Prior belief = belief before seeing the evidence
Posterior belief = belief after seeing the evidence

Bayes Theorem With Discrete

Let u, v be discrete random variables

$$P(u=z \mid v=z) = \frac{P(v=z \mid u=z)\, P(u=z)}{P(v=z)}$$

More generally:

$$P(u=m \mid v=n) = \frac{P(v=n \mid u=m)\, P(u=m)}{P(v=n)}$$

ex:

Let x be the change in gaze (in degrees) over 3 seconds after a sound is played.

P(can hear the sound) $= \frac{3}{4} \rightarrow y \sim \text{Bern} = 1$

I observe $x=0$, what is the prob the baby can hear the sound

$$P(y \mid x) = \frac{P(x=0 \mid y=1)\, P(y=1)}{P(x=0)}$$

= Since we don't have $P(x=0)$

$$= \frac{P(x=0 \mid y=1)\, P(y=1)}{P(x=0 \mid y=1)\, P(y=1) + P(x=0 \mid y=0)\, P(y=0)} = \frac{3}{8}$$

Inference With Continuous

ex:

Q: At birth, girl elephant weights are distributed as a Gaussian with mean = 160kg, std = 7kg. At birth, boy elephant weights are distributed as a Gaussian with mean = 165kg, std = 3kg. All you know about a newborn elephant is that it is 163kg. What is the probability that it is a girl?

$$P(s|w) = \frac{P(w|s)\,P(s)}{P(w)} = \frac{P(w|s)\,P(s)}{P(w|s)P(s) + P(w|s^c)P(s^c)}$$

We have to say that the probability $P(w=163 \mid G=1)$ is 0 because we are speaking about continuous random variables. But we can use ε so:

$$P(s|w) = \frac{\varepsilon \cdot f(w|s) \cdot P(s)}{\varepsilon \cdot f(w)}$$

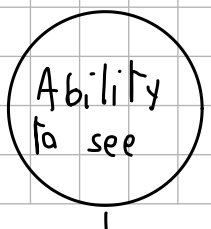
Given that we use ε for this case to take the "instant" probability we can use the PDF of a normal distribution instead of the Φ for the CDF

From here on when there is a question like how good or how bad something is i don't have to give a precise number but a probability distribution.
I can give an expectation but never a precise number

Dictionaries can also be used to represent a discretization of a continuous random var

ex:

As a Causal Model



A is a non-parametric distribution (dictionary)

Fam size is fixed

Correct y

$P(Y=1 \mid A=\alpha)$ is a Bernoulli random variable where the parameter p is a known function of α and fam size s.

Q: What is $P(A=\alpha \mid Y=o)$

Likelihood: Item response theory

If student i attempts problem j, the likelihood they answer it correctly is

$$p_{i,j} = \sigma(\alpha_i - d_j)$$

- $p_{i,j}$: prob correct
- σ: squashing function
- α_i: ability of student i
- d_j: difficulty of problem j

The ability - difficulty isn't necessary a prob but also a number, then inside the squashing function it becomes a probability.

Modeling

Real world scenario with bunch of random variables.

Bayesian Networks

I have 4 health-related problems

Flu — Undergrad

Fever — Tired

In this case I have 4 joint probabilities.

I need to describe these using causality. For example flu causes fever. Even if it isn't deterministic it changes a lot the probability

Constructing a Bayesian Network

Almost all the models that define causality are wrong but if I am able to build something with the causality I would be able to reconstruct the reality.

So the first step is creating the joint distribution using causality.

The more causality relations we introduce the more the model grows exponentially.

The second step is giving the conditional probability between each event.

If I link everything in my network i'll finish to have a joint probability table.

The third step is to implicitly assume independences.

This means that I don't need to calculate relations that don't exist.

The model/graph must be acyclic.

$$P(\text{Joint}) = \prod_i P(X_i = x_i \mid \text{parents of } X_i)$$

ex:

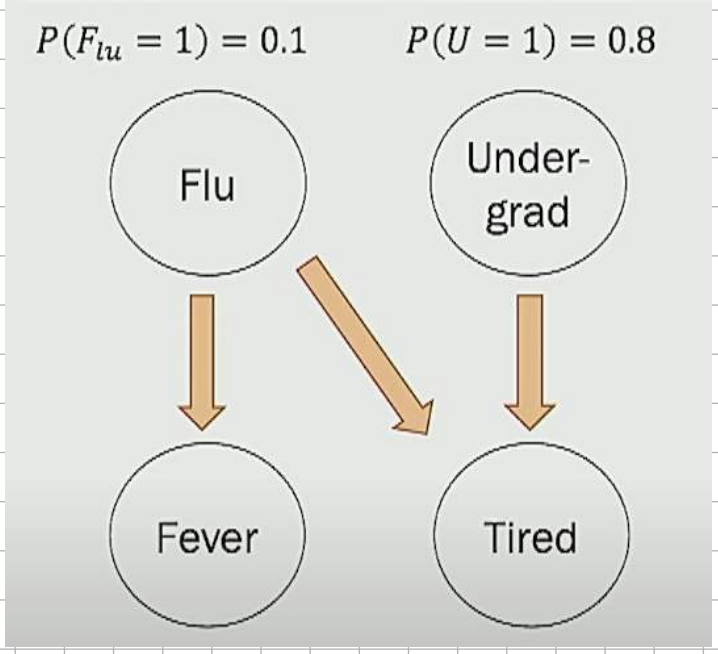
$P(F_{lu}=0, U=1, F_{ev}=0, T=1)$?

$P(U=1)$ · 0,8

$P(F_{lu}=0)$ · 0,9

$P(F_{ev}=0 \mid F_{lu}=0)$ · 1 − 0,05

$P(T=1 \mid F_{lu}=0\ U=1)$ 0,8



$P(F_{ev}=1|F_{lu}=1)=0.9$
$P(F_{ev}=1|F_{lu}=0)=0.05$

$P(T=1|F_{lu}=0,U=0)=0.1$
$P(T=1|F_{lu}=0,U=1)=0.8$
$P(T=1|F_{lu}=1,U=0)=0.9$
$P(T=1|F_{lu}=1,U=1)=1.0$

Independence in RV

Recalling the independence of RV:

$$P(EF) = P(E)P(F)$$

2 discrete random variables x and y are independent if:

For all x, y:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Intuitively: Knowing the value of X tells us nothing about the distribution of Y (and vice versa)

If 2 variables are not independent, they are called dependent

Can i discover independence from data?

The independence can give me lot of information about conditional probabilities

Covariance

Say X and Y are arbitrary random variables.
Covariance of X and Y:

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

It's the measure on how much 2 random variables vary together

- X and Y independent, $E[XY] = E[X]E[Y] \rightarrow Cov(X, Y) = 0$
- $Cov(X, Y) = 0$ DOES not imply X and Y independent.

An other way to calculate covariance:

$$\begin{aligned} Cov(X, Y) &= E[XY - E[X]Y - XE[Y] + E[Y]E[X]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

I can actually discover the independence between random variables by covarying them.

The memory for a joint table is $O(n^n)$

Make a generative model

A good probabilistic model is generative. It explains the process through which the joint is created.

Covariance of 0 doesn't mean independence

X and Y are random variables:

X is -1, 0 or 1 with equal probability:

$$Y = \begin{cases} 0 & \text{if } X \neq 0 \\ 1 & \text{otherwise} \end{cases}$$

X and Y are random variables with PMF:

Y \ X	-1	0	1	$p_Y(y)$
0	1/3	0	1/3	2/3
1	0	1/3	0	1/3
$p_X(x)$	1/3	1/3	1/3	1

$$Y = \begin{cases} 0 & \text{if } X \neq 0 \\ 1 & \text{otherwise} \end{cases}$$

$$E[X] = -1(1/3) + 0(1/3) + 1(1/3) = 0$$

$$E[Y] = 0(2/3) + 1(1/3) = 1/3$$

$$Cov(X,Y) = E[XY] - E[X]E[Y] = 0$$

The value of the covariance is 0 but the 2 random variables are not independent

Cauchy Schwarz Inequality.

$$-std(x)std(y) \leq Cov(x,y) \leq std(x)std(y)$$

Correlation is just normalized covariance

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$$

- We divide by this number because given the inequality before we always know that the covariance will be less.

ex:

$$P(Flu = 1 \mid U = 1, T = 1) = 0,1$$

1 - Compute all the joint probabilities

2 - Definition of conditional probabilities

$$\frac{\sum_{y} P(Flu = 1, U = 1, Feu = y, T = 1)}{\sum_{x}\sum_{y} P(Flu = x, U = 1, Feu = y, T = 1)} = 0,122$$

I need to check all the possible combinations. I can do this by doing marginalisation

With thousands of R.V. is not possible to do like this.

It's not scalable doing like this.

Rejection Sampling Algorithm

Step 0: Define precisely a Bayesian Network.

In this case instead of make all the calculations. I create lot of fake samples and then I count all the scenarios in which the characteristics I want are satisfied.

In order to do this I need to give all the needed probabilities from the graph.

Sampling could be a good idea. When I run a method in python all the values that they give me back are coming from sampling

step 1: Rejecting

After the first part i have a lot of samples.

Here i go through all the sample that i create and i throw away all the samples that are not consistent with the observation that i make.

From the smaller dataset that i now have now i count all the samples that respect a certain condition.
Considering the example used before:

$$P(Flu = 1 \mid U = 1, T = 1) = \frac{10\,000}{30\,000}$$

At the beginning i have 100 000 samples, from 100 000 I have 30 000 with $U=1$ and $T=1$ and from 30 000 the samples with the fever are 10 000

Why this approximation makes sense?

What is $P(flu = 1 \mid U=1, T=1)$?

$$\text{probability} \approx \frac{\#\text{Samples with } (Flu = 1, U = 1, T = 1)}{\#\text{ Samples with } (U=1, T=1)}$$

Recalling our definition of probability as a frequency: $P(E) = \lim_{n \to \infty} \frac{n(E)}{n}$

$n(E)$ = trials with E occurred

When i'm sampling and i have random variables with different distributions i just need to sample from different distributions

Markov Chain Monte Carlo

It's a way to sample with conditioned variables fixed
We just fix one random variable during sampling

Flip a coin with unknown probability

Flip a coin $(n+m)$ times, comes up with n heads.
- We don't know the probability X that coin comes up heads.

