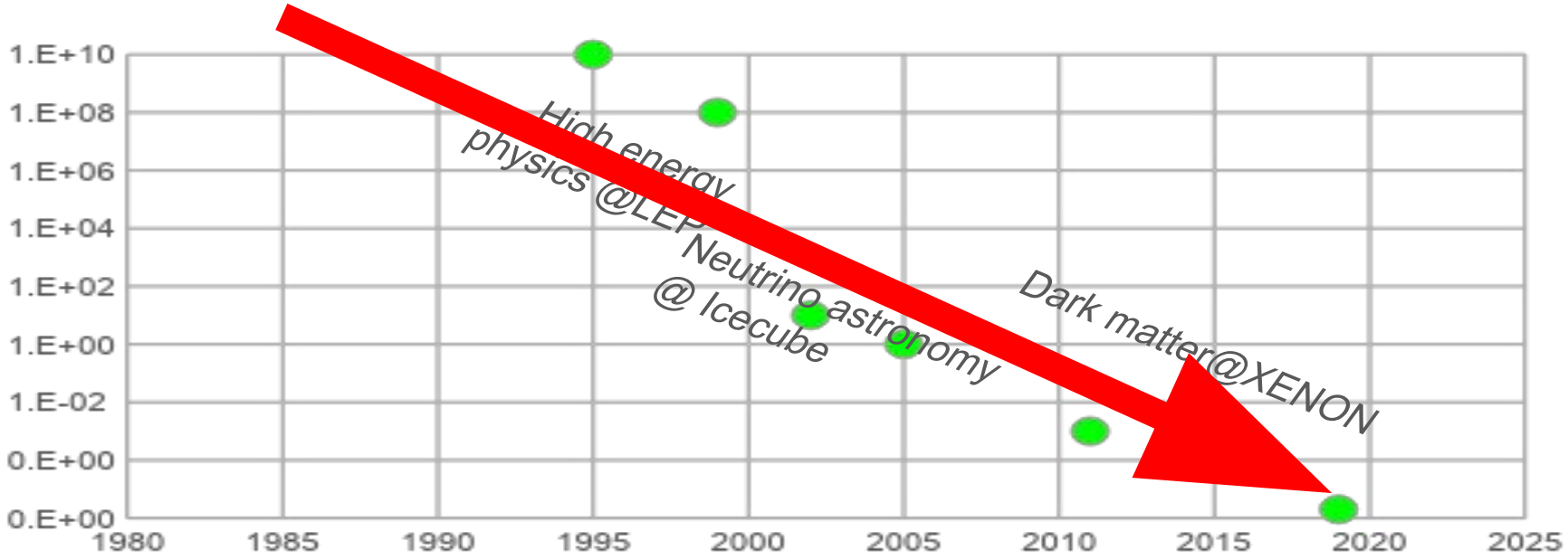
Statistics underground

PART I

Fortuna - Roman goddess of luck, chance and statistical inference

SoUP
The INFN School on Underground Physics

(; My career



WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

What I would like to do here

There are excellent books/lectures/blogs/sites on everything we will discuss, and more.

E.g.:

- R.J. Barlow, Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences, Wiley, 1989
- G. Cowan, Statistical Data Analysis, Clarendon, Oxford, 1998
  Also online lectures and talks https://www.pp.rhul.ac.uk/~cowan/
- G.D'Agostini, Bayesian Reasoning in Data Analysis: A Critical Introduction, World Scientific Publishing 2003.
- PDG, Statistics summary: https://pdg.lbl.gov/2020/reviews/rpp2020-rev-statistics.pdf
- Gelman et al: Bayesian data analysis: http://www.stat.columbia.edu/~gelman/book/
- https://telescoper.wordpress.com

What I would like to do here

There are excellent books/lectures/blogs/sites on everything we will discuss, and more.

With our limited time together, I'd like to:

- Introduce basics of statistical analysis and survey some of the ways we use statistical analysis in UG physics
- Review basic and advanced concepts in probability, decision making and inference
- Zoom into some delicate points relevant to us, underground people
- Introduce tools and techniques that helped me in understanding and doing statistics

In the end of our sessions I hope you will see the beauty and importance of statistics, and be encouraged to further investigate, simulate, read and do

Work Plan

Intro: Statistical bloopers

Part 1: Probability:

Part 2: Distributions and sampling

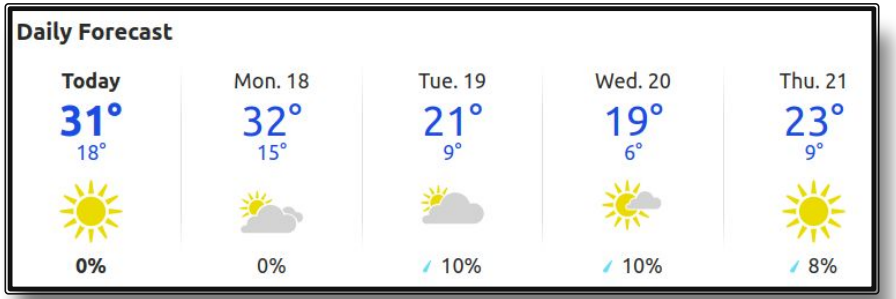
Part 3: Estimation

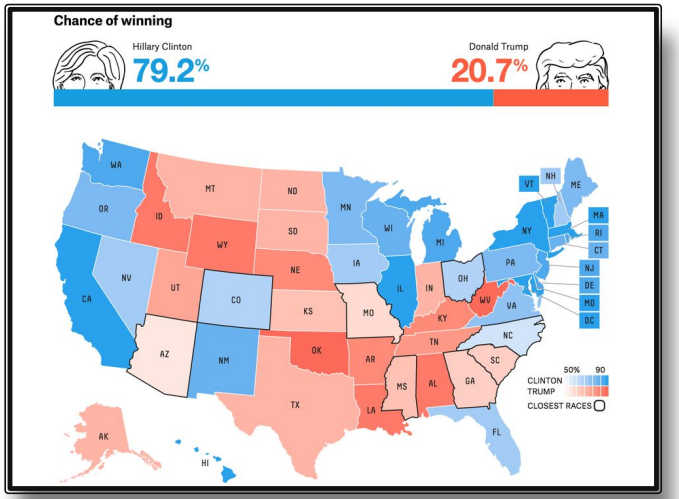
Part 4: Inference

Introduction: Statistical bloopers

The case of the wavy hand

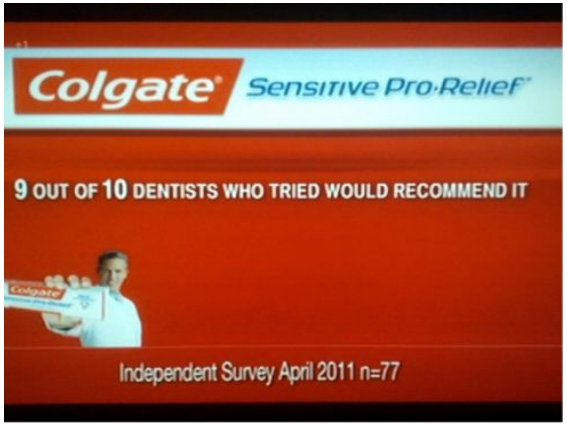
But what does it mean?





The case of the wavy hand

But what does it mean?



Hamlin et al., Am J Dent. 2009 Mar;22 Spec No A:16A-20A.

Schiff er al., Am J Dent. 2009 Mar;22 Spec No A:8A-15A.

The case of the wavy hand

But what does it mean?

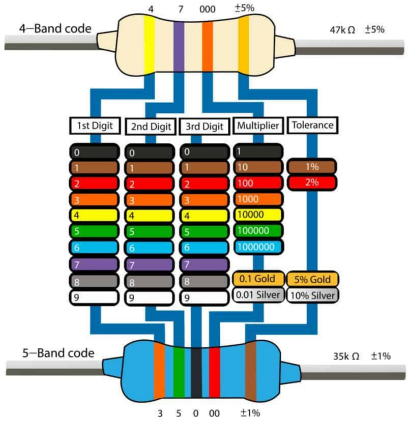
Brown = 1%,

Red = 2%,

Gold = 5%,

Silver = 10 %

None = 20%



What does it mean ?

For my birthday I got a box with 1000 resistors. 100Ω each, with 5% tolerance

Q: How is the tolerance defined? 1 sigma? 3 sigma?

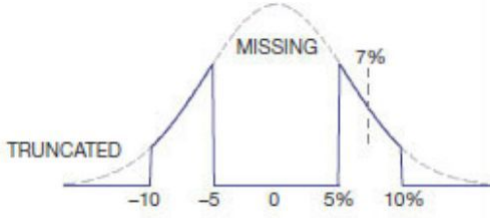
A: “only” 95-105 ohm resistance. By construction.

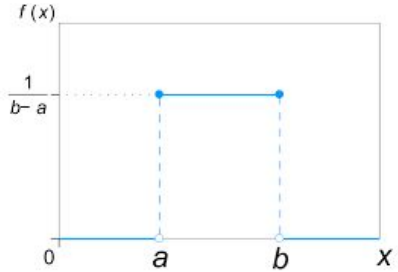
Q: What is the probability distribution of the resistivity?

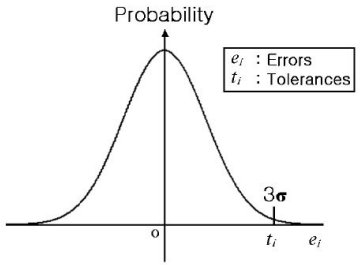
A: Yes

Q: Was the selection done by Design or by choice?

A: ahhhhh?







Is the treatment effective?

100 patient tested, 50 young and 50 less young

Researcher A looked at the results and concluded that…

Young patients:

	Recovered	Not recovered	Total	% recovery
Treated	19	21	40	47.5 %
Not treated	5	5	10	50 %
	24	26	50	

Old patients:

	Recovered	Not recovered	Total	% recovery
Treated	1	9	10	10 %
Not treated	11	29	40	27.5 %
	12	38	50	

Is the treatment effective?

100 patient tested, 50 young and 50 less young

Researcher B looked at the overall results and concluded that…

	Recovered	Not recovered	Total	% recovery
Treated ✓	20	30	50	40%
Not treated	16	34	50	32%
	36	64	100	

Is the treatment effective?

Researcher A concluded:
"Treatment is effective"

Young patient

	R	NR	Tot	% R
T	19	21	40	47.5%
NT	5	5	10	50%
	24	26	50	

Old patient

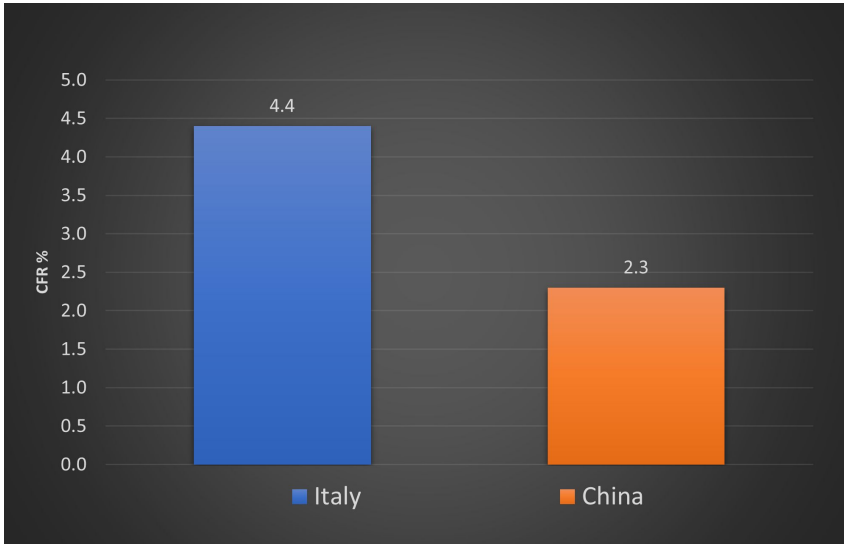
	R	NR	Tot	% R
T	1	9	10	10%
NT	11	29	40	27.5%
	12	38	50	

Researcher B concluded:
"Treatment is not effective"

	Recovered	Not recovered	Total	% recovery
Treated	20	30	50	40%
Not treated	16	34	50	32%
	36	64	100	

