Summary of Statistical modeling and design for experiments

# What is the probability?

THe probability is a numerical measure related to the occurrence of an event, it is used for example to predict the future or to make decisions. Theoretically, the probability is a random phenomenon where we do experiments, for example the toss of a coin or a cube.

In these two cases, it is not certain what is the outcome, but we can predict it in base of the distributions!

THe probability of an outcome is the relative frequency in the long-term period, in fact if we do few experiments, we can’t estimate precisely, the probability, but if we do more of these, we’ll get a better result.

# Definitions

A random experiment is an experiment that give an outcome as a result, it can be repeated in the same manner every time!

A trial is a single performance of a random experiment, an outcome instead is a possible result of a trial. The individual outcomes of a random experiment are not certain, there is a regular distribution in the long run.

# What is the inference?

The inference is the combined result of the claim, the evidence and the reasoning, so it is the process we use to get an information related an event or an experiment.

# What are the possible outcomes?

To answer to this question, we have to find the probability of each outcome, how?

* We identify the sample space, it is the set all possible outcomes and is denotated by S;
* We assign the probabilities, to do this there are some methods from the lottery to the relative frequency to de experience-based.
* We identify the events of interest, they are specific outcomes in the sample space;
* We compute the desired probabilities.

From the event, we can define the definition of simple event, it is the most basic outcome of an experiment. Moreover, there are two special event, the certain and the impossible one, where the probability is 1 or 0 respectively.

We can calculate a theoretical probability like the ratio between the number the favorable outcomes and the total number of outcomes.

# Probability scale

The probability scale is a method for measuring how an event/outcome is certain. Because of the probability is a number between 0 and 1 (so, it can’t be negative!), we can say that an event is certain when it has p=1 and impossible when p=0. If this event is event change, its probability is 0.5, but in the other cases? We can say that the event is Likely if p>0.5 and is unlikely if p<0.5.

If two events haven’t common outcomes, the probability of occurrence of one event is the sum of the individual probabilities. Moreover, the probability of an event can be expressed as 1 minus the probability of not occurrency of the event.

## Empirical probability

The empirical probability is the ratio between the number of occurrences of an event E and the total of experiments we do.

## Law of large numbers

The law of large numbers tells that, the more we repeat an experiment, the more the result is closed to the theoretical probability of that event.

As we can see in the image on the right, the toss of a coin is unbalanced at the beginning, but it rebalance itself when the number of experiments we do belongs to infinite.

## Combining probabilities

To explain this, we do an example: what is the probability of rolling a five with a die and tossing a tail with a coin? How many outcomes are there?

The number of outcomes are all the possible combinations between the two events, in this case there are 12 outcomes because the toss of a coin has 6 outcomes and the toss of the coin has 2 ones.

Which is the probability of each outcome? It is the multiplication of the probabilities of each simple event:

P{5,T}=P{5}\*P{T}=⅙ \*½=1/12

P{3,H}=P{3}\*P{H}=⅙ \*½=1/12

## Birthday problem

Given the probability p that two people in a group have the same birth date, how many people are needed in order to that probability is equals to p?

Imagine that p=0.5, how many pairs of date do we have? We have 365\*365 date!

So, to solve this problem we can do the dual reasoning: we calculate the probability the people don’t share the birth date:

## Algebra of events

We can describe all the things with the algebra of sets, in fact:

* We can consider the events A, B and C and sets;
* If the sets we define previous are the only ones in our problem, we can say that the sample space S is the union of these ones.
* Also, we can define the complement of a set as the set of occurrences that is not in the set A.

We can combine sets to get more complex sets, how? We use:

* the intersection, in this way we can get all the occurrences that is in each set;
* the union, an operation that allow us to get a set containing all the elements of the previous ones.

## Properties

* Commutativity: 𝐴∪𝐵=𝐵∪𝐴and 𝐴∩𝐵=𝐵∩𝐴
* Associativity: 𝐴∪(𝐵∪𝐶)=(𝐴∪𝐵)∪𝐶. Same for the union.
* Distributivity: 𝐴∪𝐵∩𝐶=𝐴∪𝐵∩(𝐴∪𝐶) and 𝐴∩𝐵∪𝐶=𝐴∩𝐵∪(𝐴∩𝐶)
* Identity: A∪∅=𝐴and 𝐴∩∅=∅.
* Complementarity: 𝐴∪ҧ 𝐴=Ωand 𝐴∩ҧ 𝐴=∅.
* De Morgan’s Law: , the reasoning is similar for the union.

