UCLouvain
Institut de recherche
en mathématique et physique

Statistics

or “How to find answers to your questions”

Pietro Vischia[1]

[1] CP3 — IRMP, Université catholique de Louvain

UCLouvain
Institut de recherche
en mathématique et physique

CP3—IRMP, Intensive Course on Statistics for HEP, 07–11 December 2020

UCLouvain
Institut de recherche
en mathématique et physique

- **Why statistics?**
- **Fundaments**
  - Set theory and measure theory
  - Frequentist probability
  - Bayesian probability
- **Random variables and their properties**
- **Causality**
  - The three levels of causal hierarchy
- **Distributions**

- Schedule: five days (Monday to Friday)
  - 2h morning lecture, virtual coffee break midway (09:30–11:45)
  - 2h (probably less) afternoon exercise session, virtual coffee break midway (13:30–15:45)
- Many interesting references, nice reading list for your career
  - Papers mostly cited in the topical slides
  - Some cool books cited here and there and in the appendix
- Unless stated otherwise, figures belong to P. Vischia for inclusion in my upcoming textbook on Statistics for HEP
  (textbook to be published by Springer in 2021)
  - Or I forgot to put the reference, let me know if you spot any figure obviously lacking reference, so that I can fix it
  - I cannot put the recordings publicly online as “massive online course”, so I will distribute them only to registered participants, and have to ask you to not record yourself. I hope you understand.
- Your feedback is crucial for improving these lectures (a feedback form will be provided at the end of the lectures)!
  - You can also send me an email during the lectures: if it is something I can fix for the next day, I'll gladly do so!

- This course provides 3 credits for the UCLouvain doctoral school (CDD Sciences)
  - If you need it recognized by another doctoral school, you have to ask to your school
  - Besides the certificate, I am available at supplying additional information (e.g. detailed schedule) or activity (exam? LoL)
- Online only: certificates will be provided by checking connection logs
  - The only way I have to check if you connected to most lectures is to check the Zoom logs
  - Make sure you connect with a recognizable email address (or let me know which unrecognizable address belongs to you)
- This course contributes to the activities of the Excellence of Science (EOS) Be.h network, https://be-h.be/

UCLouvain
Institut de recherche
en mathématique et physique

- I will pop up every now and then some questions
- I will open a link, and you'll be able to answer by going to www.menti.com and inserting a code
- Totally anonymous (no access even for me to any ID information, not even the country): don't be afraid to give a wrong answer!
  - The purpose is making you think, not having 100% correct answers!
- First question of the day is purely a logistics matter
  Question time: ROOT
  - The direct links are accessible to me only: you'll see in your screens the code in a second :)
- The slides of each lecture will be available one minute after the end of the lecture
  - To encourage you to really try answering without looking at the answers

General outline of the lectures

- **Lesson 1 - Fundaments**
  - Bayesian and frequentist probability, theory of measure, correlation and causality, distributions
- **Lesson 2 - Point and Interval estimation**
  - Maximum likelihood methods, confidence intervals, most probable values, credible intervals
- **Lesson 3 - Advanced interval estimation, test of hypotheses**
  - Interval estimation near the physical boundary of a parameter
  - Frequentist and Bayesian tests, CLs, significance, look-elsewhere effect, reproducibility crysis
- **Lesson 4 - Commonly-used methods in particle physics**
  - Unfolding, ABCD, ABC, MCMC, estimating efficiencies
- **Lesson 5 - Machine Learning**
  - Overview and mathematical foundations, generalities most used algorithms, automatic Differentiation and Deep Learning

Why statistics?

Statistics is all about answering questions...

- What is the chance of obtaining a 1 when throwing a six-faced die?
- What is the chance of tomorrow being rainy?

Statistics is all about answering questions...

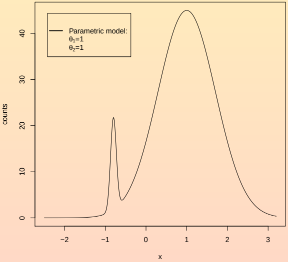
- What is the chance of obtaining a 1 when throwing a six-faced die?
  - We can throw a dice 100 times, and count how many times we obtain 1
- What is the chance of tomorrow being rainy?

UCLouvain
Institut de recherche
en mathématique et physique

- What is the chance of obtaining a 1 when throwing a six-faced die?
  - We can throw a dice 100 times, and count how many times we obtain 1
- What is the chance of tomorrow being rainy?
  - We can try to give an answer based on the recent past weather, but we cannot – in general – *repeat tomorrow* and count

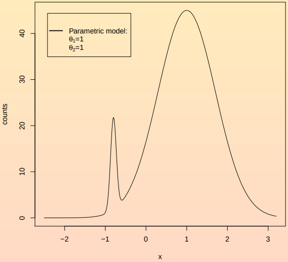
Image from "The Tiger Lillies" Facebook page

- **Theory**
  - Approximations
  - Free parameters

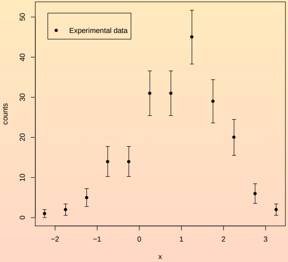


Where does statistics live

- **Theory**
  - Approximations
  - Free parameters

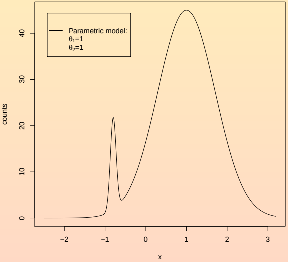


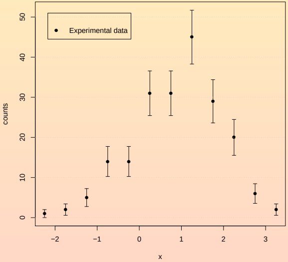
- **Experiment**
  - Random fluctuations
  - Mismeasurements (detector effects, etc)



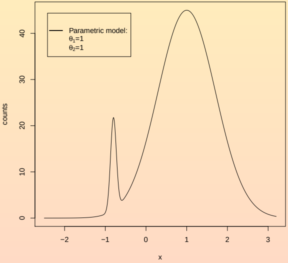
Where does statistics live

- **Theory**
  - Approximations
  - Free parameters
- **Statistics!**
- **Experiment**
  - Random fluctuations
  - Mismeasurements (detector effects, etc)

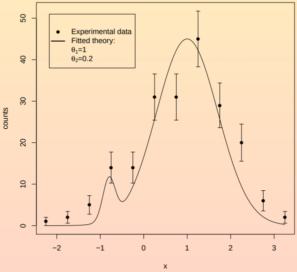




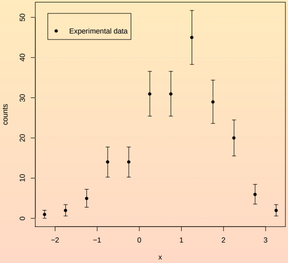
- **Theory**
  - Approximations
  - Free parameters



- **Statistics!**
  - Estimate parameters
  - Quantify uncertainty in the parameters estimate
  - Test the theory!



- **Experiment**
  - Random fluctuations
  - Mismeasurements (detector effects, etc)



Fundaments

- Ω: set of all possible elementary (exclusive) events X_i
- Exclusivity: the occurrence of one event implies that none of the others occur
- Probability then is any function that satisfies the *Kolmogorov axioms*:
  - $P(X_i) \geq 0, \forall i$
  - $P(X_i \; or \; X_j) = P(X_i) + P(X_j)$
  - $\sum_{\Omega} P(X_i) = 1$

Andrey Kolmogorov.

UCLouvain
Institut de recherche
en mathématique et physique

- Cox postulates: formalize a set of axioms starting from reasonable premises
  - doi:10.1119/1.1990764
- Notation
  - $A|B$ the plausibility of the proposition A given a related proposition B
  - $\sim A$ the proposition “not-A”, i.e. answering “no” to *“is A wholly true?”*
  - $F(x, y)$ a function of two variables
  - $S(x)$ a function of one variable
- The two postulates are
  - $C \cdot B|A = F(C|B \cdot A, B|A)$
  - $\sim V|A = S(B|A)$, i.e. $(B|A)^m + (\sim B|A)^m = 1$
- Cox theorem acts on propositions, Kolmogorov axioms on sets
- Jaynes adheres to Cox’ exposition and shows that formally this is equivalent to Kolmogorov theory
  - Kolmogorov axioms somehow arbitrary
  - A proposition referring to the real world cannot always be viewed as disjunction of propositions from any meaningful set
  - Continuity as infinite states of knowledge rather than infinite subsets
  - Conditional probability not originally defined

UCLouvain
Institut de recherche
en mathématique et physique

- Theory of probability originated in the context of games of chance
- Mathematical roots in the theory of Lebesgue measure and set functions in $\mathbb{R}^n$
- Measure is something we want to define for an interval in $\mathbb{R}^n$
  - 1D: the usual notion of length
  - 2D: the usual notion of area
  - 3D: the usual notion of volume
- Interval $i = a_\nu \leq x_\nu \leq a_\nu$

$$L(i) = \prod_{\nu=1}^{n} (b_\nu - a_\nu).$$

  - The length of degenerate intervals $a_\nu = b_\nu$ is $L(i) = 0$; it does therefore not matter the interval is closed, open, or half-open;
  - We set to $+\infty$ the length of any infinite non-degenerate interval such as $]25, +\infty]$ or $[-\infty, 2]$.
- But do we connect different intervals?

