STATISTICS

Giacomo Castagnetti - Genomics - 2021/2022

INDEX:

- Algebra
- Measure

- Combinatorics
    - Binomial coefficient
    - Fundamental theorem of counting
    - General approach
    - Conditional probability
        - Independence
        - Law of total probabilities
        - Bayes theorem
- Random variables
    - Characterisation of random variables
        - Discrete
            - Probability mass function
            - Cumulative distribution function
        - Continuous
            - Probability density function
            - Cumulative distribution function
    - Constants that describe location, variability and shape
        - Expected value
        - Variance
        - Standard deviation
    - Standardised random variables
    - Moments of a random variable
    - Quantile and percentile
- Interesting random variables
    - Bernoulli random variable
    - Binomial random variable
    - Poisson random variable
    - Gaussian (normal) random variable
    - Chi-squared random variable
    - Student’s T random variable
- Function of a random variable
    - Discrete case
    - Continuous case
        - Distribution
        - Density
- Bivariate random variables
    - Discrete random variables
        - Joint probability mass function

    - Some commonly used systems of hypothesis
  - Taking the decision
    - Empirical evidence
  - Measure of distance from the threshold
    - P value
    - Region of acceptance
  - Type I and II errors
  - Hypothesis testing on the mean
  - Test the difference of two means
  - Testing the association/dependence in contingency tables
    - Independence table
    - Chi-square test

INTRODUCTION

Statistics

Statistics is the art of learning from data

Statistical thinking help us to do scientific and rational decisions, it is the key to understand reality

Statistics is *fundamental to all the areas of science*, every aspect of reality has *uncertainty*, statistics help us to *deal with it*.

Random

How can we define random?

- **Deterministic phenomena**: phenomena that follow precisely a mathematical law, they are predictable, without error.
  Ex. eclipse: it follows a physical law and it is predictable with a negligible error
- **Random phenomena**: phenomena that cannot be predicted with certainty
  Ex. tossing a coin

There are still some strange phenomena which don't respect this distinction:

Ex. **weather forecast**

Weather prediction models exist, they have hundreds of variables but they are coherent and practical equations.

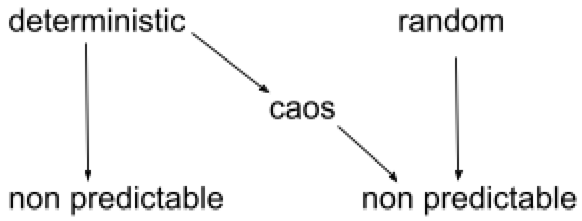
Maybe could exist also a law that can foresee the outcome of the toss of a coin but we are too ignorant to get it, maybe it isn't random

The reality is that:

- Weather forecast is faithful only if it is considered in the near future, only short term predictions are valuable.
- Weather forecast is a deterministic model

In the model that we have described before there is no space for a phenomena which is both deterministic and non-predictable, that's because nature is more complex than what we think.

The true *classification of phenomena* is:



- **Chaotic phenomena**
  A chaotic phenomena is a deterministic phenomena that acts like a random phenomena i.e.(id est) it is not predictable
  This is due to the **butterfly effect**:
    It happens when we get in touch with equations that are noticeably dependent on initial conditions, in those case we would:
    1. Have a perfect equation that describes the phenomena
    2. Put into the equation the data coming from our observations
    3. Get a prediction that heavily depends on our observation

    The tiny error that is committed is exponentially amplified resulting in an unpredictability of the phenomena (it doesn't matter how precise is the instrumentation that is use the measure, the error will still be amplified exponentially)

  The unpredictability of chaotic systems is not due to the ignorance of the equation but it is due to the fact that we are humans and our observations are affected by error.

In the case of a coin toss we could still don't know the law that describe that chaotic behaviour, *we cannot distinguish a random system from a chaotic one with extreme certainty*, maybe it exists a law or maybe the law that we have does not work on the phenomena due to a huge butterfly effect, in those cases *we decide to consider that phenomenon a random one*.
We need probability to deal with the problems generated by the phenomena that we assume to be random

FUNDAMENTALS OF STATISTICS

Introduction

Art of learning from data, study of phenomena that appear to us as random, useful to not get fooled by nature.

Ex: is the vaccination campaign effective?

A patient dies after receiving a vaccine, has this been truly caused by it?

Spurious correlation: is a relation between two events that aren't causally connected;

If the whole population is vaccinated we need a way to not consider the cases of death that would have happened even without the vaccine, statistics help us in this case.

Learn from data:

1. Collect data
2. describe/summarise data
3. Analyse data
   3.5 diagnostics
4. Interpret results
5. Draw conclusions

Each step requires knowledge and experts, there doesn't exist an algorithm that does everything, at the base of algorithms there are theoretical assumptions.

We are never completely sure about our assumptions and models, through diagnostics we can check if the assumptions are correct.

Types of statistics:

- **Descriptive statistics:**
  Description and summarization of the data, conclusions drawn from it are only valid for the specific data set and cannot be generalised;
  During the description we just count something about the phenomenon, without using assumption or probability.

- **Inferential statistics:**
  Allows to draw conclusions from data that can be generalised to the whole population or phenomenon of interest using a mathematical model to comprehend the reality
  Probability is the bridge between descriptive and inferential statistics, it is based on the description of random phenomena through mathematical models, in this way we can understand the behaviour of reality in all the cases that respect the model:
  Probability model: mathematical representation of a random phenomenon.

  Ex.compute the probability of having at least one head in two coin tosses
  We use probability to solve this problem because we are not working on physical tries, we imagine a hypothetical situation, with a unbiased coin, that has the same probability on head and tail:
  $P(H)=½$
  $P(T)=½$
  We take these assumptions as valid and we study their implications and finally we draw conclusions based on such assumptions.

In probability there is no true model, all the models are a partial refiguration of reality, there are only models that are good for the purpose they are created for.

↓

This means that if the purpose changes also the model changes

Purposes:

- ➢ Prediction/classification: you don't want to understand the nature of the model, you just need to take information from a single phenomenon.
- ➢ Description: you want to explain the nature of the phenomenon and create a model useful for similar phenomena
- ➢ Monitoring: you want to know how a phenomenon evolves in time

Ex. of statistical inference:
I have a coin in my pocket and I want to know if it is biassed or not.
We need an idealised coin (mathematical model of a coin) to study a real coin, the model is a benchmark we can use as a comparison to make our assessment.

- **Theoretical statistics:** devoted to the development, derivation and proof of statistical theorems and methods

- **Applied statistics:** deals with the application of such theorems and methods to solve real-world problems

Descriptive statistics

dictionary:

Statistical unit: subject or object about which the information is collected.

Variable: characteristic or phenomenon under investigation, it is indicated using X_n and every statistical unit has a different variable;

There are different **types of variable**, each one requires different analyses:

- **Quantitative:** measured numerically
  - Discrete: it is countable
  - Continuous: non-countable, it can assume any numerical value over a certain interval
- **Qualitative:** not measured numerically
  - Ordinal: the values can be ordered (education level, rating…)
  - Nominal: the values cannot be ordered (blood type…)

Observed sample (data set): set of the observed realisation of the variable, it is indicated using x_n.

The data can be classified according to the period over which they are collected, each type requires a different statistical analysis:

- **Cross-section:** data collected on different units, each taken at one time (time doesn't matter here).

  Ex. taking data from multiple families

  The data matrix contains one variable for each column, every row contains the

information of a single unit.
Ex.

	X1(annual income)	X2(number of children)	X3(father’s education level)
family1	x11	x12	x13
family2	x21	x22	x23

- **Time-series data:** data collected on the same unit at different times (we are interested in the evolution of the phenomenon)
  Ex. annual income of a family over the years

	X1(family1)	X2(family2)
Jan 2021	x11	x21
Feb 2021	x12	x22

<u>Sample size</u>: number of observations *n* (cardinality)

<u>Raw data</u>: sequence of collected data before they are grouped or ranked

Organising and graphing data
In order to better visualise the features of the data we can represent and describe it by using different tables and graphics, according to the type of data that we are studying.

<u>Categorical and discrete numerical data</u>
Tables:

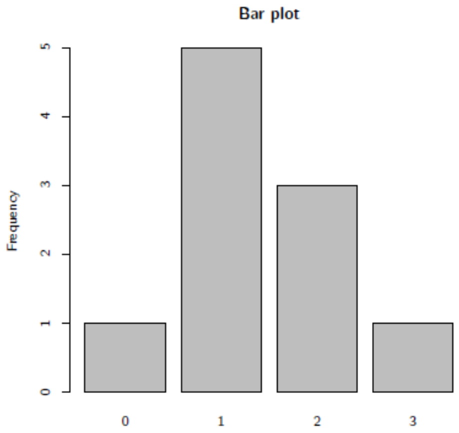
- ➢ Frequency distribution: lists all the categories and the number of elements that belong to each category
- ➢ Relative frequency distribution: lists relative frequencies of each category (frequenci of the category/sum of all frequencies)
- ➢ Percentage distribution: relative frequencies multiplied by 100

Plots:

- ➢ Bar graph (or bar chart): each category is represented by a bar (the height of each bar represents the frequency of the corresponding category)
- ➢ Pie chart: each category is represented by a portion of a circle, in order to represent the relative frequencies.

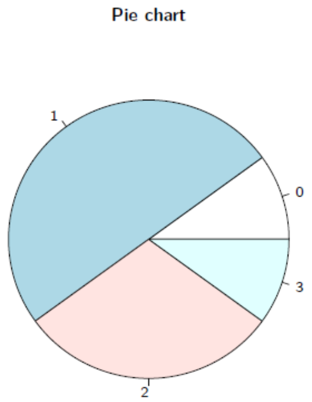
Ex.... number of children per family
Variable: X=number of children
Units: families
Raw data:

family	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
N of children	1	3	1	1	2	2	2	1	1	0



Frequency distribution table:

X	Absolut freq.	Relative freq.	precentage
0	1	0.1	10%
1	5	0.5	50%
2	3	0.3	30%
3	1	0.1	10%
tot.	10	1	100%



Continuous numerical data

Continuous numerical data can be grouped in classes, a class is an interval that includes all the values that are between two numbers that are called the lower and upper limits
The number of classes depends upon the size of the data set and it is preferable to have the same width for all classes (max value - min value / n of classes).
Now the continuous numerical data can be treated just like discrete data, the classes are the categories.

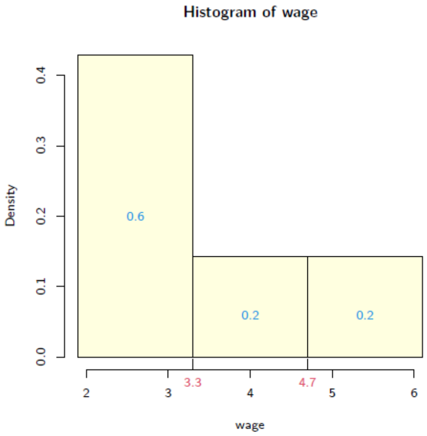
Ex. wage
Variable: X=wage
Units: employees

Raw data:

employee	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
wage	3.1	2.9	5.5	3.5	6.1	2.2	1.9	4.3	2.7	2.5

Frequency distribution table:

class	Abs. freq.	Rel. freq.	perc.
(0,3.3]	6	0.6	60
(3.3,4.7]	2	0.2	20
(4.7,inf]	2	0.2	20
sum	10	1.0	100



Plot:

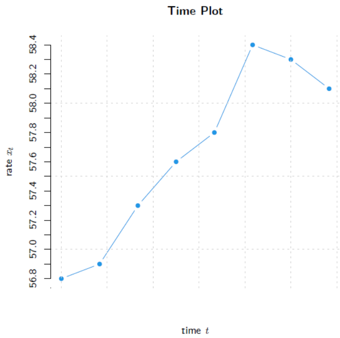
- Histogram: every class is represented by several adjacent rectangles, the length of the rectangle base is equal to the class width and the rectangle area is equal to the class frequency; the height of the rectangle corresponds to the class density (height = area / base = class frequency / class width = class density); it is the ratio between the number of elements that are part of the class and the dimension of the class, in the case of classes with the same amplitude the class density is proportional to the class frequency.

Time-series data

Plots:

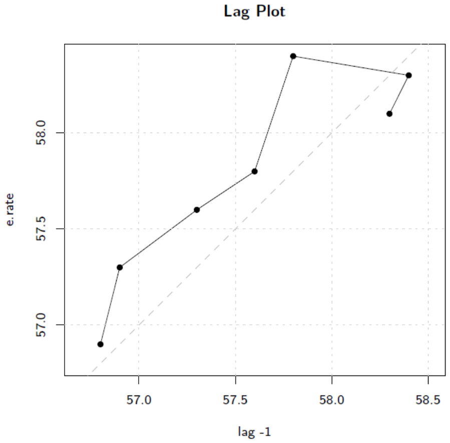
- Time plot (x_t $versus$ t): the horizontal axis corresponds to time and the vertical axis is for the observations, data points are generally connected with straight lines
  Ex. employment rate
  Variable: X=employment rate
  Raw data:

Time t	jan	feb	mar	apr	may	jun	jul	aug
x_t	56.8	56.9	57.3	57.6	57.8	58.4	58.3	58.1



- Lag plot (x_t $versus$ $X_t - 1$): both axes are for the observations but at two different times, it is a scatter plot that highlights the relationship of the phenomenon with the past; it is useful to discover repetitive patterns.

months	e. rate	Lag (e. rate -1)
jan	56.8	NA
feb	56.9	56.8
mar	57.3	56.9
apr	57.6	57.3
may	57.8	57.6
jun	58.4	57.8
jul	58.3	58.4
aug	58.1	58.3
sep	NA	58.1



Numerical descriptive measures

Useful to identify several features of the data, they are called summary measures (summarization of the data).

Divided in:

- **Measures of location**
  - Sample mean: $\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$

    It measures the location because it tells you more or less were the data is concentrated.

    It is not an observable value

    ...Ex... average number of children in a family

    $\overline{x} = \frac{1+3+...+0}{10} = \frac{14}{10} = 1,4$
  - Sample median: It represents the medium value when the data is ranked in order from the smallest to the largest.

    If the cardinality of the sample space is even, the median is the mean between the 2 central values.

    ...ex...median in the number of babies

    0 1 1 1 **1 1** 2 2 2 3 (1+1)/2=1

    If the elements of the sample space are odd, the median is the central value
  - Sample mode: the most frequently observed value

    ...ex... Mode between the number of babies = 1

- **Measures of variability (spread of the data)**
  - Sample variance:average distance of all the observations from the sample mean

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

    It measures the concentration of the data around the mean.

    ...ex… $S^2 = \frac{(1-1.4)^2 + ... + (0-1.4)^2}{10-1} = 0.71$
  - Sample standard deviation: the standard deviation is the $\sqrt{}$ of the sample variance, it is useful because it has the same measurement unit of the data

    $S = \sqrt{S^2}$

    ...ex. $S = \sqrt{0.71} = 0.843$

    We cannot tell if the sample variance is big or not, to be able to do it we should know the maximum and minimum value that it can take, in our case the measure is not normalised.

PROBABILITY

There is not a unique way to describe it, in the history it was associated with dice, the gambling and the first mathematical results were in the 19th century; Kolmogorov laid the foundation of probability theory.

Pragmatic definition of probability

Probability = degree of certainty associated with the result of a random experiment.
Random experiment experiment with two or more outcomes that cannot be predicted, the outcomes are called events:

- ➢ **Elementary event** ω: one of the possible outcomes of the experiment
- ➢ **Compound event**: an event that can be decomposed in terms of elementary events

Probability of an event = number between 0 and 1 that quantifies the uncertainty of an event.
The set of all the possible elementary events ω is the **sample space** Ω.
Ex. tossing two coins...
Ω={(H,H);(H,T);(T,H);(T,T)}
every sample space has a **cardinality**, that corresponds to the number of possible outcomes: $|\Omega| = 4$
Elementary event: ω={(H,H)} $\omega \in \Omega$
Compound event = A = {at least one head} = {(H,H);(H,T);(T,H)} $A \subseteq \Omega$
Every possible event is a subset of Ω.

Other definitions:

- Classic definition of probability:
  If A is an event of Ω (assuming that Ω is countable) then $P(A) = \frac{|A|}{|\Omega|}$
  The probability of an event A is given by the ratio of the favourable outcomes over the total number of possible outcomes, provided that each outcome is equally likely.
  This definition was introduced by laplace and bernoulli, the major problem of it is that all the possible events need to have the same probability; we are trying to define probability using the concept of probability: it's a logical fallacy.
  ...Ex…
  $P(A) = \frac{3}{4}$
- Frequentist definition of probability:
  Based on the idea of relative frequency, the probability of an event (A) is the long-term proportion of times that the event occurs when the experiment is repeated many times (*n*): $P(A) = \lim_{n \to \infty} \frac{nA}{n}$
  The problem of this definition is that it is based on a postulate that cannot be proved: we cannot be sure that carrying on an experiment many times in the physical world will get us closer to the theoretical probability.
- Subjective probability:
  Based on the idea that many experiments cannot be replicated and that everyone states his probability as a consequence of the personal information that possesses.

The probability of an event is the degree of belif that a rational individual attaches to the outcome of the event, based on the information set at his/her disposal.

- Logical probability:
  Degree of support of hypothesis H on bases of event E.

We will stick to the pragmatic definition of Kolmogorov, he assumed that probability was a measure; we need to measure the certainty of an event happening.
We just need measure theory (mathematical theory).

Set theory
We will treat events like sets.

Set world	Event world
Ω	Event that is certain
Ø	Impossible event
A	Event
$\overline{A}$	Event A doesn’t occur
$A \cap B$	Both A and B occur
$A \cup B$	Either A or B or both occur
$A \subseteq B$	If A occurs, then also B occurs
$A \cap B = 0$	If A occurs, then B cannot occur simultaneously

...Ex…
$\overline{A}$:{(H,H)}
$A \cup \overline{A} = \Omega$

Probability space
Defining an algebra
All possible events are subsets of Ω (we want to measure them)
We still don’t have a definition of probability, our objective is to **assign a probability** to all **possible events**.
To do that we need to distinguish two cases:

- Ω has a **finite** cardinality
  We can take all the possible subsets of Ω! through the power set.
  Power set of Ω: set that contains all the possible subsets of Ω.
  The cardinality of the power set is: |$\mathscr{P}(\Omega)$|=$2^{|\Omega|}$
  ...Ex
  $\mathscr{P}(\Omega)$ ={Ω,Ø,(H,H),(T,T)...} |$\mathscr{P}(\Omega)$|=16

Postulate: all the possible subsets form an algebra, we call $\mathcal{A}$ the set of all possible events associated to Ω; If Ω is countable $\mathcal{A} \equiv \mathcal{P}(\Omega)$.

- Ω has an **infinite** cardinality

  $\mathcal{A} \neq \mathcal{P}(\Omega)$, $\mathcal{A}$ contains only the sets to which can be assigned a probability, these are called good sets.

  Not all subsets of Ω are good sets, so $\mathcal{A}$ forms an algebra of the good subsets of Ω.

  A collection of events $\mathcal{A}$ is an algebra if:

  1. $\Omega \in \mathcal{A}$
  2. If $A, B \in \mathcal{A}$ then $A \cup B \in \mathcal{A}$
  3. If $A \in \mathcal{A}$ then $\overline{A} \in \mathcal{A}$

  An algebra $\mathcal{A}$ is an σ-**algebra** if for each **countable** collection of **sets** $\{A_n\}$, $n \in N$, $A_n \in \mathcal{A}$ we have: $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.

  The σ-algebra is a special set with good properties.

  But, which are the not countable/not good sets?

  Suppose that $\Omega \equiv [0, 1]$, how could we define the probability of all the subsets of Ω?

  We have $x \in [0, 1]$, with the same probability for all the elements, so:

  $P(x) = \frac{1}{|[0,1]|} = \frac{1}{\infty} = 0$ that's not a good path, the only solution is ignoring the single elements and considering only the intervals of the set.

  The good sets that compose the σ-algebra are intervals → σ-algebra is not $\mathcal{P}(\Omega)$

Defining the measure

Now that we have defined which elements we can measure we need to assign them a measure: The couple (Ω,$\mathcal{A}$) forms a measurable space (space of the measurable elements); through Kolmogorov's axioms we can associate a number, that represent the probability, to each event: $P = A \in \mathcal{A} \to [0, 1]$ this relation determines a probability measure and creates a probability space: (Ω,$\mathcal{A}$,*P*).

