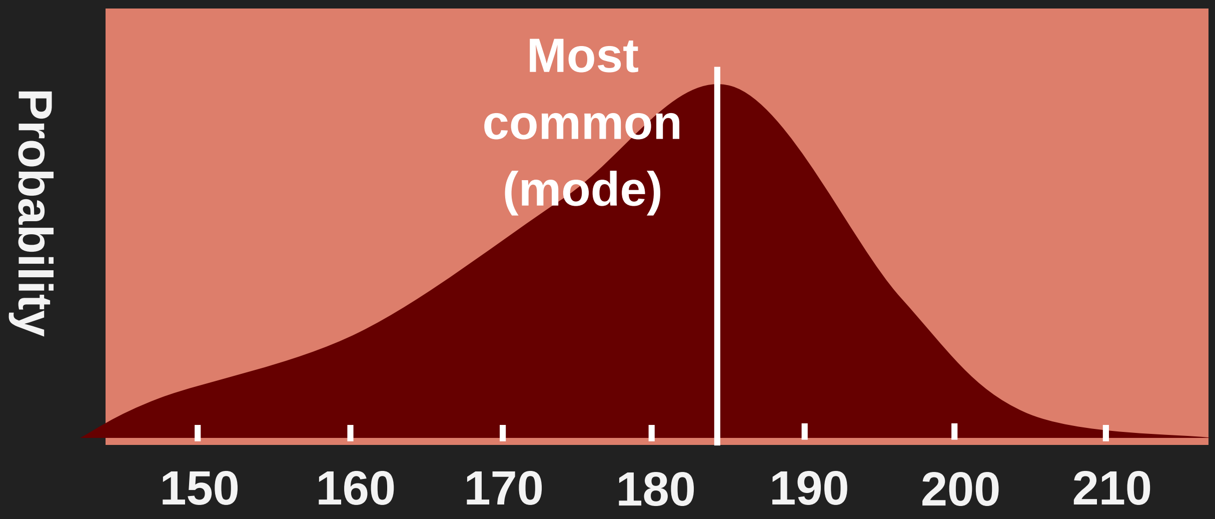
Probability distributions

Height of adults in cm



Probability distributions

Height of adults in cm



Bayes’ Theorem

$$P(w \mid m) = \frac{P(m \mid w)\ P(w)}{P(m)}$$

Bayes’ Theorem

$$P(w \mid m) = \frac{P(m \mid w)\ \overset{\textbf{prior}}{\boxed{P(w)}}}{P(m)}$$

Bayes’ Theorem

$$P(w \mid m) = \frac{\overset{\textbf{likelihood}}{\boxed{P(m \mid w)}}\; P(w)}{P(m)}$$

Bayes’ Theorem

posterior

$$\boxed{P(w \mid m)} = \frac{P(m \mid w)\ P(w)}{P(m)}$$

Bayes' Theorem

$$P(w \mid m) = \frac{P(m \mid w)\ \ P(w)}{\boxed{P(m)}}$$

marginal likelihood

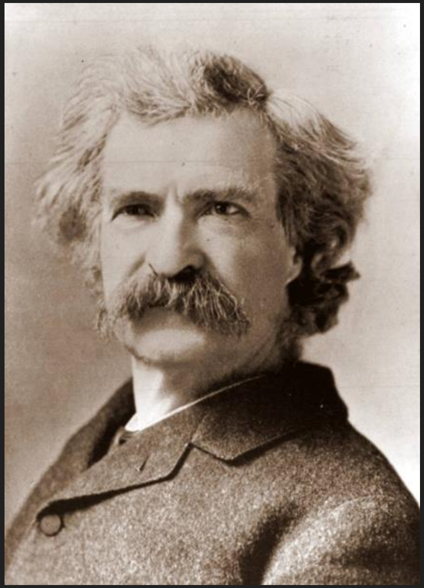
Why Bayesian inference makes us nervous

We’re not always aware of what we believe.

Putting what we believe into a distribution correctly is tricky.

We want to be able to be surprised by our data.
Inaccurate beliefs can make it hard or impossible to learn.