The Borel lemma and the additivity of length

- In $\mathbb{R}^1$, an interval $[a, b]$ has length:

$$L(i) = b - a$$
$$L(a, a) = 0$$
$$L(\infty) = \infty.$$

- Disjoint intervals (no common point with any other)

$$i = i_1 + ... + i_n, \qquad (i_\mu i_\nu = 0 \text{ for } \mu \neq \nu);$$

- Define the sum as $L(i) := L(i_1) + ... + L(i_n)$
    - Extendable to an enumerable sequence of intervals (crucial for defining continuous density functions)
- **Borel lemma:** we consider a finite closed interval $[a, b]$ and a set of Z intervals such that every point of $[a, b]$ is an inner point of at least one interval belonging to Z.
    - Then there is a subset Z' of Z containing only a finite number of intervals, such that every point of $[a, b]$ is an inner point of at least one interval belonging to Z'.
- Generalizable to N dimensions, with $L(i)$ additive function of i: $i = \sum i_n \Rightarrow L(i) = \sum L(i_n)$

- $L(i)$ is a non-negative additive function (finite- or infinite-valued): a measure
- Definition extendable from intervals to complex sets:
  - $L(S) \geq 0$
  - If $S = S_1 + \ldots + S_n$, where $S_\mu S_\nu = 0$ for $\mu \neq \nu$ then $L(S) = L(S_1) + \ldots + L(S_n)$
  - If S is an interval i, then the set function $L(S)$ reduces itself to the interval function $L(i)$, $L(S) = L(i)$
- True only for Borel sets
  - In layman's terms, sets that can be constructed by taking countable unions or intersections (and their respective complements) of open sets
- $L(S)$ is a measure and it's called Lebesgue measure
  - The extension from $L(i)$ to $L(S)$ is unique (the only set function defined on the whole $\mathcal{B}_1$ satisfying the properties above)
  - Extension to $\mathbb{R}^n$ is immediate: $L_n(S)$
-

- Generalization of $L_n(S)$: the P-measure
  1. $P(S)$ is non-negative, $P(S) \geq 0$;
  2. $P(S)$ is additive, $P(S_1 + ... + S_n) = P(S_1) + ... + P(S_n)$ where $S_\mu S_\nu = 0$ for $\mu \neq \nu$;
  3. $P(S)$ is finite for any bounded set (crucial to define the usual probability in the domain $[0, 1]$
-
- Associate to any $P(S)$ a point function $F(\boldsymbol{x}) = F(x_1, ..., x_n)$

$$F(\boldsymbol{x}) = F(x_1, ..., x : n) := P(\xi_1 \leq x_1, ..., \xi_n \leq x_n).$$

  - Trivial in one dimension. $P(S)$ must have an upper bound!
  - Map $F(a) = F(b)$ to set of null P-measure, $P(a < x \leq b) = 0$
- $F(\boldsymbol{x})$ is in each point a non-decreasing function everywhere-continuous to the right

$$P(a < x \leq a + h) = \Delta F(a) = F(a + h) - F(a),$$

- Consider a class of non-negative additive set functions $P(S)$ such that $P(\mathbb{R}^n) = 1$; then

$$
\begin{aligned}
F(\boldsymbol{x}) = F(x_1, ..., x_n) &= P(\xi \leq x_1, ..., \xi_n \leq x_n) \\
0 \leq F(\boldsymbol{x}) &\leq 1 \\
\Delta_n F &\geq 0 \\
F(-\infty, x_2, ..., x_n) &= ... = F(x_1, ..., x_n - 1, -\infty) = 0 \\
F(+\infty, ..., +\infty) &= 1.
\end{aligned}
$$

- We interpret $P(S)$ and $F(\boldsymbol{x})$ as distribution of a unit of mass over $\mathbb{R}^n$
  - Each Borel set carries the mass $P(S)$
  - Interpret $(\boldsymbol{x}$ as the quantity of mass allotted to the infinite interval $(\xi_1 \leq x_1, \ldots, \xi_n \leq x_\nu)$.
  - Defining the measure in terms of $P(S)$ or $F(\boldsymbol{x})$ is equivalent
- Usually $P(S)$ is called probability function, and $F(\boldsymbol{x})$ is called distribution function
- σ-field: a space Ω equipped with a collection of subsets containing Ω, closed by complement and by under countable union
  - The original Kolmogorov approach is expressed via a σ-field built on the space of elementary propositions (sets)

UCLouvain
Institut de recherche
en mathématique et physique

- Discrete mass point $\boldsymbol{a}$; a point such that the set $\{\boldsymbol{x} = \boldsymbol{a}\}$ carries a positive quantity of mass.

$$P(S) = c_1 P_1(S) + c_2 P_2(S)$$

or

$$F(\boldsymbol{x}) = c_1 F_1(\boldsymbol{x}) + c_2 F_2(\boldsymbol{x})$$

where

$$c_\nu \geq 0, \qquad c_1 + c_2 = 1,$$

- c_1: component with whole mass concentrated in discrete mass points. c_2: component with no discrete mass points
- $c_1 = 1$, $c_2 = 0$: $F(\boldsymbol{x})$ is a step function, where the whole mass is concentrated in the discontinuity points
- $c_1 = 0$, $c_2 = 1$, then if $n = 1$ then $F(\boldsymbol{x})$ is everywhere continuous, and in any dimension no single mass point carries a positive quantity of mass.

- Consider the n-dimensional interval $i = \{x_\nu - h_\nu < \xi_\nu \leq x_\nu + h_\nu ; \nu = 1, ..., n\}$
- Average density of mass: the ratio of the P-measure of the interval—expressed in terms of the increments of the point function—to the L-measure of the interval itself

$$\frac{P(i)}{L(i)} = \frac{\Delta_n F}{2^n h_1 h_2 ... h_n}.$$

- If partial derivatives $f(x_1, ..., x_n) = \frac{\partial_n F}{\partial x_1 \ldots \partial x_n}$ exist, then $\frac{P(i)}{L(i)} \to f(x_1, ..., x_n)$ for $h_\nu \to 0$
  - Density of mass at the point $\boldsymbol{x}$
  - f is referred to as probability density or frequency function

Marginal distributions

- Take a distribution function $F(x_1, ..., x_n)$
- Let $x_\mu \to \infty$, $\mu \neq \nu$
- It can be shown that $F \to F_\nu(x_\nu)$, and that itself is a distribution function in the variable x_ν
  - e.g. $F_1(x_1) = F(x_1, +\infty, ..., +\infty)$.
- $F_\nu(x_\nu)$ is one-dimensional, and is called the marginal distribution of x_ν.
  - It can be obtained by projection starting from the $n-$dimensional distribution
  - Shift each "mass particle" along the perpendicular direction to x_ν until collapsing into the x_ν axis
  - This results in a one-dimensional distribution which is the marginal distribution of x_ν.
  - There are infinite ways of arriving to the same x_ν starting from a generic n-dimensional distribution function
- Marginal distributions can be also built with respect to subsets of variables.

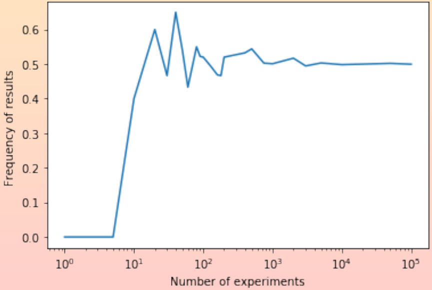
Random experiment

UCLouvain
Institut de recherche
en mathématique et physique

- Repeat a random experiment ξ (e.g. toss of a die) many times under uniform conditions
  - As uniform as possible
  - $\vec{S}$: set of all a priori possible different results of an individual measurement
  - S: a fixes subset of $\vec{S}$
- If in an experiment we obtain $\xi \in S$, we will say the event defined by $\xi \in S$ has occurred
  - We assume that S is simple enough that we can tell whether ξ is in it or not
- Throw a die: $\vec{S} = \{1, 2, 3, 4, 5, 6\}$
  - If $S = \{2, 4, 6\}$, then $\xi \in S$ corresponds to the event in which you obtain an even number of points
- Repeat the experiment: among n repetitions the event has occurred ν times
  - Then $\frac{\nu}{n}$ is the frequency ratio of the event in the sequence of n experiments
- Question time: Frequency Ratio

Random experiment

- Repeat a random experiment ξ (e.g. toss of a die) many times under uniform conditions
  - As uniform as possible
  - $\vec{S}$: set of all a priori possible different results of an individual measurement
  - S: a fixes subset of $\vec{S}$
- If in an experiment we obtain $\xi \in S$, we will say the event defined by $\xi \in S$ has occurred
  - We assume that S is simple enough that we can tell whether ξ is in it or not
- Throw a die: $\vec{S} = \{1, 2, 3, 4, 5, 6\}$
  - If $S = \{2, 4, 6\}$, then $\xi \in S$ corresponds to the event in which you obtain an even number of points
- Repeat the experiment: among n repetitions the event has occurred ν times
  - Then $\frac{\nu}{n}$ is the frequency ratio of the event in the sequence of n experiments
- Question time: Frequency Ratio
- This afternoon: obtain the answer by simulation!



- The most familiar one: based on the possibility of repeating an experiment many times
- Consider one experiment in which a series of N events is observed.
- n of those N events are of type X
- Frequentist probability for any single event to be of type X is the empirical limit of the frequency ratio:

$$P(X) = \lim_{N\to\infty} \frac{n}{N}$$

UCLouvain
Institut de recherche
en mathématique et physique

- The experiment must be repeatable in the same conditions
- The job of the physicist is making sure that all the *relevant* conditions in the experiments are the same, and to correct for the unavoidable changes.
  - Yes, *relevant* can be a somehow fuzzy concept
- In some cases, you can directly build the full table of frequencies (e.g. dice throws, poker)
- What if the experiment cannot be repeated, making the concept of frequency ill-defined?

Hand	Distinct Hands	Frequency	Probability	Cumulative probability	Odds	Mathematical expression of absolute frequency
Royal flush	1	4	0.000154%	0.000154%	649,739 : 1	$\binom{4}{1}$
Straight flush (excluding royal flush)	9	36	0.00139%	0.0014%	72,192 : 1	$\binom{10}{1}\binom{4}{1}-\binom{4}{1}$
Four of a kind	156	624	0.0240%	0.0256%	4,164 : 1	$\binom{13}{1}\binom{12}{1}\binom{4}{1}$
Full house	156	3,744	0.1441%	0.17%	693 : 1	$\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}$
Flush (excluding royal flush and straight flush)	1,277	5,108	0.1965%	0.367%	508 : 1	$\binom{13}{5}\binom{4}{1}-\binom{10}{1}\binom{4}{1}$
Straight (excluding royal flush and straight flush)	10	10,200	0.3925%	0.76%	254 : 1	$\binom{10}{1}\binom{4}{1}^5-\binom{10}{1}\binom{4}{1}$
Three of a kind	858	54,912	2.1128%	2.87%	46.3 : 1	$\binom{13}{1}\binom{4}{3}\binom{12}{2}\binom{4}{1}^2$
Two pair	858	123,552	4.7539%	7.62%	20.0 : 1	$\binom{13}{2}\binom{4}{2}^2\binom{11}{1}\binom{4}{1}$
One pair	2,860	1,098,240	42.2569%	49.9%	1.37 : 1	$\binom{13}{1}\binom{4}{2}\binom{12}{3}\binom{4}{1}^3$
No pair / High card	1,277	1,302,540	50.1177%	100%	0.995 : 1	$\left[\binom{13}{5}-10\right]\left[\binom{4}{1}^5-4\right]$
Total	7,462	2,598,960	100%	—	0 : 1	$\binom{52}{5}$

fnrs LA LIBERTÉ DE CHERCHER
UCLouvain
Institut de recherche
en mathématique et physique

- Based on the concept of degree of belief
  - $P(X)$ is the subjective degree of belief on X being true
- De Finetti: operative definition of subjective probability, based on the concept of coherent bet
  - We want to determine $P(X)$; we assume that if you bet on X, you win a fixed amount of money if X happens, and nothing (0) if X does not happen
  - In such conditions, it is possible to define the probability of X happening as

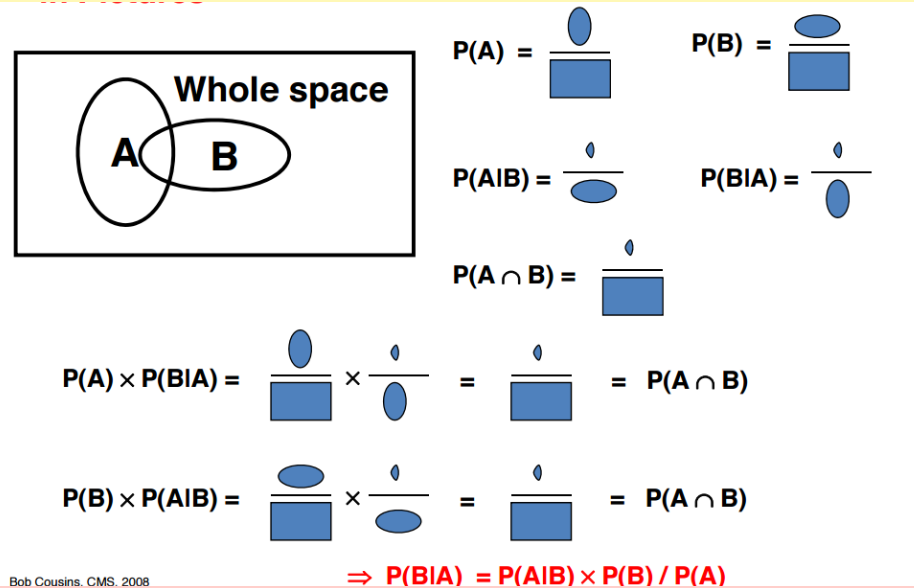
$$P(X) := \frac{\text{The largest amount you are willing to bet}}{\text{The amount you stand to win}} \tag{1}$$

- Coherence is a crucial concept
  - You can leverage your bets in order to try and not loose too much money in case you are wrong
  - Your bookie is doing a Dutch book on you if the set of bets guarantees a profit to him
  - You are doing a Dutch book on your bookie if the set of bets guarantees a profit to you
  - A bet is coherent if a Dutch book is impossible
- This expression is mathematically a Kolmogorov probability!
- Subjective probability is a property of the observer as much as of the observed system
  - It depends on the knowledge of the observer prior to the experiment, and is supposed to change when the observer gains more knowledge (normally thanks to the result of an experiment)

Book	Odds	Probability	Bet	Payout
Trump elected	Even (1 to 1)	$1/(1+1) = 0.5$	20	$20+20=40$
Clinton elected	3 to 1	$1/(1+3) = 0.25$	10	$10+30=40$
		$0.5+0.25=0.75$	30	40

Conditional probabilities: Bayes theorem

- Interestingly, Venn diagrams were the basis of Kolmogorov approach (Jaynes, 2003)



A word of advice about conditional probabilities

- Conditional probabilities are not commutative! $P(A|B) \neq P(B|A)$
- Example:
  - *speak English*: the person speaks English
  - *have TOEFL*: the person has a TOEFL certificate
- The probability for an English speaker to have a TOEFL certificate, $P(have\ TOEFL|speak\ English)$, is very small ($<< 1\%$)
- The probability for a TOEFL certificate holder to speak English, $P(speak\ English|have\ TOEFL)$, is (hopefully) $>>>>> 1\%$ ☺

Understanding conditioning can help even in marketing campaigns

From https://www.reddit.com/r/dataisugly/comments/boo6ld/when_venn_diagram_goes_wrong/

- Suppose you're on a game show, and you're given the choice of three doors
  - Behind one door is a car;
  - behind the others, goats.
- You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, which has a goat.
- He then says to you, "Do you want to pick door No. 2?"
- Is it to your advantage to switch your choice?
  Question time: Monty Hall

- Suppose you're on a game show, and you're given the choice of three doors
  - Behind one door is a car;
  - behind the others, goats.
- You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, which has a goat.
- He then says to you, "Do you want to pick door No. 2?"
- Is it to your advantage to switch your choice?
  Question time: Monty Hall
- The best strategy is to always switch!
- The key is the presenter knows where the car is $\rightarrow$ he opens different doors
  - The picture would be different if the presenter opened the door at random

UCLouvain
Institut de recherche
en mathématique et physique

- Suppose you're on a game show, and you're given the choice of three doors
  - Behind one door is a car;
  - behind the others, goats.
- You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, which has a goat.
- He then says to you, "Do you want to pick door No. 2?"
- Is it to your advantage to switch your choice?
  Question time: Monty Hall
- The best strategy is to always switch!
- The key is the presenter knows where the car is $\rightarrow$ he opens different doors
  - The picture would be different if the presenter opened the door at random
  - For the unconvinced: this afternoon we'll build a small simulation to check your answer!

fnrs LA LIBERTÉ DE CHERCHER
UCLouvain
Institut de recherche
en mathématique et physique

- Suppose you're on a game show, and you're given the choice of three doors
  - Behind one door is a car;
  - behind the others, goats.
- You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, which has a goat.
- He then says to you, "Do you want to pick door No. 2?"
- Is it to your advantage to switch your choice?
  Question time: Monty Hall
- The best strategy is to always switch!
- The key is the presenter knows where the car is $\rightarrow$ he opens different doors
  - The picture would be different if the presenter opened the door at random
  - For the unconvinced: this afternoon we'll build a small simulation to check your answer!

Behind 1	Behind 2	Behind 3	If you keep 1	If you switch	Presenter opens
Car	Goat	Goat	Win car	Win goat	2 or 3
Goat	Car	Goat	Win goat	Win car	3
Goat	Goat	Car	Win goat	Win car	2

UCLouvain
Institut de recherche
en mathématique et physique

- Bayes Theorem (1763)[1]:

$$P(A|B) := \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

- Valid for any Kolmogorov probability
- The theorem can be expressed also by first starting from a subset B of the space
- Decomposing the space S in disjoint sets A_i (i.e. $\cap A_i A_j = 0 \forall i, j$), $\cup_i A_i = S$ an expression can be given for B as a function of the A_is, the Law of Total Probability:

$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i) \tag{3}$$

- where the second equality holds only for if the A_is are disjoint
- Finally, the Bayes Theorem can be rewritten using the decomposition of S as:

$$P(A|B) := \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)} \tag{4}$$

[1] Actually the Bayesian approach has been mainly developed and popularized by Pierre Simon de Laplace

A Diagnosis problem

UCLouvain
Institut de recherche
en mathématique et physique

- The Bayes theorem permits to "invert" conditional probabilities, and can be applied to any Kolmogorov probability, therefore in particular to both frequentist and Bayesian defintions
- Let's consider a mortal disease, and label the possible states of the patients
  - D: the patient is diseased (sick)
  - H: the patient is healthy
- Let's imagine we have devised a diagnostic test, characterized by the possible results
  - +: the test is positive to the disease
  - -: the test is negative to the disease
- Imagine the test is very good in identifying sick people: $P(+|D) = 0.99$, and that the false positives percentage is very low: $P(+|H) = 0.01$
- You take the test, and the test is positive. Do you have the disease? Question time: Testing a Disease

A Diagnosis problem

- The Bayes theorem permits to “invert” conditional probabilities, and can be applied to any Kolmogorov probability, therefore in particular to both frequentist and Bayesian defintions
- Let’s consider a mortal disease, and label the possible states of the patients
  - D: the patient is diseased (sick)
  - H: the patient is healthy
- Let’s imagine we have devised a diagnostic test, characterized by the possible results
  - +: the test is positive to the disease
  - -: the test is negative to the disease
- Imagine the test is very good in identifying sick people: $P(+|D) = 0.99$, and that the false positives percentage is very low: $P(+|H) = 0.01$
- You take the test, and the test is positive. Do you have the disease? Question time: Testing a Disease
- By the Bayes Theorem:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|H)P(H)} \tag{5}$$

UCLouvain
Institut de recherche
en mathématique et physique

- The Bayes theorem permits to "invert" conditional probabilities, and can be applied to any Kolmogorov probability, therefore in particular to both frequentist and Bayesian defintions
- Let's consider a mortal disease, and label the possible states of the patients
  - D: the patient is diseased (sick)
  - H: the patient is healthy
- Let's imagine we have devised a diagnostic test, characterized by the possible results
  - +: the test is positive to the disease
  - -: the test is negative to the disease
- Imagine the test is very good in identifying sick people: $P(+|D) = 0.99$, and that the false positives percentage is very low: $P(+|H) = 0.01$
- You take the test, and the test is positive. Do you have the disease? Question time: Testing a Disease
- By the Bayes Theorem:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|H)P(H)} \tag{5}$$

- We need the incidence of the disease in the population, $P(D)$! Back to question time: Testing a Disease

A Diagnosis problem

UCLouvain
Institut de recherche
en mathématique et physique

- The Bayes theorem permits to “invert” conditional probabilities, and can be applied to any Kolmogorov probability, therefore in particular to both frequentist and Bayesian defintions
- Let’s consider a mortal disease, and label the possible states of the patients
  - D: the patient is diseased (sick)
  - H: the patient is healthy
- Let’s imagine we have devised a diagnostic test, characterized by the possible results
  - +: the test is positive to the disease
  - -: the test is negative to the disease
- Imagine the test is very good in identifying sick people: $P(+|D) = 0.99$, and that the false positives percentage is very low: $P(+|H) = 0.01$
- You take the test, and the test is positive. Do you have the disease? Question time: Testing a Disease
- By the Bayes Theorem:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|H)P(H)} \tag{5}$$

- We need the incidence of the disease in the population, $P(D)$! Back to question time: Testing a Disease
  - It turns out $P(D)$ is a very important to get our answer
  - $P(D) = 0.001$ (very rare disease): then $P(D|+) = 0.0902$, which is fairly small
  - $P(D) = 0.01$ (only a factor 10 more likely): then $P(D|+) = 0.50$, which is pretty high
  - $P(D) = 0.1$: then $P(D|+) = 0.92$, almost certainty!

UCLouvain
Institut de recherche
en mathématique et physique

- Frequentist and Subjective probabilities differ in the way of interpreting the probabilities that are written within the Bayes Theorem
- Frequentist: probability is associated to sets of data (i.e. to results of repeatable experiments)
  - Probability is defined as a limit of frequencies
  - Data are considered random, and each point in the space of theories is treated independently
  - An hypothesis is either true or false; improperly, its probability can only be either 0 or 1. In general, $P(hypothesis)$ is not even defined
  - "This model is preferred" must be read as "I claim that there is a large probability that the data that I would obtain when sampling from the model are similar to the data I already observed"[2]
  - We can only write about $P(data|model)$
- Bayesian statistics: the definition of probability is extended to the subjective probabilty of models or hypotheses:

$$P(H|\vec{X}) := \frac{P(\vec{X}|H)\pi(H)}{P(\vec{X})} \tag{6}$$

[2] Typically it's difficult to estimate this probability, so one reduces the data to a summary statistic $S(data)$ with known distribution, and computes how likely is to see $S(data_{sampled}) = S(data_{obs})$ when sampling from the model

UCLouvain
Institut de recherche
en mathématique et physique

$$P(H|\vec{X}) := \frac{P(\vec{X}|H)\pi(H)}{P(\vec{X})} \tag{7}$$

- $\vec{X}$, the vector of observed data
- $P(\vec{X}|H)$, the likelihood function, which fully summarizes the result of the experiment (experimental resolution)
- $\pi(H)$, the probability of the hypothesis H. It represents the probability we associate to H before we perform the experiment
- $P(\vec{X})$, the probability of the data.
  - Since we already observed them, it is essentially regarded as a normalization factor
  - Summing the probability of the data for all exclusive hypotheses (by the Law of Total Probability), $\sum_i P(\vec{X}|H_i) = 1$ (assuming that at least one H_i is true).
  - Usually, the denominator is omitted and the equality sign is replaced by a proportionality sign

$$P(H|\vec{X}) \propto P(\vec{X}|H)\pi(H) \tag{8}$$

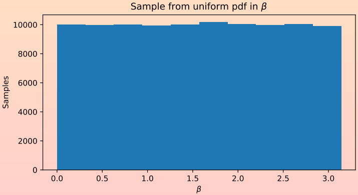
- $P(H|\vec{X})$, the posterior probability; it is obtained as a result of an experiment
- If we parameterize H with a (continuous or discrete) parameter, we can use the parameter as a proxy for H, and instead of writing $P(H(\theta))$ we write $P(\theta)$ and

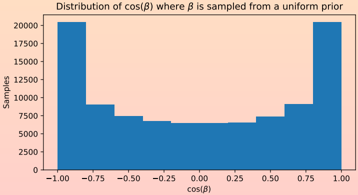
$$P(\theta|\vec{X}) \propto P(\vec{X}|\theta)\pi(\theta) \tag{9}$$

- The simplified expression is usually used, unless when the normalization is necessary
  - "Where is the value of θ such that $\theta_{true} < \theta_c$ with 95% probability?"; integration is needed and the normalization is necessary
  - "Which is the mode of the distribution?"; this is independent of the normalization, and it is therefore not necessary to use the normalized expression

Choosing a prior in Bayesian statistics; in theory... 1/

- There is no golden rule for choosing a prior
- Objective Bayesian school: it is necessary to write a golden rule to choose a prior
  - Usually based on an invariance principle
- Consider a theory parameterized with a parameter, e.g. an angle β
- Before any experiment, we are Jon Snow about the parameter β: we know nothing
  - We have to choose a very broad prior, or better uniform, in β
- Now we interact with a theoretical physicist, who might have built her theory by using as a parameter of the model the cosine of the angle, $cos(\beta)$
  - In a natural way, she will express her pre-experiment ignorance using an uniform prior **in** $cos(\beta)$.
  - This prior is not constant in β!!!
  - In general, there is no uniquely-defined prior expressing complete ignorance or ambivalence in both parameters (β and $cos(\beta)$)
- We can build a prior invariant for transformations of the parameter, but this means we have to postulate an invariance principle
  - The prior already deviates from our degree of belief about the parameter ("I know nothing")





UCLouvain
Institut de recherche
en mathématique et physique

- Two ways of solving the situation
  - Objective Bayes: use a formal rule dictated by an invariance principle
  - Subjective Bayes: use something like elicitation of expert opinion
    - Ask an expert her opinion about each value of θ, and express the answer as a curve
    - Repeat this with many experts
    - 100 years later check the result of the experiments, thus verifying how many experts were right, and re-calibrate your prior
    - This corresponds to a IF-THEN proposition: "IF the prior is $\pi(H)$, THEN you have to update it afterwards, taking into account the result of the experiment"
- Central concept: update your priors after each experimient

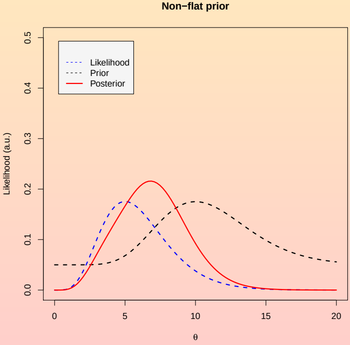
UCLouvain
Institut de recherche
en mathématique et physique

- In particle physics, the typical application of Bayesian statistics is to put an upper limit on a parameter θ
  - Find a value θ_c such that $P(\theta_{true} < \theta_c) = 95\%$
- Typically θ represents the cross section of a physics process, and is proportional to a variable with a Poisson p.d.f.
- An uniform prior can be chosen, eventually restricted to $\theta \geq 0$ to account for the physical range of θ
- We can write priors as a function of other variables, but in general those variables will be linked to the cross section by some analytic transformation
  - A prior that is uniforme in a variable is not in general uniform in a transformed variable; a uniform prior in the cross section implies a non-uniform prior (not even linear) on the mass of the sought particle
- In HEP, usually the prior is chosen uniform in the variable with the variable which is proportional to the cross section of the process sought

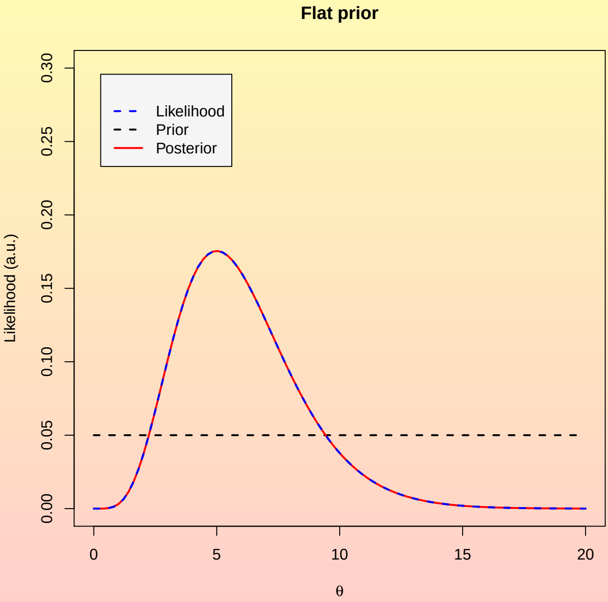
- Uniform priors must make sense
  - Uniform prior across its entire dominion: not very realistic
  - It corresponds to claimng that $P(1 < \theta \leq 2)$ is the same as $P(10^{41} < \theta \leq 10^{41} + 1)$
  - It's irrational to claim that a prior can cover uniformly forty orders of magnitude
  - We must have a general idea of "meaningful" values for θ, and must not accept results forty orders of magnitude above such meaningful values
- A uniform prior often implies that its integral is infinity (e.g. for a cross section, the dominion being $[0, \infty]$
  - Achieving a proper normalization of the posterior probability would be a nightmare
- In practice, use a very broad prior that falls to zero very slowly but that is practically zero where the parameter cannot meaningfully lie
  - This does not guarantee that it integrates to 1—it depends on the speed of convergence to zero
  - Improper prior