Axiomes (actual measures):

- $P(A) \geq 0$
- $P(\Omega) = 1$
- $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$
- For each sequence of events $\{A_n\}$ such that $A_i \cap A_j = \emptyset$, for all $i \neq j$

  $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Further implied properties:

- $P(\emptyset) = 0$
- $P(A) \leq 1$
- $P(\overline{A}) = 1 - P(A)$
- $P(\overline{A} \cap B) = P(B) - P(A \cap B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ otherwise we count the middle section twice
- $if\ A \subset B\ then\ P(A) \leq P(B)$

To assign the probability to a single event we count: $\frac{favourable\ outcomes}{total\ outcomes}$, **combinatorics** can help us in this job.

COMBINATORICS

Binomial coefficient

$(n, k)\ =\frac{n!}{k!(n-k)!}\ \ where\colon\ n\ \geq\ k\ \ \in\ N$

- Permutations
  n!=number of permutations of n objects (i.e. all possible ordered displays of the elements)
  n!=1*2*3…*n
  o!=1
  Ex. 3!=1*2*3=6 if we have the word ABC the total possible permutations of the letters are 6, all possible orders.

If we take a set w={a,b,c} we only care about the elements that compose the set, not their order; we want to know how many subsets of cardinality 2 is possible to draw from w:
{a,b},{a,c},{b,c} they are three
n=|set|=3
k=|subset|=2
$(3, 2)\ =\frac{3!}{2!*(3-2)!}=\frac{3*2}{2}=\ 3$ combinations of n elements taken k at a time
A set of cardinality n has (n,k) subsets of cardinality k

Ex... 52 cards: 13 numbers, 4 suits
Number of possible poker hands in a deck (5 cards per hand): (52, 5)=2598960=sample space Ω.

Fundamental theorem of counting

Assume a job with k separate tasks, assume that each task can be performed in n ways, then the whole job can be performed in n*n*n ways
...Ex…
Probability of having a poker of aces:
The entire job can be considered as 5 separate tasks:

$$\underbrace{\underline{A}\ \ \underline{A}\ \ \underline{A}\ \ \underline{A}}\ \ \underline{x}$$

for the first 4 tasks there is only one possibility to draw the 4 aces
The last card can be anything except an ace, so 48 cards are remaining and only one space (subset) so the possibilities of x are (48,1).
The overall probability is calculated using the fundamental theorem of counting:
1*48=48possibilities
P(Apoker)=48/(52, 5)

Probability of having 4 of a kind cards
The first 4 cards must be equal between each other, so we have 13 possibilities, the last card is free, but we have to exclude the 4 cards that have already been taken.
P(4 of a kind)=13*(52-4)/(52, 5)

Probability of having exactly a pair (i.e. two equal cards and three different)
We can split the job in four tasks

1. Denomination of the pair: 13 ways
   Two cards must be equal, so we can have 13 possible numbers
2. 2 suits out of four: for each pair we can have all the combinations of the 4 suits
   n=4 k=2 (two available spaces)
   (4, 2)=6
3. The three remaining cards must be different
   We have 12 denominations remaining, we have to take 3 different denominations; we consider all the combinations of 12 elements in 3 available spaces.
   (12, 3)=220
4. Each denomination has 4 suits: for each f the selected denominations we can also have 4 different suits
   4*4*4=64

$P(exactly\ one\ pair) = \frac{13*(4,2)*(12,3)*4^3}{(52,5)} = 0.42$

To obtain all these results it is necessary that the deck is properly shuffled; a statististician who was a magician found out a mathematical theorem that proves that the perfect number of shuffles to maintain the randomness of a deck of cards is 7.

General approach

When we are approaching a probability exercise we have to care about 2 important factors:

1. Sample scheme:
   a. With replacement
   b. Without replacement
2. Order
   a. Ordered
   b. Unordered

	Without replacement	With replacement
ordered	$\frac{n!}{(n-k)!}$	n^k
unordered	(n, k)	(n+k-1, k)

Ex. genetic alphabet={A,C,T,G}

1. We draw all possible triplets
   In this case we consider the elements both ordered and with replacement (AAA exists and AAC is different from CAA)
   In the genetic alphabet we have 4 elements and 3 available slots:
   $n^k = 4^3 = 64$
2. We draw triplets with 3 different letters
   We just have to not consider the replacement (AAC is not considered, CTG is different from GTC):
   $\frac{n!}{(n-k)!} = \frac{4!}{(4-3)!} = 24$
3. We take <u>subsets</u> of 3 letters
   In this case we take <u>unordered</u> and <u>without replacement</u> sets (AAA is not considered, CTG is equal to GTC):
   (n, k)=(4, 3)=4
4. We take subsets of 3 letters, with replacement
   (AAA is considered, AAC is equal to CAA)
   (n+k-1, k)=(4+3-1, 3)=(6, 3)=20
5. Possibility of all the triplets with 2 equal letters
   The first 2 letters can have 4 different outcomes, the last one is free but it has to be different from the others, so we have 3 possibilities.
   We also have to consider all the possible orders of these letters:
   XXY, XYX, YXX 3 other possibilities
   P(2 equal letters)=4*3*3=36
6. Possibility of palindromic triplets
   We have to consider just the triplets that have the side letters equal and the central one of any kind.
   4 possibilities for both the external letters together and 4 possibilities for the central one:
   P(palindromic triplets)=4*4=16

Conditional probability

The conditional probability is the probability that an event A occurs given that an event B has already occurred.

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

Dim.

Suppose we have a probability space (Ω,$\mathcal{A}$,P)

Suppose we have to measurable (good) events $A, B \in \mathcal{A}$

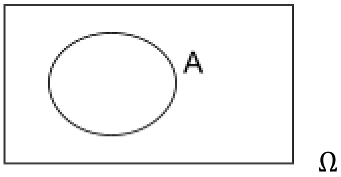
P(A|B)=probability that A occurs given B has already occurred

From the Kolmogorov axioms:

P(Ω)=1, that’s the normalisation of the probability measure, we set a maximum value

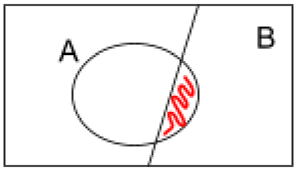
P(A)<1, P(A)=$\frac{P(A)}{P(\Omega)}$

this can be visualised through sets:



The probability of A is the ratio between the area of A and the area of Ω.
We define the area of Ω as 1 so it disappears in the formula.

Now suppose that B occurs:

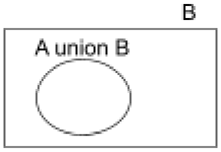


We can redesign the formula of the event:

- The P(A) will now be $P(A \cap B)$
- P(Ω) now is P(B)

So the probability of A now is:

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



P(B) is the new normalisation of the probability measure. q.e.d.

Ex. rolling two dice sequentially
Compute the probability that the sum of the two scores is 10 given that the first roll is a 4.
Ω={(1,1),(1,2)....}
|Ω|=6*6=36
P(ω)=1/36
A, B compound events
A=the sum of two rolls is 10={(4,6),(5,5),(6,4)}
|A|=3
P(A)=3/36
B=the first roll is 4={(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)}
|B|=6
P(B)=6/36
P(A|B)=P(A∩B)/P(B)=1/36 / 6/36=1/6

Joint occurrence of A and B (super important)
$P(A \cap B) = P(A|B) * P(B)$ condition on B
$P(A \cap B) = P(B|A) * P(A)$ condition on A

Let's consider 3 events:
$P(A \cap B \cap C) = P(A|B \cap C) * P(B \cap C) = P(A|B \cap C) * P(B|C) * P(C)$
General formula:
Given A1,A2,...,An ∈ $\mathcal{A}$
P(A1 ∩ A2 ∩ … ∩ An)=P(A1|A2 ∩ … ∩ An)*P(A2 ∩ … ∩ An)=
P(A1|A2 ∩ … ∩ An)*P(A2|A3 ∩ … ∩ An)*P(A3 ∩ … ∩ An)*...*P(An-1|An)*P(An)

...Ex
P(A)=3/36
P(B)=6/36

P(A|B)=1/6
P(A∩B)=P(A|B)*P(B)=1/6*1/6=1/36
P(B|A)=P(A∩B)/P(A)=1/36 / 3/36=1/3

Independence

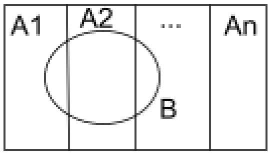
$P(A \cap B) = P(A|B) * P(B)$

If A and B are stochastically independent, when B occurs A doesn’t care, P(A) does’t change.

$P(A \cap B) = P(A) * P(B)$

The probability of A and B is the product of the probability of the single events, there is no added information when considering the joint probability since you can compute the joint from the single.

Law of total probabilities



(A1, A2,..., An) ∈ $\mathcal{A}$ in a partition of Ω
Ai ∩ Aj = 0 for i≠j
$\bigcup_{i=n}^{n}$ Ai = Ω

B = (B∩A1)∪(B∩A2)∪...∪(B∩An)
P(B) = P(B∩A1)+P(B∩A2)+...+P(B∩An)
From one of the axiomes we know that if 2 events are disjoint, the probability of their union is the sum of their probabilities.

= $\sum_{i=1}^{n}$ P(B∩Ai)= $\sum_{i=1}^{n}$ P(B|Ai)*P(Ai)

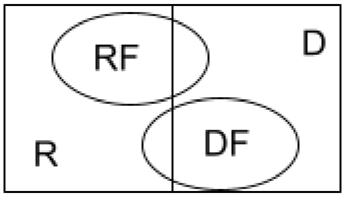
Ex. poker
Possible hands=(52, 5)
Let’s compute the probability of a hand with 4 aces using conditional probability.
A: draw 1 ace
N: drawing another card
P(4 aces)=P(A,A,A,A,N) the comma means ∩
We have a joint probability, we can expand it:
P(A1,A2,A3,A4,N)=P(A1|A2,A3,A4,N)*P(A2,A3,A4,N)=
=P(A1)*P(A2|A1)*P(A3|A2,A1)*P(A4|A3,A2,A1)*P(N|A4,A3,A2,A1)
=4/52 * 3/51 * 2/50 * 1/49 * 48/48=$\frac{4*3*2*1}{52*51*50*49}$
We have to consider also:
P(A,A,A,A,N)
P(A,A,A,N,A)
P(A,A,N,A,A)
P(A,N,A,A,A)
P(N,A,A,A,A)
All these possibilities have the same probability so P(4aces)=5*$\frac{4*3*2*1}{52*51*50*49}$

Ex. weather forecast

1. On 83% of rainy days rain has been forecast
2. On 83% of dry days no rain has been forecast
3. It rains 40% of the days

Compute the probability that rain is forecast
Events:
RF: rain forecast
DF: dry forecast
R: rainy day
D: dry day
data:
P(RF|R)=83%
P(DF|R)=17%
P(DF|D)=83%
P(RF|D)=17%
P(R)=40%
P(D)=60%

To compute the probability of the forecasted rain we need to have a clear understanding of the phenomena, in order to reach this it can be useful to do a graphic representation:

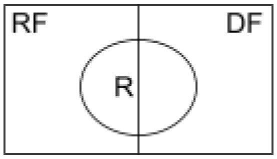


P(RF)=P(RF∩R)+P(RF∩D)
P(RF∩R)=P(RF|R)*P(R)
P(RF∩D)=P(RF|D)*P(D)

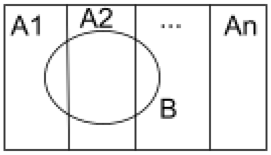
P(RF)=P(RF|R)*P(R)+P(RF|D)*P(D)=0.83*0.4+0.17*0.6=0.434=43.3%

Bayes theorem

Typically we are interested in the probability that will actually rain after that rain has been forecast:
P(R|RF) = probability that rains after that rain is forecast



To compute this we need the Bayes theorem:
Bayes was a priest who founded the bayesian statistics, which is opposed to frequentist statistics; the Bayes theorem is reached through the following procedure:



(A1, A2,..., An) ∈ $\mathcal{A}$ in a partition of Ω

P(Ai|B) = $\frac{P(Ai\cap B)}{P(B)} = \boxed{\frac{P(B|Ai)*P(Ai)}{P(B)}} = \frac{P(B|Ai)*P(Ai)}{\sum_{j=1}^{n} P(B|Aj)*P(Aj)}$

Generally:
B=prior probability
P(B|Ai)=evidence, experiment
P(Ai|B)=posterior probability, updated prior
By repeating the experiment using the updated prior we can reach a more precise result

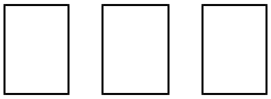
…ex
$P(R|RF) = \frac{P(RF|R)*P(R)}{P(RF)} = \frac{P(RF \cap R)}{R(RF)}$
P(R)=0.4 probability of rain, prior
P(RF|R)=0.83 evidence
P(R|RF)=0.765 posterior
Before the experiment we were 40% sure it would rain, now, after a single forecaster said it will rain, we are 76% sure it will rain.
Suppose that we ask to another forecaster and he tells us that it will rain, the total probability that we can reach with a bayesian reasoning is given by the same formula that we used before but we have to insert the accuracy of the new forecast (evidence) and our as prior probability the updated prior.
Now we can be much more sure that it will rain, that's the way of seeing the world of Bayes.

Ex. disease
A disease D is observed in 4 cases every 1000 people, in order to diagnose it we devise a clinical test T.
To test results positive in 98% of individuals that are affected by the disease and T result positive in 5% of those who are not affected.
Compute the probability that a given individual is affected by D given that the test T result positive.
D=sick rate=0.004=prior prob.
N=healthy rate=1-0.004
T+=test result positive
T-=test result negative
P(T+|D)=0.98=evidence=true positive
P(T-|D)=1-0.98=false negative
P(T+|N)=0.05=evidence=false positive
P(T-|N)=1-0.05=evidence=true negatives

P(D|T+)=$\frac{P(D \cap T+)}{P(T+)}$=?
=$\frac{P(T+|D)*P(D)}{P(T+)}$=
=$\frac{P(T+|D)*P(D)}{P(T+|D)*P(D)+P(T+|N)*P(N)}$=
=0,073
It is counterintuitively very small, that's because the disease is very rare, P(D) near to 0, so to compensate we should need a P(T+|D) near to 0.

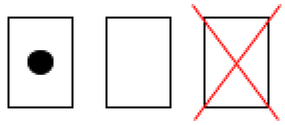
Ex. game of the 3 cards

1 card out of three is an ace, if you pick the ace you win.



Suppose you bet on the first card: P(win)=⅓

Now we get to know that the third card is not an ace

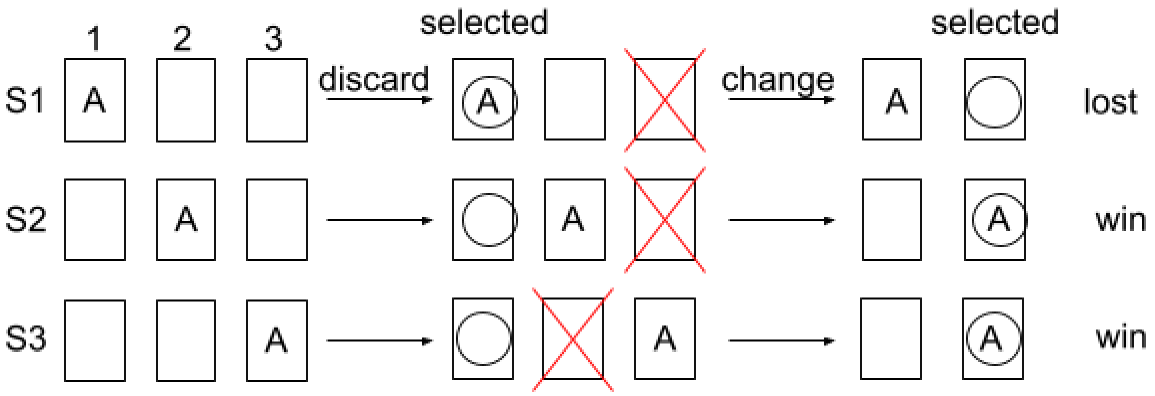


Now we get asked if we want to change our bet.

We have to accept it because if we change the card P(win)=⅔

The counter intuitive element is the fact that someone knows where the ace is and discards a card that is not an ace.

We can examine all the cases:



RANDOM VARIABLES

Ex. tossing two coins

$\Omega=\{(T,T);(H,H);(T,H);(H,T)\}$

$P(\omega)=¼ \qquad \omega_i \in \Omega$

$(\Omega,\mathcal{A},P)$ probability space

Now suppose we want to count the number of heads out of two tosses.

X=number of H in two tosses

X is a new sample space: X={0,1,2}

X is a function that associates a number to every element of Ω

$X: \Omega \rightarrow \{0, 1, 2\}$

P_x=new probability measure $(\{0,1,2\}, \mathscr{P}(\{0,1,2\}),Px)$

$P_x(x=1)$=?

To find the probability we just count:

$\omega \rightarrow \{0, 1, 2\}$

(H,H) → 2

(H,T) → 1

(T,H) → 1

(T,T) → 0

Now we just go back to and use P:

$P_x(x=1)$=Probability that we have 1 head=¼+¼=½

$X^{-1}(1) = \{(H,T), (T,H)\}$

$P_x(x=1)=P(x^{-1}(1))=P(H,T)+P(T,H)=½$

$P_x(x=0)=P(x^{-1}(0))=P(T,T)=¼$

$P_x(x=2)=P(x^{-1}(2))=P(H,H)=¼$

We have defined the probability over x by using the probability over Ω, this is the informal definition of **measurable function**.

More rigorously:

Let $(\Omega,\mathcal{A},P)$ be a probability space.

The function $x: \Omega \rightarrow E \subseteq R$ is such that (E,ξ) is also a measurable space.

Given $c \in \xi$ then $x^{-1}(c) \in \mathcal{A}$**.**

$P_x(c)=P(x^{-1}(c))$ for all c ∈ ξ

$(\Omega,\mathcal{A},P) \rightarrow (E,\xi,Px)$

In everyday life x represents a **random phenomenon** that takes numeric values with a certain probability, we work with the mathematical models that describe those phenomena.

A way to represent x is the probability mass function (PMF), which is a table that sum up all the relations between x and Ω, every point of Ω has a probability mass associated.

x	0	1	2
Px(X=x)	¼	½	¼

Random variables are defined according to the set of variables that they can take; we can distinguish:

- *discrete random variables*: |E| is finite or countable
  Ex. number of defective item in a
  Number of children in a household
  Presence/absence of a given snp(single nucleotide polymorphism) in a gene (yes/no variable)
- *Continuous random variables*: |E| not countable
  Ex. waiting time in a queue
  Proportion of a given gene mutation in the population

Characterization of random variables

Discrete:

- **Probability mass function**

  $P_x(X=x)=P(X^{-1}(x))$

  $\sum_x P_x(X = x) = 1$

  $P_x(X=x) \geq 0 \;\; \forall x$

- **Cumulative distribution function (CDF)**

  $F_x(x)=P_x(X\leq x)$

  It is called cumulative because it is the sum of all the probabilities of the values of X smaller of x

  $F_x(x) = \sum_{\omega\leq x} P_x(x = \omega)$

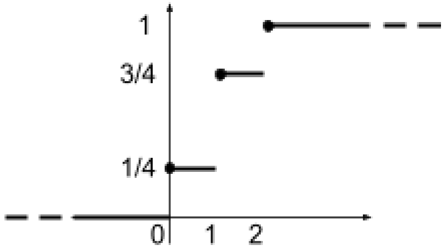
  Obviously we can link the two function together:

x	0	1	2
Px(X=x)	¼	½	¼
Fx(x)	¼	¾	1

  $F_x(1)=P_x(x\leq1)=P_x(x=0)+P_x(x=1)=¾$

  Fx is always defined for the whole real line, so to really represent it we have to display the function:

$$F_X(x)\begin{cases}0 & x<0\\ \frac{1}{4} & 0\leq x<1\\ \frac{3}{4} & 1\leq x<2\\ 1 & x\geq 2\end{cases}$$



Properties:

1. $F_x(x)$ is a non decreasing function
   If x1<x2 ⇒ $F_x(x1) \leq F_x(x2)$
2. $\lim_{x \to -\infty} F_x(x) = 0$
   $\lim_{x \to \infty} F_x(x) = 1$
3. $F_x(x)$ is right continuous
   $\lim_{x \to x_0^+} F_x(x) = F_x(x)$
   Getting closer from the right the limit and the function have the same value, while from the left the limit has a different value.
4. $P(a \leq x \leq b) = P(x \leq b) - P(x \leq a) = F_x(b) - F_x(a)$

Continuous:

$X:\Omega \rightarrow R$

It is not meaningful to speak about Px(X=x) (it is 0); the good measurable elements are intervals of R they form an “algebra of borel”, it contains just intervals.