## Mutual exclusion

Two events A and B are mutual exclusive if their intersection is null, so they haven’t common elements.

In case of mutual exclusion, the probability of an union is the sum of the probabilities:

## But in general?

In general the the probability of the union is the sum of probabilities minus the intersection:

Why do we substract the intersection? Because they are common elements of the starts events. Without this substraction, we consider the common elements twice.

## Conditional probability

The conditional probability is the probability of an event conditioned by another, we can get it with this formula:

From this definition, we can define the indipendence: two event A and B are indipendent if:

So:

## Bayes’ Theorem

Is there a relation between P{A|B} and P{B|A} ? Yes and it is the Bayes’ theorem! In fact we can get one from the other with this formula:

This equation is the basis for the Bayesian statistical analysis.

# How to select a set

### Variation

Given n=#elements and r=#elementInAGroup (with r<n), V(n,r) is the variation, the number of r ordered groups of different elements we can formed with n elements.

If we do repetitions, we use the following formula:

### Combination

The combination C(n,r) is the number of r groups of different elements we can form with n elements.

With repetition, we use this formula:

### Permutation

The permutation P(n) is the number of different way we can sort n elements.

And with repetition? We use this formula:

Each ri indicates the number of repetition at time i.

# What is a random variable?

A random variable is a numeric description of an outcome of an experiment, we can consider it like a function X:S→R where S is the sample space and R is the set of outcomes.

A random variable can be:

* discrete when it has a countable number of outcomes;
* continuous if this number is uncountable.

In base of the type of the random variable, how is its distribution?

* In case of discrete variables, we have a probability pi associated to every outcome;
* For continuous variables instead, we calculate the probability using an integral.

## Probability functions

A probability function is a function p(x) mapping the possible values of x to the respective probabilities. Every p(x) is a number between 0 and 1 and it is one-sum.

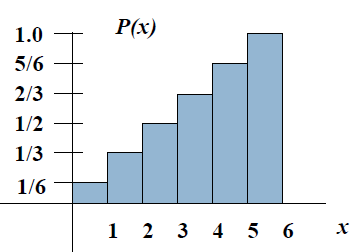
## Cumulative distribution function (CDF)

The CDF is a distribution function defined in the following way:

## Discrete distribution

In discrete distribution, we enumerate every outcome with a number, then we assign to each number a probability. In most of the cases, the probability is the same for each outcome but this is not a rule!

To do an example: every outcome of a die has the same probability that is ⅙.

What can we say if we apply the CDF on this example?

The graph on the right show the CDF of the example, we can see that F(x) belongs to 1 if we go to the highest value.

With the CDF, we can calculate the probability of various events, for example:

* roll a 3 or less:
* roll a 5 or higher:

Essentially, what are we doing? we are summing the probabilities of the interested outcomes!

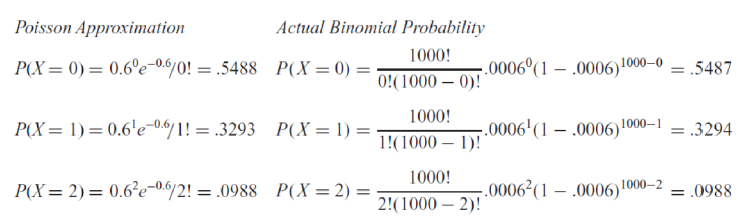
### Which discrete distributions are important?

* The binomial distribution is about binary outcomes and it describe the probability of getting exactly k success in n trials;
* The Poisson distribution indicates the probability of x occurrences in a given area.

When n is large and p is small, the Poisson distribution can approximate the binomial one!

For example: 1000 women are screened for a rare type of cancer, its incidence is about 6 cases per 10.000.

Which is the probability to find at most two cases?