Choosing a prior in Bayesian statistics; in practice... 3/

- Associating parametric priors to intervals in the parameter space corresponds to considering sets of theories
  - This is because to each value of a parameter corresponds a different theory
- In practical situations, note (Eq. 9) posterior probability is always proportional to the product of the prior and the likelihood
  - The prior must not necessarily be uniform across the whole dominion
  - It should be uniform only in the region in which the likelihood is different from zero
- If the prior $\pi(\theta)$ is very broad, the product can sometimes be approximated with the likelihood, $P(\vec{X}|\theta)\pi(H) \sim P(\vec{X}|\theta)$
  - The likelihood function is narrower when the data are more precise, which in HEP often translates to the limit $N \to \infty$
  - In this limit, the likelihood is always dominant in the product
  - The posterior is indipendent of the prior!
  - The posteriors corresponding to different priors must coincide, in this limit

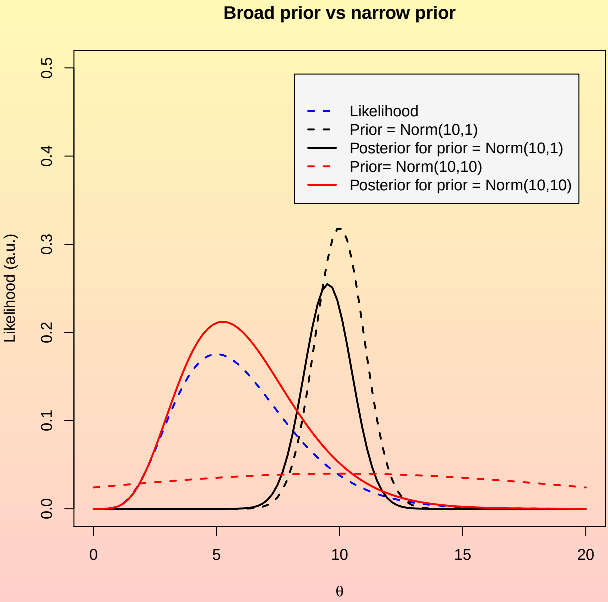


Flat prior

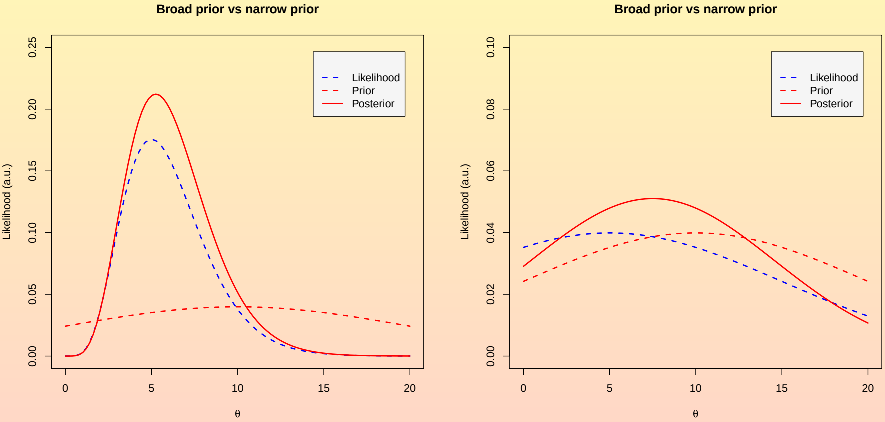


Flat prior
Likelihood
Prior
Posterior
Likelihood (a.u.)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
5
10
15
20
θ

Broad vs narrow non-flat priors



Broad prior and narrow-vs-peaked likelihood



UCLouvain
Institut de recherche
en mathématique et physique

- The authors of STAN maintain a nice set of recommendations for choosing a prior distribution https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations
  - It is supposed to present a balance between strongly informative priors (judged often unrealistic) and noninformative priors
- Deeply empirical recommendations
  - Give attention to computational constraints
  - A-priori dislike for invariance-principles based priors and Jeffreys priors
- Not necessarily applicable to HEP without debate, but many rather reasonable perspectives
  - Weakly/Strongly informative depends not only on the prior but also on the question you are asking *"The prior can often only be understood in the context of the likelihood"*
  - Weak $==$ for a reasonably large amount of data, the likelihood will dominate (a "weak" prior might still influence the posterior, if the data are weak)
  - Hard constraints should be reserved to true constraints (e.g. positive-definite parameters) (otherwise, choose weakly informative prior on a larger range)
  - Check the posterior dependence on your prior, and perform prior predictive checks doi:10.1111/rssa.12378

UCLouvain
Institut de recherche
en mathématique et physique

- Frequentists are restricted to statements related to
  - $P(data|theory)$ (kind of deductive reasoning)
  - The data is considered random
  - Each point in the "theory" phase space is treated independently (no notion of probability in the "theory" space)
  - Repeatable experiments
- Bayesians can address questions in the form
  - $P(theory|data) \propto P(data|theory) \times P(theory)$ (it is intuitively what we normally would like to know)
  - It requires a prior on the theory
  - Huge battle on subjectiveness in the choice of the prior goes here - see §7.5 of James' book

Drawing some histograms

UCLouvain
Institut de recherche
en mathématique et physique

- **Random variable:** a numeric label for each element in the space of data (in frequentist statistics) or in the space of the hypotheses (in Bayesian statistics)
- In Physics, usually we assume that Nature can be described by continuous variables
  - The discreteness of our distributions would arise from scanning the variable in a discrete way
  - Experimental limitations in the act of measuring an intrinsically continuous variable)
- Instead of point probabilities we'll work with probabilities defined in intervals, normalized w.r.t. the interval:

$$f(X) := \lim_{\Delta X \to 0} \frac{P(X)}{\Delta X} \tag{10}$$

- Dimensionally, they are densities and they are called probability density functions (p.d.f. s)
- Inverting the expression, $P(X) = \int f(X)dX$ and we can compute the probability of an interval as a definite interval

$$P(a < X < b) := \int_a^b f(X)dX \tag{11}$$

fnrs LA LIBERTÉ DE CHERCHER
UCLouvain
Institut de recherche
en mathématique et physique

- Extend the concept of p.d.f. to an arbitrary number of variables; the joint p.d.f. $f(X, Y, ...)$
- If we are interested in the p.d.f. of just one of the variables the joint p.d.f. depends upon, we can compute by integration the marginal p.d.f.

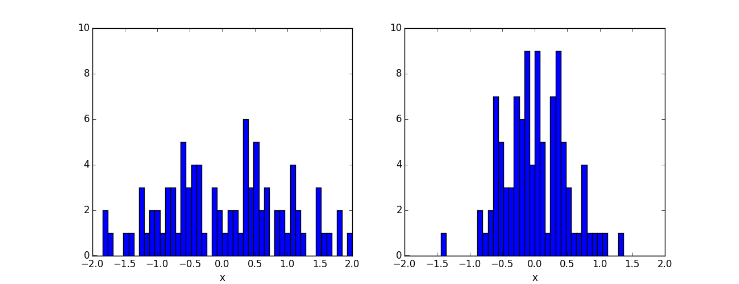
$$f_X(X) := \int f(X,Y)dY \tag{12}$$

- Sometimes it's interesting to express the joint p.d.f. as a function of one variable, for a particular fixed value of the others: this is the conditional p.d.f. :

$$f(X|Y) := \frac{f(X,Y)}{f_Y(Y)} \tag{13}$$

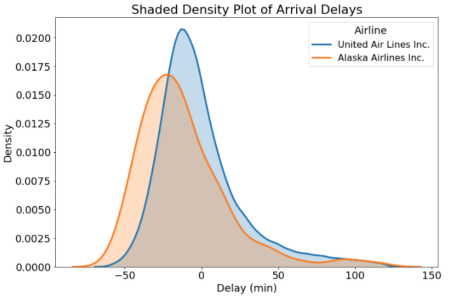
UCLouvain
Institut de recherche
en mathématique et physique

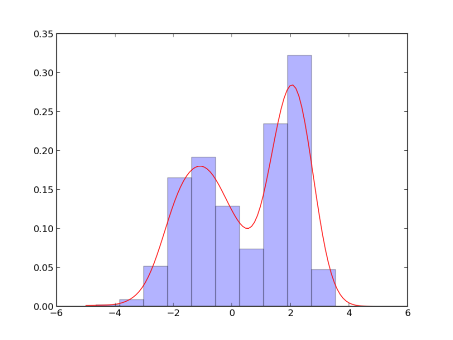
- Repeated experiments usually don't yield the exact same result even if the physical quantity is expected to be exactly the same
  - Random changes occur because of the imperfect experimental conditions and techniques
  - They are connected to the concept of dispersion around a central value
- When repeating an experiment, we can count how many times we obtain a result contained in various intervals (e.g. how often $1.0 \leq L < 1.1$, how often $1.1 \leq L < 1.2$, etc)
  - An histogram can be a natural way of recording these frequencies
  - The concept of dispersion of measurements is therefore related to that of dispersion of a distribution
- In a distribution we are usually interested in finding a "central" value and how much the various results are dispersed around it



Distributions... or not?