For instance if X: $\Omega \rightarrow [0, 1]$ uniformly

We cannot compute the probability of ⅓ but we can compute the probability of the interval [0,⅓]

Px(0≤x≤⅓)=λ([0,⅓])=⅓

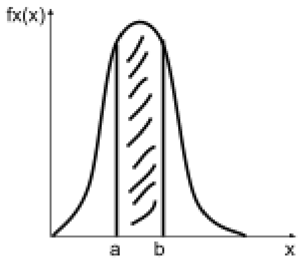
λ is just the length of the interval, it is called lebesgue measure:

λ([a,b])=b-a

We will use this length to compute probability, the function that allows us to use the lebesgue measure for the probability is the probability density function

- **Probability density function**

  f(x) is useful to compute the probability of intervals:



Px(a≤x≤b)=$\int_a^b fx(x)dx$ this is just a lebesgue measure in two dimensions

Properties:

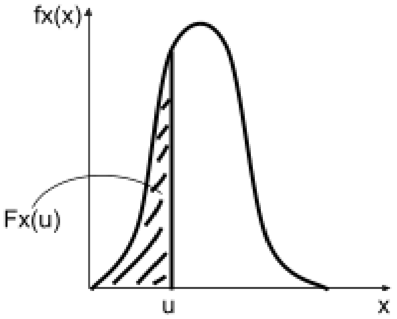
1. $f_x(x) \geq 0 \quad \forall x$
   The integral of a negative function doesn’t have a geometrical interpretation

2. $\int_{-\infty}^{+\infty} f_x(x) = 1$

   The whole area is 1

3. $f_x(x) \neq P_x(X=x)$ has no sense

4. $f_x(x) = \frac{dF_x(x)}{dx}$



   $F_x(u) = \int_{-\infty}^{u} f_x(x)dx$

   $\frac{dF_x(x)}{dx} = \frac{d\int_{-\infty}^{u} f_x(x)dx}{dx}$

   $\frac{dF_x(x)}{dx} = f_x(x)$

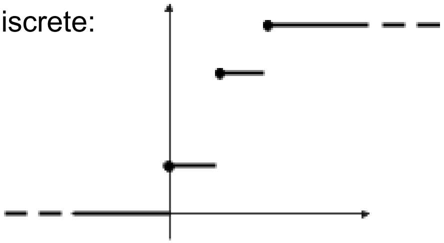
   $dF_x(x) = f_x(x)dx$

   $P(x \in [a,b]) = P(a \leq x \leq b) = \int_{a}^{b} f_x(x)dx = F_x(b) - F_x(a)$

If x is continuous then F is *continuous* and *increasing*.

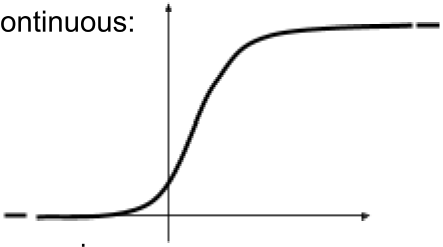
- **Cumulative distribution function:**
  always defined in the whole real line, it is a common representation for both discrete and continuous r.v.

Discrete:



Non decreasing

continuous:



increasing

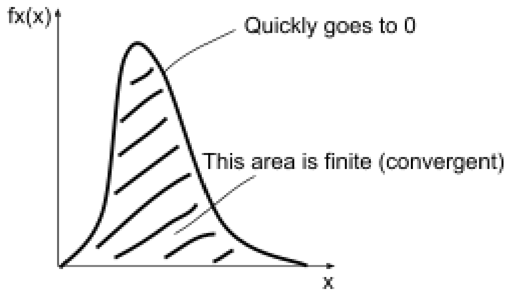
To study the random phenomenon we can use one of the three representations.

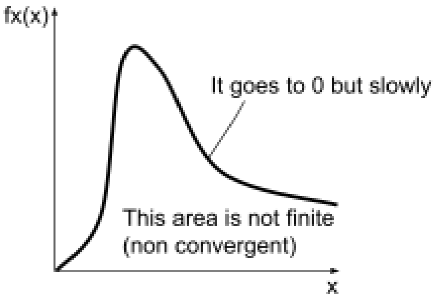
Constants that describe location, variability and shape

- **Expected value (mean) - Location**

$$E(g(x)) = \begin{cases} \sum_x xP_X(X=x)) & if \quad x \quad is \quad discrete \\ \int_{\mathbb{R}} xf_X(x)dx) & if \quad x \quad is \quad continuous \end{cases}$$

Caveat:
E(X) exists provided that the above sum/integral is convergent (finite).



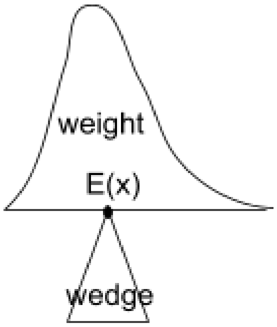


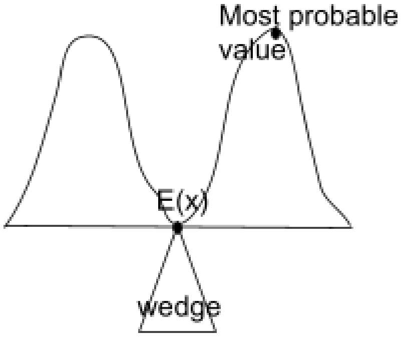
Light tailed density

heavy tailed density
it doesn't have an expected value

What is E(X)?

It is not the most probable or frequent value, it is the centre of gravity of the distribution:





Properties:

1. Given 2 rvs x,y, with E(X) and E(Y)
   E(X+Y)=E(X)+E(X)
   in general: if z=x+y, it is not easily derived from F(X) and F(y).
2. Given n rvs x1,x2,...,x
   and n real constants a1,a2,...,an ∈ R
   then :
   $Sn = \sum_{i=1}^{n} x_i$ Sn is a random variable whose distribution is not known
   E(Sn)=$\sum_{i=1}^{n} E(x_i) = E(x_1) + ... + E(x_n)$
   Linear combination of random variables
   $E(a_1 x_1 + ... + a_n x_n) = a_1 E(x_1) + ... + a_n E(x_n)$
   $If\ a \in R \quad E(ax) = aE(x)$
   The expected value is a linear operator.
3. If X is a random variable with E(x) and Y=g(x)
   In general $F_Y(y)$ cannot be derived from $F_X(x)$
   Suppose Y is continuous:

E(Y)=$\int y * fy(y)dy$ NO

=$\int g(x) * fx(x)dy$

$$E(g(x)) = \begin{cases} \sum_x g(x)P_X(X=x)) & if \quad x \quad is \quad discrete \\ \int_{\mathbb{R}} g(x)f_X(x)dx) & if \quad x \quad is \quad continuous \end{cases}$$

Ex. tossing two coins
x=number of heads

x	0	1	2
Px(X=x)	¼	½	¼

E(x)=$\sum_x xp_x(X = x) = 0 * 1/4 + 1 * 1/2 + 2 * 1/4 = 1$
Y=g(x)=2x-1

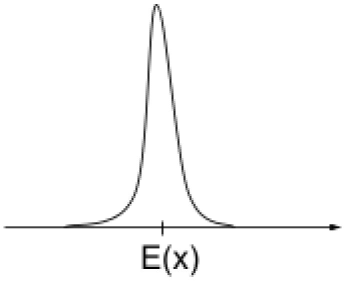
x	0	1	2
Y=g(x)	-1	1	3

E(y)=$\sum_x g(x)P_x(X = x) =- 1 * 1/4 + 1 * 1/2 + 3 * 1/4 = 1$

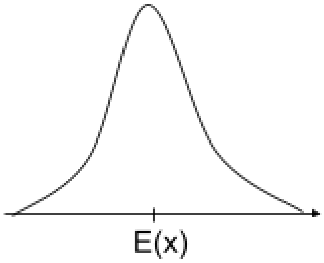
- **Variance - Variability, shape**

$$V(x) = \begin{cases} \sum_x [x - E(X)]^2 P_X(X=x)) & x \quad discrete \\ \int_{\mathbb{R}} [x - E(X)]^2 f_X(x)dx) & x \quad continuous \end{cases}$$

The variance is essentially the expected value of a function of x ($E[(x - E(x))^2]$)
It represents the average distance that x has form the expected value



small variance



large variance

Another **important way in which we can express the variance** is:
V(x)=E(x²)-[E(x)]²
This is useful for computations.

Ex.

E(x)=1

V(x)=1*¼+0+1*¼=½

Properties:

1. Given 2 independent rvs X, Y
   V(X+Y)=V(X)+V(Y)
2. Given n independent vrs $x_1, x_2,..., x_n$ and $a_1, a_2,..., a_n \in R$
   $$V(a_1x_1 + ... + a_nx_n) = a_1^2V(x_1) + ... + a_n^2V(x_n)$$
3. In general, if a ∈ R $V(ax) = a^2V(x)$

- **Standard deviation**
  It is the $\sqrt{\ }$ of the variance
  Sd(x)=$\sqrt{V(x)}$
  This is useful because it has the same measurement unit of X.

Standardised random variables

Let x be a r.v. with:

E(x)=u

V(x)=σ^2

Sd(x)=σ

$Y = \frac{x-u}{\sigma}$ Y is the standardised version of x

$$E(Y) = E(\frac{x-u}{\sigma}) = \frac{E(x-u)}{\sigma} = \frac{E(x)-E(u)}{\sigma} = \frac{u-u}{\sigma} = 0$$

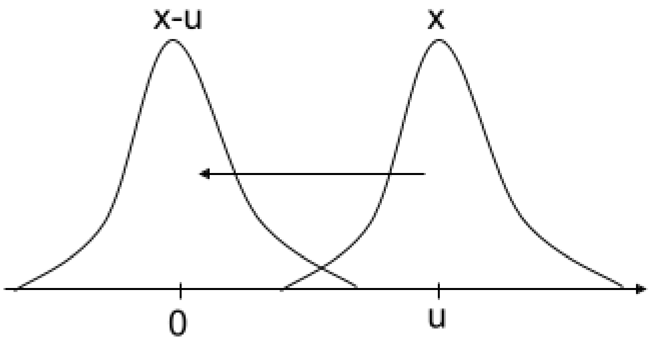
The expected value of a constant is the constant itself

$$V(Y) = V(\frac{x-u}{\sigma}) = \frac{V(x-u)}{\sigma^2} = \frac{V(x)-V(u)}{\sigma^2} = \frac{\sigma^2-0}{\sigma^2} = 1$$

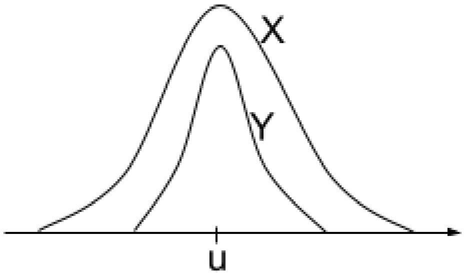
The variance of a constant is zero

Standardising the variable is equal to:

- Centering: with the standardisation we subtract the mean, the r.v. results shifted toward the zero, its mean is zero.



- Rescaling: multiplying X by a scale factor.
  By rescaling we affect the variance, so to obtain a variance equal to one we have to divide by the variance itself:



$\frac{V(X)}{\sigma^2} = 1$ $\quad V(\frac{X}{\sigma}) = 1$ for the properties of variance

$Y = \frac{X}{\sigma}$
Y has no measurement unit

Moments of a random variable

- Random moments of order r

$$\mu'_r = E[X^r] = \begin{cases} \sum_x x^r P_x(X = x) & x \quad discrete \\ \int_{\mathbb{R}} x^r f_x(x)dx & x \quad continuous \end{cases}$$

- Central moments of order r

$$\mu_r = E[|X - E(X)|^r] = \begin{cases} \sum_x |x - E(X)|^r P_x(X = x) & x \quad discrete \\ \int_{\mathbb{R}} |x - E(X)|^r f_x(x)dx & x \quad continuous \end{cases}$$

What information do they give us?
Notable cases:

- r=1 $\mu'_1 = E(X)$
- r=2 $\mu_2 = E[|X - E(x)|] = V(X)$

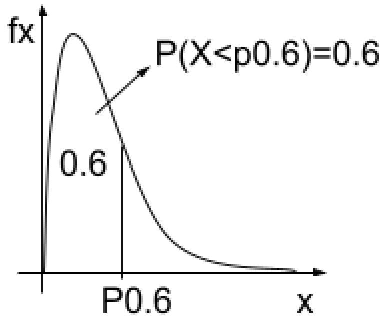
Moments provide an alternative way to characterise a random variable X

- r=3 third order moments give information on the symmetry of the distribution of X
- r=4 fourth order moments give information on the shape of the distribution of X

Quantile and percentile

If we want a portion of the probability density we need

- Quantiles:
  Given a random variable the α quantile is the value P_α such that :
  $F(P_\alpha) = P(X \leq P_\alpha) = \alpha \qquad 0 \leq \alpha \leq 1$

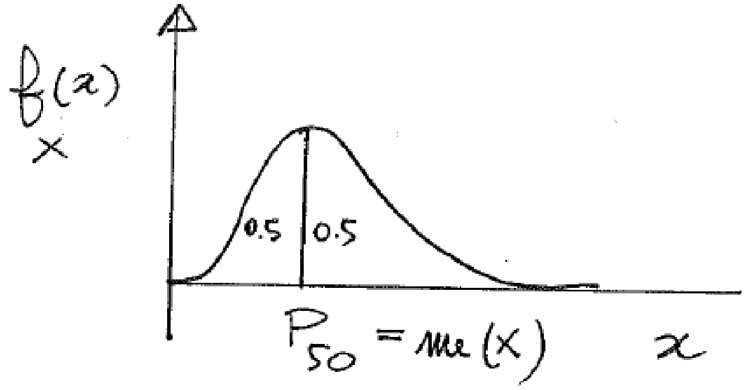


In this case α (quantile) is 0.6, so the P value of α is a certain value x so that P(X<value of x)=0.6

- Percentiles:
  $i = \alpha * 100$
  Given a random variable the i percentile is the value P_i such that:
  $F(P_i) = P(X \leq P_i) = \frac{i}{100}$ $\qquad 1 \leq i \leq 100$

  P0.6 (quantile) is also the P60 (percentile).

There is one P_α (or P_i) that is important
Assume a density with this shape:



P50 is the median me(X)

P(X>me(x))=P(X<me(x))=1/2

The median is a measure of location, it is the value that splits the distribution in two halves

INTERESTING RANDOM VARIABLES

There are random variables that describe typical phenomena, they have the name of who discovered them.

Bernoulli random variable

$X \sim Ber(\pi)$

It provides a model for the experiments with 2 possible outcomes, typically it is denoted with ‘,, failure, or 1, success.

$$X = \begin{cases} 1 & P_X(X=1) = \pi \\ 0 & P_X(X=0) = 1-\pi \end{cases}$$

$P(X = x) = \pi^{x}(1 - \pi)^{1-x} \quad x \in \{0, 1\}$

- x=1: $\pi^{1}(1 - \pi)^{1-1} = \pi$
- x=0: $\pi^{0}(1 - \pi)^{1-0} = 1 - \pi$

E(X)=π

V(X)=$\pi(1 - \pi)$

Binomial random variable

$X \sim Bin(n, \pi)$

Useful when we repeat multiple times a random experiment with two outcomes.

The B.R.B allows to assign a probability to the number of successes in n bernoulli trials

X={0,1,n}

$P(X = x) = (n, x)\pi^{x}(1 - \pi)^{n-x} \qquad x \in \{0, 1, ..., n - 1, n)$

Ex.

n=3

$P(X = 1) = (3,\ 1)\,\pi^{1}\,(1 - \pi)^{3-1} = \frac{3!}{1!\,(3-1)!} * \pi(1 - \pi)^{2} = 3 * \pi(1 - \pi)^{2}$

3 represents the number of favourable combinations: 1 0 0, 0 1 0, 0 0 1

$\pi(1 - \pi)^{2}$ is equal to the probability of a generic combination with 1 success and 2 failures

1 0 0

π $\quad 1 - \pi$ $\quad 1 - \pi$ $\quad$ P(1,0,0)=$\pi(1 - \pi)^{2}$

E(x)=nπ

V(x)=(1-π)

Properties of binomial coefficient:

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

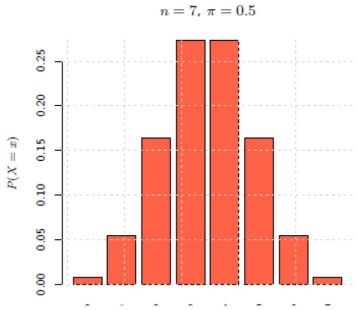
1. $(n,\ k) = (n,\ n - k)$
2. $(n, 0) = (n,\ n) = 1$

   $(n,\ 1) = (n,\ n - 1) = n$

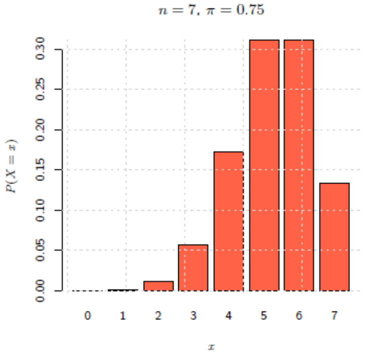
3. $\sum_{k=0}^{n} (n,\ k) = 2^{n}$

4. $(a + b)^{n} = \sum_{k=0}^{n} (n,\ k)\, a^{k}\, b^{n-k}$

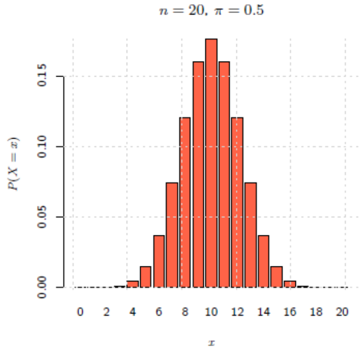
P.M.F of a binomial r.v.



Supposing we flip an unbiased coin ($\pi = 0.5$) 7 times, if we record the number of heads (successes) we found a distribution like this, the most probable result is to obtain 3 or 4 successes, while the others have a smaller probability and they are symmetrical.



If we are working with a biased coin, the probability of success could be higher ($\pi = 0.75$) as a consequence the graph loses its symmetry, the probability of not having any success will be lower than having all successes.



If the number of tries increases the P.M.F. gets closer to the density of a continuous gaussian

Poisson random variable

It is a discrete random variable over N={0,1,...}

$P(X = x) = \frac{\lambda^{x} e^{-\lambda}}{x!} \qquad \lambda \in R^{+},\ x \in N$

$x \sim poi(\lambda)$ where λ is called intensity

This random variable provides a model for the occurrence of **rare events** in a time interval [0,t] (we are observing how many rare events happen in an interval).

Under the following assumptions:

1. The time interval can be divided in sub-intervals such that the probability of observing 1 event in each sub-interval is constant.
2. The probability of observing more than 1 event in each sub interval is 0 (we are dealing with rare events).
3. The occurrence of these events is independent

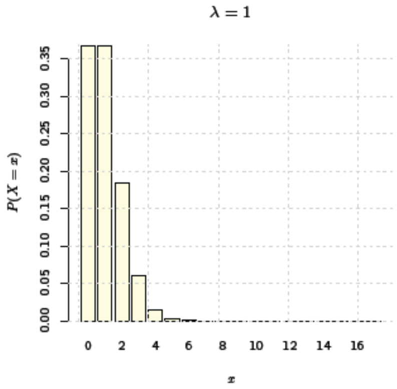
Ex. car accidents, hearth quakes, number of people struck by lightning

Actuarial sciences: professions that study statistical models for insurance companies , they use the same probability models that are used by bookmakers to protect from rare events (hedging, this branch of statistics is called extreme statistics and deals with rare events)

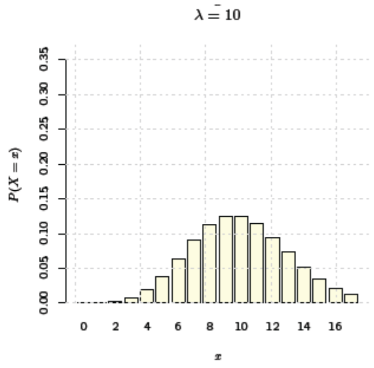
Poisson variables count the number of successes in a small interval, we can consider each interval as a Bernoulli model.
In each sub-interval we have Bernoulli, but what is the difference?
Poisson can be derived from the binomial:
$x \sim Bin(n, \pi)$

E(x)=$n\pi$
When:
$\pi \to 0$
$n \to +\infty$
We assume that $n\pi$ is constant, so the expected value in the time we are observing is finite and constant

In this particular case the phenomenon can be described by a poisson variable:
$x \sim Poi(\lambda = n\pi)$
PMF:



The intensity of the phenomenon is very low, there is a big probability of not observing the phenomenon



It’s unlikely that it doesn’t happen

Convergence

If $X \sim poi(\lambda)$

$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$ x=0,1,2…

$\sum_x P(X = x) = \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = 1$

proof:

$e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$ (it corresponds to taylor series expansion: $e^x = \sum_{x=0}^{\infty} \frac{x^k}{k!}$)

$\sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1$

Gaussian (normal) random variable

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

$f_X(\lambda) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

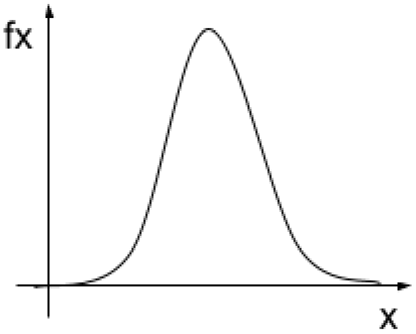
It is a continuous random variable over R.