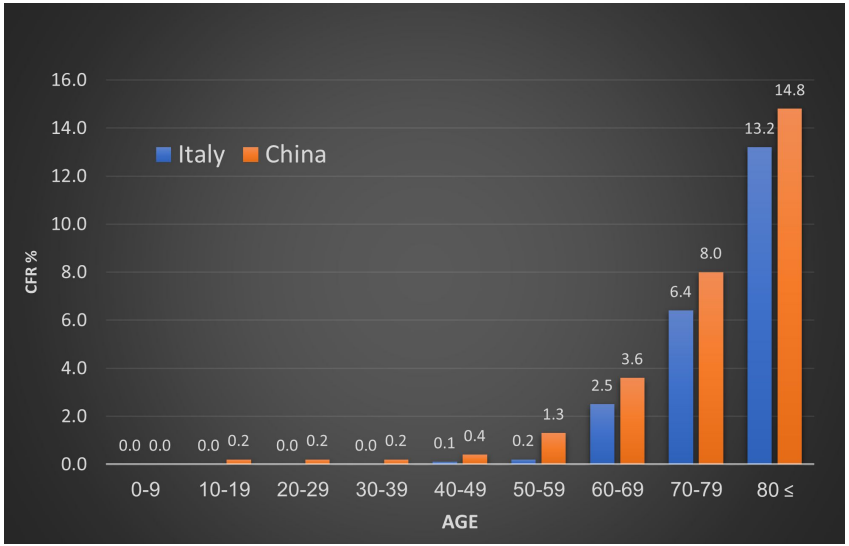
Case fatality rates in China & Italy

Overall



Italy > China

By age

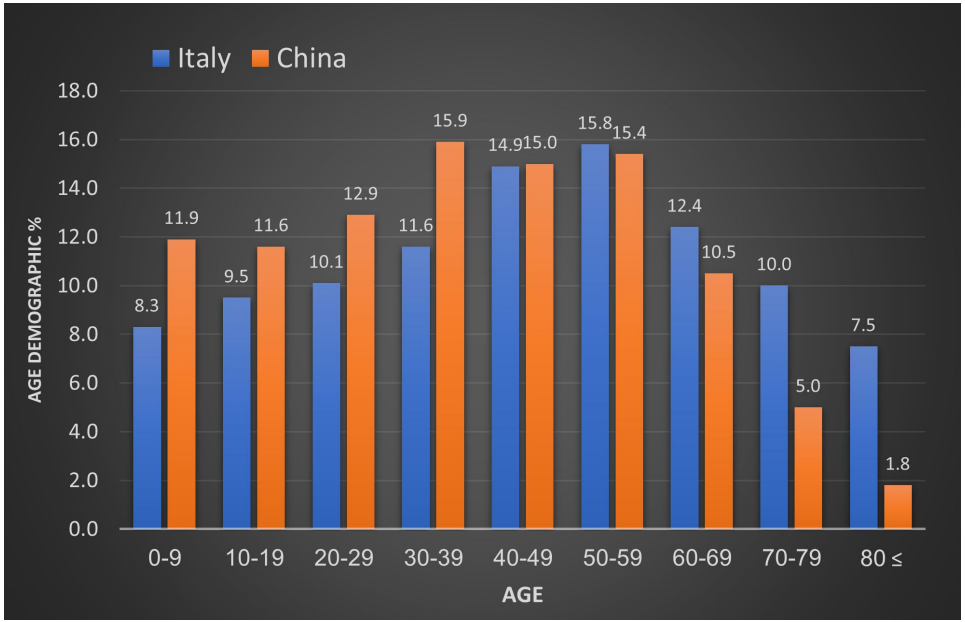


China > Italy

https://arxiv.org/pdf/2005.07180.pdf

The case of the illusive variable

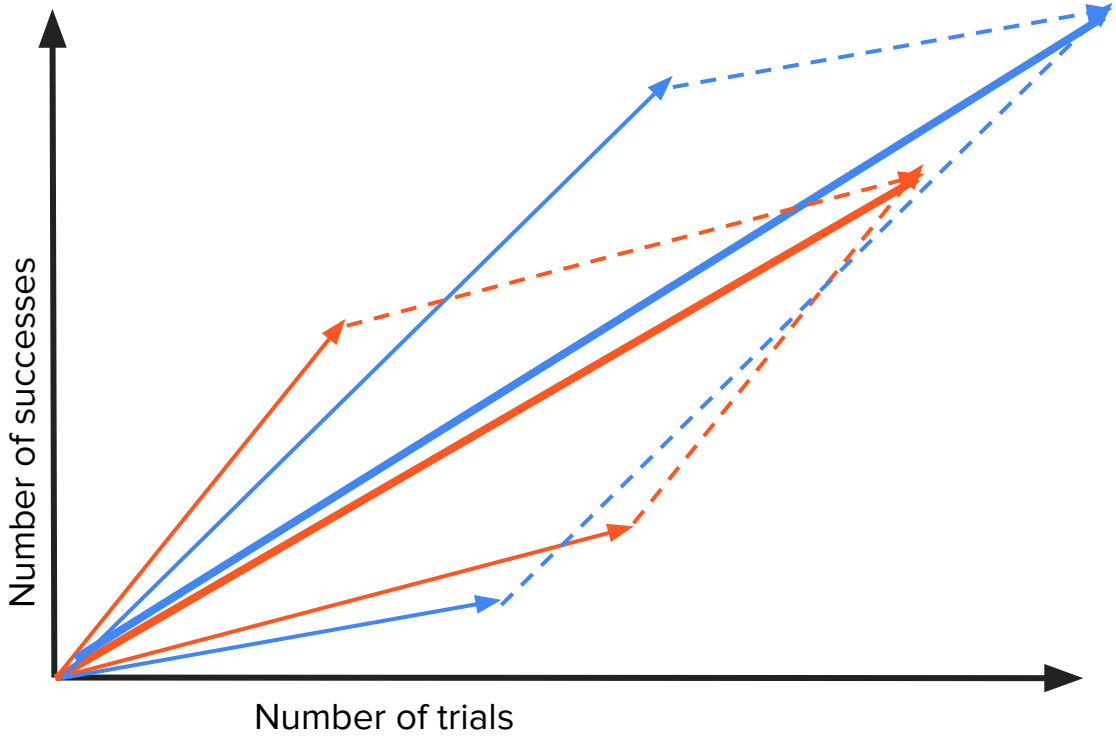
Simpson's Paradox



Simpson's paradox

Graphical illustration

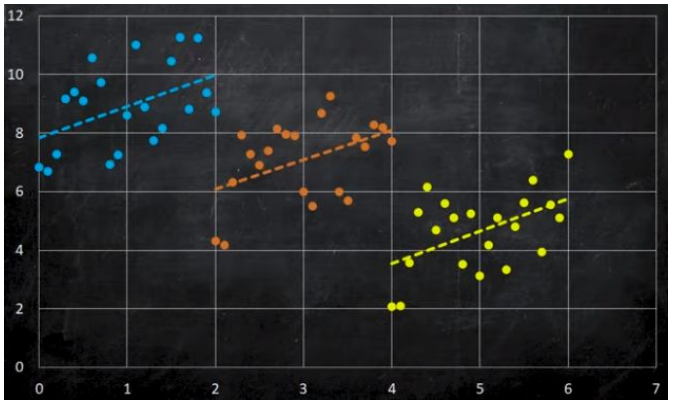
- Slope = Fraction of successes
- Steeper vector = more success
- The orange lines have a higher success rate than the blue ones
- However...The sum of the orange lines have a lower success rate than the blue on.

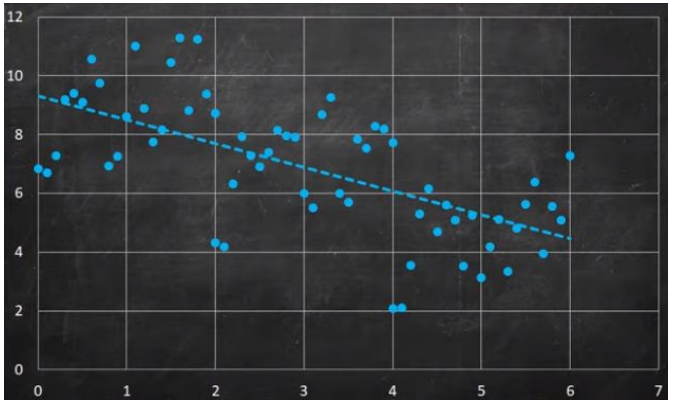


The case of the illusive variable

Simpson's Paradox

Drawing two different conclusions from the same data, depending on how you divide things up



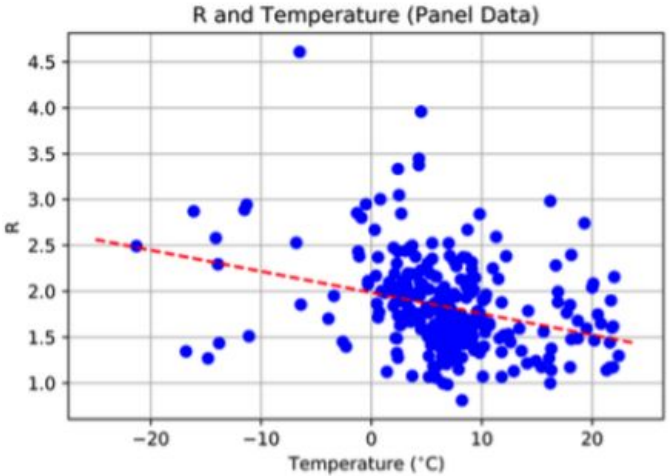


We don't know what we don't know...

https://www.youtube.com/watch?v=t-Ci3FosqZs&ab_channel=Dr.TreforBazett

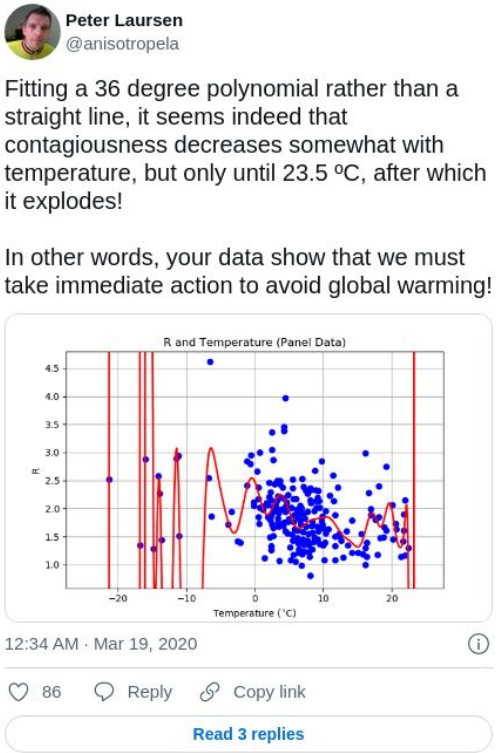
The case of the bad fit

Good news! COVID-19 is less contagious at higher temperatures...



March 2020

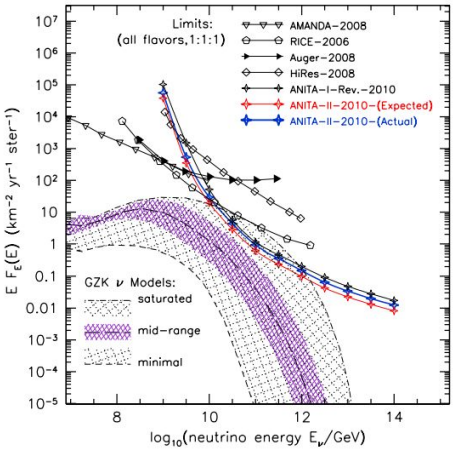
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3551767



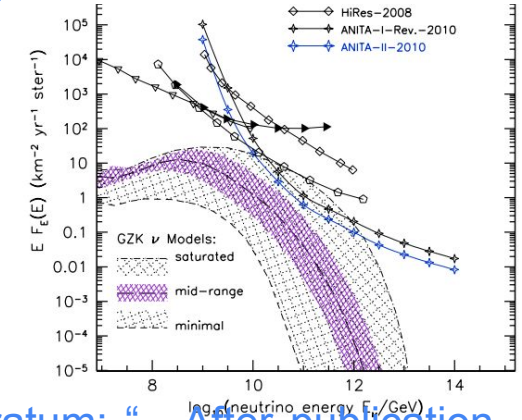
https://arxiv.org/abs/1011.5004

The case of the clerical error

Observational Constraints on the Ultra-high Energy cosmic Neutrino Flux from the Second Flight of the ANITA Experiment



March 2010: “. In a blind analysis, we find **2 surviving** events on a background, mostly anthropogenic, of 0.97±0.42 events”



Nov 2010 Erratum: “ . After publication, we subsequently determined that due to a clerical error **one of the two** surviving events, Event 8381355, **was actually one of the inserted pulser events.** The fact that this event survived its subsequent scrutiny we consider as a demonstration that the blinding procedure was truly valid”

The case of the mysterious signal

High statistics search for ultrahigh energy γ-ray emission from Cygnus X-3 and Hercules X-1

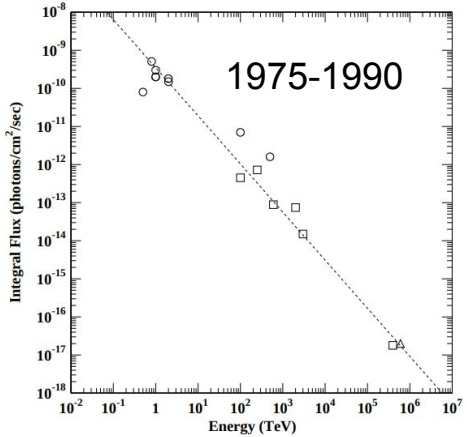


Figure 1: Published results from ground-based experiments indicating evidence for gamma-ray emission from Cygnus X-3 during the period 1975-1990. The circles indicate results from atmospheric Cherenkov telescopes, the squares show data taken by air shower arrays, and the triangle indicates the result from the Fly's Eye experiment. The dashed curve is an approximate power law fit to the data with a slope of −1.1. Not shown in this figure are several upper limits to the flux of gamma-rays from Cygnus X-3 during this same epoch. The two points at extremely high energy (5×10^5 TeV) have been slightly displaced from each other for clarity.

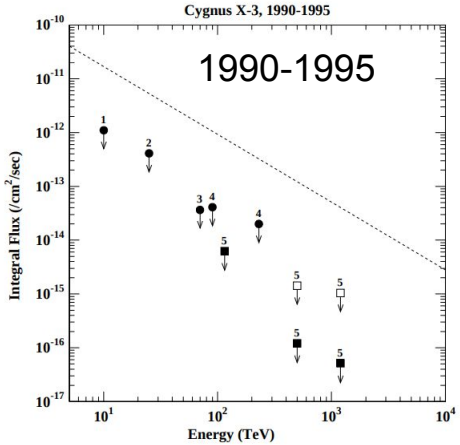


Figure 12: Flux limits reported between 1990 and 1995 on the steady emission of particles from Cygnus X-3. The squares (5) represent the results of this work. Open squares indicate limits on the emission of any neutral particle that creates air showers. Filled squares indicate limits on the emission of gamma-rays. The circles (1-4) represent results from other experiments: 1. Tibet [92], 2. HEGRA [95], 3. CYGNUS [93], and 4. EAS-TOP [94]. The dashed curve is the approximate power law fit to early results (reproduced from Figure 1).