> “It ain't what you don't know that gets you into trouble.
> It's what you know for sure that just ain't so.”
>
> - Mark Twain



Believe the impossible, at least a little bit

Leave room for believing the unlikely. Leave a non-zero probability unless you are absolutely certain.

> “When you have excluded the impossible, whatever remains, however improbable, must be the truth”
>
> - Sherlock Holmes (Sir Arthur Conan Doyle)

Believe the impossible, at least a little bit

"Alice laughed: "There's no use trying," she said; "one can't believe impossible things."

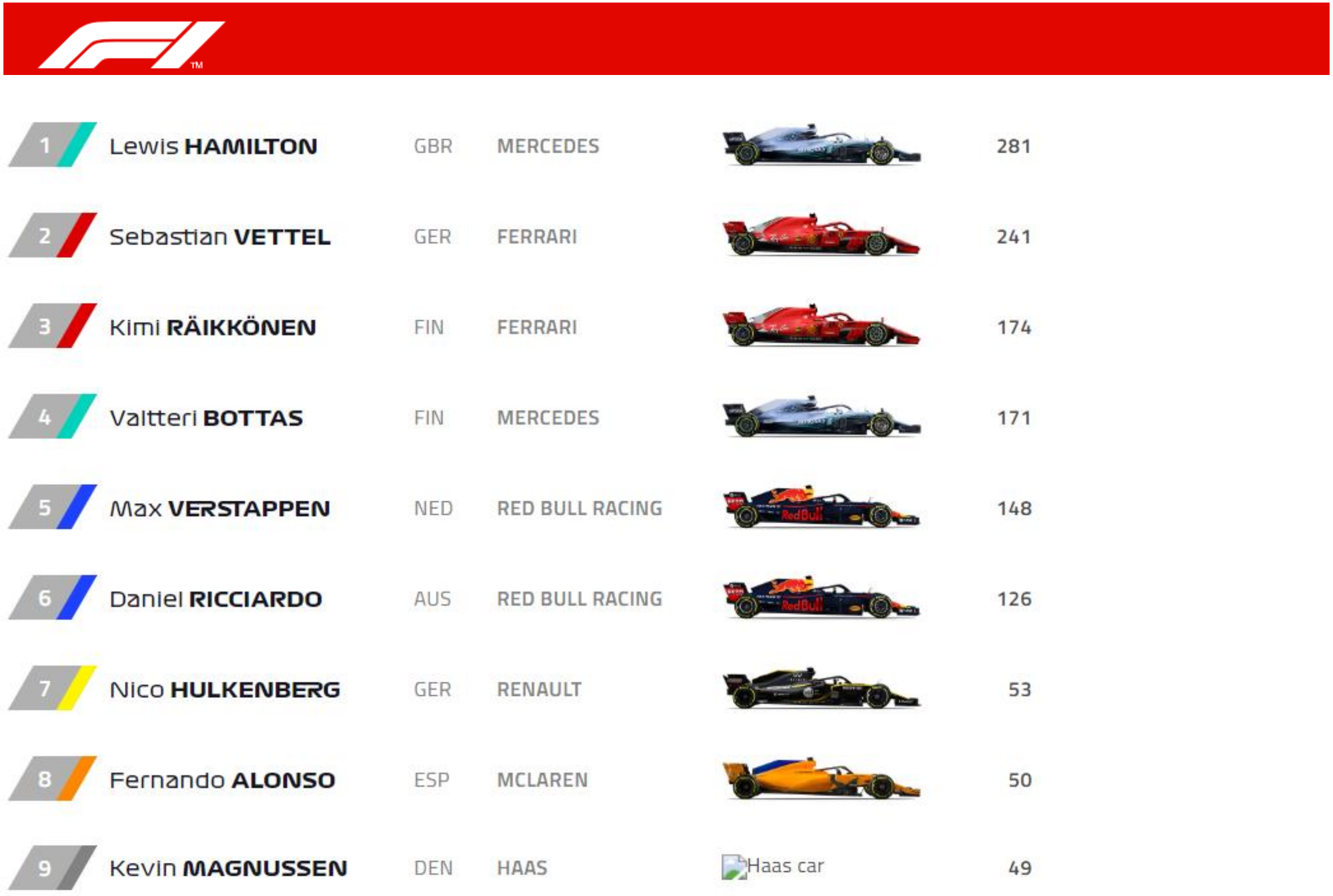
"I daresay you haven't had much practice," said the Queen. "When I was younger, I always did it for half an hour a day. Why, sometimes I've believed as many as six impossible things before breakfast."