It is the most important random variable because it fits a lot o f phenomena.

E(x)=μ

V(x)=σ^2

Shape:



Changing the mean, the distribution gets shifted.

Changing the variance, the density around the mean gets changed.

Properties:

1. **A linear combination of independent gaussian random variables is also a gaussian**
   That's super important because usually: Xrv + Yrv = we don't know
2. A linear transformation of independent gaussian random variables is also a gaussian

Gaussianity is a paradigm for linearity, linear phenomena are likely to be gaussian, complex phenomena are not gaussian

Standard normal random variable

Let's standardise the gaussian:

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

$Z = \frac{X-\mu}{\sigma} \quad is \quad Z \sim N(0, 1)$

Whose PDF is:

$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \qquad z \in R$

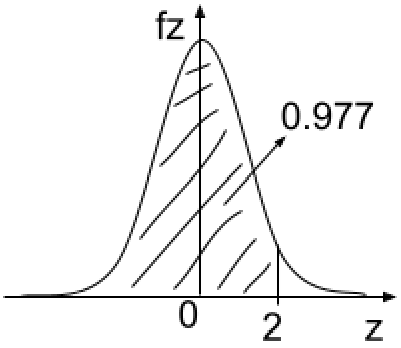
Distribution function of $Z \sim N(0, 1)$

$\Phi(x) = P_Z(Z \leq z) = \int_{-\infty}^{z} f_Z(u) du$ (we use “u” because z is already used in the integral it is a number, we need a new variable)

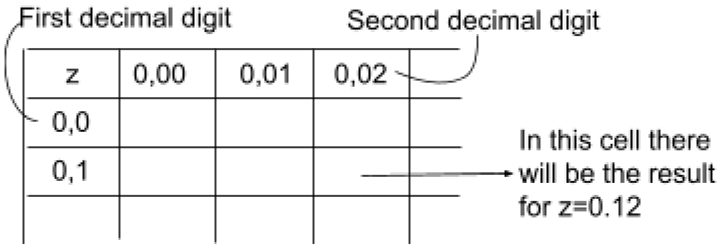
This equation has no analytic solution, we need to solve it numerically

Ex. $\Phi(2) = P_Z(Z \leq 2) = 0.997$

Graphic interpretation:



Considering that we are working with a standardised variable we have tables with all the results for each value of x



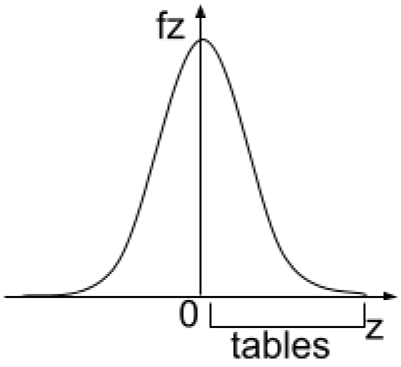
First decimal digit

Second decimal digit

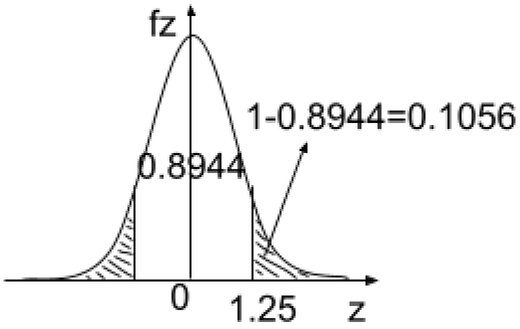
z	0,00	0,01	0,02	
0,0				
0,1				

In this cell there will be the result for z=0.12

Tables only have positive values of x:



What if we want a negative value?
Ex.
$\Phi(1.25) = P(z \leq 1.25) = 0.8944$
$P(z > 1.25) = 1 - \Phi(1.25) = 0.1056$



The gaussian distribution is symmetrical:
$\Phi(-1.25) = P(z \leq -1.25) = 1 - \Phi(1.25) = 0.1056$
(The gaussian random variable has light tails, it goes to 0 quickly)
In general, if z is negative:
$\Phi(z) = P(Z \leq z) = 1 - \Phi(|z|)$

Now let's compute $P(a \leq z \leq b) = \Phi(b) - \Phi(a)$
For instance:
$P(-1.96 \leq z \leq 1.96) = \Phi(1.96) - \Phi(-1.96) = \Phi(1.96) - (1 - \Phi(1.96)) = 2\Phi(1.96) - 1 =$
$= 2 * 0.975 - 1 = 0.95$

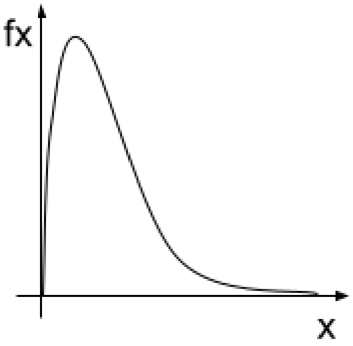
Reverse statistical tables
Another way to use statistical tables is to use them the other way around; namely, fund z such that P(Z≤z)=0.99 (z is the $P_{0.99}$ percentile)
$P_{0.99} = 2.33$

When we are normalising a gaussian we are simply applying the first property of gaussians; the linear transformation of a gaussian is a gaussian.
Ex.
$X \sim N(1, 4)$
$P_X(x \leq 2) = ?$ to find this value we can normalise the variable
$P_X(x \leq 2) = P_X(\frac{x-1}{\sqrt{4}} \leq \frac{2-1}{\sqrt{4}})$ we need to put x inside the formula of z
$P_Z(z \leq \frac{1}{2}) = \Phi(\frac{1}{2}) = 0.6915$
In general, for a gaussian random variable, with a generic E and a generic V:
$F_X(x) = P_X(X \leq x) = P_X(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}) = P_Z(z \leq \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma})$

Chi-squared random variable χ^2

It is a continuous random variable, it is defined over R^+ and it depends just upon a single parameter: ν = degrees of freedom (df).

It is skewed (asymmetric)



$X \sim \chi^2(\nu)$ $X_{\nu\to\infty} \sim N$

As ν diverges the χ^2 tends to a gaussian

E(x)=ν

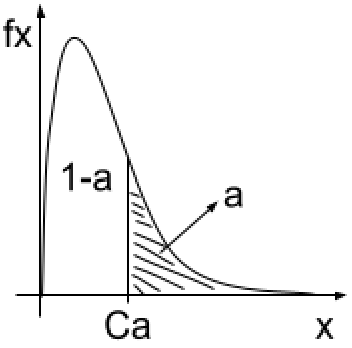
V(x)=2ν

If we have Z1,Z2,...,Zn which are independent and identically distributed random variables (i.i.d.) and $Zi \sim N(0, 1)$

Then $X = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$ the chi-square is just the sum of multiple squared gaussian

There is no standard chi-square because it is obtained by elevating to the second power, for each value of x we have a different distribution, so we display on the tables just some important percentiles: the critical values.

Critical values(C_α): they are another way of looking at quantiles/perce



C_α is such that P(x>C_α)=α

In other words C_α=$P_{(1-\alpha)} = q_{1-\alpha}$

Ex. derive the critical values at level $\alpha = 0.01$ for ν=3 ($X \sim \chi^2(3)$).

we can call it $\chi^2_{0.01;3}$

$P(X > \chi^2_{0.01;3}) = 0.01$

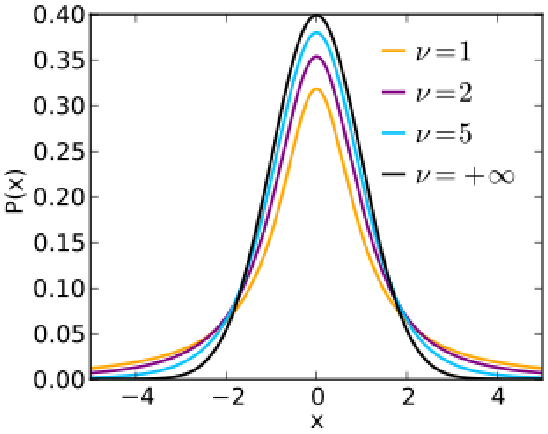
$\chi^2_{0.01;3} = 11.345$

$\alpha = 0.99$

$P(X > \chi^2_{0.99;3}) = 0.99$

$\chi^2_{0.99;3} = 0.115$

$P(0.115 \le X \le 11.345) = 0.98$



Student's T random variable

It is defined over R

$X \sim T(\nu)$

E(x)=0

V(x)=$\frac{\nu}{\nu-2}$ if $\nu > 3$ otherwise the variance doesn't exists

Student's has heavy tails, especially for small ν.

While, for a huge number of degrees of freedom we have: $T(\nu)_{\nu\to\infty} \to N(0, 1)$

Connection with the gaussian:

If $Z \sim N(0, 1)$ and $Y \sim \chi^2(\nu)$

$X = \frac{Z}{\sqrt{\frac{Y}{\nu}}} \sim T(\nu)$ (the gaussian is divided for something else, the first property of gaussians doesn't apply)

When $\nu \ge 40$ then $T(\nu) \approx> N(0, 1)$

Student's T tables have only critical values, actually only the first 40 values because considering more than that would be equal to considering a gaussian distribution.

It also presents the value for infinite degrees of freedom that corresponds actually to a gaussian.

The common trait of chi-squared and student's T is that they are connected to the gaussian.

FUNCTION OF A RANDOM VARIABLE

If we have a random variable and we transform it, the result will be another random variable, with unknown distribution.

Discrete case

Given a r.v. X

X: (Ω,𝒜,P) → (E,ξ,Px)

Y=g(x) is a new random variable (it has different density, distribution, expected value and so on).

Starting from X, having Px, fx, Fx can we derive Py, fy, Fy?

We have to reason on transformations:

Let X be a discrete uniform r.v. over E={-1,0,1,2}

PMF:

x	-1	0	1	2
$P_X(X = x)$	1/4	1/4	1/4	1/4

Now let's add a transformation:

Y=g(x)=x^2

To find out the characteristics of the new random variable we need to:

1. Find the domain of y

   We have to associate to each x the correspondent y:

x	-1	0	1	2
y	1	0	1	4

   Py={0,1,4}

   Now we can write down the PMF of y, to do so we need to find the reverse function, we need to go back from each value of y to the probability of all the corresponding x.

   X=$g^{-1}(y) = \sqrt{y}$

y	0	1	4
$g^{-1}(y) = x$	0	{-1,1}	2

   Now we compute the probability of each y by summing the probabilities of the corresponding x:

2. Derive $P_y(Y = y)$

   $P_y(y = 0) = P_x(x = g^{-1}(0)) = P_x(x = 0) = 1/4$

   $P_y(y = 1) = P_x(x = g^{-1}(1)) = P_x(x \in \{-1, 1\}) = 1/2$

$P_y(y = 4) = P_x(x = g^{-1}(4)) = P_x(x = 2) = 1/4$

Eventually we have PMF:

y	0	1	4
$P_X(X = x)$	1/4	1/2	1/4

In general:
Given X with px
X defined over the probability space
(Ω,𝒜,P) → (E,ξ,Px)
Y=g(x): (E,ξ,Px) → (F,𝓕,Py)
Py(A)=Px($g^{-1}(A)$)
We have to go back, $g^{-1}(A) \in$ ξ

Continuous case

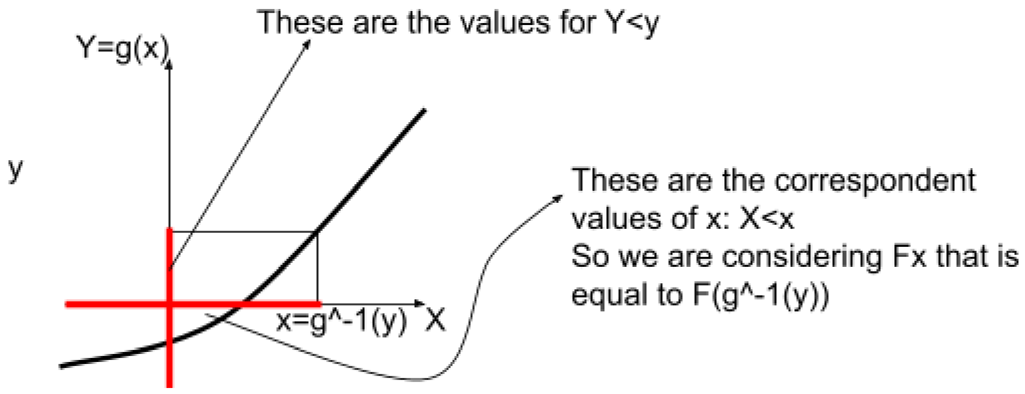
X is a continuous r.v.
X: (Ω,𝒜,P) → (R,β_R,Px)
If g(x): R → R is **invertible** (we will need this characteristic to go back)
An invertible function is obviously monotone, it can be increasing or decreasing

Distribution

a. If g is increasing

Then $F_Y(y) = F_x(g^{-1}(y))$

Proof:

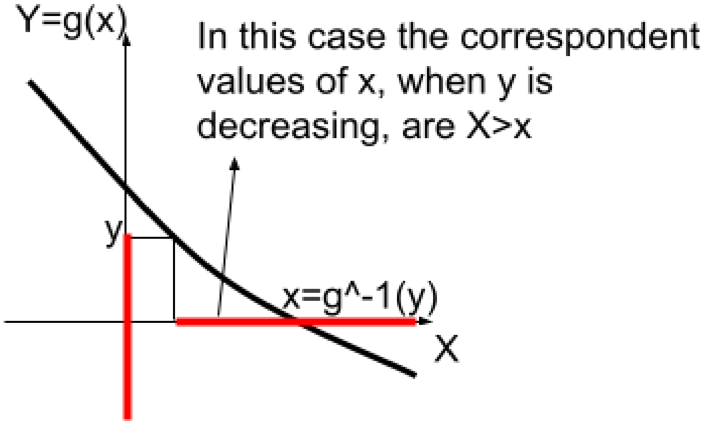

These are the values for Y<y
Y=g(x)
y
These are the correspondent values of x: X<x
So we are considering Fx that is equal to F(g^-1(y))
x=g^-1(y)
X

$F_Y(y) = P_Y(Y \le y) = P_X(X \le g^{-1}(y)) = P_x(X \le x) = F_X(x) = F_X(g^{-1}(y))$

If y decreases (we are considering the values of the distribution so all the values lower than a threshold) x decreases.

b. If g is decreasing

Then: $F_Y(y) = 1 - F_x(g^{-1}(y))$


Y=g(x)
In this case the correspondent values of x, when y is decreasing, are X>x
y
x=g^-1(y)
X

$F_Y(y) = P_Y(Y \leq y) = P_X(X \geq g^{-1}(y)) = 1 - P_X(X \leq g^{-1}(y)) = 1 - F_X(g^{-1}(y))$

We can do these proofs considering x as $g^{-1}(y)$ and without writing it, as above.
In this case, when y decreases x increases.

Density

If f is a continuous r.v. with $f_X(x)$

and $g(x)$ is invertible and differentiable with continuous derivative
Then $Y = g(x)$ has the following density

$f_Y(y) = f_x(g^{-1}(y)) * |\frac{dg^{-1}(y)}{dy}|$ (both for increasing and decreasing distributions)

BIVARIATE RANDOM VARIABLES

A bivariate random variable allows us to study the joint behaviour of 2 random phenomena.

Discrete random variables

Ex. tossing 2 coins

Ω={(H,H),(T,H),(H,T),(T,T)}

P(ωi)=¼ $\omega i \in \Omega$

X:

x=number of heads

$\omega \rightarrow X(\omega)$

Dx={0,1,2}

Now we add another variable, for each outcome of Ω we will associate 2 values.

Y:

$$y = \begin{cases} 1 & equal \quad outcomes \\ 0 & different \quad outcomes \end{cases}$$

$\omega \rightarrow (X(\omega), Y(\omega))$

(H,H)→(2,1)

(T,H)→(1,0)

(H,T)→(1,0)

(T,T)→(0,1)

Domain (bidimensional):

D(x,y)=Dx*Dy={0,1,2}*{0,1}

We can represent it with a two dimensional grid:

X\Y	0	1
0		
1		
2		

Joint probability mass function

We need to create a mass function and assign a probability to all these events

Objective: Pxy(X=x,Y=y)

All the other points have probability 0 because we have already reached a mass of 1:

Joint probability mass function = $\sum_x \sum_y P_{xy}(X = x, Y = y) = 1$

So we have obtained something like this:

X\Y	0	1
0	0	¼
1	½	0
2	0	¼

1

Now we can define all put properties in two dimensions, it is just needed a math effort

Joint distribution function

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) = \sum_{u\leq x}\sum_{v\leq y} P(x = u, y = v)$$

Ex.

$$F_{xy}(1, 0) = P(x \leq 1, y \geq 0) = \frac{1}{2} + 0 = \frac{1}{2}$$

X\Y	0	1
0	0	¼
1	½	0
2	0	¼

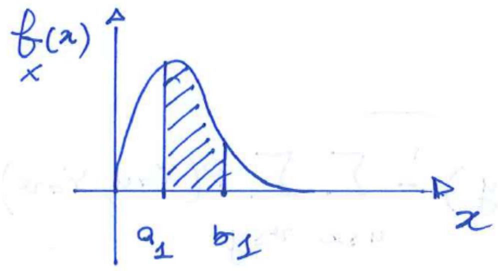
Continuous random variables

Now we consider continuous cases, we will deal with density and probability in intervals

Joint probability density function

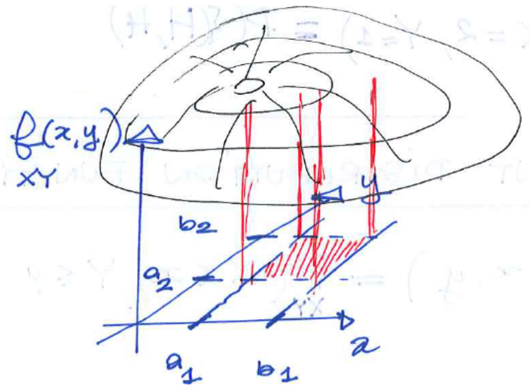
$$f_{xy}(x, y) = P(a_1 \leq x \leq b_1,\ a_2 \leq x \leq b_2) = \int_{a_1}^{b_1}\int_{a_2}^{b_2} f_{xy}(x, y)dxdy$$

In the univariate case we had:



This is an area

now we have:



this is a volume

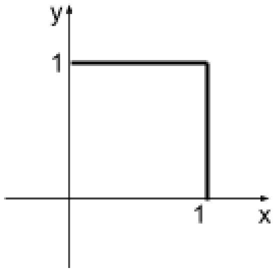
$$\int_R \int_R f_{xy}(x, y)dxdy = 1$$

Joint distribution function

$$F_{xy}(x, y) = P(X \le x, Y \le y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{xy}(u, v)dudv$$

Ex. joint bivariate rv over {0,1]*[0,1]