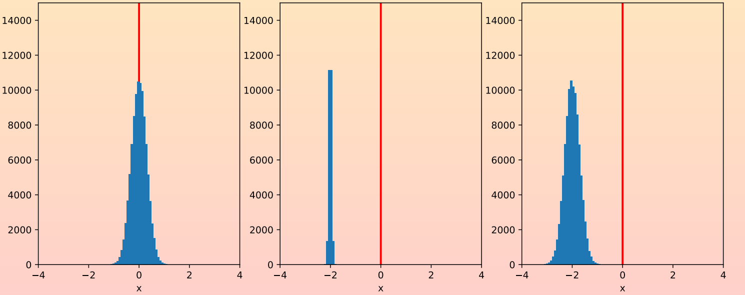
- HEP uses histograms mostly historically: counting experiments
- Statistics and Machine Learning communities typically use densities
  - Intuitive relationship with the underlying p.d.f.
  - Kernel density estimates: binning assumption $\rightarrow$ bandwidth assumption
  - Less focused on individual bin content, more focused on the overall shape
  - More general notion (no stress about the limited bin content in tails)
- In HEP, if your events are then used “as counting experiment” it’s more useful the histogram
  - But for some applications (e.g. Machine Learning) even in HEP please consider using density estimates





Plots from TheGlowingPython and TowardsDataScience

- Two fundamentally different kinds of uncertainties
  - Error: the deviation of a measured quantity from the true value (bias)
  - Uncertainty: the spread of the sampling distribution of the measurements
- **Random (statistical) uncertainties**
  - Inability of any measuring device (and scientist) to give infinitely accurate answers
  - Even for integral quantities (e.g. counting experiments), fluctuations occur in observations on a small sample drawn from a large population
  - They manifest as spread of answers scattered around the true value
- **Systematic uncertainties**
  - They result in measurements that are simply wrong, for some reason
  - They manifest usually as offset from the true value, even if all the individual results can be consistent with each other



Expected values of a random variable

UCLouvain
Institut de recherche
en mathématique et physique

- We define the expected value and mathematical expectation

$$E[X] := \int_{\Omega} X f(X) dX \tag{14}$$

- In general, for each of the following formulas (reported for continuous variables) there is a corresponding one for discrete variables, e.g.

$$E[X] := \sum_{i} X_i P(X_i) \tag{15}$$

UCLouvain
Institut de recherche
en mathématique et physique

- Extend the concept of expected value to a generic function $g(X)$ of a random variable

$$E[g] := \int_{\Omega} g(X)f(X)dX \tag{16}$$

- The previous expression Eq. 14 is a special case of Eq. 16 when $g(X) = X$
- The mean of X is:

$$\mu := E[X] \tag{17}$$

- The variance of X is:

$$V(X) := E[(X-\mu)^2] = E[X^2] - (E[X])^2 = E[X^2] - \mu^2 \tag{18}$$

- Mean and variance will be our way of estimating a "central" value of a distribution and of the dispersion of the values around it

Let's make it funnier: more variables!

UCLouvain
Institut de recherche en mathématique et physique

- Let our function $g(X)$ be a function of more variables, $\vec{X} = (X_1, X_2, ..., X_n)$ (with p.d.f. $f(\vec{X})$)
  - Expected value: $E(g(\vec{X})) = \int g(\vec{X}) f(\vec{X}) dX_1 dX_2 ... dX_n = \mu_g$
  - Variance: $V[g] = E\left[(g - \mu_g)^2\right] = \int (g(\vec{X}) - \mu_g)^2 f(\vec{X}) dX_1 dX_2 ... dX_n = \sigma_g^2$
- **Covariance:** of two variables X, Y:
  $V_{XY} = E\left[(X - \mu_X)(Y - \mu_Y)\right] = E[XY] - \mu_X \mu_Y = \int XY f(X, Y) dX dY - \mu_X \mu_Y$
  - It is also called "error matrix", and sometimes denoted $cov[X, Y]$
  - It is symmetric by construction: $V_{XY} = V_{YX}$, and $V_{XX} = \sigma_X^2$
  - To have a dimensionless parameter: correlation coefficient $\rho_{XY} = \frac{V_{XY}}{\sigma_X \sigma_Y}$

- V_{XY} is the expectation for the product of deviations of X and Y from their means
- If having $X > \mu_X$ enhances $P(Y > \mu_Y)$, and having $X < \mu_X$ enhances $P(Y < \mu_Y)$, then $V_{XY} > 0$: positive correlation!
- ρ_{XY} is related to the angle in a linear regression of X on Y (or viceversa)

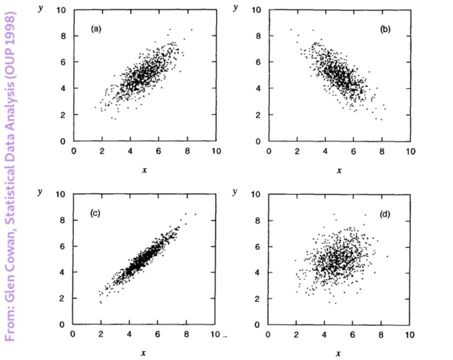


Fig. 1.9 Scatter plots of random variables x and y with (a) a positive correlation, $\rho = 0.75$, (b) a negative correlation, $\rho = -0.75$, (c) $\rho = 0.95$, and (d) $\rho = 0.25$. For all four cases the standard deviations of x and y are $\sigma_x = \sigma_y = 1$.

Let's make it funnier: more variables!

UCLouvain
Institut de recherche
en mathématique et physique

- Let our function $g(X)$ be a function of more variables, $\vec{X} = (X_1, X_2, ..., X_n)$ (with p.d.f. $f(\vec{X})$)
  - Expected value: $E(g(\vec{X})) = \int g(\vec{X}) f(\vec{X}) dX_1 dX_2 \ldots dX_n = \mu_g$
  - Variance: $V[g] = E\left[(g - \mu_g)^2\right] = \int (g(\vec{X}) - \mu_g)^2 f(\vec{X}) dX_1 dX_2 \ldots dX_n = \sigma_g^2$
- **Covariance:** of two variables X, Y:
  $V_{XY} = E\left[(X - \mu_X)(Y - \mu_Y)\right] = E[XY] - \mu_X \mu_Y = \int XY f(X, Y) dX dY - \mu_X \mu_Y$
  - It is also called "error matrix", and sometimes denoted $cov[X, Y]$
  - It is symmetric by construction: $V_{XY} = V_{YX}$, and $V_{XX} = \sigma_X^2$
  - To have a dimensionless parameter: correlation coefficient $\rho_{XY} = \frac{V_{XY}}{\sigma_X \sigma_Y}$

- V_{XY} is the expectation for the product of deviations of X and Y from their means
- If having $X > \mu_X$ enhances $P(Y > \mu_Y)$, and having $X < \mu_X$ enhances $P(Y < \mu_Y)$, then $V_{XY} > 0$: positive correlation!
- ρ_{XY} is related to the angle in a linear regression of X on Y (or viceversa)
  - It does not capture non-linear correlations
    Question time: CorrCoeff

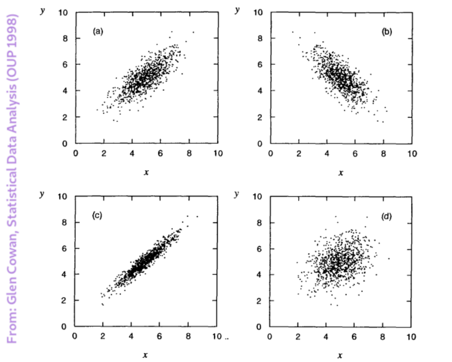


Fig. 1.9 Scatter plots of random variables x and y with (a) a positive correlation, $\rho = 0.75$, (b) a negative correlation, $\rho = -0.75$, (c) $\rho = 0.95$, and (d) $\rho = 0.25$. For all four cases the standard deviations of x and y are $\sigma_x = \sigma_y = 1$.

What can you get from ρ_{XY}

- Informs on the direction (co-increase, increase-decrease, none) of a linear correlation
- Does NOT inform on the slope of the correlation
- Several non-linear correlations yield ρ_{XY}

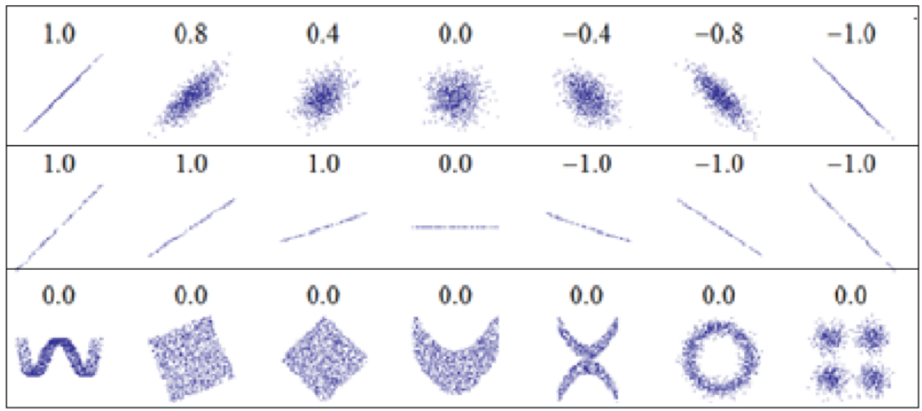


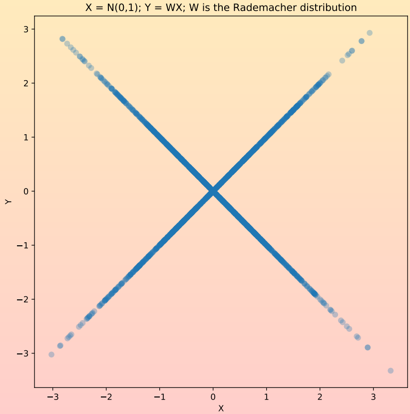
Figure from BND2010

Take it to the next level: the Mutual Information

- Covariance and correlation coefficients act taking into account only linear dependences
- Mutual Information is a general notion of correlation, measuring the information that two variables X and Y share

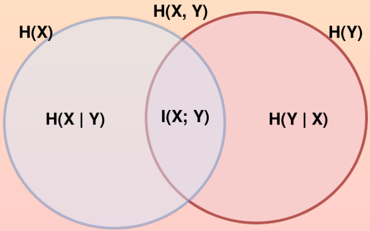
$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \left(\frac{p(x,y)}{p_1(x) p_2(y)} \right)$$

- Symmetric: $I(X;Y) = I(Y;X)$
- $I(X;Y) = 0$ if and only if X and Y are totally independent
  - X and Y can be uncorrelated but not independent; mutual information captures this!

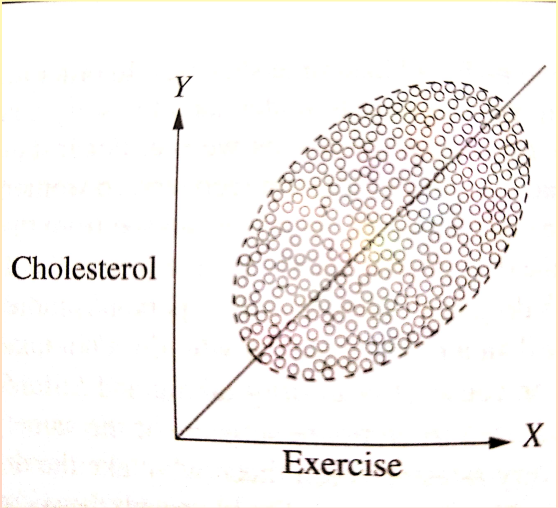


- Related to entropy

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$



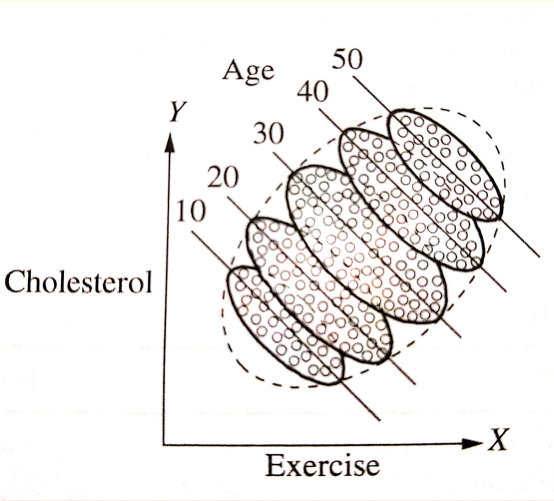
Does cholesterol increase with exercise?



- Question time: Cholesterol

Does it, though?

UCLouvain
Institut de recherche
en mathématique et physique



Should we prescribe the drug?

- If we know the biological sex[3], then prescribe the drug
- If we don't know the biological sex, then don't prescribe the drug

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)

- Question time: DrugEffectiveness

[3] Biological sex: *anatomy of an individual's reproductive system, and secondary sex characteristics*. Gender: *either social roles based on the sex of the person (gender role) or personal identification of one's own gender based on an internal awareness* (https://en.wikipedia.org/wiki/Sex_and_gender_distinction)

Should we prescribe the drug?

UCLouvain
Institut de recherche
en mathématique et physique

- If we know the biological sex[3], then prescribe the drug
- If we don't know the biological sex, then don't prescribe the drug

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)

- Question time: DrugEffectiveness
- Imagine we know that estrogen has a negative effect on recovery
  - Then women less likely to recovery than men
  - Table shows women are significantly more likely to take the drug

[3] Biological sex: *anatomy of an individual's reproductive system, and secondary sex characteristics*. Gender: *either social roles based on the sex of the person (gender role) or personal identification of one's own gender based on an internal awareness* (https://en.wikipedia.org/wiki/Sex_and_gender_distinction)

Should we prescribe the drug?

- If we know the biological sex[3], then prescribe the drug
- If we don't know the biological sex, then don't prescribe the drug

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)

- Question time: DrugEffectiveness
- Imagine we know that estrogen has a negative effect on recovery
  - Then women less likely to recovery than men
  - Table shows women are significantly more likely to take the drug
  - Consult the separate data to decide on the drug, in order not to mix effects

[3] Biological sex: *anatomy of an individual's reproductive system, and secondary sex characteristics*. Gender: *either social roles based on the sex of the person (gender role) or personal identification of one's own gender based on an internal awareness* (https://en.wikipedia.org/wiki/Sex_and_gender_distinction)

Should we prescribe the drug?

- BP = Blood Pressure

	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)

- Question time: DrugEffectiveness

Should we prescribe the drug?

UCLouvain
Institut de recherche
en mathématique et physique

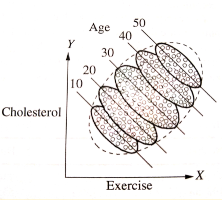
- BP = Blood Pressure

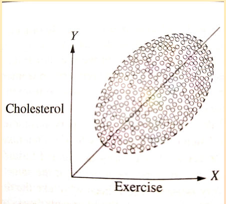
	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)

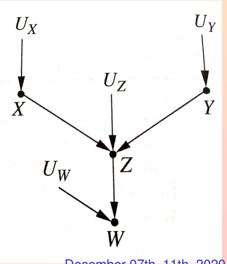
- Question time: DrugEffectiveness
- Same table, different labels; here we must consider the combined data
  - Lowering blood pressure is actually part of the mechanism of the drug effect

The Simpson paradox: correlation is not causation

- Correlation alone can lead to nonsense conclusions
  - If we know the biol.sex, then prescribe the drug
  - If we don't know the biol.sex, then don't prescribe the drug
- Imagine we know that estrogen has a negative effect on recovery
  - Then women less likely to recovery than men
  - Table shows women are significantly more likely to take the drug
- Here we should consult the separate data, in order not to mix effects
- Same table, different labels; must consider the combined data
  - Lowering blood pressure is actually part of the mechanism of the drug effect
- Same effect in continuous data (cholesterol vs age)
- The best solution so far are Bayesian causal networks
  - Graph theory to describe relationship between variables







Figures from Pearl, 2016

First level of causal hierarchy: seeing

- X and Y are marginally dependent, but conditionally independent given Z
  - Same concept we have seen (with a more dramatic effect) in the cholesterol example
- Conditioning on Z *blocks the path*

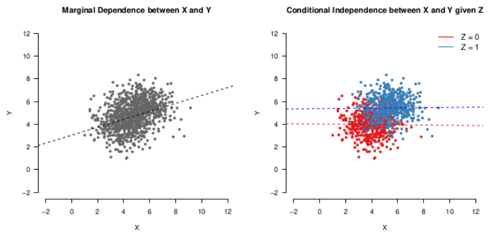


Figure 2. Left: Shows marginal dependence between X and Y. Right: Shows conditional independence between X and Y given Z.

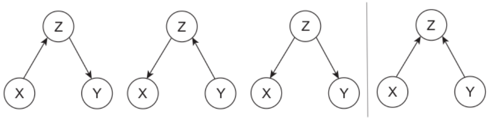


Figure 3. The first three DAGs encode the same conditional independence structure, $X \perp\!\!\!\perp Y \mid Z$. In the fourth DAG, Z is a collider such that $X \not\perp\!\!\!\perp Y \mid Z$.

Figures from Dablander, 2019

First level of causal hierarchy: seeing

- X and Y are marginally independent, but conditionally dependent given Z
  - Z is called a *collider* (not the particle physics one ☺)
- Conditioning on Z induces *collider bias*

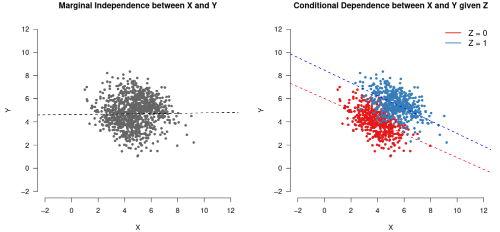


Figure 4. Left: Shows marginal independence between X and Y. Right: Shows conditional dependence between X and Y given Z

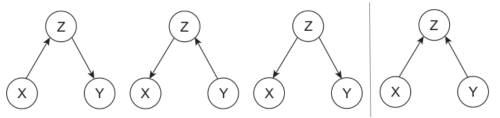


Figure 3. The first three DAGs encode the same conditional independence structure, $X \perp\!\!\!\perp Y \mid Z$. In the fourth DAG, Z is a collider such that $X \not\perp\!\!\!\perp Y \mid Z$.

Figures from Dablander, 2019

Second level of causal hierarchy: doing

- Interventionist approach (Pearl, 2016) (not everyone agrees with this formal approach)
  - X has a causal influence on Y if changing X leads to changes in (the distribution of) Y
- Setting (by intervention) $X = x$ cuts all incoming causal arrows
  - The value of X is determined only by the intervention
  - Must be able to do intervention: not mere conditioning (seeing): from $P(Y|X = x)$ to $P(Y|do(X = x))$
  - Difficult in social sciences
- Intervention discriminates between causal structure of different diagrams
  - Assuming that there is no unobserved confounding (i.e. all causal relationships are represented in the DAG)

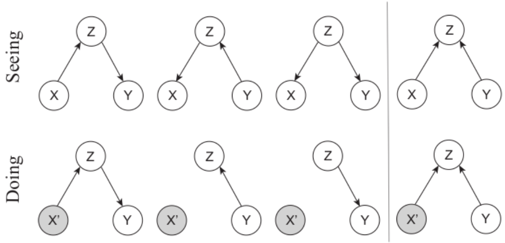


Figure 6. *Seeing*: DAGs are used to encode conditional independencies. The first three DAGs encode the same associations. *Doing*: DAGs are causal. All of them encode distinct causal assumptions.

Figures from Dablander, 2019

“Doing” applied to Simpson’s paradox

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)

	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)

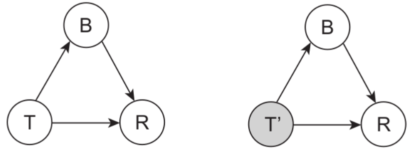


Figure 8. Underlying causal DAG of the example with treatment (T), blood pressure (B), and recovery (R).

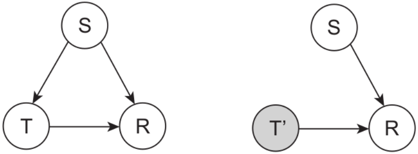


Figure 7. Underlying causal DAG of the example with treatment (T), biological sex (S), and recovery (R).

Figures from Dablander, 2019

"Doing" is for populations

- Good predictors can be causally disconnected from the effect!
- The *do* operator operates on distributions defined on populations

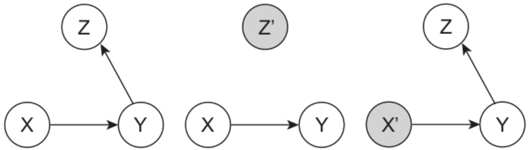


Figure 9. An excellent predictor (Z) need not be causally effective.

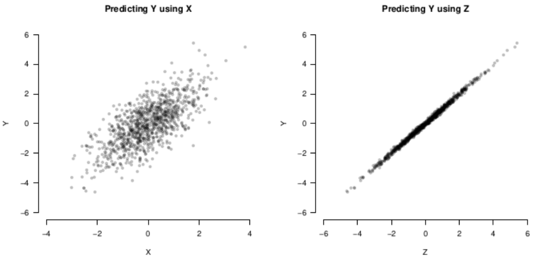


Figure 10. X is a considerably worse predictor of Y than Z.

Figures from Dablander, 2019

Third level of causal hierarchy: imagining

- The strongest level of causality acts on the individual
  - "As a matter of fact, humans constantly evaluate mutually exclusive options, only one of which ever comes true; that is, humans reason counterfactually."
- Structural Causal Models relate causal and probabilistic statements
  - $Treatment := \epsilon_T \sim N(0, \sigma)$
  - $Response := \mu + \beta Treatment + \epsilon$
  - Measure $\mu = 5$, $\beta = -2$, $\sigma = 2$
- Causal effect obscured by individual error term ϵ_i for each patient: if determined, model fully determined
- Can determine response for individual treatment!

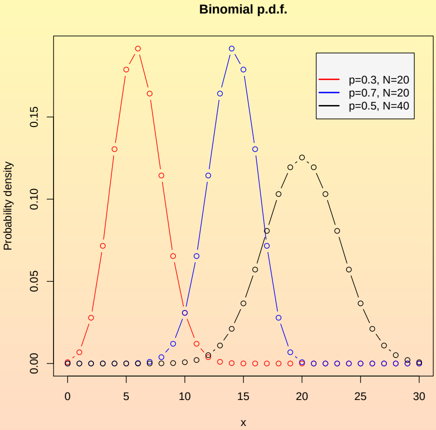
Table 4

Data simulated from the SCM concerning grandma's treatment of the common cold.

Patient	Treatment	Recovery	ε_k
1	0	5.80	0.80
2	0	3.78	-1.22
3	1	3.68	0.68
4	1	0.74	-2.26
5	0	7.87	2.87

Figures and quote from from Dablander, 2019

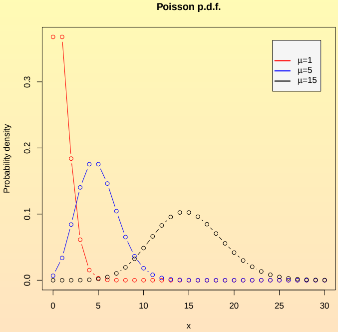
- **Binomial**
  - Discrete variable: r, positive integer $\leq N$
  - Parameters:
    - N, positive integer
    - p, $0 \leq p \leq 1$
  - Probability function:
    $P(r) = \binom{N}{r} p^r (1-p)^{N-r}$, $r = 0, 1, \ldots, N$
  - $E(r) = Np$, $V(r) = Np(1-p)$
  - Usage: probability of finding exactly r successes in N trials. The distribution of the number of events in a single bin of a histogram is binomial (if the bin contents are independent)



- Example: which is the probability of obtaining 3 times the number 6 when throwing a 6-faces die 12 times?
- $N = 12$, $r = 3$, $p = \frac{1}{6}$
- $P(3) = \binom{12}{3}\left(\frac{1}{6}\right)^3 (1 - \frac{1}{6})^{12-3} = \frac{12!}{3!9!}\frac{1}{6^3}\left(\frac{5}{6}\right)^9 = 0.1974$

UCLouvain
Institut de recherche
en mathématique et physique

- **Poisson**
  - Discrete variable: r, positive integer
  - Parameter: μ, positive real number
  - Probability function: $P(r) = \frac{\mu^r e^{-\mu}}{r!}$
  - $E(r) = \mu$, $V(r) = \mu$
  - Usage: probability of finding exactly r events in a given amount of time, if events occur at a constant rate.
- Example: is it convenient to put an advertising panel along a road?



- Probability that at least one car passes through the road on each day, knowing on average 3 cars pass each day
  - $P(X > 0) = 1 - P(0)$, and use Poisson p.d.f.

$$P(0) = \frac{3^0 e^{-3}}{0!} = 0.049787$$

  - $P(X > 0) = 1 - 0.049787 = 0.95021$.
- Now suppose the road serves only an industry, so it is unused during the weekend; Which is the probability that in any given day exactly one car passes by the road?

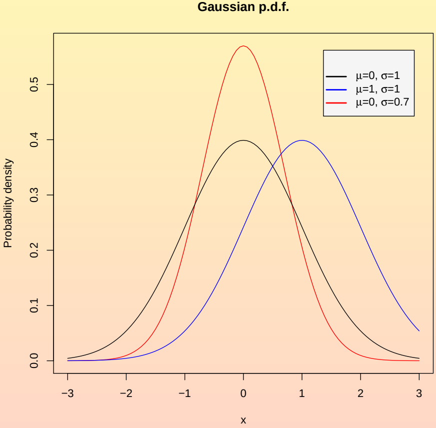
$$N_{avg\ per\ dia} = \frac{3}{5} = 0.6$$

$$P(X) = \frac{0.6^1 e^{-0.6}}{1!} = 0.32929$$

- **Gaussian** or Normal distribution
  - Variable: X, real number
  - Parameters:
    - μ, real number
    - σ, positive real number
  - Probability function:

$$f(X) = N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\frac{(X-\mu)^2}{\sigma^2} \right]$$

  - $E(X) = \mu$, $V(X) = \sigma^2$
  - Usage: describes the distribution of independent random variables. It is also the high-something limit for many other distributions



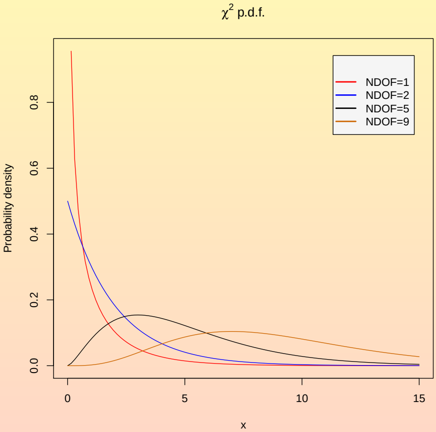
- Parameter: integer $N > 0$ degrees of freedom
- Continuous variable $X \in \mathcal{R}$
- p.d.f., expected value, variance

$$f(X) = \frac{\frac{1}{2}\left(\frac{X}{2}\right)^{\frac{N}{2}-1} e^{-\frac{X}{2}}}{\Gamma\left(\frac{N}{2}\right)}$$

$$E[r] = N$$

$$V(r) = 2N$$

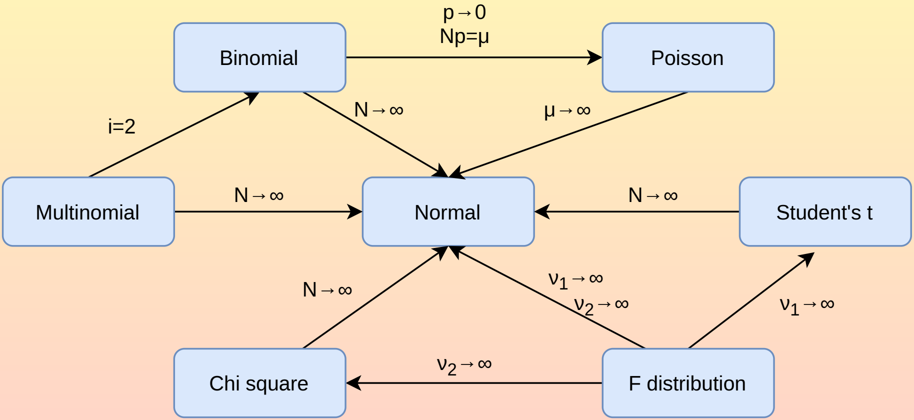
- It describes the distribution of the sum of the squares of a random variable, $\sum_{i=1}^{N} X_i^2$



Reminder: $\Gamma() := \frac{N!}{r!(N-r)!}$

UCLouvain
Institut de recherche
en mathématique et physique

- It is often convenient to know the asymptotic properties of the various distributions



- Minimal packages needed besides standard ones: `numpy`, `matplotlib`
- Optional (for fancy table): `pandas`
- `random` should be a base package

- Code available at: https://github.com/vischia/intensiveCourse_public
  - You can either download the raw version of the scripts
  - or better do `git clone https://github.com/vischia/intensiveCourse_public.git` in your shell
- Once you have the code in a directory, go to that directory and run, depending on your system,
  - `ipython notebook` or `ipython3 notebook`
  - or `jupyter notebook` or `jupyter3 notebook`

End of Lesson 1

Why statistics?

Fundaments
- Set theory and measure theory
- Frequentist probability
- Bayesian probability

Random variables and their properties

Causality
- The three levels of causal hierarchy

Distributions

THANKS FOR THE ATTENTION!

Backup