Fuequentist

$$X = \lim_{n+m \to +\infty} \frac{n}{n+m} \approx \frac{n}{n+m}$$

X is often a single value

Bayesian (Prior is guess)

$$f_{X|N}(x|n) = \frac{P(N=n \mid X=x) f_X(x)}{P(N=n)}$$

X is a random variable. Leads to a belief distribution which captures confidence

Our belief before flipping coins is that: $X \sim Unif(0,1)$
Let N = number of heads.
Given $X = x$, coin flips independent: $(N|x) \sim Bin(n+m, x)$

$$f_{X|N}(x|n) = \frac{P(N=n \mid X=x) f_X(x)}{P(N=n)}$$

$$= \frac{\binom{n+m}{n} x^n (1-x)^m}{P(N=n)}$$

$\downarrow$

K to normalize the result

In this case we put $f(x) = 1$ because we choose a Beta with all the possibilities equally likely.

If you start with a $x \sim Uni(0,1)$ prior over probability and observe:
- n successes
- m failures

My new belief about the probability is:

$$f_x(x) = \frac{1}{c} \cdot x^n (1-x)^m$$

where:

$$c = \int_0^1 x^n (1-x)^m$$

Equivalently

If you start with a $X \sim Uni(0,1)$ prior over probability and observe:

a = num of successes + 1
b = num of failures + 1

My new belief about probability is:

$$F_X(x) = \frac{1}{c} x^{a-1} (1-x)^{b-1}$$

Where c:

$$c = \int_0^1 x^{a-1} (1-x)^{b-1}$$

Beta Random Variable

X is a beta random variable: $x \sim Beta(a,b)$
PDF where $a, b > 0$:

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} \\ 0 \quad \text{otherwise} \end{cases} \qquad B(a,b) = \int_0^1 x^{a-1} (1-x)^{b-1} \, dx$$

Basically with a Beta variable I have a certain prior

and i update it based on the observed data.
for example if i have:

$$P(p \mid H=9, T=1) \propto p^9 (1-p)^1$$

(p = probability of getting heads)

$$p \sim Beta(\alpha + H, \beta + T) = Beta(\alpha + H, \beta + T)$$

If i assume a uniform prior Beta(1, 1) Then the posterior would be Beta(10, 2)

> In Bayesian statistics, before we see any data, we express our initial belief about a probability using a **prior distribution**. For a probability parameter p (e.g., the probability of getting heads in a coin flip), a common choice is the **Beta distribution**:
>
> $$p \sim \text{Beta}(\alpha, \beta)$$
>
> where:
>
> - α is the number of **pseudo-observed** successes (e.g., heads).
> - β is the number of **pseudo-observed** failures (e.g., tails).
>
> The **Beta(1,1) distribution** is special because it is a **uniform prior**, meaning that before seeing any coin flips, we assume all values of p (from 0 to 1) are equally likely. This corresponds to saying, *"We don't favor any particular probability of heads."*

It's symmetric when a = b:

$$E[X] = \frac{a}{a+b} \qquad Var(X) = \frac{ab}{(a+b)^2 (a+b+1)}$$

Beta variables can come from experience

The Beta variable express how unconfident i am with my probability

What if the prior is Beta?

What if our prior belief about X was beta:

$$F(X = x) = \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1}$$

What is our posterior belief about x after observing n heads

Because we put our prior belief as Uni before, now we put in the posterior part as a Beta random variable.

$$f(X=x \mid N=n) = ?$$

so in this case i will not have $f(x)=1$ any more

$$f(X=x \mid N=n) = \frac{P(N=n \mid X=x)\, f(X=x)}{P(N=n)}$$

$$= \frac{\binom{m+n}{n} x^n (1-x)^m \, f(X=x)}{P(N=n)}$$

$$= \frac{\binom{m+n}{n} x^n (1-x)^m \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1}}{P(N=n)}$$

$$= K_3 \, x^n (1-x)^m x^{a-1} (1-x)^{b-1}$$

$$= K_3 \, x^{n+a-1} (1-x)^{m+b-1}$$

If my prior is a Beta also my belief is a Beta

$$= x \mid N \sim \text{Beta}(a+n, b+m)$$

This is called conjugate prior.

A bit understanding

If the prior distribution X (before seeing the data) is Beta

Then also the posterior is Beta

Beta is a conjugate distribution for Bern
- practically, conjugate means easy update
  → add number of heads and tails seen to Beta parameters.

Difference between Bayesian and Frequentist.

Approach	Frequentist 🏛	Bayesian 🎲
Definition of Probability	Probability is the **long-run frequency** of an event occurring.	Probability is a **measure of belief** or degree of certainty.
What is p? (e.g., coin flip probability)	p is a **fixed but unknown number.**	p is a **random variable** that follows a probability distribution.
How to Estimate p?	Use **sample proportions:** $\hat{p} = \frac{\text{heads}}{\text{total flips}}$.	Use **Bayes' theorem** to update a probability distribution for p.
Uncertainty Representation	Uses **confidence intervals** based on sample data.	Uses a **posterior distribution** to represent uncertainty.
Prior Knowledge	Assumes no prior knowledge—only data matters.	Incorporates **prior beliefs**, which are updated with data.
Example: Is a Coin Fair?	Flip the coin many times. If the fraction of heads is 0.5, conclude it's fair.	Start with a belief (e.g., **Beta(1,1)**), then update with coin flip data.
Use Case	A/B testing, clinical trials, classic hypothesis testing.	Machine learning, Bayesian inference, small-sample learning.