- Domain

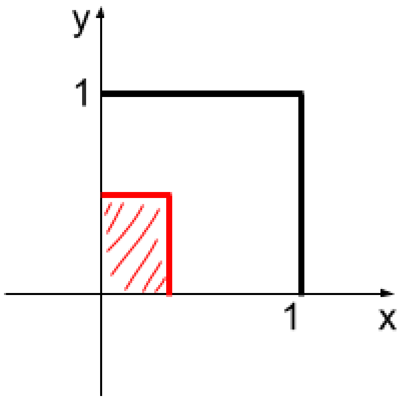


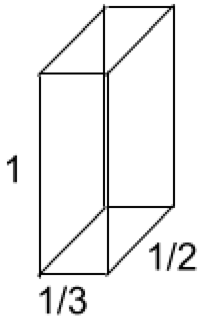
0≤x≤1
0≤y≤1
$f_{xy} = 1$, this is derived from the domains, it is a uniform cube

- $F_{xy}(\frac{1}{3}, \frac{1}{2}) = P_{xy}(x \le \frac{1}{3}, y \le \frac{1}{2}) = \int_0^{\frac{1}{3}} \int_0^{\frac{1}{2}} 1dxdy = \frac{1}{3} * \frac{1}{2} * 1 = \frac{1}{6}$

  side x side y height

  That's actually the volume of a cube





Marginal distributions from joint distribution

…ex. Tossing two coins
x=number of heads

$$y = \begin{cases} 1 & equal \quad outcomes \\ 0 & different \quad outcomes \end{cases}$$

X\Y	0	1	
0	0	¼	
1	½	0	
2	0	¼	
			1

Can we recover the marginal distributions of x and y from the joint distribution?
In order to find the probability of x=0 we to sum the probabilities of x without caring about y:

$$P_X(x = 0) = \sum_y P_{XY}(x = 0, Y = y) = 0 + \frac{1}{4} = \frac{1}{4}$$

We have just summed the row for x=0

More in general
Probability of x:

$P_X(X = x) = \sum_y P_{XY}(X = x, Y = y) =$ row sum

x is fixed
Probability of y:

$P_Y(Y = y) = \sum_x P_{XY}(X = x, Y = y) =$ column sum

y is fixed

X\Y	0	1	
0	0	¼	¼
1	½	0	½
2	0	¼	¼
	½	½	1

Marginal densities from joint density

$f_X(x) = \int_R f_{XY}(x, y)dy$ x is fixed

$f_Y(y) = \int_R f_{XY}(x, y)dx$ y is fixed

Instead of summing we are integrating

Conditional probability

Joint probability of 2 events A and B:

$$P(A \cap B) = P(A|B) * P(B)$$
$$= P(B|A) * P(A)$$

Conditional PMF of X|Y

The joint probability of (X,Y):

$$P_{XY}(X = x, Y = y) = P_{X|Y}(X = x|Y = y) * P_Y(Y = y)$$

joint pmf conditional pmf

$P_{X|Y}(X = x|Y = y) = \frac{P_{XY}(X=x,Y=y)}{P_Y(Y=y)}$ the numerator corresponds to a single cell in the joint PMF

while the denominator corresponds to the sum of a column

$$P_{X|Y}(X = 0|Y = 0) = \frac{P_{XY}(X=0,Y=0)}{P_Y(Y=0)} = \frac{0}{1/2} = 0$$

$$P_{X|Y}(X = 0|Y = 1) = \frac{P_{XY}(X=0,Y=1)}{P_Y(Y=1)} = \frac{1/4}{1/2} = 1/2$$

X|Y	0	1	
0	0	½	¼
1	1	0	½
2	0	½	¼
	1	1	

We have a distribution for each value of Y.

If x and y were independent this conditional probability mass function should be equal to the joint probability mass function (previous page)

Conditional PMF of Y|X

X|Y	0	1	
0	0	1	1
1	1	0	1
2	0	1	1
	½	½	

$$P_{X|Y}(Y = y|X = x) = \frac{P_{XY}(Y=y,X=x)}{P_X(X=x)}$$

$$P_{X|Y}(Y = y|X = x) = \frac{P_{XY}(Y=y,X=x)}{P_X(X=x)} = \frac{1/4}{1/4} = 1$$

We have one PMF for each value of x.

Conditional distribution is important because it tells us how variables interact.

$P_{XY}(X = x, Y = y) = P_{X|Y}(X = x|Y = y) * P_{Y}(Y = y)$

If x and y are independent P(X) is not affected by the value of y so:

$P_{XY}(X = x, Y = y) = P_{x}(X = x) * P_{Y}(Y = y)$

The joint probability is the product of marginal probabilities.

In this case we can use the same formula for density and distribution:

$f_{XY}(x, y) = f_{X}(x) * f_{Y}(y)$ if (x,y) continuous

$F_{XY}(x, y) = F_{X}(x) * F_{Y}(y)$

Conditional densities

$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_{Y}(y)}$

$f_{Y|X}(y|x) = \frac{f_{xy}(x,y)}{f_{X}(x)}$

In general:

$$\begin{aligned} f_{XY}(x, y) &= f_{X|Y}(x|y) * f_{Y}(y) \\ &= f_{Y|X}(y|x) * f_{X}(x) \end{aligned}$$

Independence table

In the joint distribution of two independent variables the joint conditional distribution should be the product of the marginals.

In our previous case we didn’t have independent variables, to investigate how much those variables are linked we can create the independence table (display of an ideal independent joint distribution) and compare it with the true joint distribution.

If x, y are independent

$P_{XY}(X = x, Y = y) = P_{x}(X = x) * P_{Y}(Y = y)$

Basing on our previous example we obtain this joint PMF:

$P_{XY}(X = 0, Y = 0) = P_{X}(X = 0) * P_{Y}(Y = 0) = 1/2 * 1/4 = 1/8$

We compute all the joint as if the variables were independent

X\Y	0	1	
0	⅛	⅛	¼
1	¼	¼	½
2	⅛	⅛	¼
	½	½	

This is the **independence table**, it is a joint probability.

If we manage to establish a distance between the independence table and the normal joint (conditional) we can state if the 2 variables are independent.

We can find the joint probability from the marginals only if they are independent.

Expected value

$E(X, Y) = [E(X), E(Y)]'$

Ex.

given a r.v in R^2 (X,Y)

With: $E(X, Y) = [E(X), E(Y)]'$

Now, given:

$g(x, Y)$: $(x, y) \rightarrow g(x, y)$

$R^2 \rightarrow R^m \qquad m \leq 2$

$$E[g(X,Y)] = \begin{cases} \sum_x \sum_y g(x,y) P_{XY}(X=x, Y=y) & discrete \quad case \\ \int_{\mathbb{R}} \int_{\mathbb{R}} g(x,y) f_{XY}(x,y) dx, dy) & continuous \quad case \end{cases}$$

Variance

$$V(X,Y) = \begin{bmatrix} V(X) & Cov(X,Y) \\ Cov(X,Y) & V(Y) \end{bmatrix}$$ variance/covariance matrix

Where:

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

$= E[XY] - E(X)E(Y)$

The **covariance** is a special instance of the variance, it measures the linear dependence between x and y that is to say how x and y interact.

(ex. If x and y are positive the covariance is positive)

If we standardise (x,y)

$X^* = \frac{X-E(X)}{\sqrt{V(X)}} \quad Y^* = \frac{Y-E(Y)}{\sqrt{V(Y)}}$

$E(X^*, Y^*) = (0, 0)'$

$$V(X^*, Y^*) = \begin{bmatrix} 1 & Corr(X,Y)) \\ Corr(X,Y) & 1 \end{bmatrix}$$

$Cov(X^*, Y^*) = E(X^*, Y^*) = Corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$

The resulting parameters are from the original r.v., that’s because **the standard covariance is equal to the correlation of the original couple of random variables**.

Correlation coefficient (K. Pearson)

$Corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}} =$ ρ(rho)

It can be seen as the covariance between the standardised version of X and Y.

Properties:

1. $\rho_{XY} \in [-1, 1]$
2. $\rho_{XY} = 0$ if X and Y are uncorrelated or linearly independent

3. $\rho_{XY} = 1$ if there is perfect positive linear dependence between X and Y

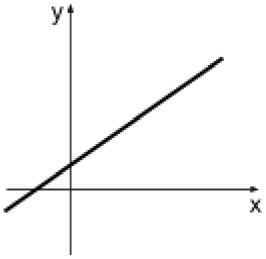
   i.e. Y=a+bx (Y is just a linear transformation of X)

4. $\rho_{XY} =- 1$ if there is perfect negative linear dependence between X and Y

   i.e. Y=a-bx

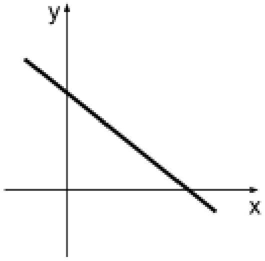
In practice:

- $\rho_{XY} = 1$



Y=a+bx b>0
All the points are perfectly overlapped to the line which has positive slope

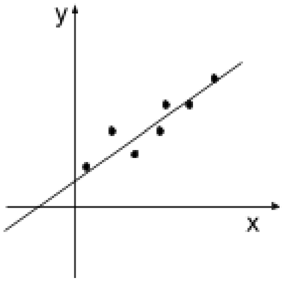
- $\rho_{XY} =- 1$

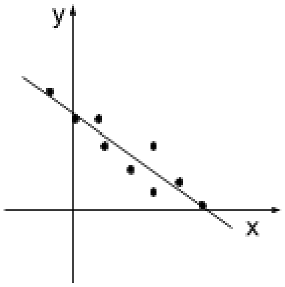


Y=a-bx b>0
Negative slope

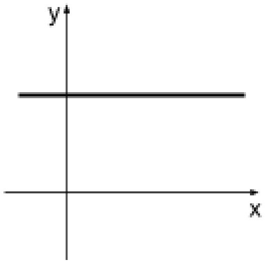
In these cases there is no randomness, no error, now let’s **consider randomness**:

- $\rho_{XY} = 0.6$ $\rho_{XY} =- 0.6$

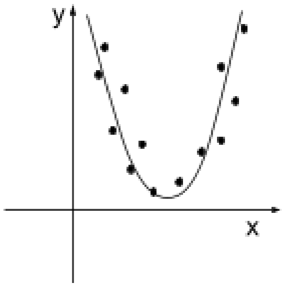




- $\rho_{XY} = 0$



X and Y are completely **independent**, even if X increases Y doesn’t change



Here instead the dependence is not linear, Y is a function of X squared.
By computing the correlation we still find 0 (the correlation is equal to sum of all the slopes), that's because our probability tools do not allow us to "measure" this, we need to embrace the complexity of our world and wait to discover other tools.

We have to be aware that **nonlinear complex dependence exists**.
The correlation coefficient corresponds to our naif idea of dependence, only dominated by linearity: If x goes up y goes up or down.
But a lot of phenomena do not follow this linearity.

Bivariate gaussian random variables

$(X, Y) \sim N(u = (u_X, u_Y);\ \Sigma = (\sigma^2_X,\ Cov(X, Y) \mid Cov(X, Y),\ \sigma^2_Y)$

Now let's focus on standardised bivariate gaussian:

$u = (0, 0)$

$\Sigma = (1,\ P_{XY} \mid P_{XY},\ 1)$

We put a dash under these elements because they are vectors/matrices.

$$f_{XY}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp\{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\}$$

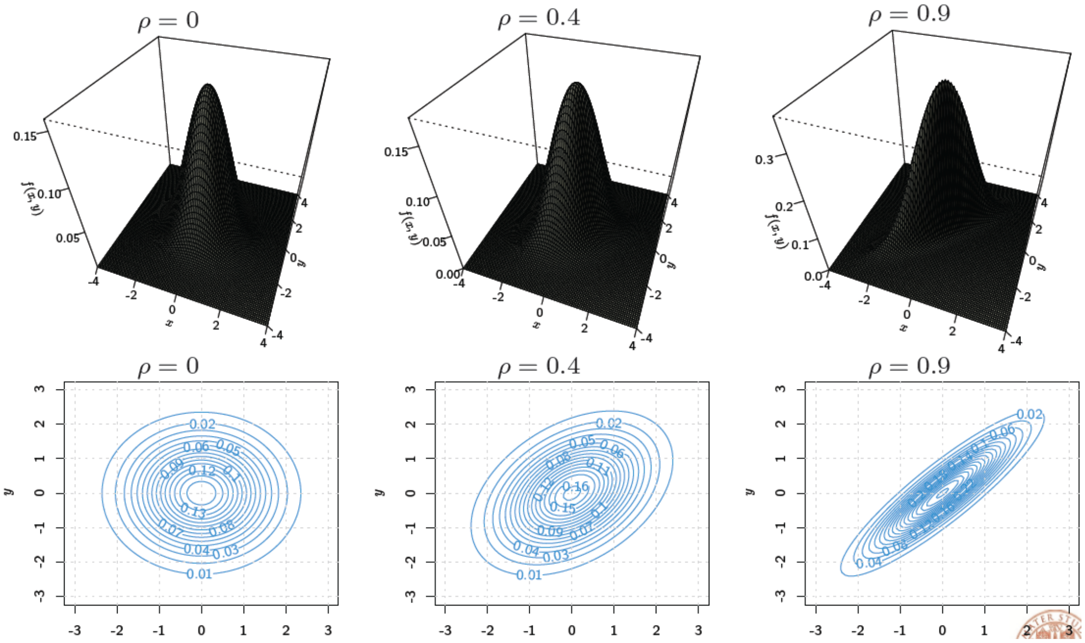
If $\rho = 0$, then:

$$f_{XY}(x, y) = \frac{1}{2\pi} exp\{-\frac{1}{2}(x^2 + y^2)\} =$$

$$= \frac{1}{2\pi} exp(-\frac{1}{2}x^2) exp(-\frac{1}{2}y^2) =$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} * \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} = f_X * f_Y$$

In this case the two variables are independent, otherwise if there is a **linear dependence** it affects the distribution in this way:



CENTRAL LIMIT THEOREM

Let: X_1, X_2, X_3, …, X_n be a sequence of i.i.d. random variables (independent and identically distributed, i.e. same type of distribution)

With:

$E(X_i) = u < \infty$

$V(X_i) = \sigma^2 < \infty$

Now define:

$S_n = \sum_{i=1}^{n} X_i$

We do not know this sum because we do not know the distribution of each X (only if they were gaussians we could know that the sum is a gaussian).

Let's consider the expected value of this new random variable:

$E(S_n) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} u = nu$

$V(S_n) = V(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} V(X_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$

(in computing the variance of the sum of multiple variables we should consider also the covariance but in this case the variables are independent so the covariance is 0, we are in a particular case)

Now we standardise S_n

$Z_n = \frac{S_n - nu}{\sqrt{n\sigma^2}} \underset{n\to\infty}{} \to N(0, 1)$

The central limit theorem tells us that **this variable converges to a standard gaussian.**

i.e.

$F_{Z_n}(z) \to \Phi(z)$ For every z on which F_{Zn} is continuous

That's the reason why we experiment the gaussian distribution in so many cases.

STATISTICAL INFERENCE

Introduction

Probability theory: studies models (inspired by observations) of random phenomena (ranodm variables)
Typical question: what is the probability of having two heads out of 2 coin tosses
This is probability because we are working with idealised models, they are not possible in reality, we are not working with physical trials.
Statistical inference: learning from data
Typical question:we have 2 real coins in my pocket, we want to know whether they are biassed or not; in other words we want to estimate the probability of getting head.
In probability this was granted we were working with idealised coins.
In statistics we have to toss the coin many times (drawing a sample) and basing on the data we try to learn from it.

Statistical inference means learning from data.
The main problems are:

- ➢ How do we collect data?
- ➢ How do we learn from data?

These 2 important problems are the subject of mathematical statistics and they need theorems.

Basic notions

We need to make a parallel between theory and practice.

Population and sample

A population is a set of individuals, or units, that posses a given feature

- Finite populations
  Ex.
  The set of italian families
  The set of genes of the human genome
- Infinite populations(it can be just theoretically infinite)
  Ex.
  Set of items produced by a factory
  Set of stars in the universe

Features of a population

The feature is the random phenomena that we are interested in studying.
Ex.
Number of children per family
GC content of a gene

Typically we denote the feature as X (parallel with the random variable), in some cases X will be called the population itself.

Parameter of a variable/population
Ex.
Average number of children in the population
Average GC content in genes

From probability theory to practice

If we are able to observe the whole population then we can compute these parameters right away.
For instance, with a census we can know the average number of children per family, exactly and without error; in this case we do not need statistical inference, we are just using descriptive statistics.

If for whatever reason we cannot observe the whole population or if the population is infinite we need to draw a sample and use the theorems of statistical inference to learn about the population.

If the population is infinite then X can be seen as a random variable

This is the key to connect probability theory with practice:
The parameters of X we are studying are exactly the parameters of a random variable.

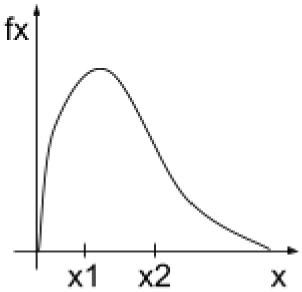
Parameters of a population → Parameters of a random variable

X: population mean E(X)=μ
population variance V(X)=σ^2
We cannot know them exactly but we can estimate them.

Sample theory

Basic idea: draw a sample from the population X randomly (randomness is the key to a good sample draw, sample theory is the science that studies how to draw samples)
Each draw has to be independent from the previous one and the population from which the elements are drawn does not have to change.

If X is the random variable that represents the population (random phenomena), we can suppose that X has a density:



We need to draw a random sample from it:
Let's consider a random sample constituted of two elements: x_1 and x_2.

Drawing a random sample of two units means drawing n=2 realisations from $f_X(x)$

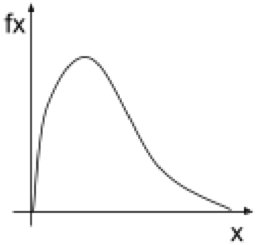
$\underline{x}$=(x_1, x_2) observed sample

With *random* we are referring to the random distribution of the variable, by drawing a lot of units we could reconstruct the distribution and each value of X has a different probability.

How can we draw multiple random elements?

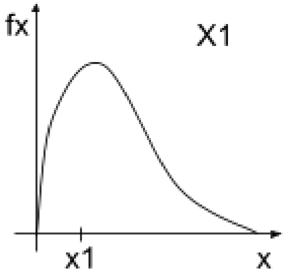
To distinguish x_1, x_2 we can think in this way:

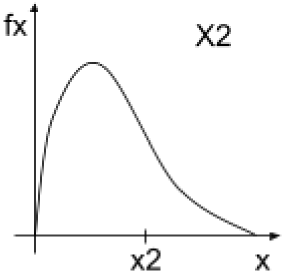
We have the population X that doesn't change:



Since we want to distinguish the 2 elements we can consider:

X1=X gives us x_1 X2=X gives us x_2





$\underline{x}$ = (x_1, x_2)

From a theoretical point of view the observed sample x is a random realisation of the theoretical sample $\underline{X}$ = (X_1, X_2)

We assume there to be 2 random variables but they are iid (independent and identically distributed) so that each draw is not influenced from the previous one and it is taken from the same population.

Random sample

Theoretical (X_1, X_2, ..., X_n) iid random copies of the population X

Observed (x_1, x_2, …, x_n)

Now that we have a random sample (once we have established that it has good properties) we can use it to understand something about the population.

Statistic

We can reason on two levels, with the iid random variables or with their realisations.
X is a random variable with:
$\mu = E(X)$ and $\sigma^2 = V(X)$
We cannot exactly know these but with a good sample we can get some information (statistical inference)

We use a statistic, namely a function of the sample: $T = f(x_1, ..., x_n)$
A statistic gives us a good approximation of the population quantities.

For instance we could have to guess the mean of a population: $\mu = E(X)$
We can estimate it with the sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, (this is a function of the sample, a statistic T)
The theoretical sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
If X is a random variable we can apply theorems of probability, the distribution of $\bar{X}$ is a function of the random sample $(X_1, ..., X_n)$ and as such it is a random variable (it could be linked to the distribution of X_i, in some way, through some theorem).

We have the same reasoning with the sample variance: $\sigma^2 = V(X)$
The observed variance is (estimation): $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$
The theoretical variance is: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$

Now the problem is, how can we state that the characteristics of X correspond to the characteristics of the population?

POINT ESTIMATION

Let X be a random variable that represents the random phenomena under study.

We assume that X is continuous with probability density function $f_X(x, \theta)$ that depends upon an unknown parameter θ.

- f is assumed to be known.
- θ is the unknown parameter (E(X) or V(X)), it is the object of inference; it is a constant and we won't ever know its true value, we will just be able to estimate it.