- Lewis Carroll (Alice’s Adventures in Wonderland)

Part 4: Lab

FORMULA 1

Formula 1



1	Lewis **HAMILTON**	GBR	MERCEDES		281
2	Sebastian **VETTEL**	GER	FERRARI		241
3	Kimi **RÄIKKÖNEN**	FIN	FERRARI		174
4	Valtteri **BOTTAS**	FIN	MERCEDES		171
5	Max **VERSTAPPEN**	NED	RED BULL RACING		148
6	Daniel **RICCIARDO**	AUS	RED BULL RACING		126
7	Nico **HULKENBERG**	GER	RENAULT		53
8	Fernando **ALONSO**	ESP	MCLAREN		50
9	Kevin **MAGNUSSEN**	DEN	HAAS	Haas car	49

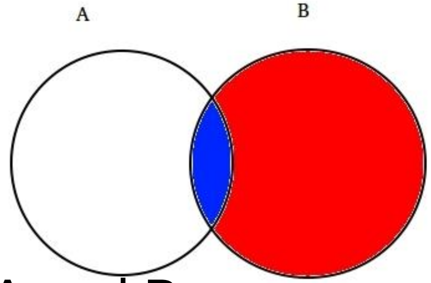
Another example: Formula 1

- Suppose, out of all the 4 championship races (F1) between Lewis Hamilton and Fernando Alonso
  - Lewis won 3 times while Fernando 1
- So, if you were to bet on the winner of next race, who would it be?

Informative `Prior`

- It rained once when Lewis won, and once when Fernando won and it is definite that it will rain on the next date
  - So, who would you bet your money on now ?

Conditional probability



- Assume two partially intersecting sets A and B
- We wish to calculate the probability of A given B has already happened
  - Represent the happening of event B by shading it with red
- Since B has happened, the part which now matters for A is the part shaded in blue
- **So, the probability of A given B =** $\frac{BlueArea}{RedArea + BlueArea}$
- Therefore, we can write the formula for event B given A has already occurred by

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- Also:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- And so:

$$P(A|B) = \frac{P(B|A) X P(A)}{P(B)}$$

Rainy F1

- Suppose, B be the *event of winning for Fernando*
- A be the *event of raining*
- P(A) =1/2, since it rained twice out of four days
- P(B) is 1/4, since Fernando won only one race out of four
- P(A|B)=1, since it rained every time when Fernando won
- Substituting the values in the conditional probability formula, we get:
  - P(B|A) = P(A|B)*P(B) / P(A) = 1 . ¼ . 2 = 1/2
- The probability is 50%, which is almost the double of 25% when rain was not taken into account!
- Further strengthened our belief of Fernando Alonso winning in the light of new *evidence* i.e rain
- Pretty amazing..

Bayes' Formula

- We have a prior belief in event A , beliefs formed by previous information, e.g., our prior belief about bugs being in our code before performing tests
- Secondly, we observe our evidence. if our code passes X tests, we want to update our belief to incorporate this. We call this new belief the posterior probability
- Updating our belief is done via the following equation, known as Bayes' Theorem, after its discoverer Thomas Bayes:

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)}$$

The formula is not unique to Bayesian inference: it is a mathematical fact with uses outside Bayesian inference

- Bayesian inference merely uses it to connect prior probabilities P(A) with updated posterior probabilities P(A|X)

More useful Formulation

- **When multiple events A_i form an exhaustive set with another event B**
- B can be written as $B = \sum_{i=1}^{n} B \cap A_i$
- So, probability of B can be written as $P(B) = \sum_{i=1}^{n} P(B \cap A_i)$
- Since $P(B \cap A_i) = P(B|A_i) \times P(A_i)$
- Replacing P(B) in the equation of conditional probability:

$$P(A_i|B) = (P(B|A_i) \times P(A_i))/(\sum_{i=1}^{n}(P(B|A_i) \times P(A_i)))$$

Why do we care about Bayesian ML?