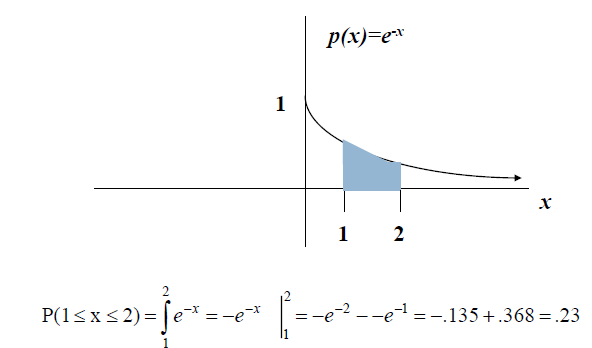
And what about the exponential distribution? With a mean arrival rate of λ per unit of time and x occurrences per t units of time, the Poisson distribution is described by this process:

So, the probability of no occurrences in t units of time is the following:

## Continuous distribution

The continuous distribution assigns the probability of the outcomes by calculating an integral. The limits of this integral depend on the problem we have to solve.

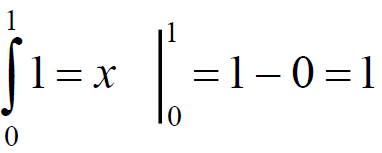
For example:



## Uniform distribution

It is a particular continuous distribution where all the outcomes occurred with the same probability, we can define the pdf in this way:

Because of this is a continuous distribution, we calculate the probabilities using an integral:



# Expected value and variance

The expected value or mean is the ponderated average of a random variable:

The variance of a random variable indicates how the outcomes are far from the mean:

A particular mean is the sample mean: the expected value of a sample of n subjects:

### About the mean

* Given two random variables X and Y, if Y=f(X), we can calculate the mean in this way:
* The mean of a uniform distribution is equal to the arithmetic mean of the limits:
* The mean results for the decision making!

### Properties of the mean

* The mean of a constant is the constant itself:
* The mean of a sum of a constant and a variable is the sum of the same constant with the mean. We can do the same reasoning for the multiplication:
* The mean of the sum of two variables is the sum of the means:

### About the variance

The variance is the expected squared distance from the mean, we can calculate it with this formula:

Like for the mean, we can define the sample variance as the variance of a sample of n subjects:

Why n-1? Because there is a loss of a degree of freedom due to the estimation of the sample mean.

### Properties of the variance

* The variance of a constant is always 0:
* The variance of a sum of a constant and a variable is the variance of the variable, why? We can consider this sum as a shift of the distribution in the space, so it doesn’t affect the shape:
* Regarding the multiplication of a constant with a variable, we can calculate the variance in this way:
* And what about the variance of the sum of variables? It depends if the variable are dependents or not:

So, if X and Y are independent, their covariance is equal to zero!

# Statistical inference

### What is an estimator? And an estimate?

An estimator is a statistic we derive from a sample for inferring the value of a population parameter, instead an estimate is the value of an estimator for a given sample.

### What is the sampling error?

The sampling error is the difference of an estimate from a given population parameter, usually this last one is unknown, so we cannot calculate it. In case there is, we can calculate the error in this way:

In case the population is normally distributed, the sample mean has a normal distribution centered on and with this deviation:

### What is a confidence interval?

A confidence interval is an interval we specify with the probability that this one will contain 𝜇.

So, by the Central Limit Theorem, if the sample is large enough that the sample mean converges to normal, the probability of an interval to contain the true mean is 1-a.

This probability is the confidence level and it is expressed as a percentage.

If we know the variance, how can we calculate the confidence interval? We calculate it using this formula:

And without the variance? We use the sample variance:

### What are the degrees of freedom?

The degrees of freedom are the number of samples that are free to vary. In case of n samples, the number of degrees of freedom is n-1.

### What is a hypothesis?

A null hypothesis H0 is a statement based on equality/similarity, it represents a research we have to prove and expresses the assumption about the value of the population parameter. It always contains an equal sign.

Instead, an alternative hypothesis Ha is a statement based on differences, so it is an assumption we have to demonstrate and it is opposite to H0. For this last thing, it never contains an equals sign about a given value of the population parameter.

#### Examples

* The average connection speed is 54 Mbps, as claimed by the internet service provider.

* The average number of users increased by 2000 this year.

* The service times have Normal distribution.

### Other definitions

The level of significance a is the probability of rejecting H0 when it is true:

A test statistic is a value we use to determine if H0 should be rejected. The critical value is the largest value of a test statistic resulting in the rejection of H0.

The p-value(p) is a probability of an evidence against H0, the smaller is this value, the more it is the evidence against H0.