Point estimation consist in the estimation of θ based on a sample or a statistic.

We draw a random sample, we assume it to be generated from n iid copies of X

$(x_1, ..., x_n)$	observed sample
$(X_1, ..., X_n)$	theoretical sample
$T = T(X_1, ..., X_n)$	statistic

T is the point estimator for the unknown parameter θ

Ex.

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

$\theta = \mu$ unknown

$\sigma^2 = 1$ known

In order to estimate μ we need a statistic, we draw a random sample of n=3 observations:

Observed sample: (x_1, x_2, x_3)

$x_1 = 0.55 \quad \rightarrow \quad X_1 \sim N(\mu,\ 1)$

$x_2 =- 0.03 \quad \rightarrow \quad X_2 \sim N(\mu,\ 1)$

$x_3 =- 0.12 \quad \rightarrow \quad X_3 \sim N(\mu,\ 1)$

We assume $\overline{X}$ to be a good approximation for μ:

In this case, $T = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\overline{X}$ is the point estimator for μ (it is a random variable)

Through the observed sample we can calculate the estimate, the observed sample mean:

$$\overline{x} = \frac{1}{3}\sum_{i=1}^{3} x_i = \frac{0.55-0.03-0.12}{3} = 0.13$$

$\overline{x}$ is a realisation of the random variable $\overline{X}$, every time we draw a random sample we will have a $\overline{x}$, it is like drawing observation from $\overline{X}$.

How can we say that $\overline{X}$ to be a good approximation for μ?

In our previous example every X_i has the same $E(X_i)$ and $V(X_i) = 1$

Based on these observations we can actually say something about $\overline{X}$ (we have studied theorems, haven't we? pepelaugh)

We know that the sum of independent gaussian random variables is also gaussian: $\overline{X} \sim N()$

So we can state the characteristic of this gaussian random variable ($\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$):

$$E(\overline{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}(E(\sum_{i=1}^{n} X_i)) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{3}\sum_{i=1}^{3} E(X_i) = \frac{\mu+\mu+\mu}{3} = \frac{3\mu}{3} = \mu$$

$$V(\overline{X}) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}(V(\sum_{i=1}^{n} X_i)) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{9}\sum_{i=1}^{3} V(X_i) = \frac{1+1+1}{9} = \frac{1}{3}$$

We can use all these properties just because all the variables of the summation are independent.

$\overline{X} \sim N(\mu, \frac{1}{3})$

0.13 is a random realisation of $\overline{X} \sim N(\mu, \frac{1}{3})$

In this case we can see that the expected value of the estimator corresponds to the characteristic that we have to find, so we could be happy with it.
In reality it actually exists a precise set of rules to state if an estimator is good to estimate a characteristic of a population:

Properties of an good estimator

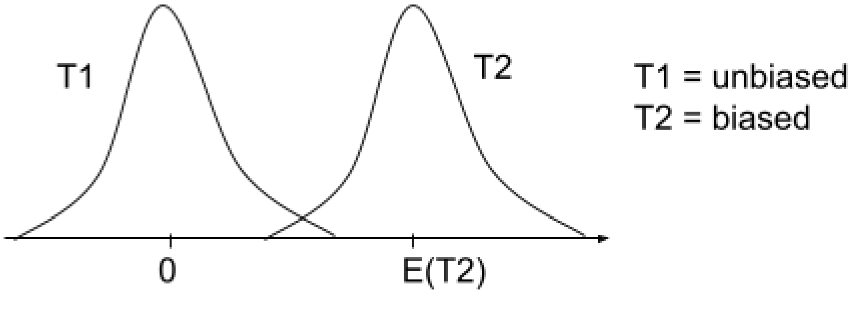
$X \sim f(X, \theta)$ f is known
θ is the unknown parameter

We draw a random sample from X: $(X_1, ..., X_n)$

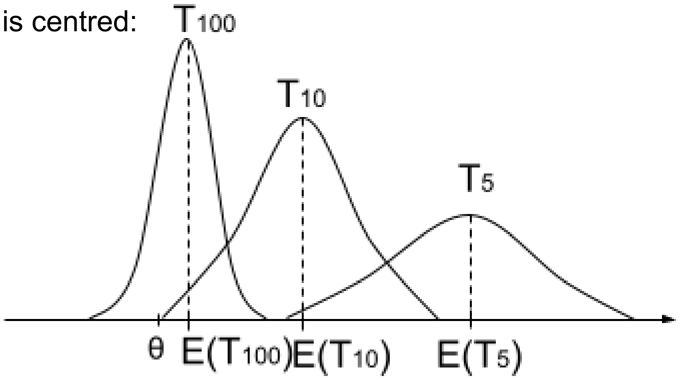
We define $T = T(X_1, ..., X_n)$ as the estimator of θ

Unbiasedness

We want the estimator to be centred on θ.
T is an unbiased estimator for θ if $E(T) = \theta$.
If not, T is biassed and the bias is defined as $B(T) = E(T) - \theta$.
The bias is like a systematic error:



Ex.
in the previous example the estimator is unbiased, it is centred:
$X \sim N(x, \mu)$
μ is unknown
$\overline{X}$ is the estimator
$E(\overline{X}) = \mu$



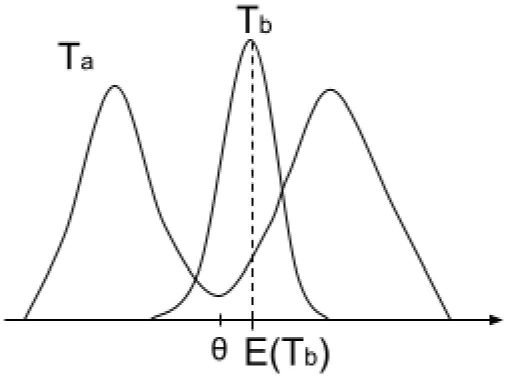
Asymptotic unbiasedness:
An estimator T_n is asymptotically unbiased if :

$B(T_n) \to 0$ with $n \to \infty$

Where n is the number of elements of the sample on which is based the estimator.

Asymptotic properties are valid only in the case of large samples.
From this we can infer that the good properties of an estimator are based on its distribution, through unbiasedness we can just measure the distance of the mean of the distribution from the unknown value, it is not referred to the distance of the whole distribution.
We need to come up with a measure of distance between the estimator (as a random variable) and the unknown parameter θ.

Mean square error (not a property)
Suppose we have 2 estimators:



T_a is unbiased

T_b is biassed

Even if T_a unbiased, its mass is concentrated away from the mean, by drawing 1 element we have a little possibility to draw something near the mean.
T_b on the other hand has a systematic error but its values are pretty near the true value.

T_b is preferable to T_a since it has lower variance and its distribution is closer to θ.
We need a measure of distance between the estimator T (which is a random variable) and the unknown parameter θ (a constant).

The mean square error can help us to have an estimation of it:

$MSE(T_n) = E[(T_n - \theta)^2]$ (we square it to make the error always positive)

n = sample size

$T_n = T(X_1, ..., X_n)$ estimator of θ

This is a more complete measure of distance and it can be used to define the property of efficiency

Efficiency

An estimator T_n is more efficient than the estimator G_n (both of them for θ) if:

$MSE(T_n) < MSE(G_n)$

For all the values of θ and n

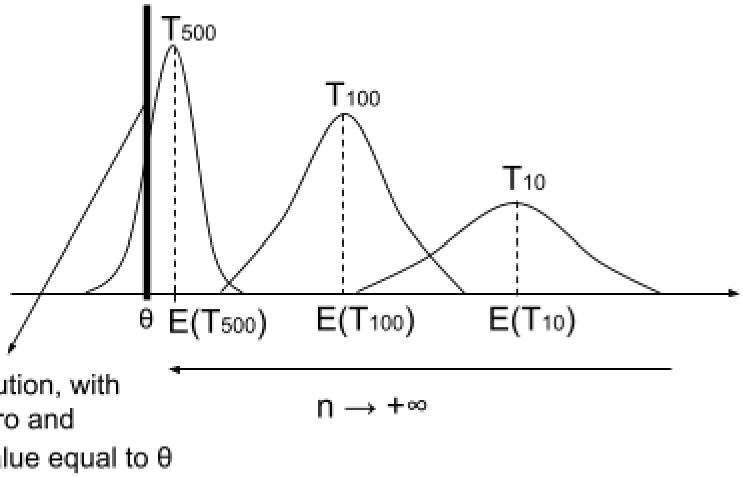
Consistency

An estimator for θ T_n is consistent if:

$MSE(T_n) \rightarrow 0$ when $n \rightarrow +\infty$

$MSE(T_n) = V(T_n) + [B(T_n)]^2$ consistency asymptotically includes unbiasedness

example:



T500
T100
T10
θ
E(T500)
E(T100)
E(T10)
Limit distribution, with variance zero and expected value equal to θ
n → +∞

Experimentally we can see that as n increases:

$B(T_n) \rightarrow 0$

$V(T_n) \rightarrow 0$

We need theorems to prove this mathematically

Consistency of $\overline{X}_n$ for μ

If $X \sim f_X(x, \mu)$ where μ is unknown

$\overline{X}_n$: estimator for μ

$MSE(\overline{X}_n) = E[(\overline{X}_n - \mu)^2]$

$= V(\overline{X}_n) + [B(\overline{X}_n)]^2$

Since we already proved that:

$E[\overline{X}_n] = \mu$ unbiased

$[B(\overline{X}_n)]^2 \to 0$

$MSE(\overline{X}_n) = V(\overline{X}_n)$

$V(\overline{X}_n) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} * V(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

$\lim_{n \to \infty} \frac{\sigma^2}{n} = 0$

The sample mean $\overline{X}_n$ is a consistent estimator for the population mean.

With an estimator we can understand something about the population but not its distribution, on the other hand we know the distribution of the estimator because of the central limit theorem.

Recap

X: population/phenomenon under investigation

$X \sim f_X(x, \theta)$

f: known

θ: unknown, it is not a random variable, it is a constant number and we will not ever know it exactly

Objective: estimate θ

We draw a random sample $(x_1, ..., x_n)$ we can think of it as multiple realisation from a set of iid random variables, the theoretical sample $(X_1, ..., X_2)$

Now we can apply functions to the sample (estimators):

Theoretical sample $(X_1, ..., X_2)$	Realisation →	Observed sample $(x_1, ..., x_n)$
↓		↓
Estimator for θ $T = T_n(X_1, ..., X_2)$	→	Estimate for θ $t = T_n(x_1, ..., x_2)$

Each estimator has to be proven to fit for its purpose, it has to have some good properties:

- Unbiasedness: $E(X) = \theta$

Mean square error: $MSE(T_n) = E[(T_n - \theta)^2] = V(T_n) + [B(T_n)]^2$

- Efficiency (comparison):
  $if: MSE(T_n) < MSE(G_n)\ then: T_n\ is\ more\ efficient\ than\ G_n$
- Consistency: $MSE(T_n) \to 0$ when $n \to +\infty$

Estimator for the proportion

Objective: find a point estimator for the proportion of units of population that possess the attribute A.

Ex. estimating the proportion of individuals that possess a certain gene mutation

$X \sim Ber(\pi)$ π: probability of mutation

The population follows the Bernoulli model because an individual can just have or not have the mutation.

π coincides with the proportion of people that have the mutation.

$P(X=1) = \pi$

Probability that a unit drawn at random from X (the population) possesses the attribute A (e.g. the gene mutation)

- If the population is finite and has 5 units
  1 2 3 4 5 red = mutation; black = no mutation.
  $\pi=⅕$: probability of drawing a unit with the mutation.

- If the population is infinite
  We assume that π doesn't change upon observing a unit (removing a unit doesn't change the probability of observing the mutation in the remaining population).
  hence, we can say: $X \sim Ber(\pi)$

  $$X = \begin{cases} 1 & if\ A\ is\ true \quad (mutated) \\ 0 & if\ A\ is\ false \quad (notmutated) \end{cases}$$

  We also know
  $E(X) = \pi$
  $V(X) = \pi(1-\pi)$
  To estimate π we need to compute the expected value.

  Since the population is infinite, we have to rely on a random sample:
  $(X_1, \ldots, X_n)$ i.i.d. copies of X
  $(x_1, \ldots, x_n)$ observed sample of 1 or 0
  By taking the sample mean (or sample proportion):

  $$\overline{X}_n = \frac{\sum_{i=n}^{n} X_i}{n} = \frac{1}{n}\sum_{i=n}^{n} I(X_i = 1) = proportion\ of\ ones\ in\ the\ sample$$

  $\sum_{i=n}^{n} X_i$ the value of X_i can be either 0 or 1, by summing them we are just "counting the 1s", then by dividing for n we obtain the proportion of the 1 in the population.

  If we use the function $I(A)$:

  $$I(A) = \begin{cases} 1 & if\ A\ is\ true \\ 0 & if\ A\ is\ false \end{cases}$$

So $\sum_{i=n}^{n} I(X_i = 1) = n_1$ is the number of individuals/units that possess A

To compute the proportion: $\overline{X}_n = \frac{n_1}{n}$

Properties of the estimator of the proportion:

- $\overline{X}_n$ is unbiased

  $$E(\overline{X}_n) = E(\frac{\sum_{i=1}^{n} X_i}{n}) = \frac{1}{n}E(\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \pi = \frac{n\,\pi}{n} = \pi$$

  X_i are iid copies of a Bernoulli r.v. so their expected value is π.

- $\overline{X}_n$ is consistent

  $$MSE(\overline{X}_n) \rightarrow 0 \quad if\, n \rightarrow + \infty$$

  $MSE(\overline{X}_n) = V(\overline{X}_n) + [B(\overline{X}_n)]^2$ $B(\overline{X}_n)$ is 0 because $\overline{X}_n$ is unbiased

  $$= V(\frac{1}{n}\sum_{i=n}^{n} X_i) = \frac{1}{n^2}V(\sum_{i=n}^{n} X_i) = \frac{1}{n^2}\sum_{i=n}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=n}^{n} \pi(1 - \pi) = \frac{n\,\pi(1-\pi)}{n^2}$$

  $$= \frac{\pi(1-\pi)}{n} \rightarrow 0 \quad if\, n \rightarrow + \infty$$

INTERVAL ESTIMATION

So far we have discussed point estimators:
Unknown parameter → estimator → estimate of the parameter
If we want to make a prediction, a statement (ex. Tomorrow it will rain with a probability of…) we have to use interval estimation.
Powerful theorems allow us to come up with predictions that most of the time are good.
We need an interval estimator of the kind $[L_1, L_2]$

$L_1 = L_1(X_1, ..., X_n)$

$L_2 = L_2(X_1, ..., X_n)$

This is called a random interval because L_1 and L_2 are random variables.

The random interval $[L_1, L_2]$ is a confidence interval for the parameter θ at level (1 - α) if:

$P(L_1 \leq \theta \leq L_2) = 1 - \alpha$

The probability that the random interval $[L_1, L_2]$ contains θ is 1 - α, where α is typically set by the researcher.
We have to set α in a way that our prediction is correct most of the time but it is also "short".

The problem with this interval is that the predictions are random variables (they change) and the unknown parameter is a constant.
We need to use a realisation of the theoretical interval, that will be slightly different each time:
$[L_1, L_2]$ → theoretical interval estimator, L_1 and L_2 are functions of the random sample $(X_1, ..., X_n)$
$[l_1, l_2]$ → estimated confidence interval for θ because l_1 and l_2 are numbers, based on the observed sample $(x_1, ..., x_n)$

Confidence interval for the mean of a gaussian population

Let $X \sim N(\mu, \sigma^2)$ be the population.
μ is unknown, it is the object of inference.
Objective: derive a confidence interval for μ

1. σ^2 is known
   We start by considering a point estimator for μ.

   $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (unbiased and consistent)

   - $E(\overline{X}_n) = \mu$
   - $V(\overline{X}_n) = \frac{\sigma^2}{n}$ n=sample size

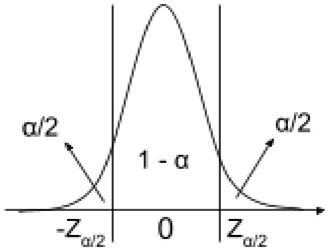
   In order to come up with an interval we need to have the distribution of the sample mean and variance.
   These functions are sums of iid copies of a gaussian, we know that the sum of a gaussian rv is:

$\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$

Now we standardise $\overline{X}_n$

$\overline{Z}_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

In this way we can use the statistical tables for mean and variance:



We can find the value of $Z_{\alpha/2}$ such that

$P(-Z_{\alpha/2} \le \overline{Z}_n \le Z_{\alpha/2}) = 1 - \alpha$

$\overline{Z}_n$ contains the unknown parameter ($\overline{Z}_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$).

$P[-Z_{\alpha/2} \le \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \le Z_{\alpha/2}] = 1 - \alpha$

$P[-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \overline{X}_n - \mu \le Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] = 1 - \alpha$

$P[-\overline{X}_n - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le -\mu \le -\overline{X}_n + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] = 1 - \alpha$

$P[\overline{X}_n - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}_n + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] = 1 - \alpha$

$L_1 = \overline{X}_n - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

$L_2 = \overline{X}_n + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

The interval estimator at level 1 - α for μ results:

$[\overline{X}_n - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X}_n + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$

We can rewrite it as:

$[\overline{X}_n \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$ standard form for interval estimators

[point estimator ± quantile * standard deviation]

The quantile ($Z_{\alpha/2}$) depends upon α and on the distribution of $\overline{X}_n$, we are capable of choosing α but we have to find a compromise:

- If α is too big the interval is not informative, it is too wide.
- If α is too small the interval is not very probable.

The standard deviation depends upon n, as n enlarges the error decreases.

Once α and n are fixed, a good interval estimator has the shortest length (more accurate interval).

2. σ^2 is unknown

   $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$

   σ^2 is unknown and has to be estimated on the sample, a good estimator (unbiased and consistent) for the population variance is the sample variance.

   $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$

   We divide by n - 1 because if we divide for n we obtain a consistent but biassed estimator.

   Now we apply the same procedure as before:

   $\overline{Z}_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ we cannot use this since σ is unknown

   $= \frac{\overline{X}_n - \mu}{S/\sqrt{n}} \sim T_{n-1}$

   We have a problem! This is a ratio of random variables, hence it doesn't result in a gaussian but it is a student's T distribution.

   The interval estimator results:

   $[\overline{X}_n \pm T_{\alpha/2;n-1} \frac{S}{\sqrt{n}}]$

   The price that we have to pay in such instance is that by using the studentìs T distribution we get longer and less accurate intervals (predictions)

3. Unknown population

   $X \sim f(\mu, \sigma^2)$ f, μ, σ^2 unknown

   $\overline{X}_n \sim ?$

   The X_i are not gaussian so, in general, also $\overline{X}_n$ will not be gaussian

   We can leverage upon the central limit theorem but the price that we have to pay is that we have to use a big sample size.

   $\overline{Z}_n = \frac{\overline{X}_n - \mu}{S/\sqrt{n}} \to N(0, 1)$ when $n \to + \infty$

   The asymptotic confidence interval results:

   $[\overline{X}_n \pm Z_{\alpha/2} \frac{S}{\sqrt{n}}]$

   Typically we say that the clt holds if n>120 but it depends a lot on the data.

Ex. from a survey on the habits of students we have a random sample of n=26 students.
X: monthly expenditure on leisure time.

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

Objective: derive a confidence interval at (1 - α) = 0.95 (α = 0.05) for the mean monthly expenditure μ in the population.

From the sample survey we get:

$\overline{X} = 120$ estimated sample mean

$S^2 = 121$ estimated sample variance

$\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$

$T_n = \frac{\overline{X}_n - \mu}{S/\sqrt{n}} \sim T_{n-1}$ student's T

$[\overline{X}_n \pm t_{\frac{\alpha}{2}; n-1} \frac{S}{\sqrt{n}}]$

n=26, n-1=25

α=0.05, 1-α=0.95, α/2=0.025

$t_{0.025;\, 25} = 2.06$

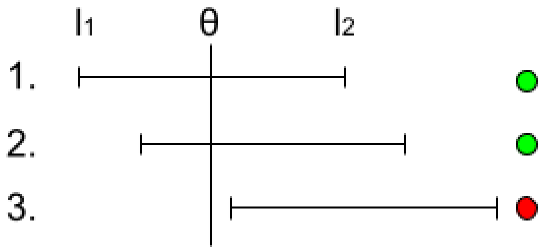
$[\overline{X}_{26} \pm t_{0.025;\, 25} \frac{S}{\sqrt{26}}] = [120 \pm 2.06 \frac{\sqrt{121}}{\sqrt{261}}] = [115.56;\ 124, 44]$

Is this interval good? We don't know, the theory just tells us that 95% of the time this is correct, but in reality we could experiment the 5% chance that it does not contain μ.

The random interval [L_1, L_2] is a confidence interval at level 1-α for θ if:

$P(L_1 \leq \theta \leq L_2) = 1 - \alpha$

$1 - \alpha$: confidence level



The interpretation is frequentist: if we were able to draw many independent samples we would observe that the prediction is correct (the confidence interval contains the value) (1- α)% of the time.

Dangers are around the corner:

Ex. US presidential election 1936, competition between Alfred Landon (republican) Franklin D. Roosevelt (democrat).

A respected magazine performed one of the most expensive polls ever conducted, with a sample size of 2.4 million citizens.

Prediction 1: Landon 57%; Roosevelt 43%

An independent poll based upon 50,000 people predicted a different outcome:

Prediction 2: Landon 38%; Roosevelt 62%

What went wrong:

- ➢ The units of the sample were taken from telephone directories and magazine subscribers.
  In 1936 only rich people owned a telephone, the sample wasn't drawn at random (SELECTION BIAS)

➢ 2.4 million units over 10 million asked.
Typically the missing data has a meaning, there is always a reason behind missingness, it is not random.
(NON RESPONSE BIAS)

Confidence interval for the variance of a gaussian population

$X \sim N(\mu, \sigma^2)$ both μ and σ^2 unknown
First of all we have to estimate these values:
μ: estimated through $\overline{X}_n$

σ^2: estimated through $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$

Both these estimators are unbiased and consistent, in this case we will focus on finding a confidence interval for the estimator of the sample variance so we'll need to work with the formula of the sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

$$(n - 1)S^2 = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{\sigma^2}$$

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}_n}{\sigma}\right)^2$$

This formula resembles a standard gaussian, maybe we will be able to threat it like:

$$\frac{(n-1)S^2}{\sigma^2} \approx \sum_{i=1}^{n} Z_i^2$$

In reality we could do this only if we had $\left(\frac{X_i-\mu}{\sigma}\right)^2 = Z_i^2 = \chi_i^2$, hence $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \approx \chi_n^2$

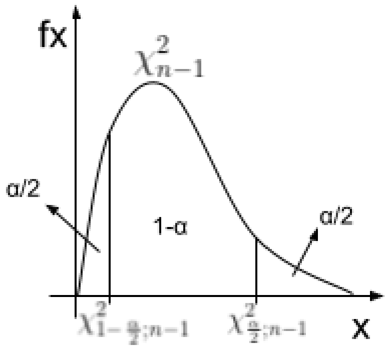
But μ is estimated and we have instead:

$$\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}_n}{\sigma}\right)^2 \approx \chi^2_{n-1}$$

The fact that we use the sample mean as the population mean causes the loss of a degree of freedom.

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

So at this point we can use the tables to infer a confidence interval for the variance of the distribution.


fx
χ^2_{n-1}
α/2
1-α
α/2
$\chi^2_{1-\frac{\alpha}{2};n-1}$
$\chi^2_{\frac{\alpha}{2};n-1}$
X

We can say that the random variable has a probability of 1 - α to stay between the values (in terms of variance)

$$P(\chi^2_{1-\frac{\alpha}{2};n-1} \le \frac{S^2(n-1)}{\sigma^2} \le \chi^2_{\frac{\alpha}{2};n-1}) = 1 - \alpha$$

Let's isolate σ^2 (the parameter we wanna find)

$$P(\frac{\chi^2_{1-\frac{\alpha}{2};n-1}}{S^2(n-1)} \le \frac{1}{\sigma^2} \le \frac{\chi^2_{\frac{\alpha}{2};n-1}}{S^2(n-1)}) = 1 - \alpha$$

$$P(\frac{S^2(n-1)}{\chi_{\frac{\alpha}{2};n-1}} \le \sigma^2 \le \frac{S^2(n-1)}{\chi_{1-\frac{\alpha}{2};n-1}}) = 1 - \alpha$$

Ex.

Random sample of 4 strains, n=4

In the population the variable fitness follows a gaussian distribution.

X: fitness of the strain with respect to the initial population after 100 days.

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

Objective: derive a confidence interval at level 95% for σ^2

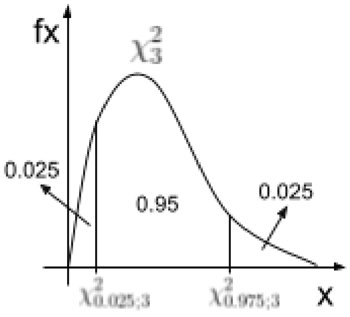
$$\frac{S^2(n-1)}{\sigma^2} = \chi^2_{n-1}$$

In our case n=4

$$\frac{S^2*3}{\sigma^2} = \chi^2_3$$

The confidence interval is: $(1 - \alpha = 0.95;\ \alpha = 0.05;\ \alpha/2 = 0.025;\ 1 - \alpha/2 = 0.975)$

$$[\frac{3*S^2}{\chi^2_{0.025;3}}; \frac{3*S^2}{\chi^2_{0.975;3}}]$$



$\chi^2_{0.025;3} = 9.348 \qquad \chi^2_{0.975;3} = 0.216$

Now we just need to compute the sample variance (the point estimator on which we will attach the interval):

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

$$\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{4} = \frac{0.063-0.062-0.043+1.340}{4} = 0.272$$

$$S^2 = \frac{1}{3}[(0.063 - 0.272)^2 +...+ (1.340 - 0.272)^2] = 0.360$$

Confidence interval:

$$[\frac{3*0.360}{9.348}; \frac{3*0.360}{0.216}] = [0.116; 5]$$

With variance we cannot use the central limit theorem, if we do not have a gaussian it is needed a more advanced method.

Determination of the sample size that provides a given level of precision

What if we need a specific length of the interval and not a specific 1 - α?

We have to play with the sample size.

If we increase the sample size the prediction gets more precise.

Let $X \sim N(\mu, \sigma^2)$ where σ^2 is known

The confidence interval at level 1 - α for μ is:

$$[\overline{X}_n \pm Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$$

we can measure the precision by looking at the length

Length: $L_2 - L_1$

$$(\overline{X}_n + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) - (\overline{X}_n - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 2\,Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = 2\delta$$

$\delta = Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$: half of the length of the interval

If α and δ are fixed then:

$$\sqrt{n} = Z_{\frac{\alpha}{2}}\frac{\sigma}{\delta} \Rightarrow n = [Z_{\frac{\alpha}{2}}\frac{\sigma}{\delta}]^2$$

Ex.

$X \sim N(\mu, \sigma = 12.81)$

1. Derive n as to obtain a confidence interval at level 95% of the kind $\overline{X} \pm 3$

   We know that:

   $\overline{X} \pm 3 \Rightarrow \delta = 3 \qquad \overline{X} - 3$ —δ— $\overline{X}$ —δ— $\overline{X} + 3$

   $$n = [Z_{\frac{\alpha}{2}}\frac{\sigma}{\delta}]^2$$

   $\alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025,\ Z_{0.025} = 1.96$

$n = [1.96 * \frac{12.81}{3}]^2 = 70$

2. What is the sample size needed to halve the length of the confidence level?

   $\delta = 1.5$

   $n = [1.96\frac{12.81}{1.5}]^2 = 280$

In general

An interval estimator is an assessment on the uncertainty of a point estimator.
The typical structure for a confidence interval for a parameter θ is:

$T_n \pm q_{\alpha/2} * SD(T_n)$

Point estimator for θ ± quantile of T_n(partially under control) * standard deviation of T_n(variability of the estimator)

Point estimators with good properties (unbiased, efficient and consistent) produce interval estimators with good properties (short length).

Once α and n are fixed we judge the interval based on its length.

HYPOTHESIS TESTING

Objective: accept/reject a statistical hypothesis with high confidence, based on a random sample of observations.

Correlated to decision making:

We have an hypothesis on the parameter and we want to test it.

We are not interested in estimating the parameter (θ) but in verifying an hypothesis about the parameter

Before seeing how it works we need to define:

Statistical hypothesis

We assume that X is the population r.v. that represents the phenomenon under investigation.

$X \sim f_x(x, \theta)$

Also, we assume that the functional form of the distribution f is known and that θ is unknown.

A statistical hypothesis is a conjecture regarding θ, we already have a previous idea about θ.

Important! The hypothesis must be formulated prior to the experiment

We define the following system of hypothesis

$$\begin{cases} H_0: & null & hypothesis \\ H_1: & alternative & hypothesis \end{cases}$$

We must decide between H_0 and H_1, for instance $X \sim N()$ then $\begin{cases} H_0: & \mu = 0 \\ H_1: & \mu \neq 0 \end{cases}$

We assume that H_0 is true and we see whether the data/empirical evidence in consistent with it.

Ex.

According to a producer of batteries the average duration is 3400h

If we want to test this statement:

$$\begin{cases} H_0: & \mu = 3400h \\ H_1: & \mu < 3400h \end{cases}$$

μ is the mean of the population (all the batteries of the producer)

Simple hypothesis and composite hypothesis

Assume $X \sim f(x, \mu)$

Where f is known, μ is unknown

- μ=0 is a simple hypothesis, since it identifies exactly a single distribution, once we assume the hypothesis is true we know everything about the population.
- μ>0 is a composite hypothesis, because it identifies an infinite collection of r.v.s.

Parametric space Θ

$\theta \in \Theta$ is the set of all possible values that θ can assume

Typically the system of hypothesis partitions the parametric space $\Theta = \Theta_0 \cup \Theta_1$

$$\begin{cases} H_0: & \theta \in \Theta_0 \\ H_1: & \theta \in \Theta_1 \end{cases}$$

For example, in the battery experiment

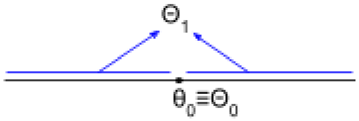
$$\begin{cases} H_0: & \mu \geq 3400h \\ H_1: & \mu < 3400h \end{cases} \qquad \mu \in \Theta \equiv R^{+}$$

$$\Theta \equiv R^{+} = [0, 3400) * [3400, \ldots)$$
$$= \quad \Theta_1 \quad \cup \quad \Theta_o$$

Some commonly used systems of hypothesis

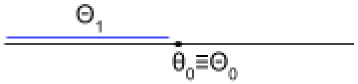
1. Two sided hypothesis

$$\begin{cases} H_0: & \theta = \Theta_0 \quad single \quad value \\ H_1: & \theta \neq \Theta_0 \quad everything \quad else \end{cases}$$



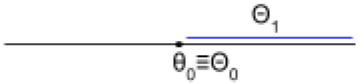
2. One sided hypothesis (left side)

$$\begin{cases} H_0: & \theta = \Theta_0 \\ H_1: & \theta < \Theta_0 \end{cases}$$



3. One sided hypothesis (right side)

$$\begin{cases} H_0: & \theta = \Theta_0 \\ H_1: & \theta > \Theta_0 \end{cases}$$



Taking the decision

The decision depends upon the random sample

$(X_1, \ldots, X_n) \rightarrow test \rightarrow H_0 \; or \; H_1$

The test statistic is a function of the random sample (random variable) whose distribution is completely known under H_0.

For example:

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

We want to test:

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases} \qquad \text{where } \mu_0 \text{ is a number}$$

We draw a sample (X_1, …, X_n) and we consider the sample mean

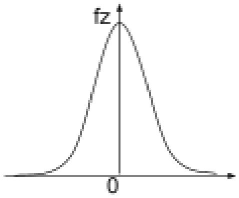
$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

If H_0 is true (we need to assume this to carry out this test) then $\overline{X}_n \sim N(\mu_0, \frac{\sigma^2}{n})$ is the hypothesis distribution (distribution of the sample mean under H_0).

The test statistic is:

$\overline{Z}_n = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$ (we standardise so that we can use tables)

If H_0 is true then $\overline{Z}_n \sim N(0, 1)$



My data should be compatible with this distribution.
Now we have to compare our empirical data with this conjecture.

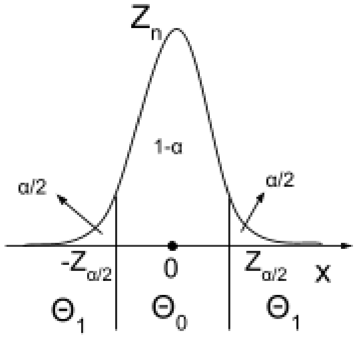
Empirical evidence

$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ (we are not working with rvs anymore, this is a number)

We standardise:

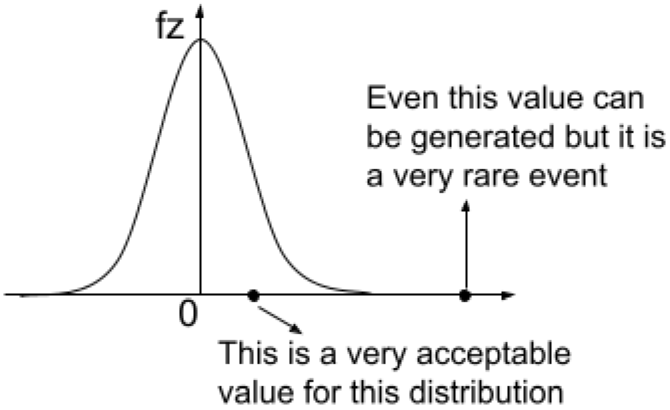
$z_n^* = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}$ (the asterisk indicates the empirical evidence)

Important! If H_0 is true then z^* is drawn from a standard gaussian $\overline{Z}_n \sim N(0, 1)$



We set α and find $\overline{Z}_{\alpha/2}$, if z^* is between thresholds we accept the hypothesis distribution (the value z^* could have been generated from the distribution described by H_0 ($\overline{Z}_n$) with a probability of 1 - α).

In summary:
If H$_0$ is true then the empirical evidence is a realisation of a standard gaussian $Z_n \sim N(0, 1)$.
But we know that Z_n can take values in the whole R, the only condition is that some values are more probable than others.


fz
0
Even this value can be generated but it is a very rare event
This is a very acceptable value for this distribution

So we set α to determine a significance level and derive the critical values
$Z_{\alpha/2}: P(Z_n > Z_{\alpha/2}) = \frac{\alpha}{2}$ **critical value**
We reject H$_0$ if:
$z^* > Z_{\alpha/2}$ or $z^* <- Z_{\alpha/2}$
Namely: $|z^*| > Z_{\alpha/2}$

Ex.
The ropes produced by a machine have an average break resistance of μ=1800N, with a standard deviation of σ=100N.
A new version of the machine is supposed to produce ropes with an improved resistance.
$X \sim N(\mu, \sigma^2 = 100^2)$
From the past we know:
$\begin{cases} H_0 & \mu = 1800N \\ H_1 & \mu > 1800N \end{cases}$ One sided hypothesis
α=0.01
We draw a random sample of 30 observations, and we get our empirical evidence: $\overline{x} = 1850N$
Now we take H$_0$ as true, and we find the random variable that has generated our empirical evidence:
$\overline{X} \sim N(1800, \sigma^2 = 100^2)$
In order to see if $\overline{x}$ is a realisation of $\overline{X}$ we can standardise both, then we will also be able to use the statistical tables.
Test statistic under H$_0$:
$\overline{Z} \sim \frac{\overline{X}-1800}{100/\sqrt{30}} \sim N(0, 1)$ null distribution

$Z_{0.01}$=2.33   threshold (0.01 = α)

P(Z>$Z_{0.01}$) = 0.01

$\Phi(Z_{0.01}) = 1 - \alpha$

$\Phi(2.33) = 0.99$

We will reject if $z^* > 2.33$

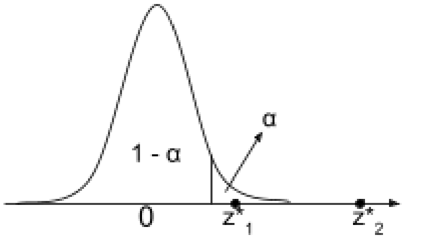
Standardisation of the empirical evidence:

$z^* = \frac{\bar{x}-1800}{100/\sqrt{30}} = \frac{1850-1800}{100/\sqrt{30}} = 2.74$

Since 2.74 is bigger than 2.33 we have to reject H_0, the new machines have improved

Measure of distance from the threshold

In the previous instances we usually set α and then rejected the values bigger than Z_α, now suppose we have an empirical evidence which is very close to the threshold but still bigger than Z_α.



Both values are over the threshold but with different distances

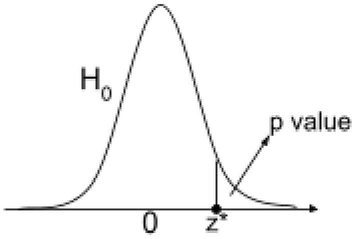
We cannot reject any value as if we were machines, we need a way to measure the distance from the threshold.

A smart way to have a valuable hypothesis testing is to let the evidence speak for itself, by introducing the p value.

P value

p value = $P(Z_n \geq z^* | H_0)$

We get rid of α and we replace z^* with the empirical evidence



The p value is essentially the probability of observing the empirical evidence or less probable events, under the null hypothesis.

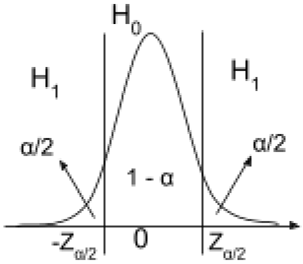
Closer the p value is to 0, stronger is the evidence against H_0.

Ex.
The exercise talked about some kind of machines, I don't care.
$\overline{Z} \sim N(0, 1)$ null distribution
$z^* = 2.74$ empirical evidence
$p\ value = P(Z \geq 2.74 | H_0) = 1 - \Phi(2.74) = 1 - 0.9969 = 0.0031$
This is the probability that, if the mean hasn’t improved (H_0 is true), we draw by chance a machine with a strength of 2.74 (we are referring to the standardised version of H_0).

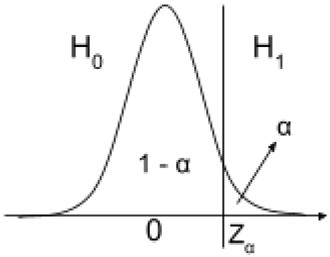
Region of acceptance
The region of acceptance/rejection depends upon the system of hypothesis:

- For α we had:

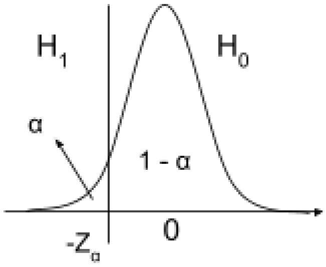
Double sided



right sided

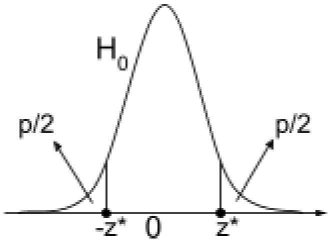


left sided

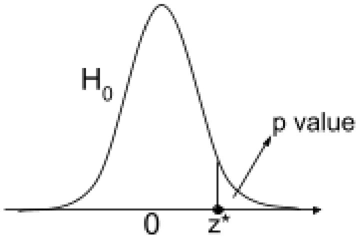


- For the p value we have:

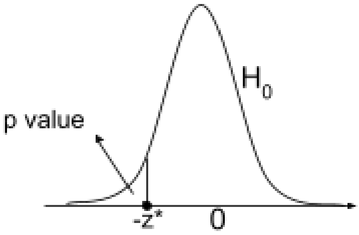
Double sided



right sided



left sided

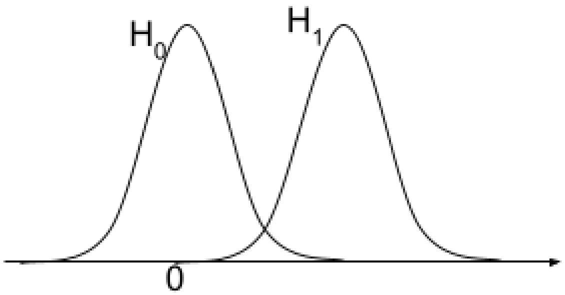


Type I and II errors

Suppose:

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu = \mu_1 \end{cases}$$

where $\mu_1 > \mu_0$

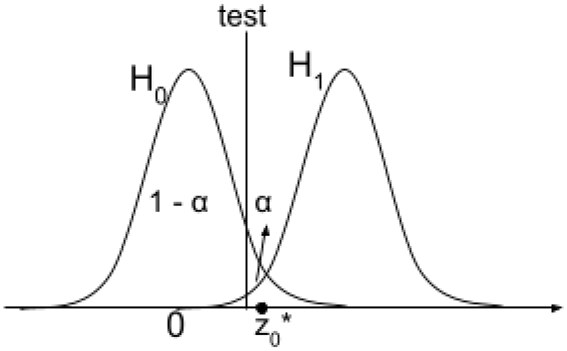


		test	
		H_0	H_1
truth	H_0	1 - α	α
	H_1	β	1 - β

If H_0 is true we shouldn’t reject the empirical value but we can have a value outside the threshold with a probability α.