- Because when it snows, that changes the equations of self-driving cars
  - **A lot of things can occur which we *don't have enough data on***
- Machine learning is a set of methods for creating models that describe or predicting something about the world
  - It does so by learning those models from data
- **Bayesian machine learning allows us to encode our prior beliefs about what those models should look like, independent of what the data tells us**
  - **This is especially useful when we don't have a ton of data to confidently learn our model**
- Because we want to know ***why*** Machines make decisions
  - If our model is **theta** and our data is **D**, then
    P(theta | D) = P(D | theta) * P(theta) / P(data)

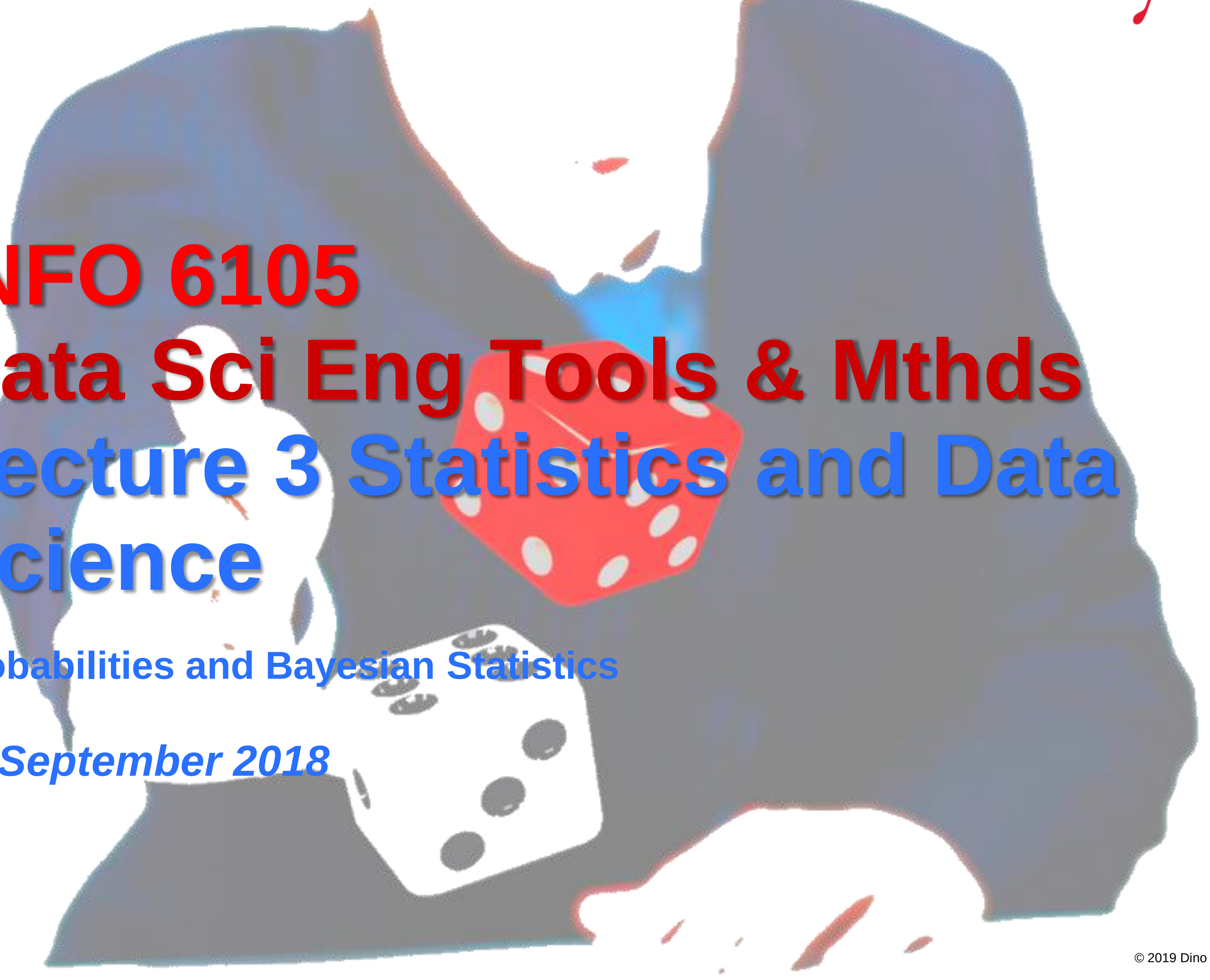
Bayesian ML

1. Have a model or distribution
2. Specify the prior belief we have about the parameters
3. Observe some data
4. Compute posterior P(θ/D) - Probability Distribution of the model obtained
5. Do this for multiple models and see which model fits best subsequent data

Part 5

REVIEW

Northeastern University



INFO 6105
Data Sci Eng Tools & Mthds
Lecture 3 Statistics and Data Science

Probabilities and Bayesian Statistics

27 September 2018

Bayes’ Formula

- We have a prior belief in event A , beliefs formed by previous information, e.g., our prior belief about bugs being in our code before performing tests
- Secondly, we observe our evidence. if our code passes X tests, we want to update our belief to incorporate this. We call this new belief the posterior probability
- Updating our belief is done via the following equation, known as Bayes' Theorem, after its discoverer Thomas Bayes:

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)}$$

The formula is not unique to Bayesian inference: it is a mathematical fact with uses outside Bayesian inference

- Bayesian inference merely uses it to connect prior probabilities P(A) with updated posterior probabilities P(A|X)

More useful Formulation

- **When multiple events A_i form an exhaustive set with another event B**
- B can be written as $B = \sum_{i=1}^{n} B \cap A_i$
- So, probability of B can be written as $P(B) = \sum_{i=1}^{n} P(B \cap A_i)$
- Since $P(B \cap A_i) = P(B|A_i) \times P(A_i)$
- Replacing P(B) in the equation of conditional probability:

$$P(A_i|B) = (P(B|A_i) \times P(A_i))/(\sum_{i=1}^{n}(P(B|A_i) \times P(A_i)))$$

Bayes’ Rule

- As Bayesians, we start with a belief, called a *prior*
- Then we obtain some data and use it to update our belief
- The outcome is called a *posterior*
- Should we obtain even more data, the old posterior becomes a new prior and the cycle repeats
- **All components are probability distributions**
- This process employs the **Bayes rule**:
  - **P(A | B) = P(B | A) * P(A) / P(B)**
- In Bayesian machine learning we use the Bayes rule to infer model parameters (theta) from data (D):
  - **P(theta | D) = P(D | theta) * P(theta) / P(data)**
  - ***This is how we can ask machines what parameters they use to make decisions***

Bayesian ML: P(theta | D) = P(D | theta) * P(theta) / P(data)

- **P(data) is something we generally cannot compute**
  - Since it’s just a normalizing constant, it doesn’t matter much. When comparing models, we’re mainly interested in expressions containing theta, because P(data) stays the same for each model
- **P(theta) is *a prior*, our belief of what the model parameters *might* be**
  - Most often our opinion in this is vague and if we have enough data, we simply don’t care
  - Inference converges to probable theta as long as it’s not 0 in the prior
  - One specifies a prior in terms of a parametrized distribution
- **P(D | theta) is called *likelihood of data given model parameters***
  - The formula for likelihood is model-specific. People often use likelihood for evaluation of models: a model that gives higher likelihood to real data is better
- **Finally, P(theta | D), *a posterior*, is what we’re after**
  - It’s a probability distribution over model parameters obtained from prior beliefs and data

Inference and Model

- Inference refers to how you learn parameters of your model. A model is separate from how you train it, especially in the Bayesian world
- In classical deep learning, you train a network using variants of Stochastic Gradient Descent
- The most important method of inference is Monte Carlo sampling

CLASS DISMISSED
iStockphoto