### Types of error

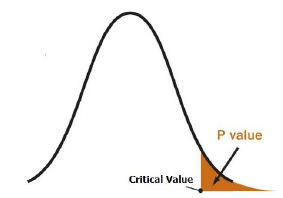
There are four types of error:

* Type 1: rejecting H0 when it is true;
* Type 2: accepting H0 when it is false;
* Type 3: solving the wrong problem;
* Type 4: incorrect interpretation of a correct rejected hypothesis.

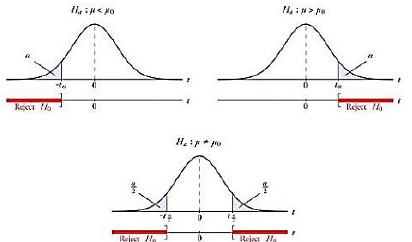
Between type 1 and type 2, which is more important? And why? The type 1 is more important because of the evidence, in fact:

* p-values are only correlated with evidence;
* it is relative. When data are more likely assuming one model is true, they provide evidence for H0.

### Step of hypothesis Testing

1. Define H0 and Ha;
2. Specify the level of significance a and the decision rules;
3. Calculate the test statistic of the hypothesis;
4. Make a decision using p-value or critical value.

### Approaches

The p-value approach uses the test statistic to calculate the p-value, if this value is at most a, we reject H0.

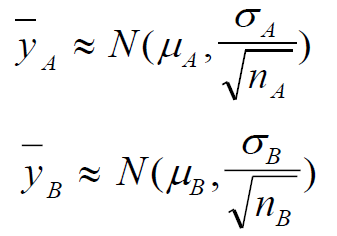
THe critical value approach determine the critical value and the rejection rule using the level of significance a, then we use them to determine if we have to reject H0 or not.

## Comparison of distributions

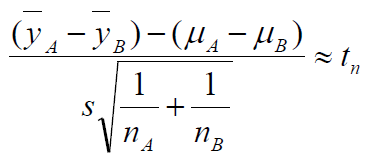
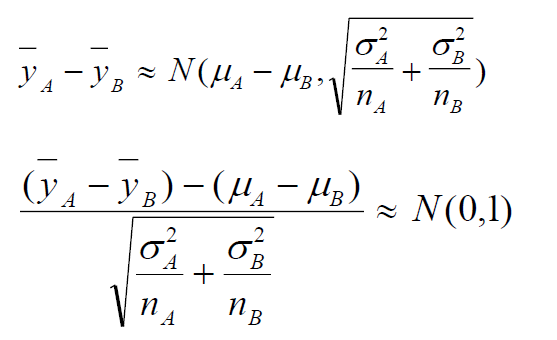
Given the distributions of the rights, what can we say?

We can do an hypothesis test:

So, with the central limit theorem we can get ths:

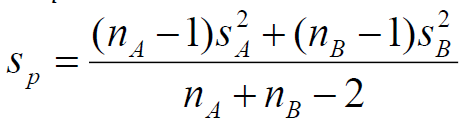
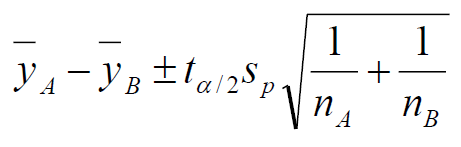


So we can deduce this and calculate the sample variance:

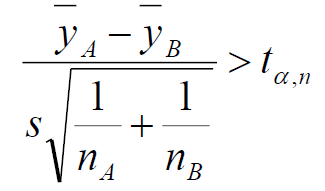


How many degrees of freedom do we have? We have n=nA+bB-2 degrees.

And in case of equal variances? We use these formulas:

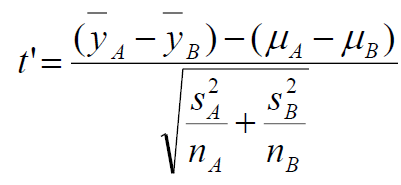


So, we do the test in this way, rejecting H0 if it is true:



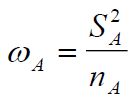
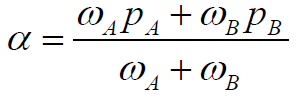
## In case of unequal variances?

In case of unequal variances, we cannot assume it, so we use the sample variances and this formula:



What about the level of significance a in this case?

* if n is equals to nA and nB, we can get the level a using the tof Student table with n-1 degrees of freedom;
* in the other cases, we use the t’ value to find different levels pA and pB, with nA-1 and nB-1 degrees of freedom. <i could write it better>



And about the confidence interval? We calculate it with this formula:

