Statistics & Probabilities

Data Science & Business Analytics

Lesson 1

Tiago Cabo

Lesson Plan

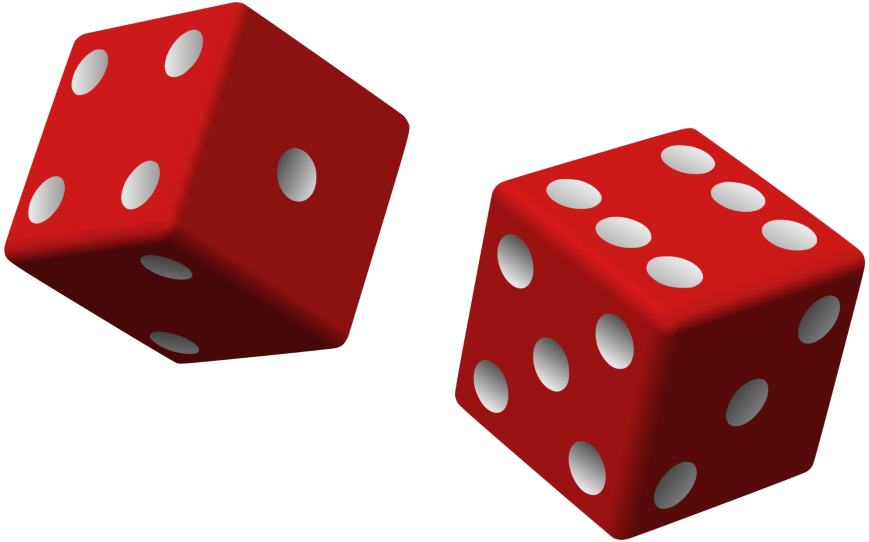
- Introduction to Probability theory
- Frequentist vs bayesian theories
- Probability distributions
- Information theory
- Exercises

What is probability theory

Probability theory is a branch of mathematics that studies probability. This is understood by the definition of a metric between 0 and 1 that represents the likelihood of some event happen in a sample space.

The two main outcomes of probability theory are:

- Law of large numbers
- Central limit theorem

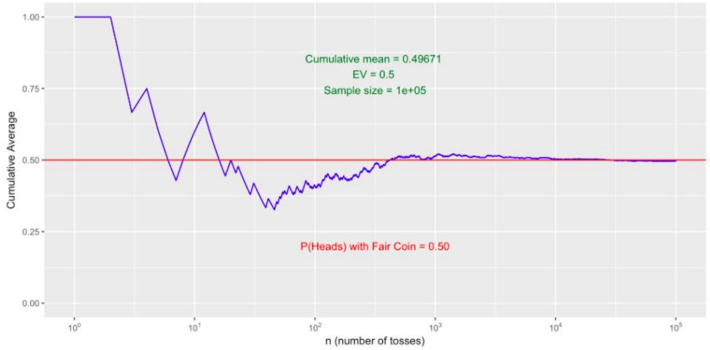


Law of large numbers

This law describes the results of the outcome of the same event a large number of times. The law states that with the ever increasing number of tries the outcome tend to the expect value.

This means, that this law is only valid for the average

$$\lim_{n\to\infty}\sum_{i=1}^{n}\frac{X_i}{n}=\overline{X}$$

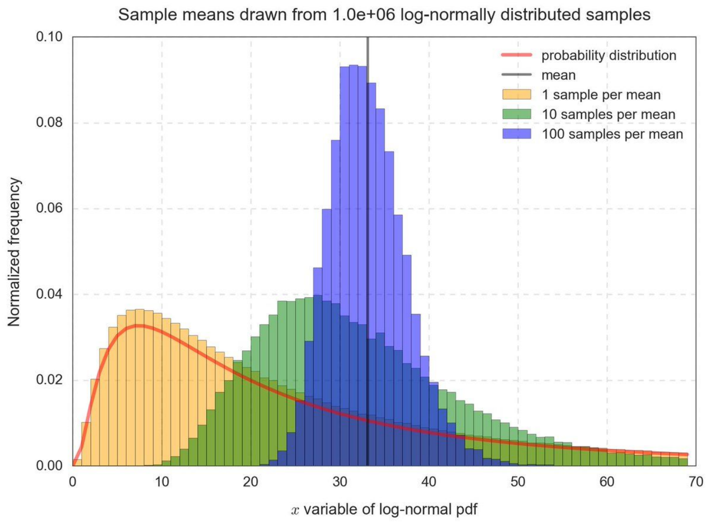


Central limit theorem

The CLT states that, for a big enough sample size, i.e. > 30, regardless of the population data distribution, The mean values of that distribution tend to the normal distribution.

This is useful to:

- Evaluate how similar are two distributions using a t-test
- Perform confidence intervals



https://www.youtube.com/watch?v=YAlJCEDH2uY&ab_channel=StatQuestwithJoshStarmer

Definition of probability

- **Frequentist theory**

Maximum Likelihood Estimation (MLE)

Given a coin flip, we expect something around 50 %

$$\mathrm{MLE} = \frac{\text{Number of successes}}{\text{Number of trials}}$$

- **Bayesian theory**

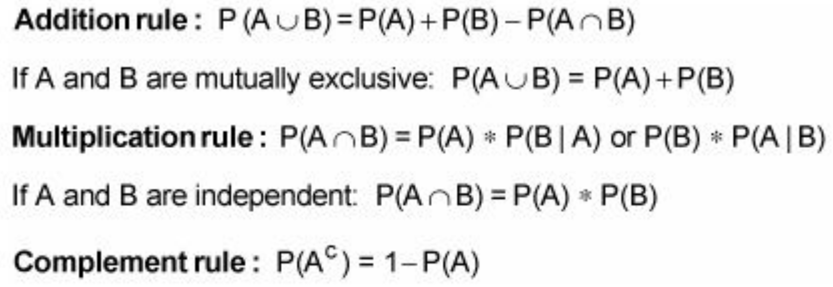
This has the Bayes' Theorem in mind.

This takes in consideration that the experiment might have a prior or condition

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

More info here: https://towardsdatascience.com/frequentist-vs-bayesian-statistics-54a197db21

Probability rules



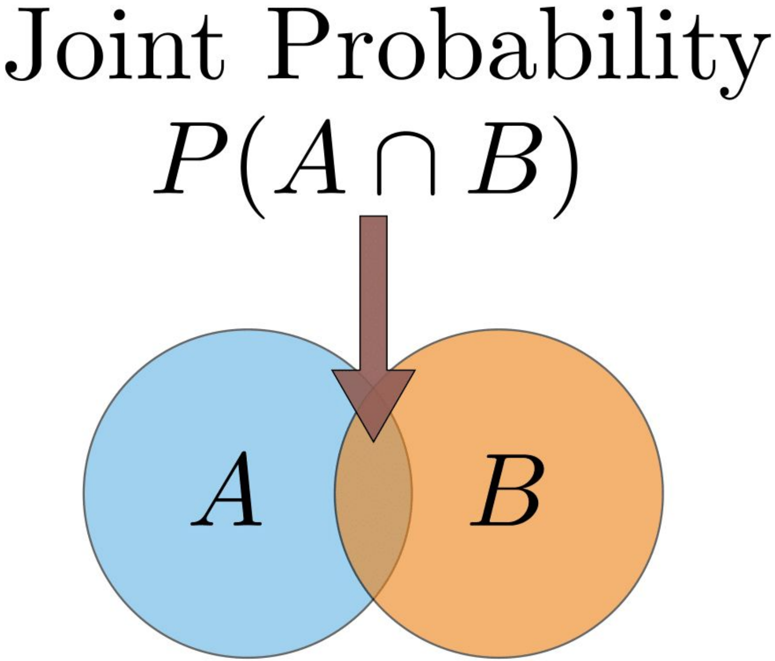
Addition rule : $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

If A and B are mutually exclusive: $P(A \cup B) = P(A) + P(B)$

Multiplication rule : $P(A \cap B) = P(A) * P(B \mid A)$ or $P(B) * P(A \mid B)$

If A and B are independent: $P(A \cap B) = P(A) * P(B)$

Complement rule : $P(A^c) = 1 - P(A)$



Probability Exercises

We flip a **fair** coin **5 times**.

1 - What is more likely to happen?

a) H T T H T

b) T T T T T

2 - What coin should follow next?

Probability Exercises - Solution

We flip a **fair** coin **5 times**.

1 - What is more likely to happen?

Both as the same probability because each independent flip has a change of 0.5

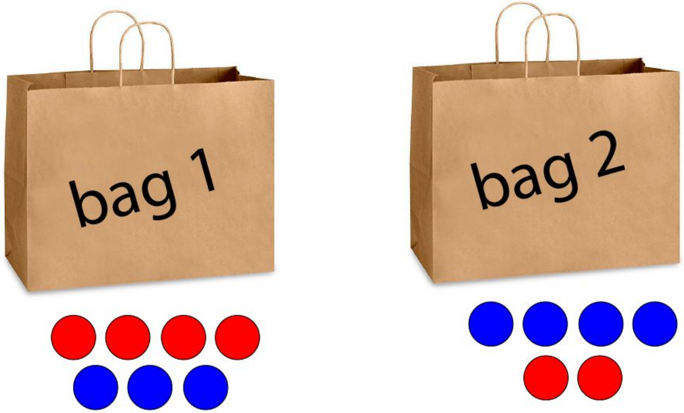
2 - What coin should follow next?

Again this is 0.5 for heads and tails, because is a independent events

Probability Exercises

We extract two balls from bag 1, one after the other, without replacement, from a box that contains 4 red balls and 3 blue balls.

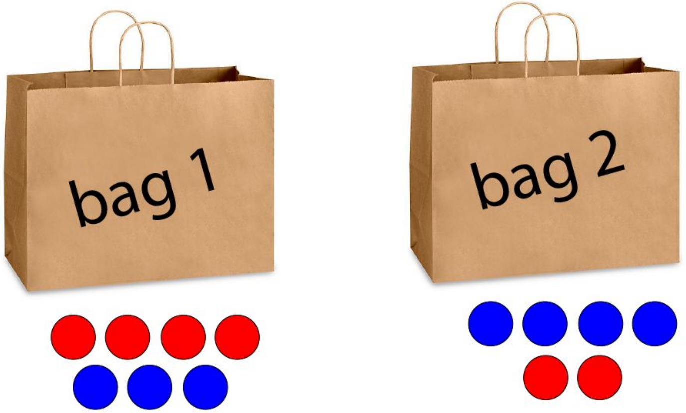
What is the probability that the two balls are red?



Probability Exercises - Solution

We extract two balls from bag 1, one after the other, without replacement, from a box that contains 4 red balls and 3 blue balls.

What is the probability that the two balls are red?



$$\frac{4}{7} \times \frac{3}{6} = \frac{12}{42} = 0.286$$

Permutations vs Combinations

Count how many different possibilities may exist if you select 5 times different letters of the alphabet.

__ __ __ __ __

Permutations: When order matters.

$$P(n,r) = \frac{n!}{(n-r)!}$$

Combinations: When does not.

$$C(n,r) = \binom{n}{r} = \frac{n!}{(r!(n-r)!)}$$

Permutations

23 * 22 * 21 * 20 * 19 = 4037880

Combinations

(23 * 22 * 21 * 20 * 19) / 5! =

4037880 / (5*4*3*2*1) = 4037880 / 120 =

= 33649

Exercise

An elevator with 3 passengers inside goes up. There are 5 floors in the building.

a) What are the assumptions that we might need to solve?

Compute the probability of:

b) Each person will leave on a different floor

c) All of them will leave on the second floor

Exercise - Solution

An elevator with 3 passengers inside goes up. There are 5 floors in the building.

a) What are the assumptions that we might need to solve?

Independent & order does not matter.

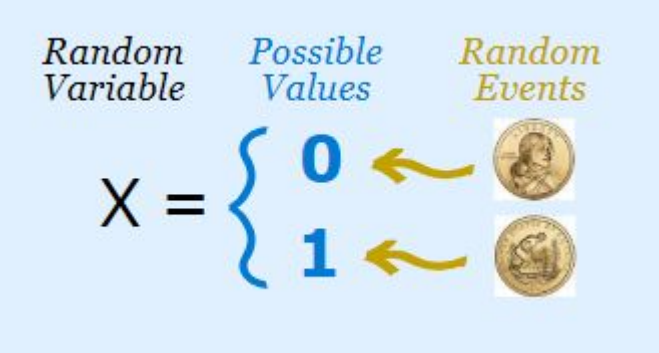
b) Each person will leave on a different floor $$\frac{5\times4\times3}{5\cdot5\cdot5} = 0.48$$

c) All of them will leave on the second floor

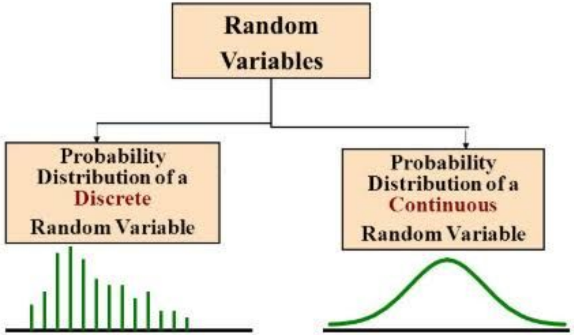
$$\frac{1}{5} \times \frac{1}{5} \times \frac{1}{5} = 0.008$$

Random Variables (RV)

- Easy way to translate probability questions into mathematical language.



- Random variables can be continuous or discrete.


Random Variables
Probability Distribution of a Discrete Random Variable
Probability Distribution of a Continuous Random Variable

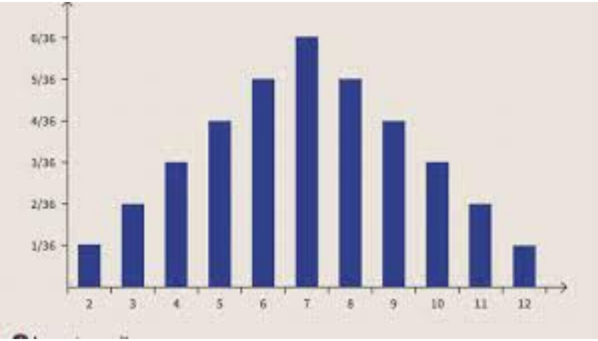
More info here: https://www.youtube.com/watch?v=dOr0NKyD31Q&ab_channel=KhanAcademy

Probability Distribution

Distribution of the probability values that the random value can have. Example of a discrete distribution.

By definition, the sum of the P(X) should be = 1.

$$\sum_{i=1}^{n} p_i = 1 \qquad \int_{-\infty}^{\infty} f(x)dx = 1$$

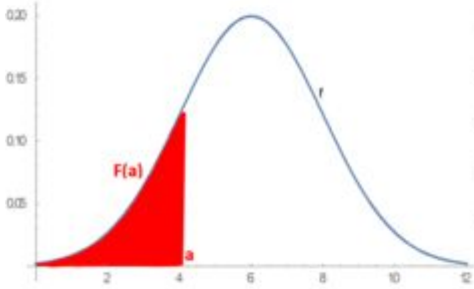


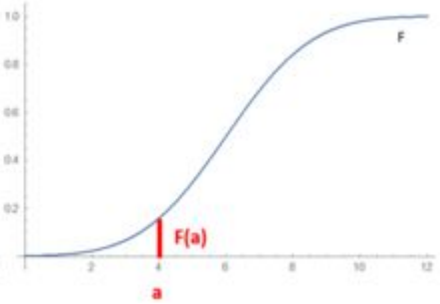
Cumulative Probability Function (cdf)

Describes the probability that a random variable does not exceed a value:

$$F_X(x) = P(X \leq x)$$

$$F_X(x) = P(X \leq x) = \sum_{y \leq x} P_X(X = y)$$





Main metrics

Mean or Expected value

$$E\big(g(X)\big) = \sum_{x} g(x)P(X = x)$$

Variance

$$V\big(g(X)\big) = \sum_{x} \big(g(x) - E\big(g(X)\big)\big)^2 P(X = x)$$

For continuous distributions

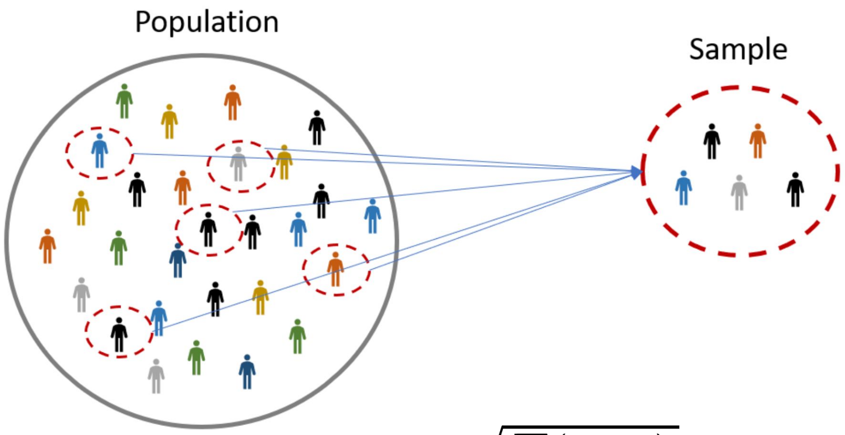
$$Mean\, \mu = \int_{-\infty}^{\infty} x f(x) dx$$

For continuous distributions

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

Sampling

When doing statistical analysis most of the time we do no have access to all population, so we need to sample and given that estimate for the general population



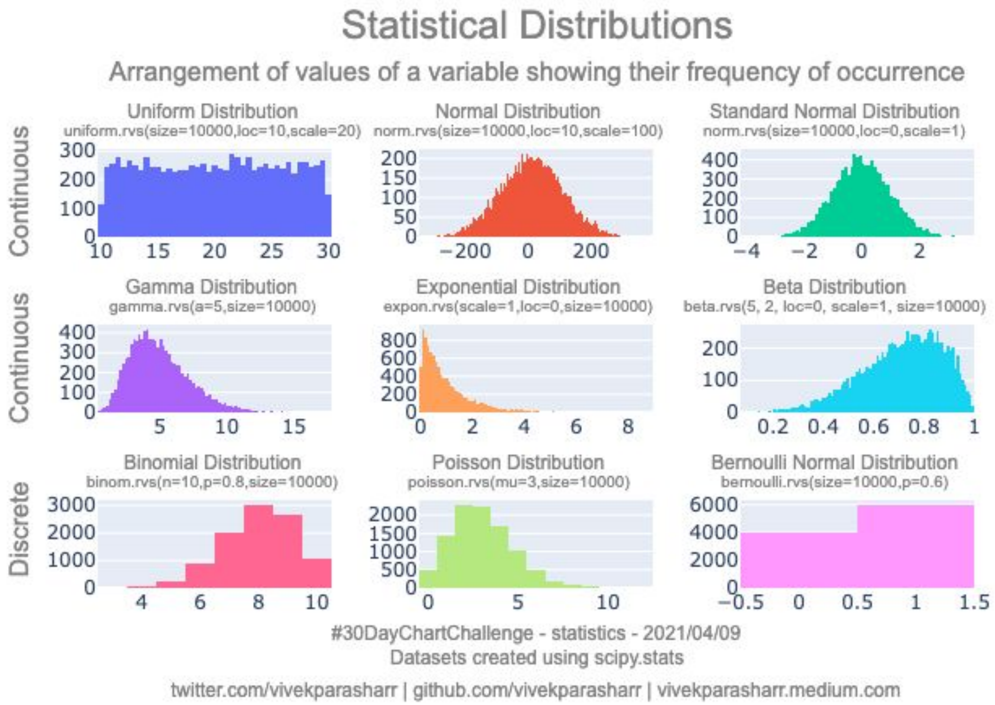
Population mean: $$\mu = \frac{\sum x_i}{N}$$

Population variance: $$\sigma = \sqrt{\frac{\sum (x_i - \mu)}{N}}$$

Sample mean: $$\bar{x} = \frac{\sum x_i}{n}$$

Sample variance: $$s = \sqrt{\frac{\sum (x_i - \bar{x})}{n-1}}$$

Probability distributions



Statistical Distributions

Arrangement of values of a variable showing their frequency of occurrence

Uniform Distribution
uniform.rvs(size=10000,loc=10,scale=20)

Normal Distribution
norm.rvs(size=10000,loc=10,scale=100)

Standard Normal Distribution
norm.rvs(size=10000,loc=0,scale=1)

Gamma Distribution
gamma.rvs(a=5,size=10000)

Exponential Distribution
expon.rvs(scale=1,loc=0,size=10000)

Beta Distribution
beta.rvs(5, 2, loc=0, scale=1, size=10000)

Binomial Distribution
binom.rvs(n=10,p=0.8,size=10000)

Poisson Distribution
poisson.rvs(mu=3,size=10000)

Bernoulli Normal Distribution
bernoulli.rvs(size=10000,p=0.6)

#30DayChartChallenge - statistics - 2021/04/09
Datasets created using scipy.stats

twitter.com/vivekparasharr | github.com/vivekparasharr | vivekparasharr.medium.com

Binomial distribution

Two possible outcomes with probabilities p and 1 − p.

Example: toss a coin. If we have N trials, and the probability of success in each is p, then the probability of obtaining n successes is:

$$P_{\text{Binomial}}(n) = \frac{N!}{n!\,(N-n)!}\, p^n\, (1-p)^{N-n}$$

Mean:

E(n)=pN, this results by extending the Bernoulli distribution

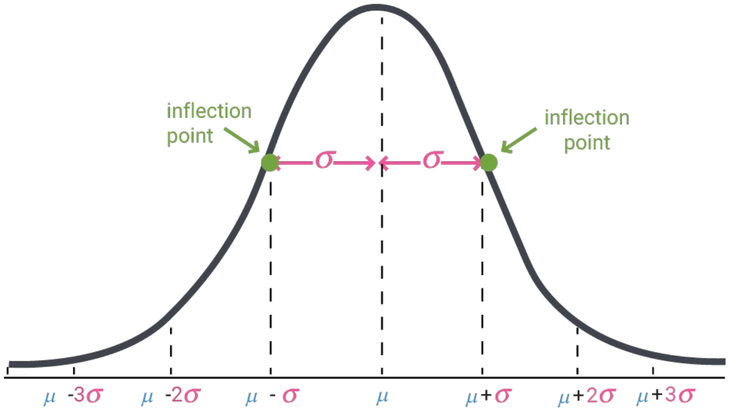
Variance:, this results by extending the Bernoulli distribution

$V(n)=Np(1-p)$

Gaussian or normal distribution

The Gaussian (or “normal”) probability distribution for a variable x, with mean μ and standard deviation σ is:

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



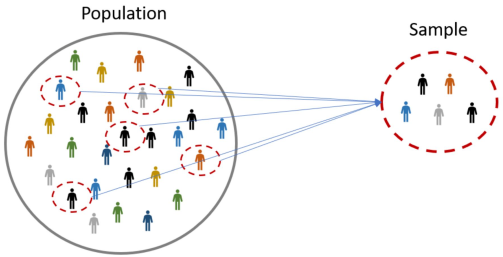
Calcworkshop.com

Confidence interval, p value, z score

One of the importance of distributions is the possibility to estimate errors. This is done using confidence intervals. The confidence interval is normally defined by $\alpha < 0.05$, this translates to 95 %.

- When assuming normal distributions

$$CI = \overline{X} \pm Z^{*} \frac{\sigma}{\sqrt{n}}$$



- Where Z transforms our sample population in a normalized one

$$Z = \frac{\bar{x} - \mu}{\sigma}$$

Importance of p-value

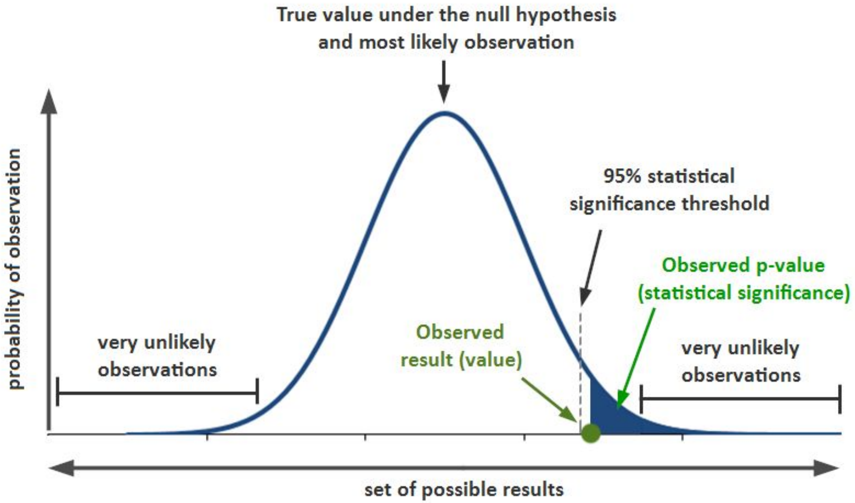
When doing an A/B test we need to setup two types of hipotesis.

H0: Average usage time remains == 20 min/day with black icon

Ha: With blue icon average goes up to 25 min/day

If we defined alfa being 0.05 when p < alfa we can reject H0, but when that does not happen, we can't. **This is very important, because this does not mean that H0 is true, is just means that we can't reject.**

P value is computed using the z score.



Confidence interval Exercise

1. The braking distances of a simple random sample of cars has n=32, $\bar{x}$=132 m

Find the margin of error and 95% confidence interval for the braking distances of cars, if the necessary assumptions are met. Also, σ is known to be 7 m

Confidence interval Exercise - Solution

1. The braking distances of a simple random sample of cars has n=32, $\bar{x}$=132 m

Find the margin of error and 95% confidence interval for the braking distances of cars, if the necessary assumptions are met. Also, σ is known to be 7 m.

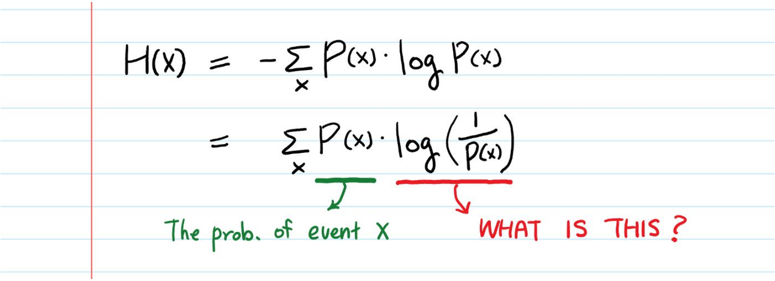
Solution

1. Find z for 95% = 1.96
2. n≥30, so the distribution of sample means is approximately normal.
3. Applying $CI = \overline{X} \pm Z^{*} \frac{\sigma}{\sqrt{n}}$
4. Interval = [129.575,134.425]

Information theory

Information theory is the scientific study of the quantification, storage, and communication of digital information.

The most common measure is entropy. For discrete random variable. Log base 2 is commonly used.



$$H(X) = -\sum_{x} P(x)\cdot\log P(x)$$

$$= \sum_{x} \underbrace{P(x)}_{\text{The prob. of event X}}\cdot\underbrace{\log\left(\frac{1}{P(x)}\right)}_{\text{WHAT IS THIS ?}}$$

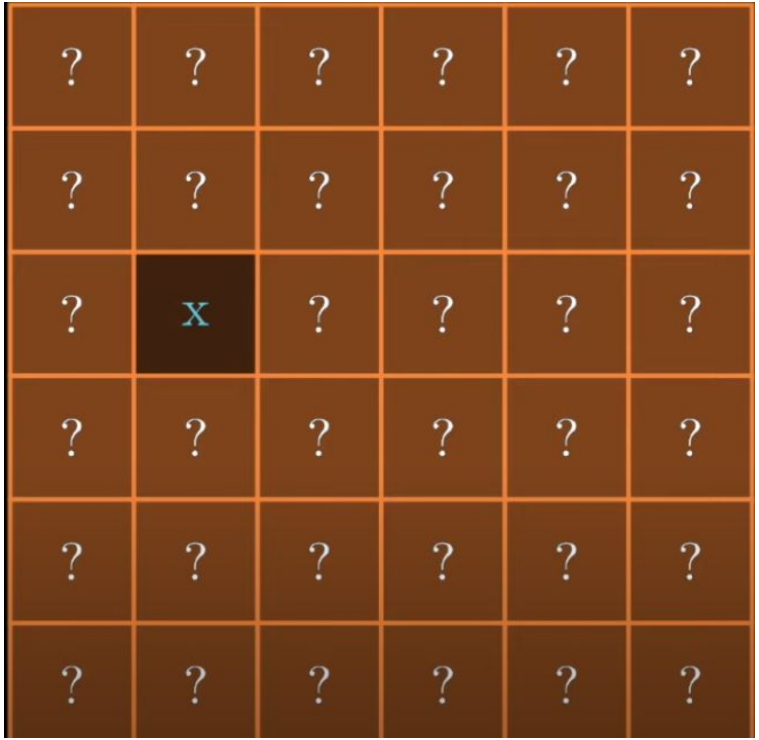
Claude Shannon, father of information theory,
More info: https://en.wikipedia.org/wiki/Claude_Shannon

Simulation time (Extra exercise)

Let's imagine we have a 36 square matrix, where in one of the squares we have a submarine.

Compute the entropy by summing the individual bits of information that you gain, as you make tries.

Video Time!



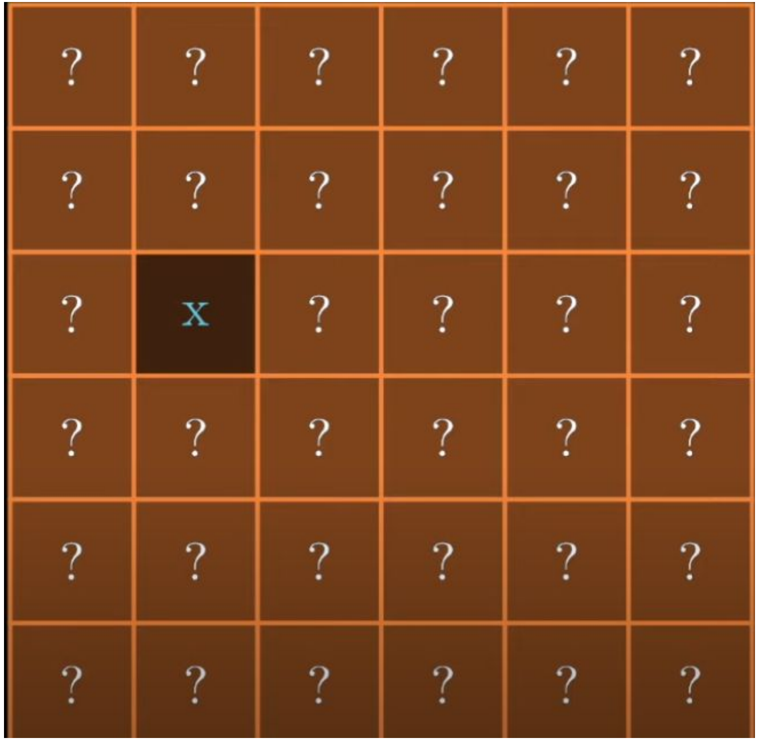
?	?	?	?	?	?
?	?	?	?	?	?
?	X	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?

Simulation time - Solution

Try to reproduce the shannon information simulation presented in the video for the worst case scenario i.e only pick the ship in the last choice. The formula is

$$h(X{=}x) = \log_2 \frac{1}{P(X{=}x)} = -\log_2 P(X{=}x)$$

https://colab.research.google.com/drive/1CucgrRyNs6vJxI9WlRBLFEiiyvrMg9ik?usp=sharing

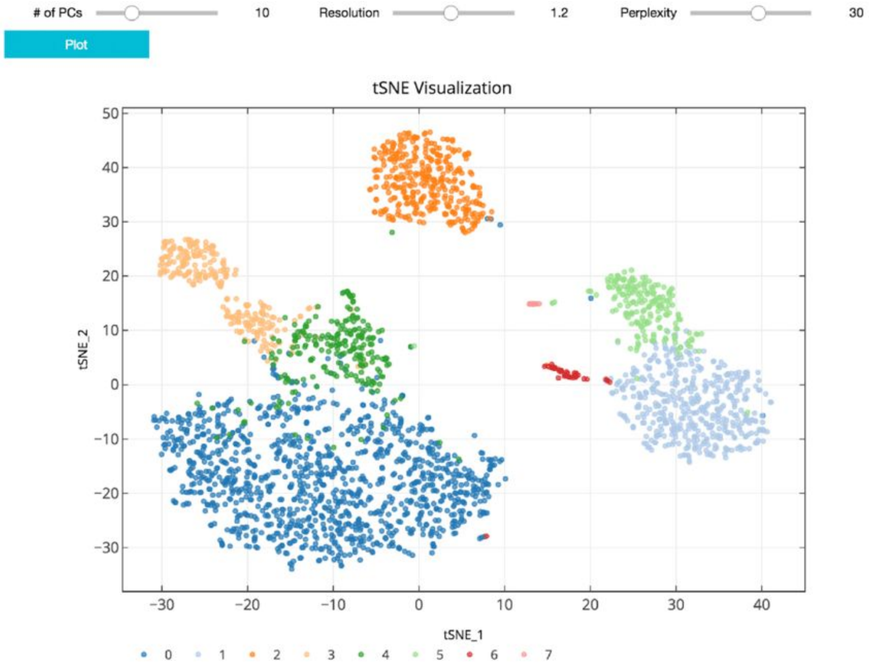


?	?	?	?	?	?
?	?	?	?	?	?
?	X	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?
?	?	?	?	?	?

Kullback–Leibler divergence

Measures of how one probability distribution Q is different from a second, reference probability distribution P.

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

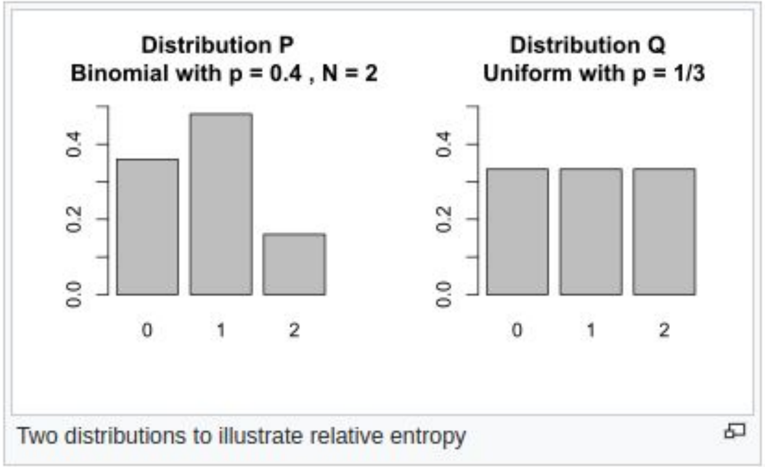


# of PCs: 10

Resolution: 1.2

Perplexity: 30

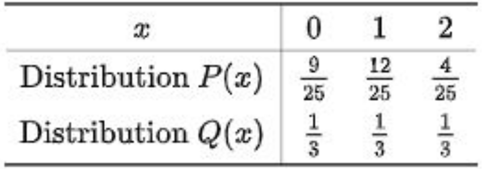
K-L divergence example (Extra exercise)



Distribution P
Binomial with p = 0.4 , N = 2

Distribution Q
Uniform with p = 1/3

Two distributions to illustrate relative entropy

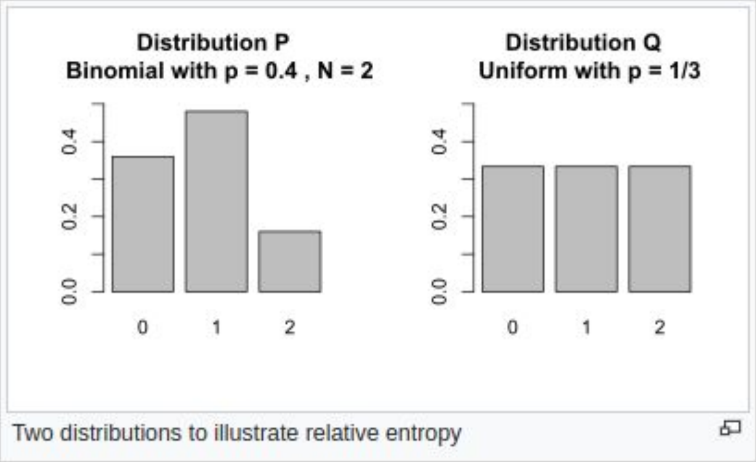


x	0	1	2
Distribution $P(x)$	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$
Distribution $Q(x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

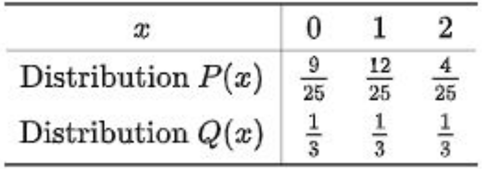
Let's implement in python?

1. Create method
2. Create variables
3. Apply method

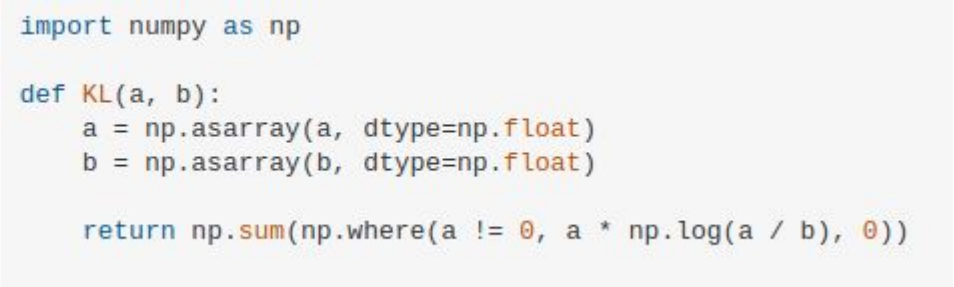
K-L divergence example - Solution



Two distributions to illustrate relative entropy



x	0	1	2
Distribution $P(x)$	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$
Distribution $Q(x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$



```python
import numpy as np

def KL(a, b):
    a = np.asarray(a, dtype=np.float)
    b = np.asarray(b, dtype=np.float)

    return np.sum(np.where(a != 0, a * np.log(a / b), 0))
```

https://colab.research.google.com/drive/1oFvrEsp0aRa2wlNNU4CF5TeJhKcWUFtM?usp=sharing

Tiago Cabo

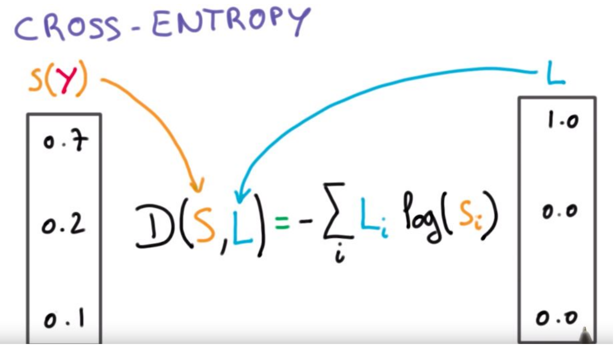
Cross-entropy

Cross-entropy between two probability distributions p and q over the same underlying set of events measures the average number of bits needed to identify an event drawn from the set. If a coding scheme used for the set is optimized for an estimated probability distribution q, rather than the true distribution p.

$$H(p,q) = -\operatorname{E}_p[\log q],$$

$$H(p,q) = H(p) + D_{\mathrm{KL}}(p \parallel q).$$

$$H(p,q) = -\sum_{x\in\mathcal{X}} p(x)\,\log q(x)$$



https://medium.com/data-science-bootcamp/understand-cross-entropy-loss-in-minutes-9fb263caee9a
https://machinelearningmastery.com/cross-entropy-for-machine-learning/

Cross-entropy and KL divergence use case (Extra exercise)

1. Create 3 normal distributed sets of data (500 samples) with the following parameters
   a. X = N(0,1)
   b. Y = N(0, 1.2)
   c. Z = N(0.1, 1.1)
2. Plot the data as if is were time series data
3. Compute PDF with 10 bins
4. Create functions for KL (using **log2**) and cross entropy
5. Let's analyse results

Cross-entropy and KL divergence use case (Extra exercise)

1. Create 3 normal distributed sets of data (500 samples) with the following parameters
   a. X = N(0,1)
   b. Y = N(0, 1.2)
   c. Z = N(0.1, 1.1)
2. Plot the data as if is were time series data
3. Compute PDF with 10 bins
4. Create functions for KL (using **log2**) and cross entropy
5. Let’s analyse results

Possible solution: https://colab.research.google.com/drive/1iCZ8Nq43uo05LCEAPqw9IQx8xb26_IiO