- α: probability that i reject the H_0 given that H_0 is true.
  $P(reject\ H_0 | H_0\ true)$ false positive

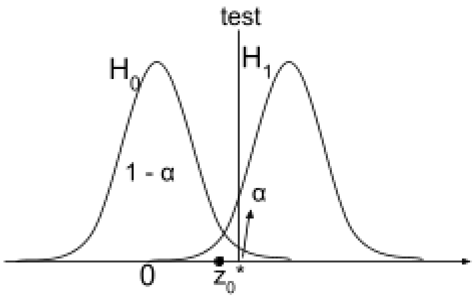
  This is the type 1 error, the test says that we have to reject the hypothesis (positive result of the test) but in reality we shouldn’t (α=probability of type 1 error).

  The empirical value is outside our threshold but it has been generated by H_0.

- 1 - α: $P(accept\ H_0 | H_0\ true)$

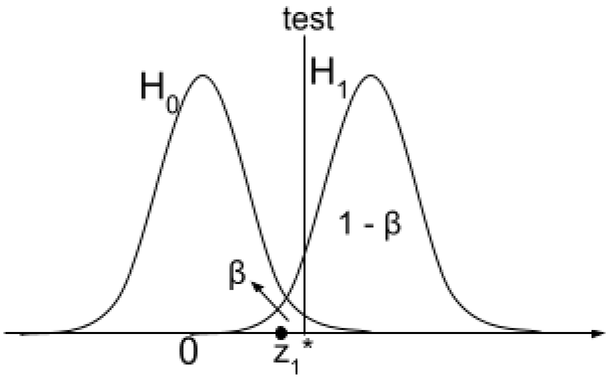
  If we want to reduce the probability of the type 1 error we could move α to the right but there are no free lunches.

  Now the empirical value is accepted under H_0 but at the same time, from this side of the threshold there are also a lot more “values” of H_1.

- β: probability that i accept H_0 given that H_0 is false
  $P(accept\ H_0 | H_0\ false)$ false negative

  This is type 2 error, it corresponds to the probability that the empirical evidence has been generated from another distribution, and by chance it is before the threshold.



In this case the considered empirical evidence has been generated from H_1 but we are considering it has a realisation of H_0.

If the threshold is very far, it could include the realisations of other distributions.

- 1 - β: $p(reject\ H_0 | H_0\ false)$ power of the test

In summary if we decrease α we increase β.

Ex.
Assume we are testing for a disease:

test

		H_0	H_1
truth	H_0		α
	H_1	β	

H_0: healthy
H_1: disease
α: probability that the test is positive but you are healthy (false positive)
β: probability that the test is negative but you are ill (false negative)
We usually set the probability of each error (by choosing the threshold) depending on the aim of our investigation.
In the case of cancer we want to avoid missing a diagnosis, it is preferable to keep the false negatives very low, it is fundamental to have an early diagnosis; while in the case of an epidemic disease maybe it is convenient to have a discrete number of false negative in order to not obstruct the sanitary system (few false positive).

Ex.
A pharmaceutical company wants to introduce a new drug

$$\begin{cases} H_0: & new\ drug\ has\ no\ effect \\ H_1: & new\ drug\ has\ a\ positive\ effect \end{cases}$$

This was one of the first statistical tests, we cal historically H_0 null hypothesis (null=no effect) test

		H_0	H_1
truth	H_0		α
	H_1	β	

α: $P(reject\ H_0|H_0\ true)\ =\ P(test\ says\ there\ is\ no\ effect\ |\ there\ is\ effect)$

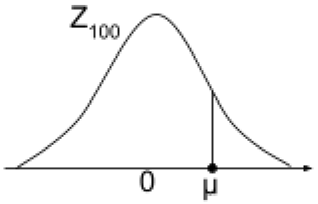
In pharmaceutical companies, before saying and investing on a new drug we have to be sure that it has an effect, we cannot accept false positive results from the test (tiny α).

β:$P(accept\ H_0|H_0\ false)\ =\ P(test\ says\ there\ is\ no\ effect\ |\ there\ is\ effect)$

Every statement or conclusion based on a statistical test should be accompanied by an assessment of the magnitude of the effect (i.e. confidence interval).

If the registered effects are tiny we don't want to invest money on the drug.

This problem is enhanced with big data:



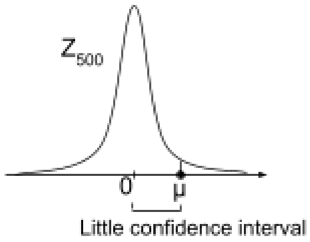
$$\begin{cases} H_0: & \mu = 0 \\ H_1: & \mu > 0 \end{cases}$$

Suppose that H_1 is true

The true mean is >0 but the null distribution will be centred around 0.

The values that we will experiment will be usually near to μ and not to 0.

If we use a large sample size the hypothetical distribution will be:



Now for the values near μ we have a small p value, we will be tempted to reject the null hypothesis (the drug has effect) but in reality the difference between the null hypothesis and the true value is so small that it is not convenient to invest on a new drug (small effect).

Hypothesis testing on the mean

1. $X \sim N(\mu, \sigma^2)$ σ^2 known

Two sided	right sided	left sided
$\begin{cases} H_0 : & \mu = \mu_0 \\ H_1 : & \mu \neq \mu_0 \end{cases}$	$\begin{cases} H_0 : & \mu = \mu_0 \\ H_1 : & \mu > \mu_0 \end{cases}$	$\begin{cases} H_0 : & \mu = \mu_0 \\ H_1 : & \mu < \mu_0 \end{cases}$

If H_0 is true then $\overline{X}_n \sim N(\mu_0, \frac{\sigma^2}{n})$

So the test statistic is:

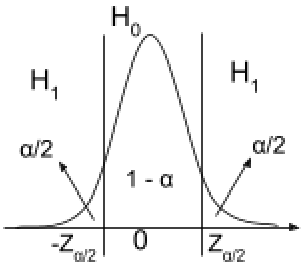
$$\overline{Z} = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

If H_0 is true thi is the standardised distribution that has generated our empirical evidence

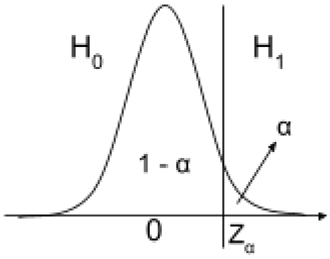
$$z^* = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}$$

We set α

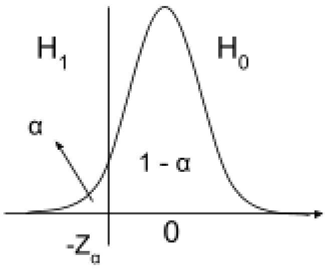
Double sided



right sided



left sided



Reject H_0 if:

Double sided	right sided	left sided
$\lvert z^* \rvert > Z_{\alpha/2}$	$z^* > Z_{\alpha}$	$z^* < - Z_{\alpha}$

P value:

Double sided	right sided	left sided
$2 * P(Z \geq \lvert z^* \rvert)$	$P(Z \geq z^*)$	$P(Z \leq z^*)$

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

We have to estimate σ^2: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

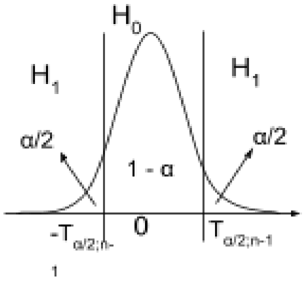
Under H_0 the statistic is:

$$T = \frac{\overline{X}_n - \mu_0}{S/\sqrt{n}} \sim T_{n-1}$$

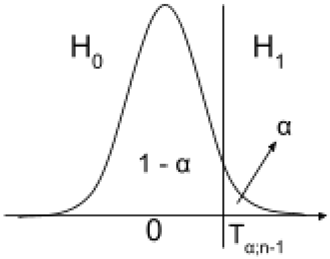
$t^* = \frac{\overline{x}_n - \mu_0}{S/\sqrt{n}}$ empirical evidence

We set α

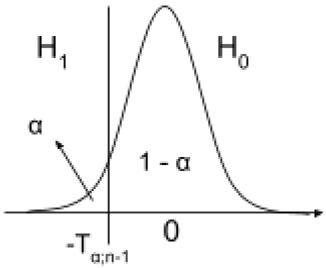
Double sided



right sided



left sided



Reject if:

$|t^*| > T_{\alpha/2}$ $\qquad t^* > T_{\alpha}$ $\qquad t^* <- T_{\alpha}$

P value:

$2 * P(T \geq |\bar{t}^*|)$ $\qquad P(T \geq \bar{t}^*)$ $\qquad P(T \leq \bar{t}^*)$

3. $X \sim f(\mu, \sigma^2)$ σ^2 known, f unknown

We have to use the central limit theorem to guess the distribution of H_0.

We must use a large sample size, to threat the hypothesis distribution as gaussian and we must estimate σ^2 with S^2

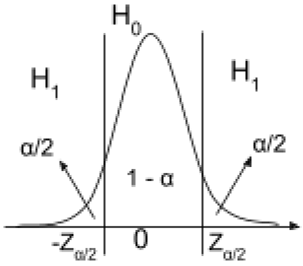
The test statistic is:

$\overline{Z} = \frac{\bar{X}_n - \mu_0}{S/\sqrt{n}} \rightarrow N(0, 1)$ when $n \rightarrow +\infty$ asymptotic null distribution

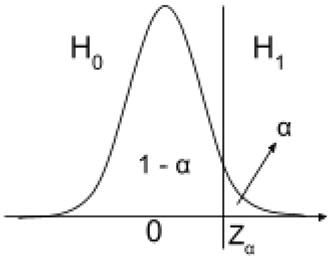
$z^* = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$ empirical evidence

We set α

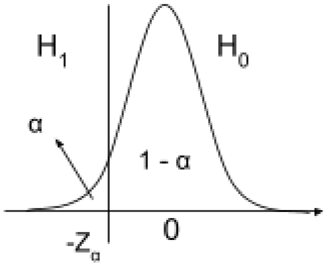
Double sided



right sided



left sided



Reject H_0 if:

$|z^*| > Z_{\alpha/2}$ $\qquad z^* > Z_{\alpha}$ $\qquad z^* <- Z_{\alpha}$

P value:

$2 * P(Z \geq |z^*|)$ $\qquad P(Z \geq z^*)$ $\qquad P(Z \leq z^*)$

Test for the difference of two means

1. Gaussian populations

   $X_1 \sim N(\mu_1, \sigma^2_1)$

   $X_2 \sim N(\mu_2, \sigma^2_2)$

   We would like to test whether $\mu_1 = \mu_2$

   H_0: $\mu_1 = \mu_2 \Rightarrow H_0$: $\mu_D = 0$ where $\mu_D = \mu_1 - \mu_2$

   a. Known variances σ^2_1, σ^2_2

      We draw random samples of size n_1, n_2.

Population	Sample size	Sample mean
$X_1 \sim N(\mu_1, \sigma^2_1)$	n_1	$\overline{X}_1 \sim N(\mu_1, \frac{\sigma^2_1}{n_1})$
$X_2 \sim N(\mu_2, \sigma^2_2)$	n_2	$\overline{X}_2 \sim N(\mu_2, \frac{\sigma^2_2}{n_2})$

H_0: $\mu_D = \mu_1 - \mu_2 = 0$

$(\overline{X}_1 - \overline{X}_2) \sim N(\mu_1 - \mu_2, \frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2})$ sum of gaussians = gaussian

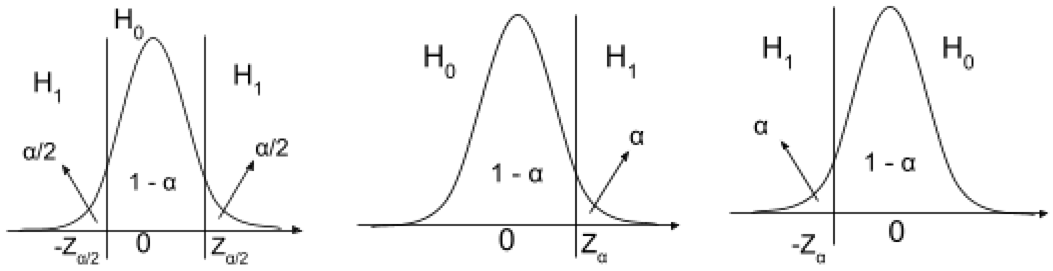
If H_0 is true the test statistic results:

$Z = \frac{(\overline{X}_1 - \overline{X}_2) - 0}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}} \sim N(0, 1)$ null distribution

$z^* = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}}$ empirical evidence

Two sided	right sided	left sided
$\begin{cases} H_0: & \mu_D = 0 \\ H_1: & \mu_D \neq 0 \end{cases}$	$\begin{cases} H_0: & \mu_D = 0 \\ H_1: & \mu_D > 0 \end{cases}$	$\begin{cases} H_0: & \mu_D = 0 \\ H_1: & \mu_D < 0 \end{cases}$
$(\mu_1 \neq \mu_2)$	$(\mu_1 > \mu_2)$	$(\mu_1 < \mu_2)$

We set α



Reject H$_0$ if:

$|z^*| > Z_{\alpha/2}$ $\qquad$ $z^* > Z_\alpha$ $\qquad$ $z^* <- Z_\alpha$

P value:

$2 * P(Z \geq |z^*|)$ $\qquad$ $P(Z \geq z^*)$ $\qquad$ $P(Z \leq z^*)$

b. σ^2_1 and σ^2_2 are unknown but equal

$\sigma^2_1 = \sigma^2_2$

Population	Sample size	Sample mean
$X_1 \sim N(\mu_1, \sigma^2_1)$	n$_1$	$\overline{X}_1 \sim N(\mu_1, \frac{\sigma^2_1}{n_1})$
$X_2 \sim N(\mu_2, \sigma^2_2)$	n$_2$	$\overline{X}_2 \sim N(\mu_2, \frac{\sigma^2_2}{n_2})$

S^2_1 estimates σ^2_1

S^2_2 estimates σ^2_2

But $\sigma^2_1 = \sigma^2_2$, hence we can build a pooled estimator for σ^2:

$$S^2_p = \frac{S^2_1(n_1-1)+S^2_2(n_2-1)}{n_1+n_2-2}$$

Under H$_0$:

$$T = \frac{\overline{X}_1-\overline{X}_2}{\sqrt{S^2_p(\frac{1}{n_1}+\frac{1}{n_2})}} \sim T_{n_1+n_2-2} \qquad \{V(\overline{X}_1 - \overline{X}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))\}$$

$$t^* = \frac{\overline{x}_1-\overline{x}_2}{\sqrt{S^2_p(\frac{1}{n_1}+\frac{1}{n_2})}}$$

Testing the association/dependence in contingency tables

Suppose we are given a random sample of americans classified according to:

- Gender (female-male)
- Belief in afterlife (yes-no)

n=2859

belief

		yes	no	marginals
gender	female	1230	357	1587
	male	859	413	1272
	marginals	2089	770	Tot = 2859

We can think of these absolute frequencies as a reflection of the bivariate probability.

Question: is there an association between belief and gender?

$$\begin{cases} H_0: & there\ is\ no\ association \quad (the\ variables\ are\ idependent) \\ H_1: & there\ is\ association/dependence \quad (a\ gender\ is\ more\ prone\ to\ belive\ in\ afterlife) \end{cases}$$

We cannot compute the correlation coefficient because we are dealing with categorical data (see “variance” paragraph of chapter “bivariate random variables”)

When 2 random variables(x, y) are independent:

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

We will rely on this to build our null hypothesis, then the empirical evidence will be compared against H_0 (the independence table).

> **Recall on independence:**
>
> X has h categories X_1, …, X_h, we will refer to these categories with the subscript i; in our example X is the gender, with categories male and female.
>
> Y has k categories Y_1, …, Y_k, we will refer to these categories with the subscript j; in our example Y is the belief, with categories yes and no.
>
> $P(X = x_i, Y = y_j) = \frac{n_{i,j}}{n} = \widehat{P}_{i,j}$
>
> $n_{i,j}$ represents the number of individuals in correspondence of the categories i and j ($n_{row,\ column}$ identifies a cell of the table), by dividing it for the total we obtain an absolute frequency.
>
> In our example:
>
> $P(G = F,\ belief = yes) = \frac{1230}{2859} = 0.43$
>
> If X and Y are independent:
>
> $P(X = x, Y = y) = P(X = x_i) * P(Y = y_j) = P_{i\bullet} * P_{j\bullet} = \frac{n_{i\bullet}}{n} * \frac{n_{j\bullet}}{n}$
>
> We use $i\ \bullet$ and $j\ \bullet$ because we are considering the marginal frequencies, we are considering all the cells with the category i (a row) or all the cells belonging to the category j (a column):
>
> In our example
>
> $P(G = F) = \frac{1587}{2859}$
>
> $P(belief = yes) = \frac{2089}{2859}$

Now we can build the independence table:

$$\begin{cases} H_0: & P_{ij} = P_{i\bullet} * P_{j\bullet} \\ H_1: & P_{ij} \neq P_{i\bullet} * P_{j\bullet} \end{cases}$$

a. Observed counts $n_{i,j}$ (empirical evidence)

   Since $\widehat{P}_{i,j} = \frac{n_{i,j}}{n} \rightarrow n_{i,j} = \widehat{P}_{i,j} * n$ (observed probability of a cell * total)

b. Theoretical counts under H$_0$

We know that $P_{i,j} = \widehat{P}_{i\bullet} * \widehat{P}_{\bullet j}$ are the joint probabilities under H$_0$.

$P_{ij}^{*} = \frac{n_{i\bullet}}{n} * \frac{n_{\bullet j}}{n} \Rightarrow n * P_{ij}^{*} = \frac{n_{i\bullet} * n_{\bullet j}}{n} = n_{ij}^{*}$ expected frequencies

We compare the theoretical frequencies n_{ij}^{*} with the observed empirical evidences n_{ij} (why the fuck is it the opposite in respect of all the information in the previous paragraphs!?)

Independence table

$n_{ij}^{*} = \frac{n_{i\bullet} * n_{\bullet j}}{n}$

		belief		
		yes	no	marginals
gender	female	1160	427	1587
	male	929	343	1272
	marginals	2089	770	Tot = 2859

Now that we have the empirical evidence and the independence table, we can use a statistical test to compare them.

Chi-square test

We have 2 bivariate distributions to compare:

- Observed $n_{i,j}$
- Theoretical $n_{i,j}^{*}$

We use the χ^2 statistic as a measure of distance between two distributions

$X^2 (capital\ chi) = \sum \frac{(O-E)^2}{E} = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(n_{i,j} - n_{i,j}^{*})^2}{n_{i,j}^{*}}$

Under H$_0$ (if the empirical are independent) $X^2 \sim \chi^2_{(n-1)*(k-1)}$ when $n \to +\infty$

At this point we can use the χ^2 tables:

$\chi^2 = \frac{(1230-1160)^2}{1160} + ... + \frac{(423-343)^2}{343} = 35.96$

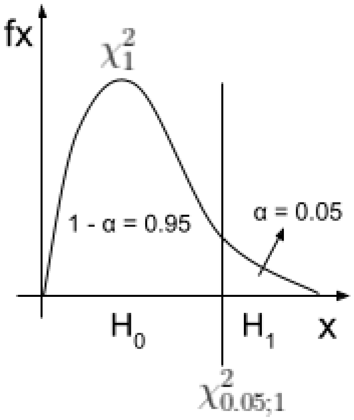
Degrees of freedom: k=h=2 (2-1)*(2-1)=1

Under H$_0$: $X^2 \sim \chi^2_1$

We set α = 0.05

$\chi^2_{0.05;1} = 3.81$

Since 35.60>3.841 we reject H$_0$ at 5% level.


fx
χ^2_1
1 - α = 0.95
α = 0.05
H$_0$
H$_1$
x
$\chi^2_{0.05;1}$