Can set $X \sim Beta(a, b)$ as prior to reflect how biased you think coin is apriori.

- This is subjective probability (aka Bayesian)!
- The prior probability X based on seeing $(a+b-2)$ 'imaginary' trials where
  $(a-1)$ of them were head
  $(b-1)$ of them were tails

- Update to get posterior probability:
  → $X \mid$ (n head and m tails) $\sim Beta(a+n, b+m)$

Laplace Smoothing (Simple prior)

Prior: $X \sim Beta(a=2, b=2)$

- a: One imagined head
- b: One imagined tail

$Beta(2,2)$ is a popular prior to start with

If the expectation is the weighted average the point with the highest probability is called Mode

if $X \sim Beta(a=19, b=8)$

$$E[X] = \frac{a}{a+b} = \frac{19}{19+8} \cong 0,70$$

$$Mode(x) = \frac{a-1}{a+b-2}$$

In the denominator i have the amount of success and it's divided by the number of total trials.

Adding Random Variables

IID Random Variables

Consider n random variables $x_1, x_2, \ldots, x_n$
x_i are all independently and identically distributed (IID)
All have the same PMF/PDF
All have the same expectations
All have the same variance

ex:

2 players
$P(X=x)$ = amount of points scored by player 1
$P(Y=y)$ = amount of points scored by player 2

What is the probability of a tie?

$$P(tie) = \sum_{i=0}^{100} P(X=i, Y=i)$$ Because they are independent

The insight of convolution proof

What is the probability that $x+y=n$ $\left(P(x+y=n)\right)$

x	y	ind
0	n	0
1	n-1	1
2	n-2	2
⋮	⋮	⋮
n	0	n

They are mutually exclusive and countable and cover all possible ways

In this case the 2 random variables are dependent so I need to think about the joint probability together

$$P(x+y)=\sum_{i=0}^{n} P(x=i, y=n-i)$$

If they are independent i need to do the summation over the product of the probabilities.

$$P(x+y=n)=\sum_{i=0}^{n} P(x=i)\,P(y=n-i)$$

if we speak about continuous random variable we would use an integral.

Sum of independent binomials

In some cases adding random variables is easy:
Let x and y be independent binomials with the same value for p.

- $X \sim Bin(n_1, p)$ and $Y \sim Bin(n_2, p)$

If we sum these 2 random variables we get $Z \sim Bin(n_1+n_2, p)$
When we sum we obtain a random variable because the result is not deterministic.

Intuition:

X has n_1 trials and Y has n_2 trials
Each trial has same success probability p
Define Z to be n_1+n_2 trials, each with successe probability p.
$Z \sim Bin(n_1+n_2, p)$ and so $Z = X+Y$

Generally, have n independent random variables
$X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots, n$

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Linear Transformation

Y is a linear transformation

$X \sim N(\mu, \sigma^2)$

$Y = X + X = 2 \cdot X$

$Y = N(2\mu, 2\sigma^2)$

ex:

Gaussian Sampling and ELO Ratings

Each team has an elo score s, calculated on its past performance
each game a team has ability $A \sim N(s, 200^2)$
The team with the higher sampled ability wins
What is the probability that Warriors win a game?

$P(A_w > A_o)$?

Warriors' $A_w \sim N(s = 1657, 200^2)$

Opponent's $A_o \sim N(s = 1470, 200^2)$

Probability of winning a game

$$P(\text{Warriors win}) = P(A_w > A_o)$$
$$= P(A_w - A_o > 0)$$

I apply a linear transformation to A_o

$-1 \cdot A_o = -A_o \sim (-1470, (-1)^2 200^2)$

The difference them would be a sum

$D = A_w + (-A_o) = (1657 - 1470, 200^2 + 200^2) = (X, 2 \cdot 200^2)$

Then i can ask about what is the probability that this is > 0

$P(D > 0) = 1 - F_D(0) \approx \gamma$

We don't need anymore to sample we can solve like this.

Discrete vs Continuous

$$P(X+Y=a) = \sum_{y=-\infty}^{+\infty} P(X=a-y)P(Y=y)$$

$$f(X+Y=a) = \int_{y=-\infty}^{+\infty} f(X=a-y)\, f(Y=y)\, dy$$

Sum of Independent uniforms

Let x and y be independent variables
$X \sim Uni(0,1)$ and $Y \sim Uni(0,1) \rightarrow f(x) = 1$ for $0 \leq x \leq 1$

If you think about it: $(1 < a < 2)$

$f(X+Y=a)$?

$$f(X+Y=a) = \int_{y=-\infty}^{+\infty} f(X=a-y)\, f(Y=y)\, dy$$

Since x and y are independent I am using the convolution formula by multiplying the CDF of one result by another probability.

$$f(X+Y=a) = \begin{cases} a & 0<x<1 \\ 2-a & 1 \leq x < 2 \\ 0 & \text{otherwise} \end{cases}$$

The sum of random variables isn't always a clean answer just look at the result coming from the sum of 2 Uni

Central Limit Theorem

Consider n independent and identically distributed variables $X_1, X_2, \ldots, X_n$ with $E[X_i] = \mu$ and $Var(X) = \sigma^2$

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2) \quad \text{As } n \rightarrow \infty$$

With a huge number of i.i.d. R.V., if I sum all of them, no matter what the result will be a normal distribution.

Proof of CLT

The Fourier transform of a PDF is called a characteristic function.
Take the characteristic function of the probability mass of the sample normalize
Shows that this approaches an exponential function in the limit as $n \to \infty$: $f(x) = e^{-x^2/2}$
This function is in turn the characteristic function of a Normal Distribution.

Average of IID Variables

Let x_i be n i.i.d variables and $\bar{x}$ be the average.

$$\bar{x} = \frac{1}{n}\sum_{i=1} x_i$$

→ Gaussian by CLT
$N(n\mu, n\sigma^2)$

By the central limit theorem the mean of IID variables are distributed normally. As $n \to \infty$

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Coming from the variance scaling property:

$$\text{Var}\left(\frac{1}{n}\sum_{i=1} x_i\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1} x_i\right)$$

Sampling Definitions

I want to know the true mean and variance of happiness in Bhutan.
- Obv i can't ask to all
- I ask to 200 people

Happiness = $\{u_1, u_2, \ldots, u_{200}\}$

$\mu = 83$

Is it the true mean of happiness?

Mathematically, what is a sample?

Consider n random variables:

The sequence of random variables is a sample from the distribution F if:

- All the R.V. are iid
- All the variables have the same distribution function F where $E[X_i] = \mu$, $Var(X_i) = \sigma^2$

Estimating Core Statistics (μ, σ^2)

How to estimate the core statistics giving few samples.

How good is our estimation of the mean for example

$$E[\bar{X}] = E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} n\mu = \mu$$

In this case this mean is called unbiased estimator of the population mean μ.

Intuition About the sample variance σ^2

For example population variance:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

μ: population mean

Now let's see the sample variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$\bar{x}$: sample mean

$\bar{x}$ and μ surely differ because $\bar{x}$ is the mean of the subset of the real population.

We put $n-1$ in the denominator because the variance we are calculating is an estimate using an estimate so it needs to be scaled a little bit more

Proof that s^2 is unbiased.

$$E[S^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \quad \Rightarrow \quad (n-1)E[S^2] = E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

$$(n-1)E[S^2] = E\left[\sum_{i=1}^{n}\left((X_i - \mu) + (\mu - \bar{X})\right)^2\right] \qquad \text{(introduce } \mu - \mu\text{)}$$

$$= E\left[\sum_{i=1}^{n}(X_i - \mu)^2 + \sum_{i=1}^{n}(\mu - \bar{X})^2 + 2\sum_{i=1}^{n}(X_i - \mu)(\mu - \bar{X})\right]$$

$$= E\left[\sum_{i=1}^{n}(X_i - \mu)^2 + n(\mu - \bar{X})^2 - 2n(\mu - \bar{X})^2\right]$$

$$= E\left[\sum_{i=1}^{n}(X_i - \mu)^2 - n(\mu - \bar{X})^2\right] = \sum_{i=1}^{n}E[(X_i - \mu)]^2 - nE[(\bar{X} - \mu)^2]$$

$$= n\sigma^2 - n\text{Var}(\bar{X}) = n\sigma^2 - n\frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

Therefore $E[S^2] = \sigma^2$

$$2(\mu - \bar{X})\sum_{i=1}^{n}(X_i - \mu)$$

$$2(\mu - \bar{X})\left(\sum_{i=1}^{n}X_i - n\mu\right)$$

$$2(\mu - \bar{X})n(\bar{X} - \mu)$$

$$-2n(\mu - \bar{X})^2$$

Standard error of the mean

$E[\bar{X}] = \mu$

We want to estimate $\frac{\sigma^2}{n}$

Def: The standard error of the mean is an estimate of the standard deviation of $\bar{x}$

$$SE = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

This measure how we think our mean estimate is.

Intuition:

s^2 is an unbiased estimate of σ^2

s^2/n is an unbiased estimate $\text{Var}(\bar{x})$

$\sqrt{s^2/n}$ can estimate $\sqrt{\text{Var}(x)}$

The estimated standard error is a measure on how much the sample mean varies across different samples (random samples = subset of the population)

The value of confidence in my claim is called the p-value

Bootstrap: randomize algorithm

This algorithm allows you to:
→ Know the distribution of statistics
→ Calculate p values.
→ Using computers.

How can i know how good my estimated variance is if i have the possibility to know the true value coming from the real life?

Since we are using a subset of the entire dataset the variance that i calculate is the one from the subset and it's called sample variance.
And I want to know how much this number change if i choose a different subset of other 200 people (This is basically the variance of the variance)
The sample variance is a random variable so it has a distribution.

The first half of the bootstrap is basically to choose 10000 samples, obtain the PMF and then take all needed conclusions.

Bootstrapping Assumption

$F \approx \hat{F}$

$\hat{F}$ → the sample distribution
F → the underlying distribution

The distribution i get from the histogram of my samples is a good approximation of the true distribution

Algorithm

1- Estimate the PMF using the sample
2- Repeat 10000 times
   a- Resample len(sample) from PMF
   b- Recalculate the stat on the resample
3- You now have a distribution of your stat

In an easier way:

We take 10000 subsets from the main set

For every subset we calculate the variance, the mean, the median.
With the 10000 values we create a distribution.

At the end I have different PMFs for different metrics.

Considering the central limit theorem after the calculation of all the means if i sum all the values i will obtain a normal distribution

Then after bootstrapping i could've know the probability that the true mean is between a range of values

To calculate the probability of this range i can also simply count the number of values in the list that are within the range.

Bootstrap tells me how wrong I might be about a value

If i get a p-value $\leq 0,05$ is accepted

How can we use bootstrapping with the null hypothesis?

Null hypothesis: There is no true difference between random sampled groups. The only differences that we observe are due to random chance / sampling error.

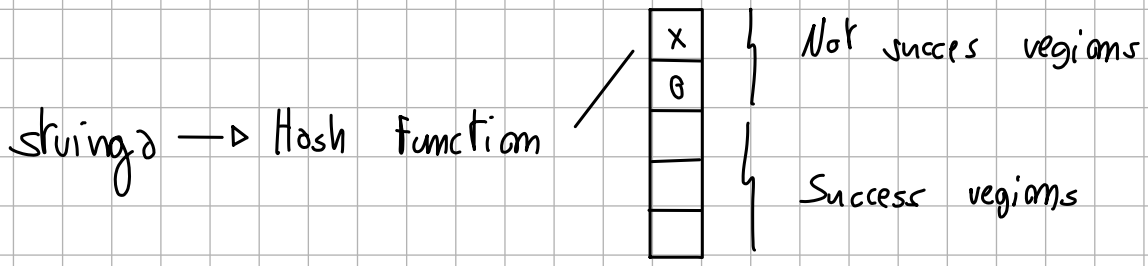
The bootstrapping in this case is useful due to the fact that we can check if the difference coming from the 2 groups is significant or just random noise

Basically the p value is the probability that our error / difference comes from noise and not due to errors.

Algorithmic Analysis

ex: Coupon Collector problem.

Consider an hash table with n buckets.
Each string equally likely to get hashed into any bucket.
Let X = # string to hash until each bucket $\geq$ 1 string
What is $E[X]$?



5 buckets

In the first attempt we need 7 rounds to fill the hash table

Let X_i = # of trials to get success after i-th success
where success is hashing string to previously empty bucket
After i buckets have ≥ 1 string, probability of hashing a string to an empty bucket is $p = \frac{(n-1)}{n}$

$$P(X_i = k) = \frac{n-i}{n}\left(\frac{i}{n}\right)^{k-i} \quad \text{equivalently} \quad X_i \sim Geo\left(\frac{n-1}{n}\right)$$

$$E[X_i] = \frac{1}{p} = \frac{n}{n-i}$$

$$X = X_0 + X_1 + \ldots + X_n \rightarrow E[X] = E[X_0] + E[X_1] + \ldots + E[X_{n-1}]$$

Every bucket could be considered as a random variable so i'll have to sum n different random variable.

$$E[X] = \underbrace{\frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} \ldots + \frac{n}{1}}_{\text{This sum is called armonic sum and the result is constant} \cdot \log n} = n\left(\frac{1}{n} + \frac{1}{n-1} + \ldots + 1\right) = O(n \log n)$$

Conditional Expectation

X and Y are jointly discrete random variables:

We define the conditional expectation of X given $Y = y$

$$E[X \mid Y = y] = \sum_x x P(X = x \mid Y = y) = \sum x p_{X|Y}(x|y)$$

For continuous random variable:

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \qquad E[X|Y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x,y)\,dx$$

ex:

Roll 2 dices

X = value of $D_1 + D_2$ $\quad Y$ = Value of D_2

What is $E[X \mid Y = 6]$

$$P(x \mid y) = \frac{P(xy)}{P(y)} = \frac{1/36}{1/6} = 1/6$$

$$E[X|Y] = \sum_{i=7}^{12} i P(X = x \mid Y = 6) = 1/6\,(7+8+9+10+11+12) = \frac{57}{6}$$

Considering now $g(Y) = E[X|Y]$, It's a function of Y.

In our case $E[X]$ is a value while $E[X \mid Y = y]$ is a function of y. Writing $E[X = 5]$ doesn't make sense.

Law of total expectation

$$E[E[X|Y]] = \sum_y E[X|Y] P(Y) \qquad g(Y) = E[X|Y]$$

$$= \sum_y \sum_x x\, p(x \mid Y = y)\, p(y)$$

$$= \sum_y \sum_x x\, p(x,y)$$

$$= \sum_x \sum_y x\, p(x,y)$$

$$= \sum_x x \sum_y p(x,y)$$

$$= \sum x\, p(x) = E[X]$$

In this passage i use the definition of marginal probability by summing over all the values of y.

Analyzing Recursive Code

```
int Recurse() {
   int x = randomInt(1, 3);  // Equally likely values

   if (x == 1) return 3;
   else if (x == 2) return (5 + Recurse());
   else return (7 + Recurse());
}
```

Let Y = value returned by Recurse().

What is E[Y]?

We can see X as a distribution so it can be considered as a random value.

$$E[Y] = p(X=1)E[Y|X=1] + p(X=2)E[Y|X=2] + p(X=3)E[Y|X=3]$$

$$E[Y|X=1] = 3$$

$$E[Y|X=2] = E[5+Y] = 5+E[Y]$$

$$E[Y|X=3] = E[7+Y] = 7+E[Y]$$

$$E[Y] = 3(1/3) + (5+E[Y])(1/3) + (7+E[Y])(1/3) = (1/3)(15+2E[Y])$$

Machine Learning Estimation
Whats are parameters

Consider different probability distributions:

Every one of them have special numbers for example the mean the std ... etc
We call these prob distributions with these numbers parametric models
I don't have to describe all the points in the curve i just need to give the value of the parameters.

Given model, parameters yield actual distribution
- Usually refer to parameters of distribution as θ
- The θ can also be a vector of parameters.

With bootstrapping we already have a method to estimate some parameters for example the variance and the mean

Likelihood Given The Parameters

$$\text{Likelihood} = \prod_{i=1}^{N} f(x_i \mid \mu = 5,5 , \sigma = 1,1)$$

N = all the dataset

It's the PDF of all the points given some values of μ and σ in this case but it's in general

The likelihood is the product between all the PDF of all the points given a certain value of mean and standard deviation

Difference between max and argmax

If i have the function $f(x) = -x^2 + 5$

$\max_x \; -x^2 + 5$ will return the maximum value that the function can return

$\arg\max_x \; -x^2 + 5$ will return 0 because it returns the maximum input used by the function to return the maximum value.

The argument that maximize a function is also the argument that maximize the log of the function
-> this is because log is monotonic so the longer the input the longer the output

We use the logarithm because computers are not very good at representing very very small number but with the log i am able to represent numbers with a little bit of sanity.

Maximum Likelihood Estimation

1. Decide on a model for the distribution of your samples. Define the PMF / PDF for your sample.
2. Write out the log likelihood function.
3. State that the optimal parameters are the argmax of the log likelihood function.
4. Use an optimization algorithm to calculate argmax

The likelihood function

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

- n i.i.d. data points
- parameters distribution

So i'm calculating the probability that certain R.V. all together have certain values.
In order to have all the values for the random variables at the same time i need to multiply their probability

The result will be how likely the data are generated given a certain model and certain parameters.

Maximum likelihood

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

$$LL(\theta) = \sum_{i=1}^{n} \log\left(f(x_i \mid \theta)\right)$$

$$\hat{\theta} = \arg\max_{x} LL(\theta)$$

MLE For Poisson

$$X \sim Poi(\lambda)$$

Now: need to find the maximum likelihood

The PMF can be written as: $f(x_i \mid \lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

The likelihood can be written: $L(\theta) = f(x_1 \dots x_n \mid \lambda) = \prod_{i=1}^{n} f(x_i \mid \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

The log likelihood $= \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right)$

$$= \sum_{i=1}^{n} -\lambda + x_i \log\lambda - \log(x_i!)$$

Now i just need to do the derivative to find the point in which the value of the function is at max.

If i put that equal to 0 i'll have $\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i$

MLE For Bernoulli

$$X \sim Bern(p)$$

$$LL(\theta) = \sum_{i=1}^{n} \log f(x_i \mid p)$$

$$f(x_i \mid p) = \begin{cases} p & \text{if } x_i = 1 \\ 1-p & \text{if } x_i = 0 \end{cases}$$

Given that this function isn't differentiable we need to find a way

Differentiable pmf for Bernoulli

Consider iid random variables $x_1, x_2, \dots, x_n$

$x_i \sim \text{Bern}(p)$

Given that the continuous form gives the same result we can write $f(x_i \mid p) = p^{x_i}(1-p)^{1-x_i}$

But this form happens to have sense just for 0 and 1 That are what we need in case of Bernulli and it's also differentiable.

$$\text{Likelihood} = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$LL(\theta) = \sum_{i=1}^{n} \log\left(p^{x_i}(1-p)^{1-x_i}\right) = \sum_{i=1}^{n} \left[x_i (\log p) + (1+x_i)\log(1-p)\right]$$

If we differentiate and set to 0:

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

$$\frac{\partial LL(\theta)}{\partial p} = Y\frac{1}{p} + (n-Y)\frac{-1}{1-p} = 0 \rightarrow \boxed{p_{MLE} = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

MLE for Gaussian

$$X \sim N(\mu, \sigma^2)$$

Let $x_i (\mu, \sigma^2)$ $\qquad f(x_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$LL(\theta) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) = \sum_{i=1}^{n}\left[-\log\left(\sqrt{2\pi\sigma}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= -\sum_{i=1}^{n} \log\left(\sqrt{2\pi\sigma} + \sum_{i=1}^{n}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]\right.$$

I can derive with respect to μ

$$\frac{\partial LL(\theta)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

I can derive with respect to σ:

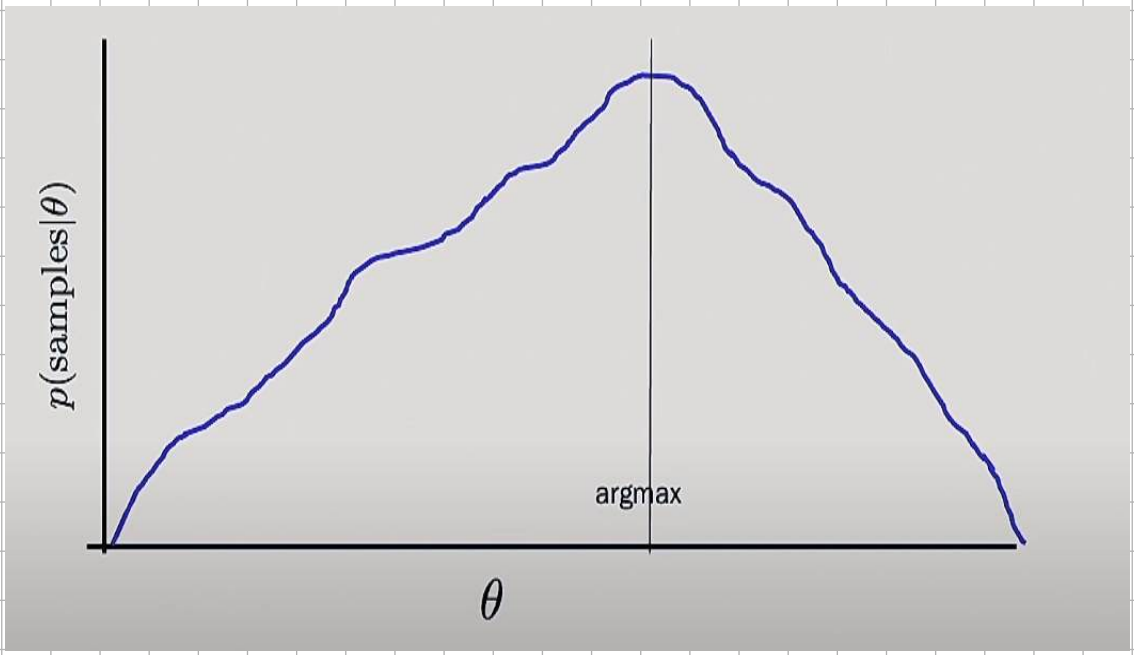
$$\frac{\partial LL(\theta)}{d\mu} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

In this case i have 2 unknowns and 2 equations

Solving for μ_{MLE}: $\frac{1}{n}\sum_{i=1}^{n} x_i$

Solving for σ_{MLE}: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{MLE})^2$

Gradient Ascent



If my step size is small enough i can walk uphill and find a local maxima

With argmax i will have lot of inputs and it can become computationally super expensive

I need to take the θ curve doing the derivative and then follow the part where the derivative is positive
If i repeat this operation i will reach the top of the curve, at that point the values of the derivative in all the direction becomes negative.
Small step size guarantee you to reach the top but it will take more time

Repeat many times:

$$\theta_j^{new} = \theta_j^{old} + \eta \frac{\partial LL(\theta^{old})}{\partial \theta_j^{old}}$$

At the beginning this is my initial guess.

η: step size constant

In every optimisation algorithm I want to use gradient descent because I don't want to find the max but I want to find the minimum of a loss. In our case here we are maximizing the likelihood but you can also minimize the negative likelihood

The problem with MLE is the fact that it overfit the data it sees it doesn't generalize the information. it doesn't think about data it hasn't seen.

Maxima A Posteriori

Basically the same as the Beta random variable but it doesn't apply only Bernoully variable but also for other random variables.

MLE Vs MAP

Data: $x_1, \ldots, x_n$

Maximum likelihood estimation:

$$\hat{\theta}_{MLE} = \arg\max f(x_{(1)} = x_1, \ldots, x_n = x_n \mid \theta)$$

Maximum a posteriori:

$$\hat{\theta}_{MAP} = \arg\max f(\Theta = \theta \mid x_1 = x_1, \ldots, x_n = x_n)$$

I will obtain the parameters that are most likely given the values of the data.
Given that the parameter is the unknown and the data are the observable we can say that this is a more natural approach

Notation Shorthand

$\theta = \Theta = \theta$

$x^i = X^i = x^i$

→ The i is in the apex because in ML the obs points could have multiple values for the same data point

$\hat{\theta}_{MAP} = \arg\max f(\theta \mid x^1, \ldots, x^n)$

Basically it returns whatever point of Theta with the highest probability.
So again basically this is used to avoid the overfitting of the data

Conjugate Distribution

MAP estimator:

$$\theta_{MAP} = \arg\max f(\theta \mid x_1, x_2, \ldots, x_n)$$

The mode of the posterior distribution of θ

The conjugate allow us to simplify the math behind the data

$$\arg\max_{\theta} f(\theta = \theta \mid Data) = \arg\max f(data \mid \theta) f(\theta)$$

↓
prior belief of my parameters
How I think my parameters are before any data is seen

Multinomial

Dirichlet $(\alpha_1, \alpha_2, \ldots, \alpha_m)$ is a conjugate for multinomial.
- It generalize betas in the same way multinomial generalize Bernulli and binomial

Prior: Saw $\sum_{i=1}^{m} \alpha_i - m$ imaginary trials, with $\alpha_i - 1$ of outcome i

Naive Bayes

MLE: $\arg\max_{\theta} \left(\sum_i \log f(x^i | \theta) \right)$

MAP: $\arg\max_{\theta} \left(\log(g(\theta)) + \sum_{i=1}^{n} \log(f(x^i | \theta)) \right)$

Training Data

Assume iid data:

$(x^1, y^1), (x^2, y^2) \ldots$

$m = |x^i|$

each data point has m features and a single output.

The output is basically if the user likes a particular element