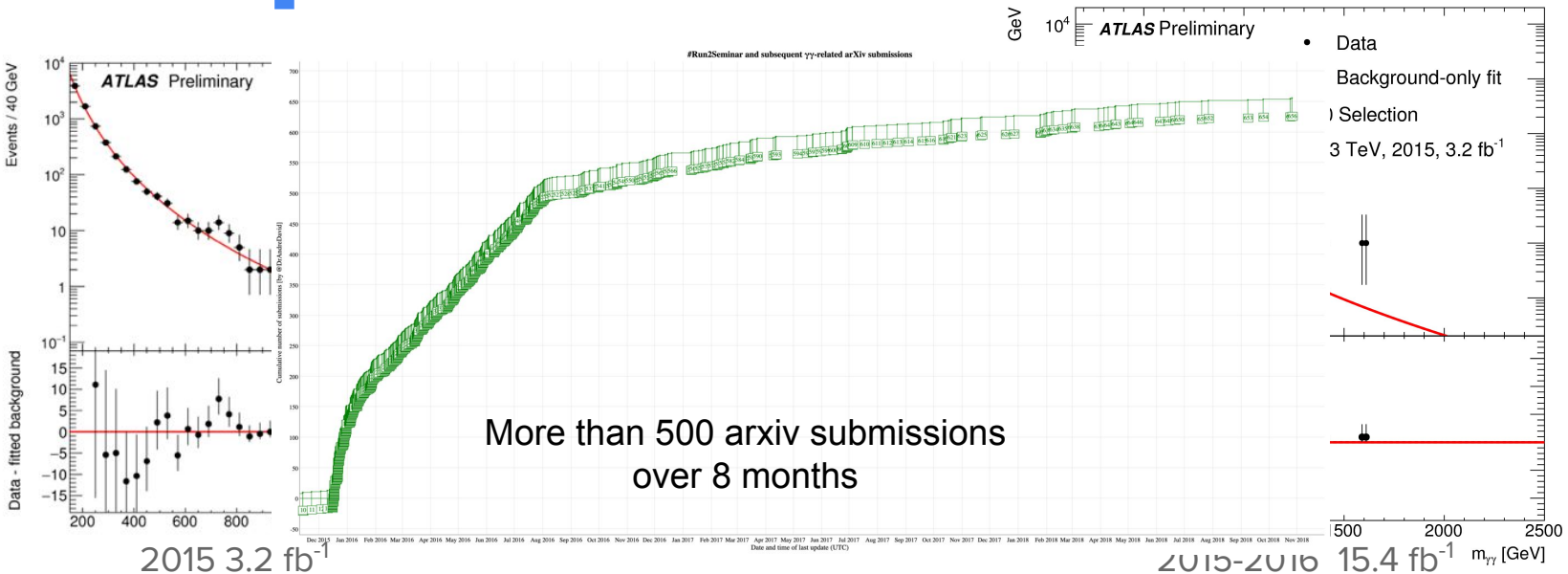
Using data taken with the CASA-MIA detector over a five year period (1990-1995), we find no evidence for steady emission from either source at energies above 115 TeV. The derived upper limits on such emission are more than two orders of magnitude lower than earlier claimed detections

Chicago Air Shower Array (**CASA**) + Michigan Muon Array (MIA), 1992-1999

https://arxiv.org/abs/astro-ph/9611117

The case of the statistical fluctuation

750 GeV diphoton excess



A deviation from the Standard Model background-only hypothesis corresponding to 3.4 standard deviations is observed in the 2015 data for a resonance mass hypothesis of 730 GeV. No significant excess at such mass over the background expectation is observed in the 2016

https://inspirehep.net/literature/1480039

The case of the illusive background

Primordial gravitational waves

Top story

Primordial gravitational wave discovery heralds 'whole new era' in physics

17 Mar 2014: Gravitational waves could help unite general relativity and quantum mechanics to reveal a 'theory of everything'

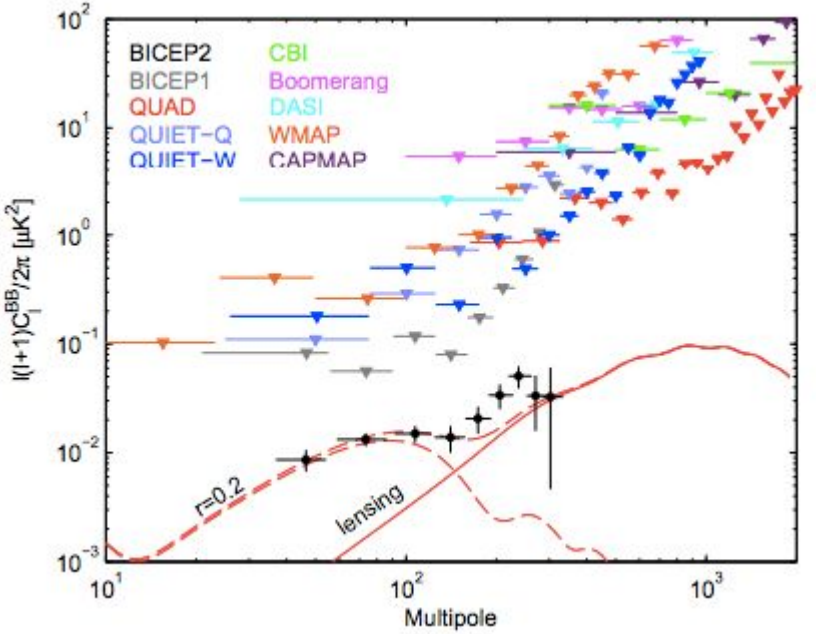
1675 comments

Most recent

Gravitational waves turn to dust after claims of flawed analysis

4 Jun 2014: Astronomers who thought they had detected echoes of the big bang may have only seen the effects of space dust

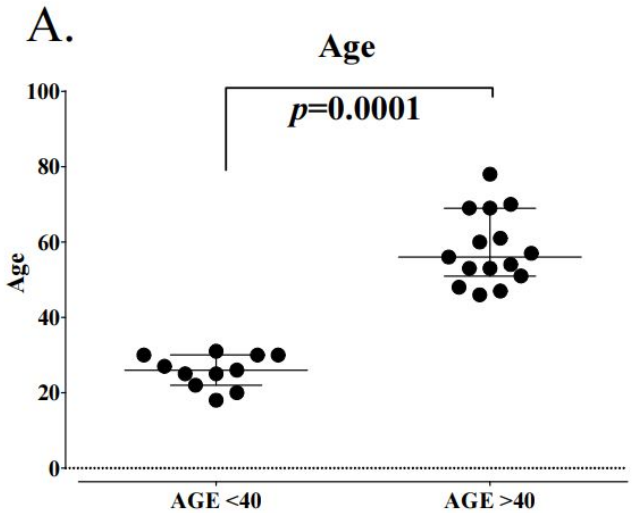
160 comments



https://arxiv.org/abs/2110.00483

The case of the mysterious plot

Analyses of 123 Peripheral Human Immune Cell Subsets: Defining Differences with Age and between Healthy Donors and Cancer Patients not Detected in Analysis of Standard Immune Cell Types



https://journals.sagepub.com/doi/pdf/10.5772/62322

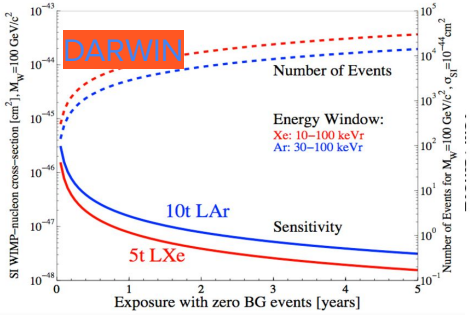
Part 1 memes conclusions

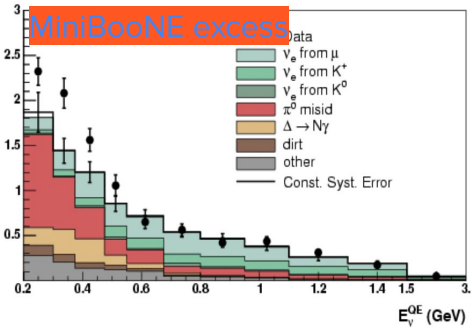
Why do we

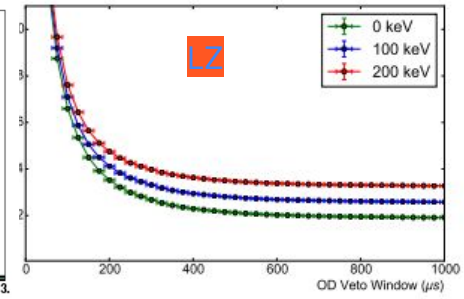
- quantify our know
  - Detector b
- Make predictio
  - Sensitivitie
- Make discoveri
- Compare, and

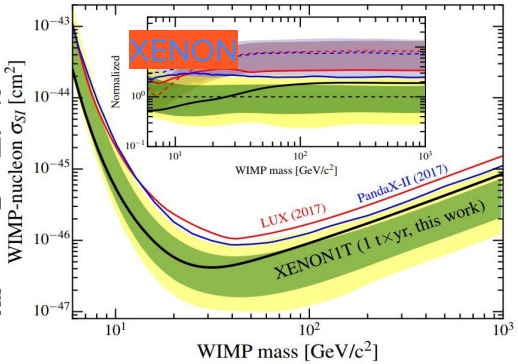
Underground challe

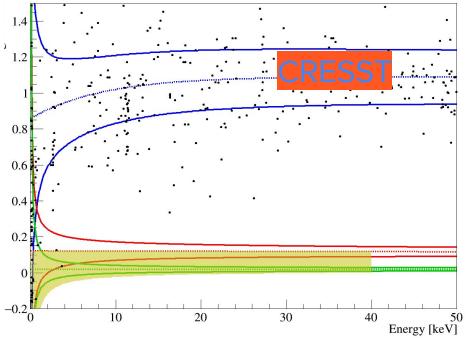
- Rare sea
- Backgro
- Low bac
- Large da
- Blind ana

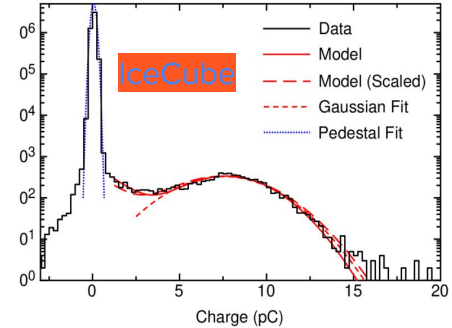


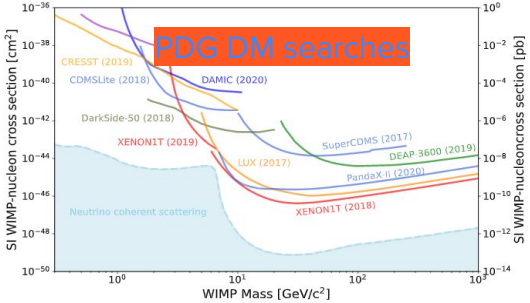


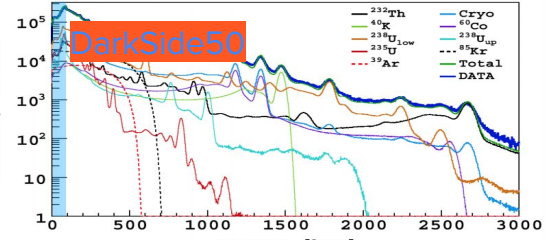


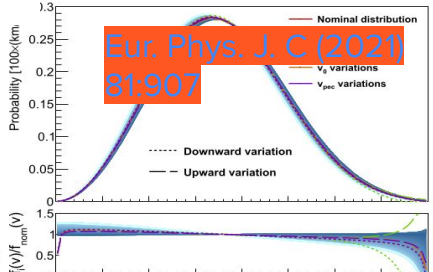












My top 10 (Paranoid) advices for doing statistics

⇒ All models are wrong, but some are useful.

⇒ Always read the fine prints (in papers…in codes….in manuals…).

⇒ Visualize the numbers. Be creative.

⇒ Black boxes are scary.

⇒ Try it yourself - Best way to understand - is to do it!

⇒ Comment everything (not just for other users, also for the future you)

⇒ Test your code often on “simpler” and “diverse” scenarios. Do sanity checks

⇒ If your code compiles on the “first trial” - beware!

⇒ Always wear sun screen

Part 1: Probability

What is probability?

Depends who you ask....

A Mathematician, two physicists and a politician walk into a bar....

The bartender asks them: “Can you tell me what probability is?”

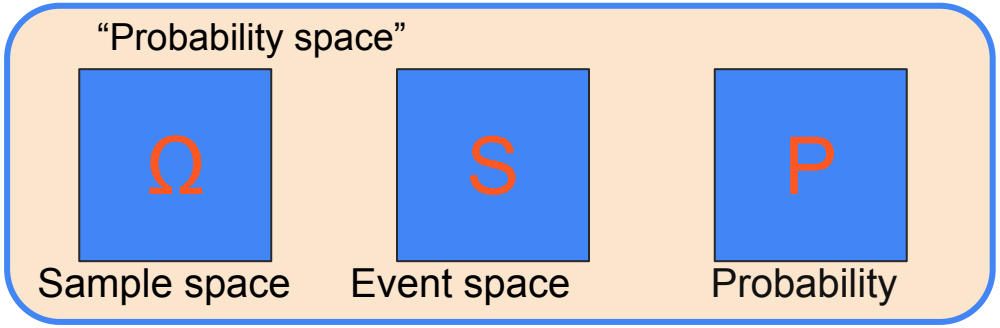
The Mathematician says : “A Number between 0-1, assigned to objects in a sample space”

The Frequentis says: “Frequency of an outcome in a repeating experiment”

The Bayesian says: “Probability is a subjective term, representing our degree of belief in a hypothesis”

The politician says: “Definitely yes!! Probably not!!”

What is probability? The building blocks



Sample space Ω: The set of all the outcomes of a random experiment. An outcome is an element in the sample space $\omega \in \Omega$.

Event space S: A set whose elements $A \in S$ (called events) are subsets of Ω (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment).

Probability measure P: A function $P : F \rightarrow R$ that maps objects in S to the interval [0, 1].

What is probability?

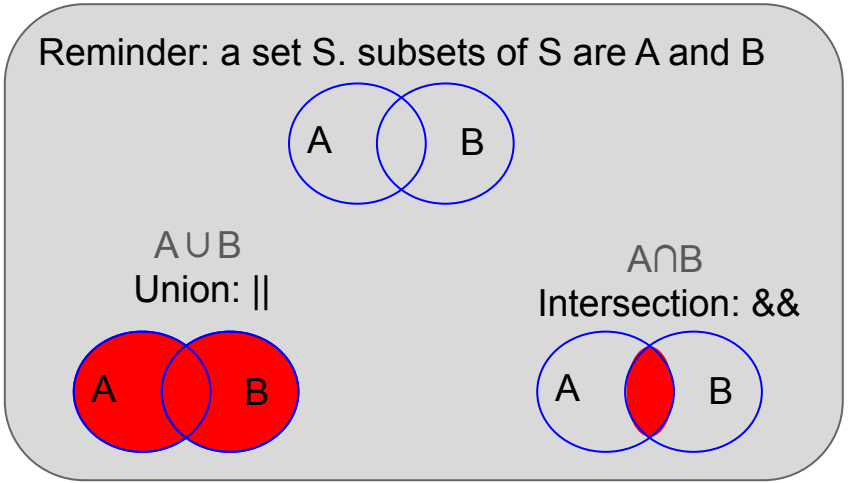
Let S denote a sample space with a probability measure P defined over it, such that probability of any event $A \subset S$ is given by P(A). Then, the probability measure obeys the following axioms:

Kolmogorov axioms: (1933)

Some set S… A,B are subsets of S.

- Non negativity: For all $A \subset S$, $P(A) \geq 0$
- Unitarity: $P(S)=1$
- Countable additivity:

  If $A \cap B=0 \Rightarrow P(A \cup B) = P(A) + P(B)$



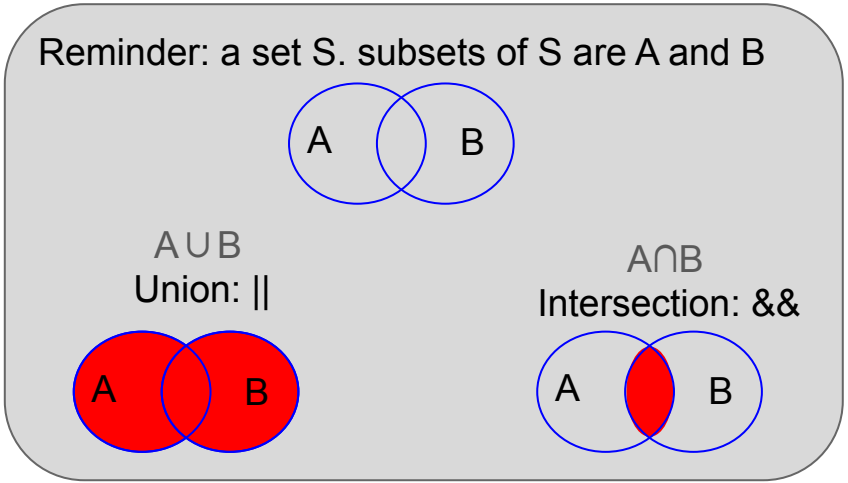
We can also deduce that

We can also deduce that:

- $P(A) = 1 - P(A)$ (A is the complement of A)
- $P(A \cup A) = 1$
- $P(\varnothing)=0$
- If $A \subset B$ then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

👆👆👆👆👆👆👆👆👆👆👆👆👆👆

Can be deduced from Kolmogorov axioms

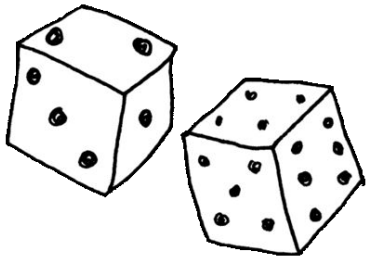


Roll two dice

The sample space:

$\Omega = \{1,2,3,4,5,6\}x\{1,2,3,4,5,6\}$

$=\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),.....,(6,4),(6,5),(6,6)\}$



The event space:

S=Various combos of outcomes, e.g.:

What is L, the event that the sum of the dice is 7::

$L=\{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$ ⟹ $P(L)=1/6$

What is M, the event that the sum of the two dice is 6:

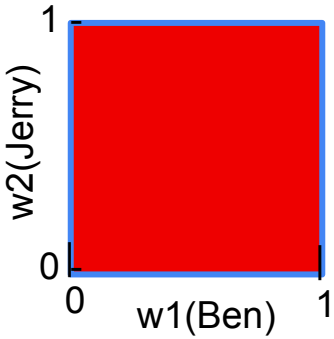
$M= \{(1,5), (2,4), (3,3), (4,2), (5,1)\}$ ⟹ $P(M)=5/36$

What is N ?

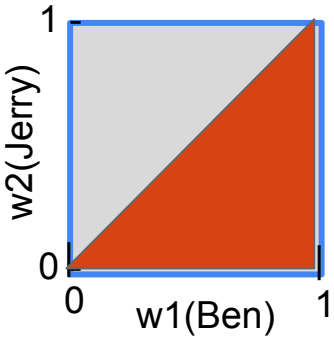
$N=\{(4,1), (4,2), (4,3), (4,4), (4,5), (4,6)\}$ ⟹ $P(N)=1/6$

Graphical representation of sample space

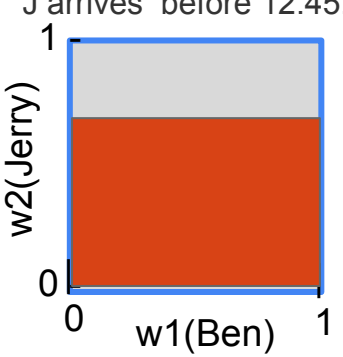
Ben & Jerry plan to meet for ice cream between noon and 1 but they are not sure of their arrival times



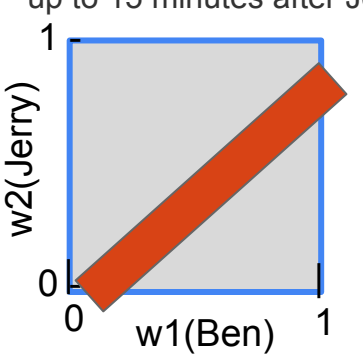
the event that
B arrives after J



the event that
J arrives before 12:45.



the event that Ben arrives
up to 15 minutes after Jerry

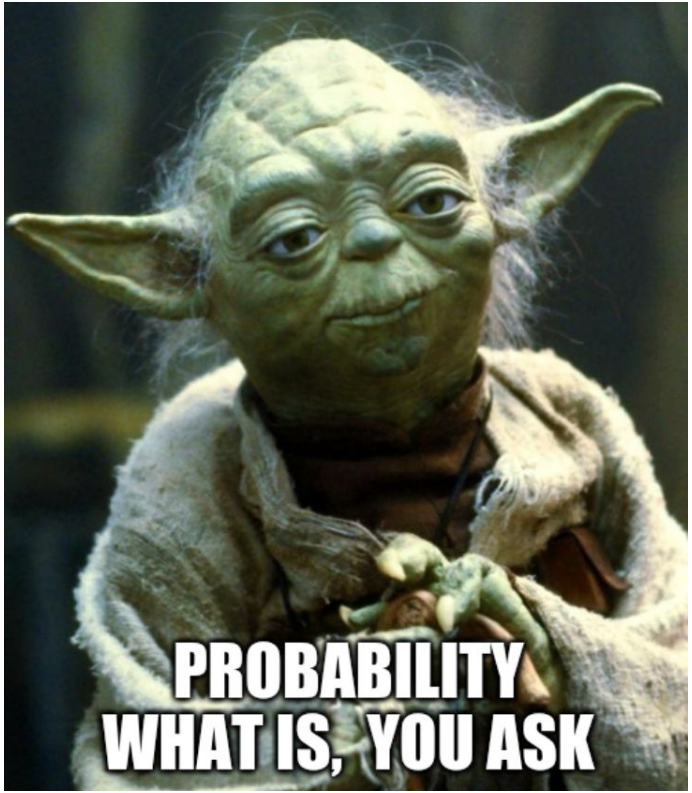


Probability Ben arrives same time or after Jerry = 0.5

Probability that Jerry arrives first and Ben arrives at most 15 minutes after Jerry= 7/32=0.21875.

Probability that Jerry arrives first and Ben arrives at most 1 minutes after Jerry =0.0165

What is probability?



Relative frequency	Subjective probability
The probability of the event is the proportion of times that the event would occur in a very large number of hypothetical repetitions of the random phenomenon. Probabilities are associated only with c	The probability is a the “strength of believe” that it is correct. Probabilities are associated with state of knowledge on parameters - given some prior ties, how should they change provided
$P(A)=\lim_{n\to\infty}\frac{\#\text{outcome A in n measure}}{n}$	egree of belief that hypothesis A is true
Elements of s = possible outcomes of repeatable measurement	of s = hypotheses that are true or false (“hypothesis space”). Mutually exclusive.
Subset A = event = corresponds to the occurrence of any of the outcomes in the subset	Subset A = set of one or more hypothesis
“Frequentist approach”	“Bayesian approach”

Conditional probability

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

P(A ∩ B)= intersection = Unconditional probability involving both events

P(A|B)= Conditional = Conditional probability of one event, given the other

P(B)= Marginal=Unconditional probability of a single event

$$P(A|B) \neq P(B|A)$$

Law of total probability

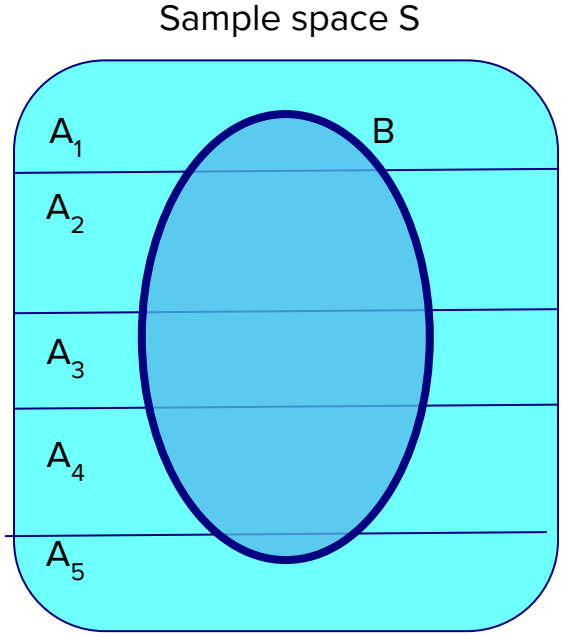
If: $A_1,\ldots,A_n$ are disjoint

&& if: $A_1 \cup A_2 \cup \ldots \cup A_k = S$

then: $$P(B) = \sum_{i=1}^{n} P(B \cap A_i)$$

(Conditional probability)

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

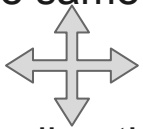


Law of total probability

Example:
The “Look over there!!!” game ["acchi muite hoi" (あっち向いてホイ.])

<u>How to play it:</u>

- Two players face each other 👫
- On the count of 3, player A points her finger up, down, left or right. 👆👇👈👉
- In the same time player B points his head up, down, left or right.



- If the directions are the same, player B wins the game. ✌
- If the directions are different they switch roles and do another round😐.

https://mangaboat.com/manga/karakai-jouzu-no-takagi-san/ch-082/

What are my chances to win if I go first?

A=event of winning the game

B=event of winning the game on the first round

B'=event of not winning the game on the first round

P(A) = probability of "me" winning the game=p

By the law of total probability:
$P(A)=P(A|B)\cdot P(B)+P(A|B')\cdot P(B')$
$p = 1 \cdot 0.25 + (1-p)\cdot 0.75$
Solve for p : $p=4/7=0.57$

How can you verify this result?

Simulate:
Program it!

Experiment:
Play it!

https://mangaboat.com/manga/karakai-jouzu-no-takagi-san/ch-082/

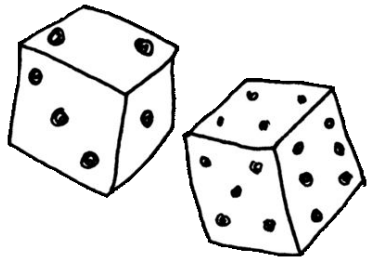
How to code the lookaway game

- The each round simulates one (point, look) pair
  - If the game ends in an even number of rounds, the other player will win
  - If the game ends in an odd number of rounds, the player who looked away will win
  - ⇒ Loop until there is a win
    - ⇒ Run the games 1000 times
      - ⇒ The probability of player X winning the game is the fraction of times the games ended with an odd X

Independence

if P(A∩B)=P(A)P(B) subsets A and B are independent. ⇒ and $P(A|B|) = \frac{P(A) \cdot P(B)}{P(B)} = P(A)$

<u>Independet events doesn't necessarily mean that A∩B = 0 :</u>



⇒ L : Getting sum of 7: {(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)} P(L)=⅙

⇒ M: Getting sum of 6: {(1,5),(2,4),(3,3),(4,2),(5,1)} P(M)=5/36

⇒ N: Getting four on first :{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)} P(N)=⅙

Are L and N independent?

L∩N = {(4,3)} = 1/36 P(L)·P(N) = ⅙ * ⅙ = 1/36 ⇒ Independent…

P(N|L)=if we know we got sum of 7, then the probability of getting 4 on first is ⅙, which is identical to the probability of getting 1:6 anyhow

Are M and N independent?

M∩N={(4,2)} = 1/36 P(M)*P(N) = 5/36*⅙=5/224 ⇒ not independent

P(N|M) = if we know we got sum of 6, then the probability of getting 4 on first is ⅕

Bayes' theorem

Using the conditional probability we can say

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

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

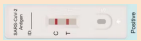
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \cap A)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$$

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

Example: An antigen test I took came positive*
What is the probability I am actually sick?

D+ :I got the disease. I am sick !

D- : I did not get the disease. I am healthy



T+ : The antigen test came poistive

T - : The antigen test came negative

True negative P(T-|D-)=99.7%=”specificity”

False negative P(T+|D-)=0.3%

True positive P(T+|D+)=53.8% = “sensitivity”

False negative P(T-|D+)=46.2%

Prevalence P(D+)=0.1%

0.1% of population sick

* Pressumably

$$P(D+ \mid T+) = \frac{P(T+|D+)\cdot P(D+)}{P(T+)}$$

P(D+)=0.001 P(T+|D+)=0.997

P(T+) = P(D+)·P(T+|D+) + P(D-)·P(T+|D-) =0.001*0.538 + 0.999*0.003 = 0.005675

P(D+|T+) = 0.15 = 15%

P(D+|T-) = … = 0.05%

For PCR tests: Specificity 99.7% ,

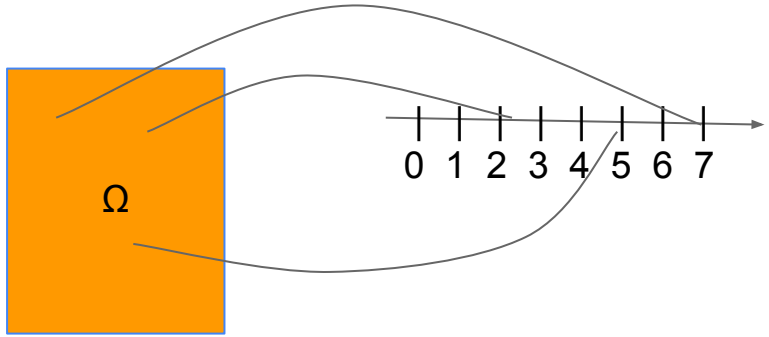
Sensitivity 95.7%

Random variables

A Random Variable - takes on a specific value for each element of the set S

Random variables can be:

- Discrete / Continuous
- Single value / vector
- Finite / infinite sample space

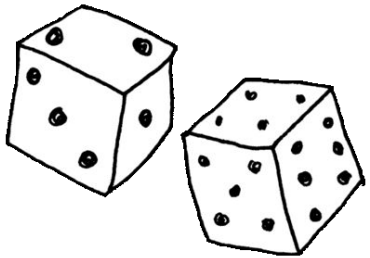


E.g.:

X is the sum of the two dice e.g.: {X=4}={(1,3),(2,2),(3,1)} is the event that the sum of the two dice is 4.

Y is the largest of two values e.g. {Y=3}={(1,3),(2,3),(3,3),(3,2),(3,1)} is the event that the largest value is 3

A Function of a random variable is also a random variable. That is, if X is a random variable and g(X) is a function, then g(X) is a random variable



From Bayes theorem to Bayesian Statistics

We can use Bayes' theorem to assign probabilities to hypothesis (H), based on assumed knowledge (I), which can be updated when data (X) become available.

Probability of a hypothesis (H) given a data (x): $$P(H \mid X) = \frac{P(X \mid H) \cdot P(H)}{P(X)}$$

P(X|H) = The likelihood - Assuming some model, what is the probability to get data?

P(H) = (actually P(H|I)) = Prior probability - before including the new data
⇒ Determination of the prior is subjective. Even a flat prior is informative.

P(H|X)= posterior probability - how the prior probability changed based on the new data

P(X)=(actually P(X|I)) = Normalization over all possible hypothesis. Estimated using the law of total probability = $\int P(X|H)P(H)dH$

What does this mean?

$$g = 10.1 \pm 0.4 \, \frac{m}{s^2}$$

$$P(H \mid X) = \frac{P(X \mid H) \cdot P(H)}{P(X)} \propto P(X \mid H) \cdot P(H)$$

Part 2: Distributions

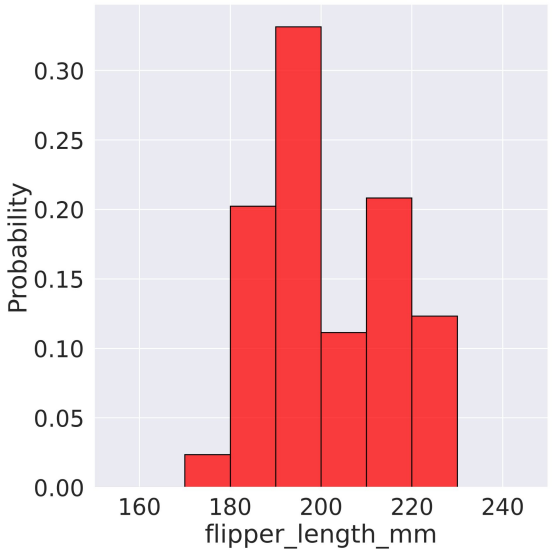
PMF - Probability Mass Function

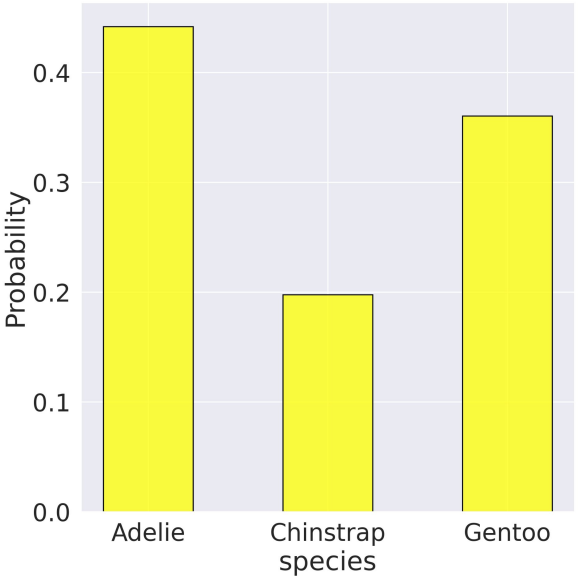
If the outcome of an experiment is discrete x_i:

$P(x_i)$ is the probability mass function

$$P(x_i) = P_i$$

$$\sum_{i=1}^{N} x_i = 1$$





PDF- Probability Density Functions

If the outcome of an experiment is continuous x:

f(x) is the probability density function (PDF)

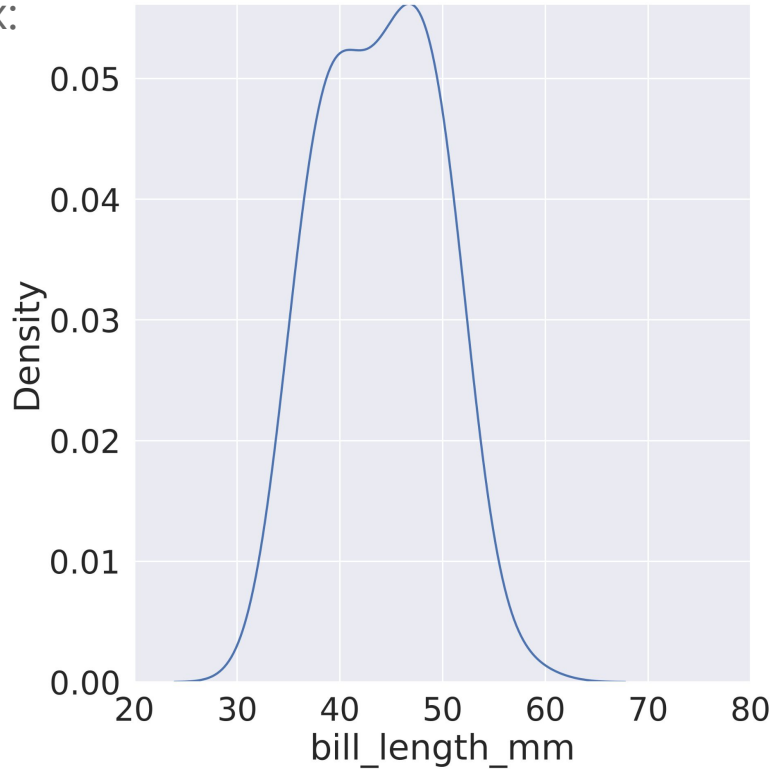
$$P(A)=P(x\in[x,x+dx])=f(x)dx$$

$$\int_a^b f(x)dx=P(a\leqslant x\leqslant b)$$

⇒ f(x) is nonnegative

⇒ PDFs are normalized $\int f(x)=1$

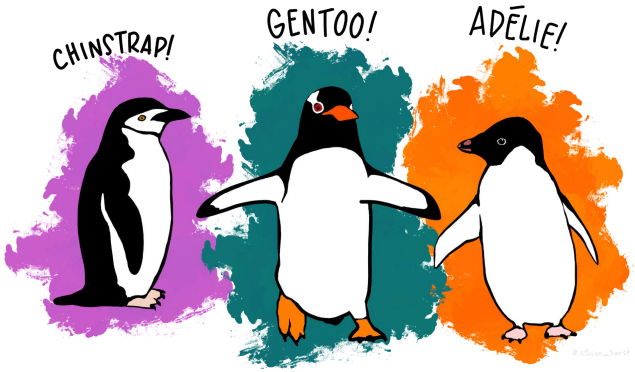
⇒ (x) is not a probability

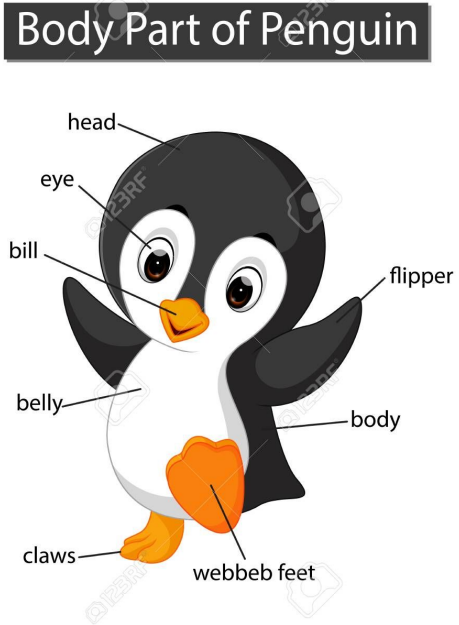


Palmer Archipelago penguins...dataset

A great intro dataset for data exploration & visualization

https://github.com/allisonhorst/penguins





PDF- Probability Density Functions

If the outcome of an experiment is continuous

f(x) is the probability density function (PDF)

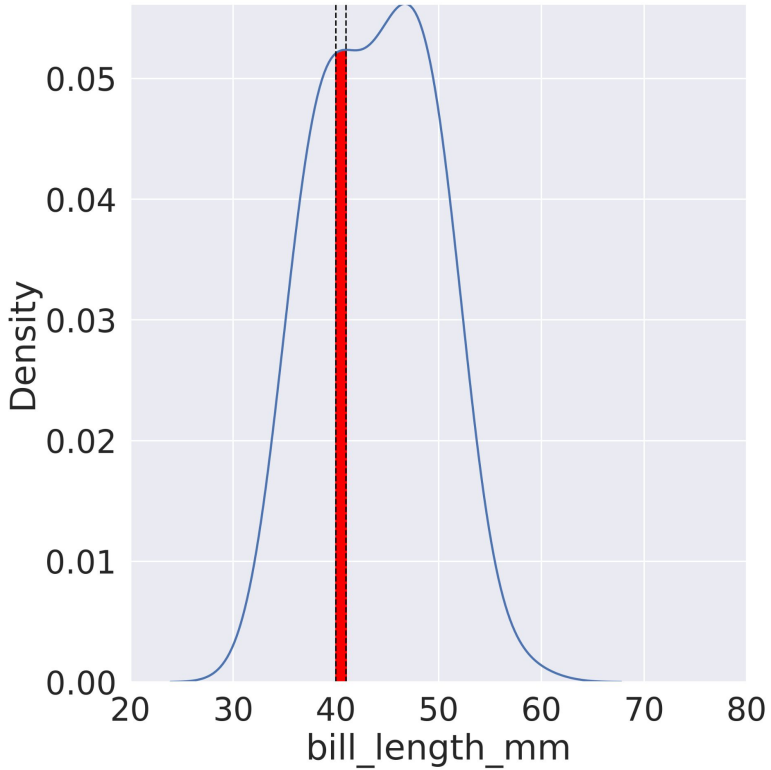
$$P(A) = P(x \in [x, x + dx]) = f(x)dx$$

$$\int_a^b f(x)dx = P(a \leq x \leq b)$$

⇒ f(x) is nonnegative

⇒ PDFs are normalized $\int_{-\infty}^{+\infty} f(x)dx = 1$

⇒ (x) is not a probability



PDF- Probability Density Functions

If the outcome of an experiment is continuous x:

f(x) is the probability density function (PDF)

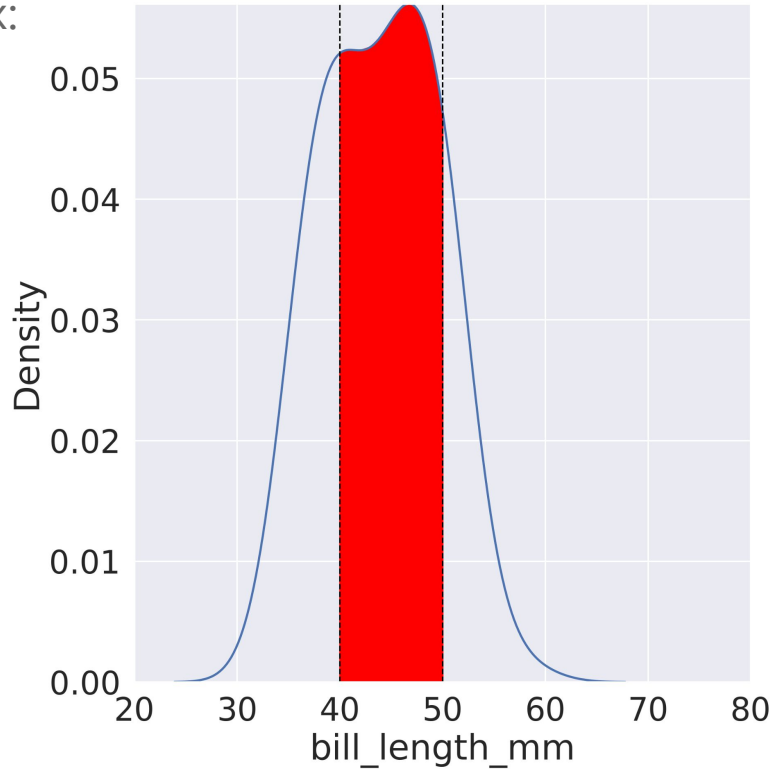
$$P(A) = P(x \in [x, x + dx]) = f(x)dx$$

$$\int_a^b f(x)dx = P(a \leq x \leq b)$$

⇒ f(x) is nonnegative

⇒ PDFs are normalized $\int_{-\infty}^{+\infty} f(x)dx = 1$

⇒ (x) is not a probability



PDF- Probability Density Functions

If the outcome of an experiment is continuous x:

f(x) is the probability density function (PDF)

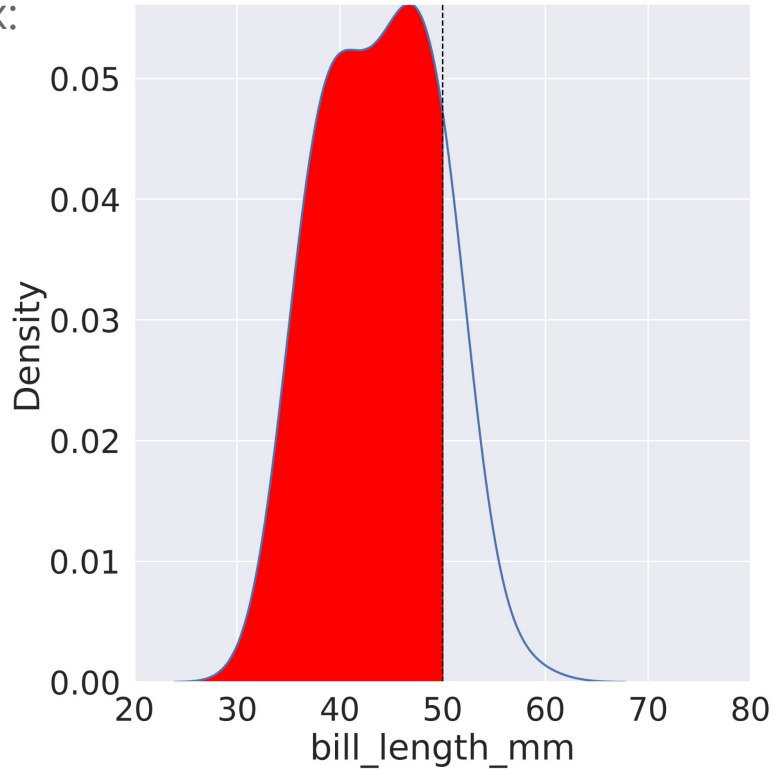
$$P(A) = P(x \in [x, x + dx]) = f(x)dx$$

$$\int_a^b f(x)dx = P(a \leq x \leq b)$$

⇒ f(x) is non-negative

⇒ PDFs are normalized $\int_{-\infty}^{+\infty} f(x)dx = 1$

⇒ (x) is not a probability



CDF - Cumulative distribution function

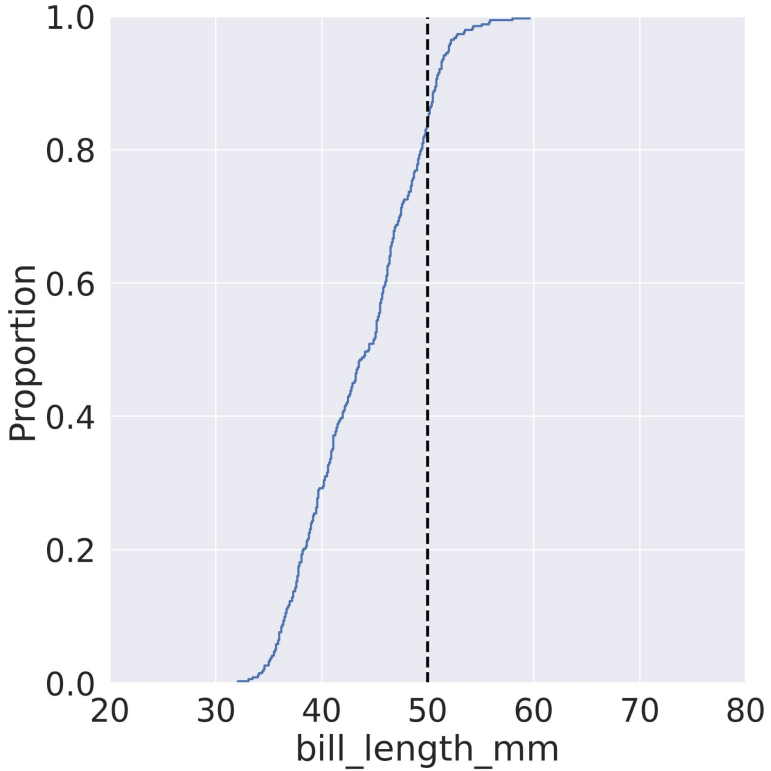
$$F(b)=P(X\leqslant b)=\int_{-\infty}^{b} f(X)dX$$

$$f(X)=\frac{dF(X)}{dX}$$

- Non-decreasing, i.e. $a\leqslant b\Rightarrow F(a)\leqslant F(b)$
- Cannot be less than 0, or more than 1

  $0\leqslant F(X)\leqslant 1$

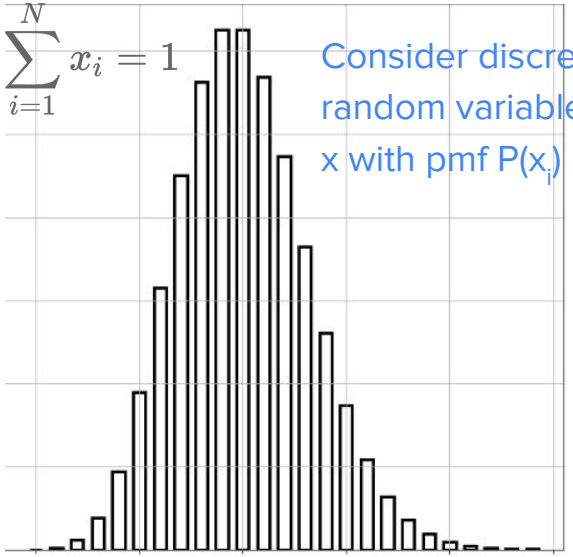
  $\lim_{X\to-\infty} F(X)=0$ $\lim_{X\to\infty} F(X)=1$
- $P(a\leqslant X\leqslant b)=F(b)-F(a)$
- Complementary CDF (1-F(x))



The fun things we can with a (to a?) PDF

$$\sum_{i=1}^{N} x_i = 1$$

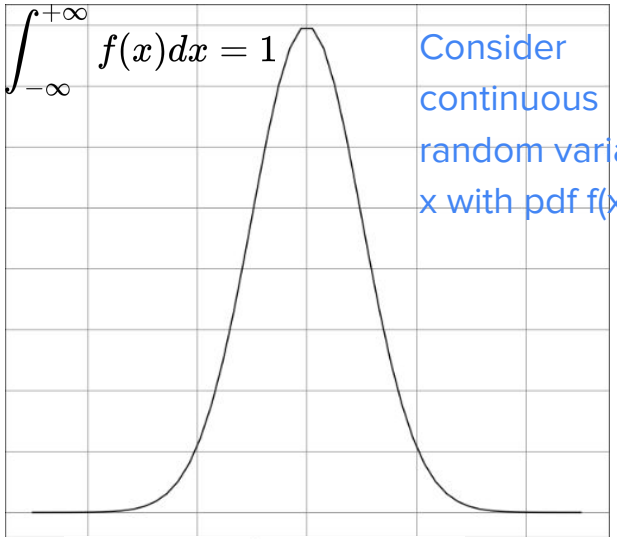
Consider discrete random variable x with pmf $P(x_i)$



$$E[r] = \sum_{r} rP(r)$$

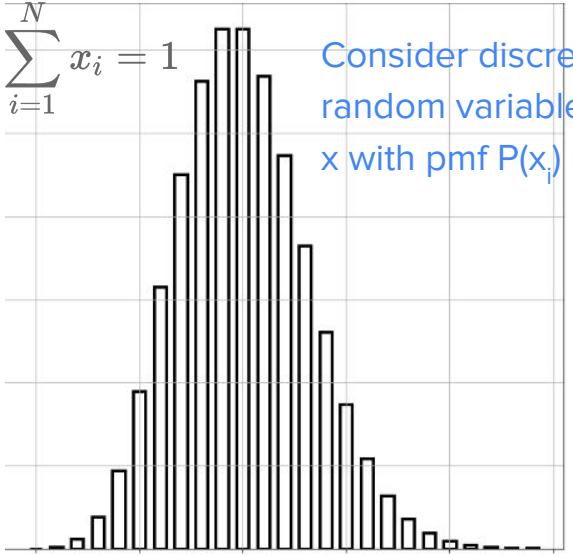
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Consider continuous random variable x with pdf f(x)

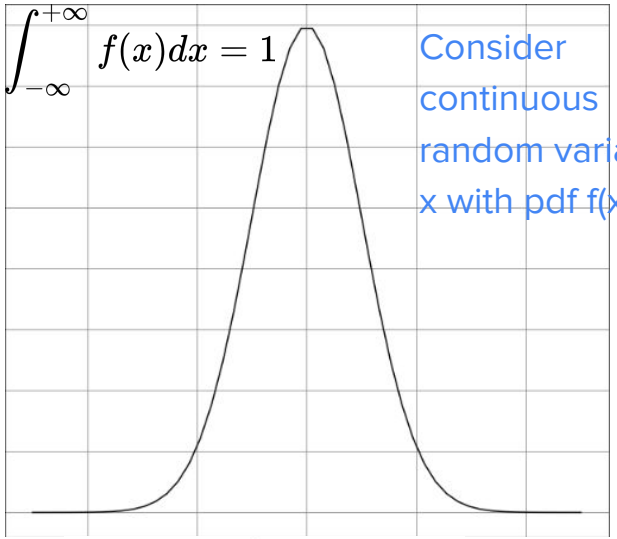


$$E[x] = \int x\, f(x)\, dx$$

The fun things we can with a (to a?) PDF
Reduce it to a number



Consider discrete random variable x with pmf $P(x_i)$



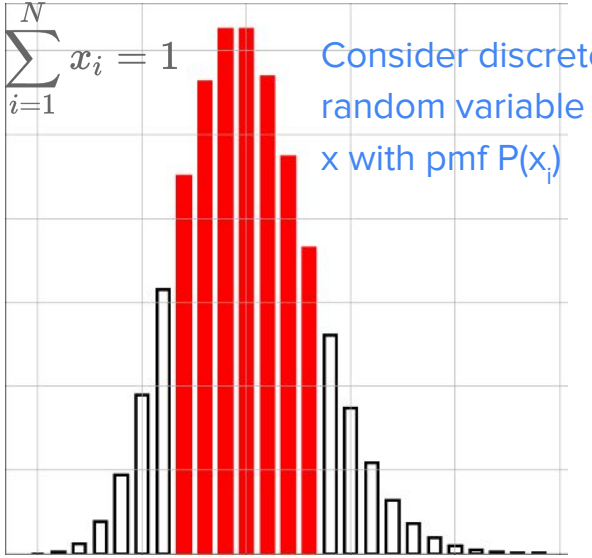
Consider continuous random variable x with pdf f(x)

Mean: $E[r] = \sum_r rP(r)$ $\quad E[x] = \int x\, f(x)\, dx$

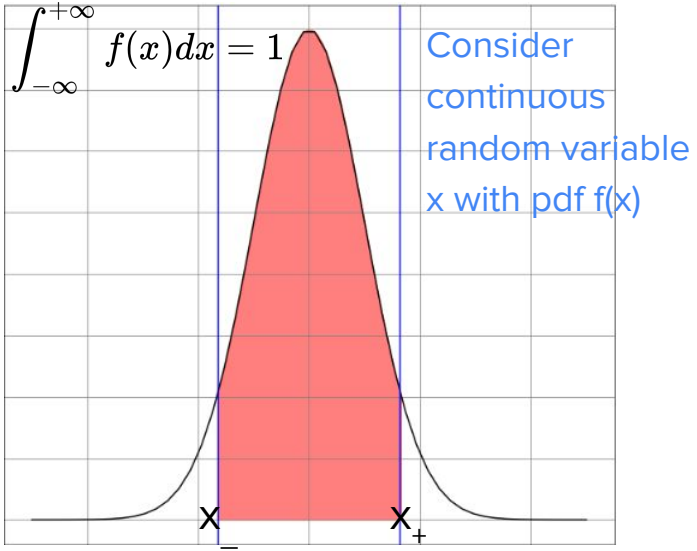
Variance: $V[r] = E[(X - E[X])^2] = E[X^2] - E[X]^2$

Standard deviation: σ=sqrt(V[x])

The fun things we can with a (to a?) PDF
Confidence Interval

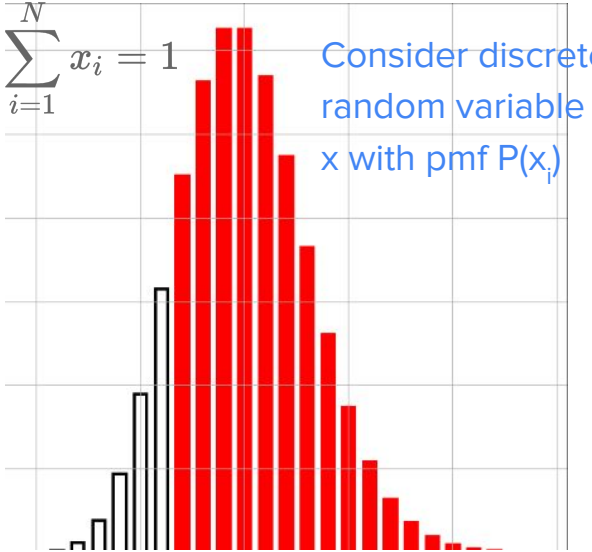


$$P(x_{i_-} \leq x \leq x_{i_+}) = \sum_{i_-}^{i_+} x_i \approx C$$

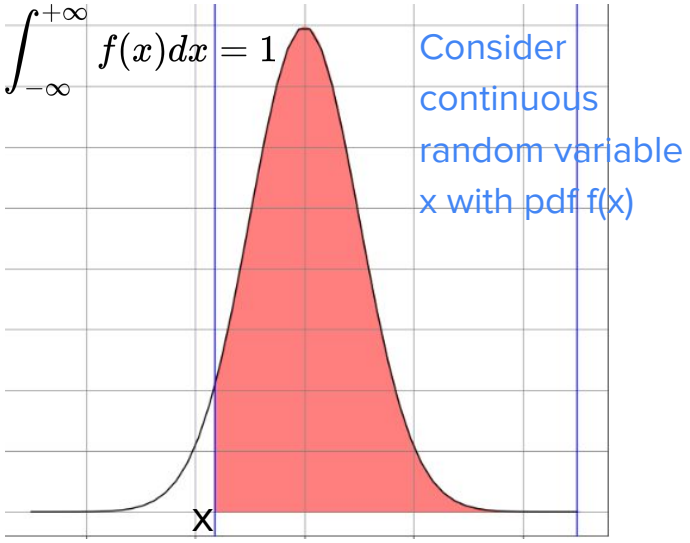


$$P(x_- \leq x \leq x_+) = \int_{x_-}^{x_+} f(x)dx = C$$

The fun things we can with a (to a?) PDF
Lower limit

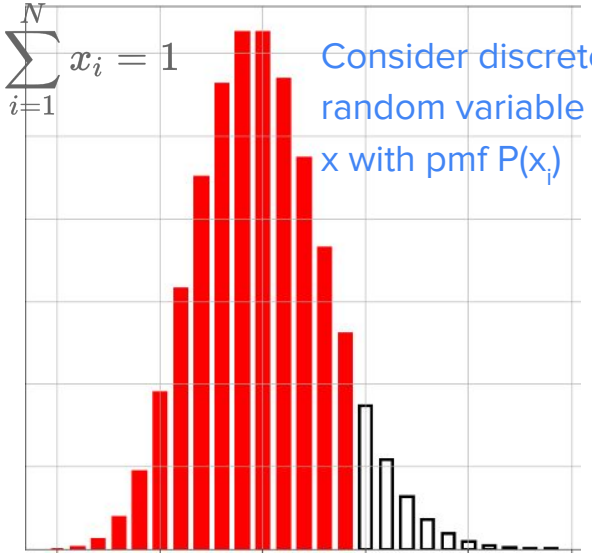


$$P(x \geq x_{i\_}) = \sum_{i\_}^{N} x_i \approx C$$

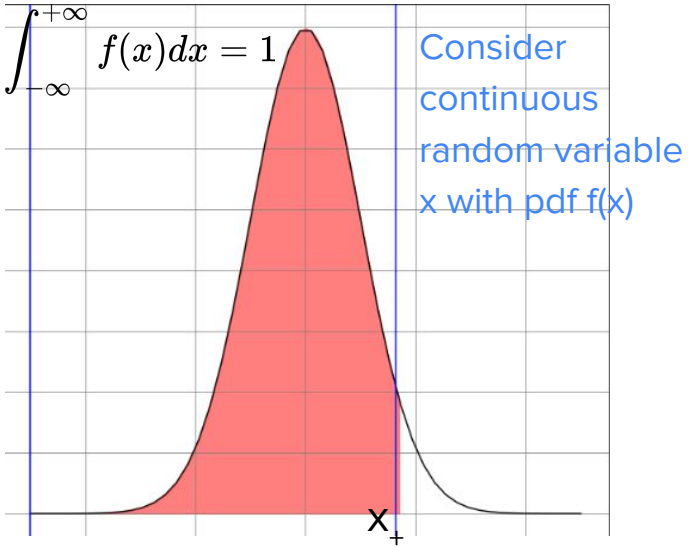


$$P(x \geq x_{\_}) = \int_{x\_}^{\infty} f(x)dx = C$$

The fun things we can with a (to a?) PDF
Upper limit



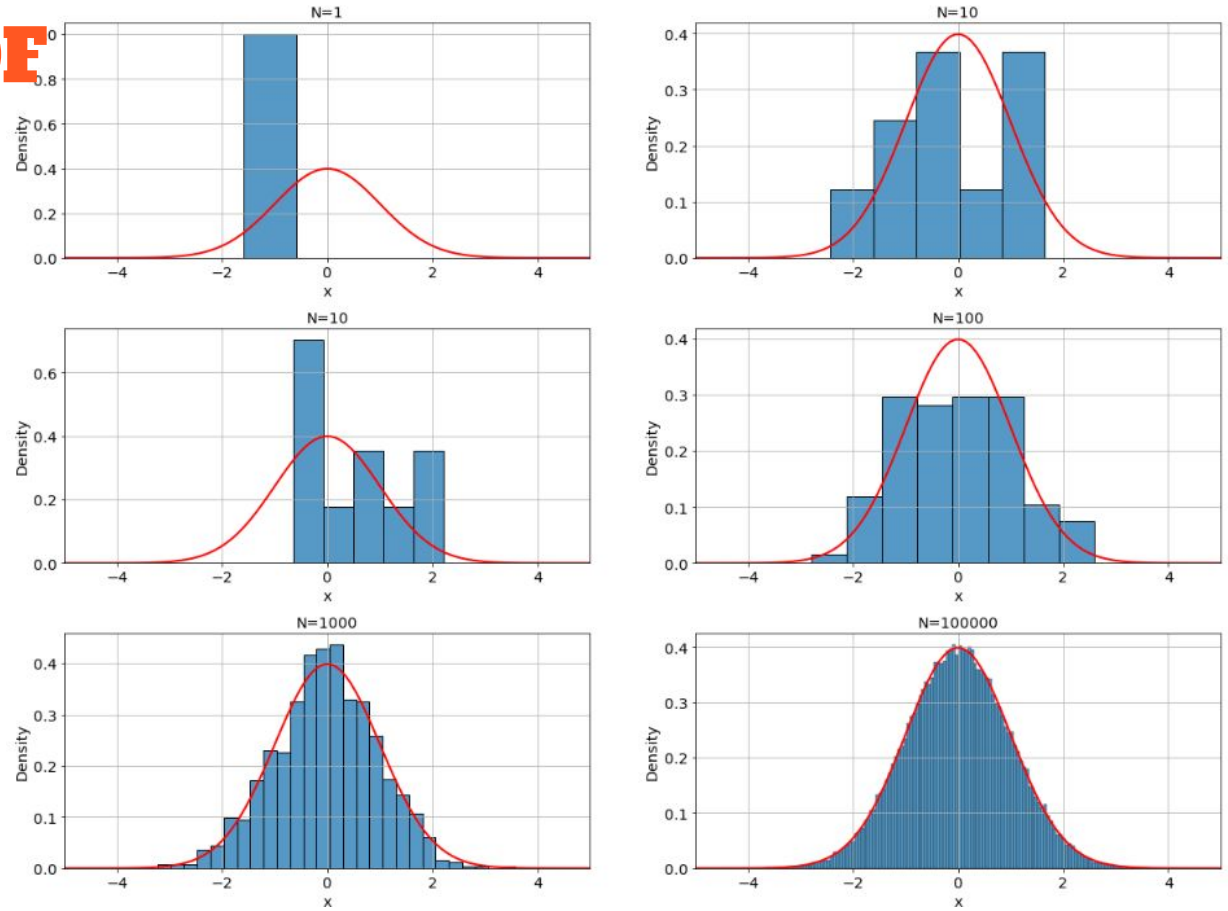
$$P(x \leq x_{i_+}) = \sum_0^{i_+} x_i \approx C$$



$$P(x \leq x_+) = \int_{-\infty}^{x_+} f(x)dx = C$$

From data to PDF

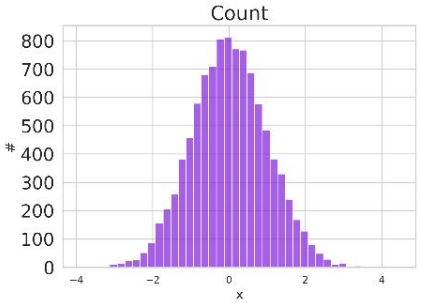
- Infinite sample size
- Zero bin width
- Normalized to 1



Normalizing, ah? Pay attention to what you use

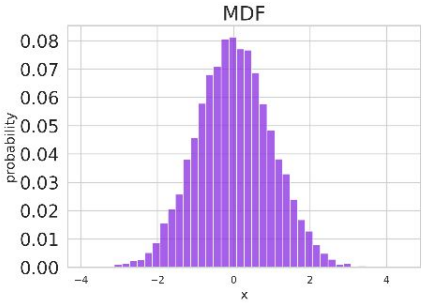
Bin width = 0.2

$\sum X_i = 10000$



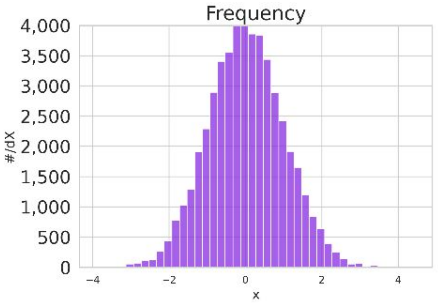
number of observations in each bin

$\sum X_i = 1$



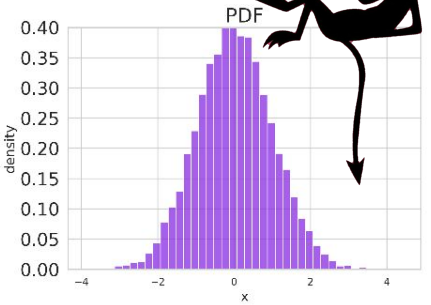
normalize such that bar heights sum to 1

$\sum X_i * w_i = 10000$



number of observations divided by the bin width

$\sum X_i * w_i = 1$

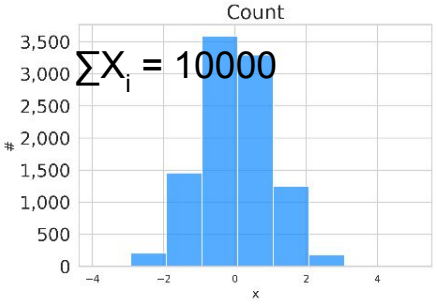


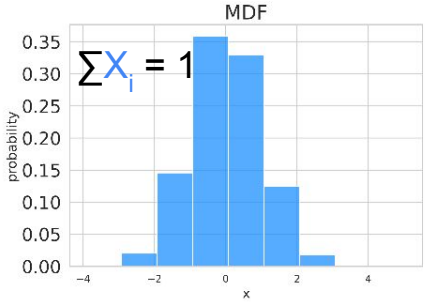
normalize such that the total area of the histogram equals 1

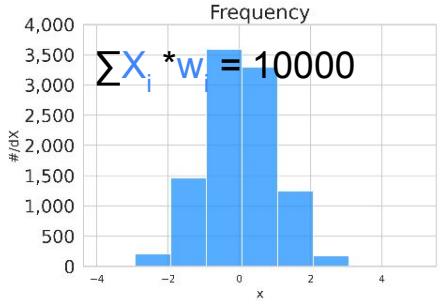
$$\sum_{i=1}^{N} x_i = 1$$

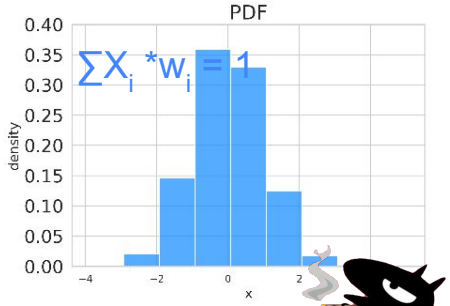
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

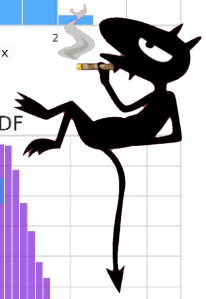
Bin width = 1.0



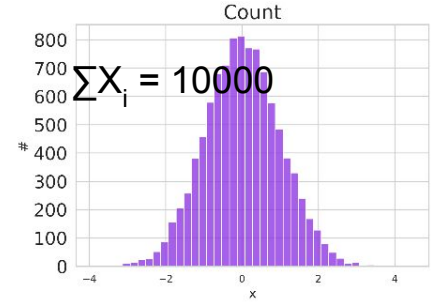


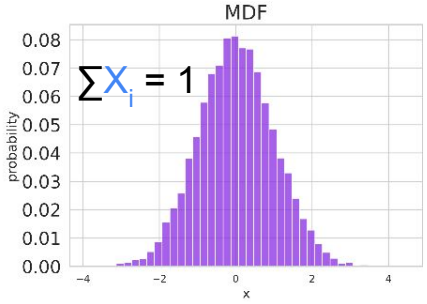


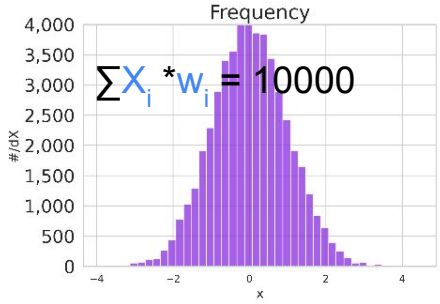




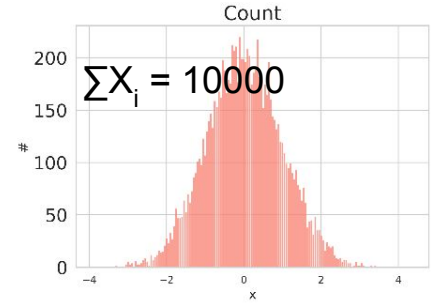
Bin width = 0.2

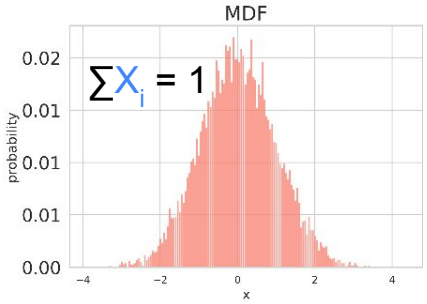


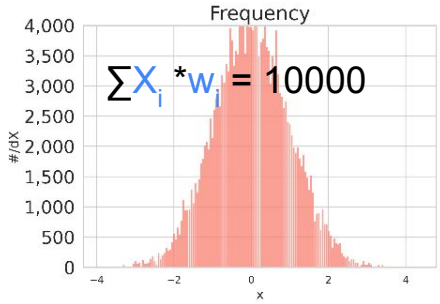


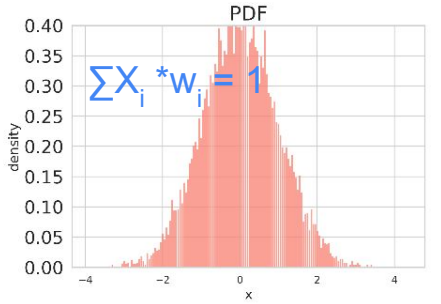


Bin width = 0.05









Bin width = 0.2

Count

$\sum X_i = 20$

MDF

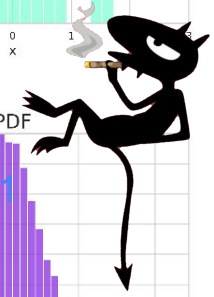
$\sum X_i = 1$

Frequency

$\sum X_i * w_i = 20$

PDF

$\sum X_i * w_i = 1$



Bin width = 0.2

Count

$\sum X_i = 10000$

MDF

$\sum X_i = 1$

Frequency

$\sum X_i * w_i = 10000$

PDF

$\sum X_i * w_i = 1$

Bin width = 0.2

Count

$\sum X_i = 1000$

MDF

$\sum X_i = 1$

Frequency

$\sum X_i * w_i = 1000$

PDF

$\sum X_i * w_i = 1$

Multivariate PDFs

PIn case there are several random variables (e.g. x and y):

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1 \qquad f(x,y) \geq 0 \qquad P(x,y \in A) = \iint^{A} f(x,y)dxdy$$

P(A∩B) = P(x found in [x,x+dx] and y found in [y,y+dy]) = f(x,y)dxdy

If the variables are independent:

P(A∩B) = P(A)P(B) ⇒ $f(x,y) = f_x(x)f_y(y)$

Multivariate PDFs - Marginalization

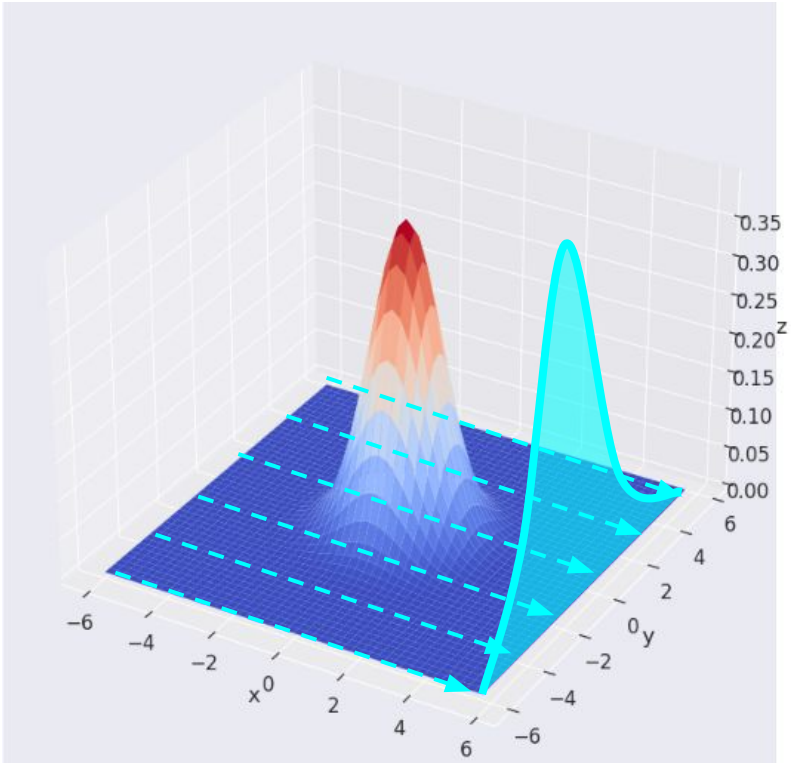
Marginalization:

Extracting information for some of the components

$$f_x(x) = \int f(x, y) dy$$

The expectation value e.g.

$$E[x] = \int \int x f(x, y)\, dx\, dy = \int x f_x(x)\, dx = \mu_x$$



Multivariate PDFs - Conditional

The probability that y equals Y, given that x=X.

$$P(y \mid x) = p(y = Y \mid x = X)$$

$$= \frac{P(x = X \text{ and } y = Y)}{P(x = X)} = \frac{joint}{\text{marginal}} = \frac{f(x,y)}{f(x)}$$

- The probability is:

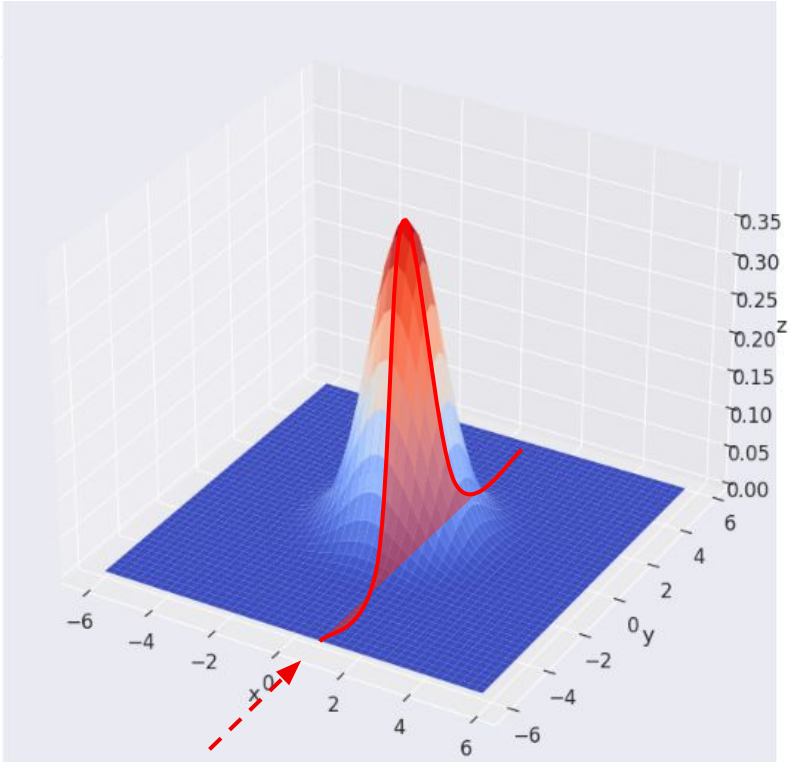
$$P(a \le y \le b \mid x = X) = \int_a^b f(y \mid x) dy$$

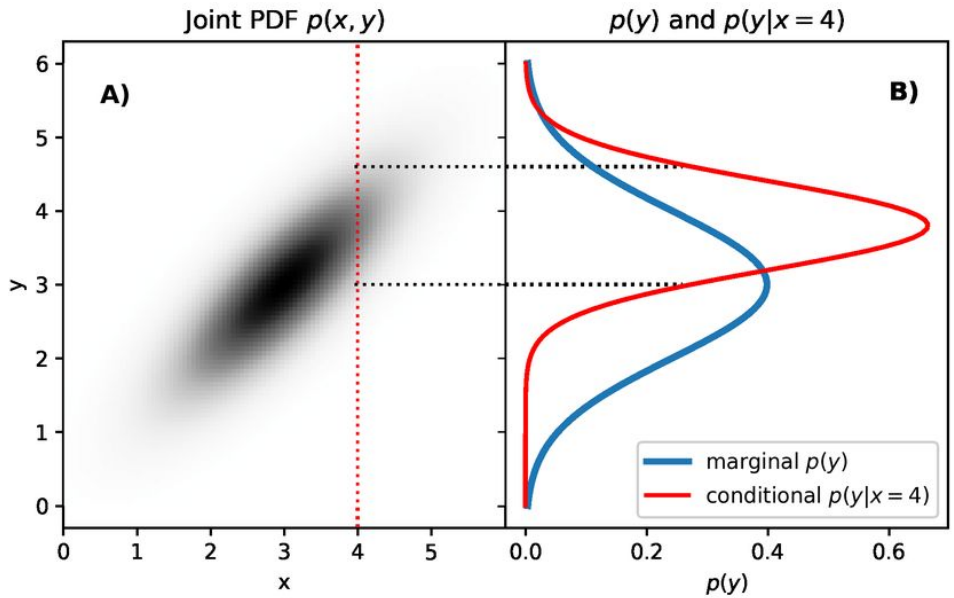
- Normalization still holds:

$$\int f(y \mid x) dy = 1$$

- If the variables are independent:

$P(A \cap B) = P(A)P(B) \Rightarrow \quad f(x,y)=f_x(x)f_y(y)$




Joint PDF $p(x, y)$
$p(y)$ and $p(y|x = 4)$
A)
B)
y
x
$p(y)$
marginal $p(y)$
conditional $p(y|x = 4)$
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6

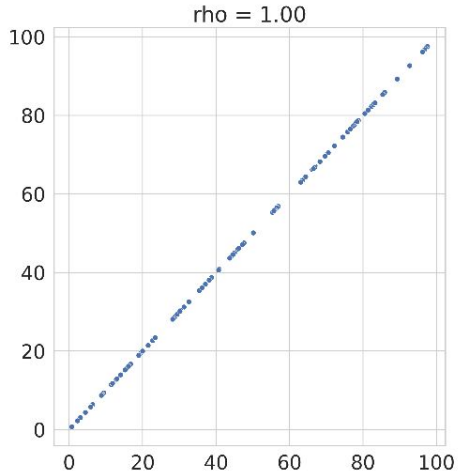
Covariance & correlation

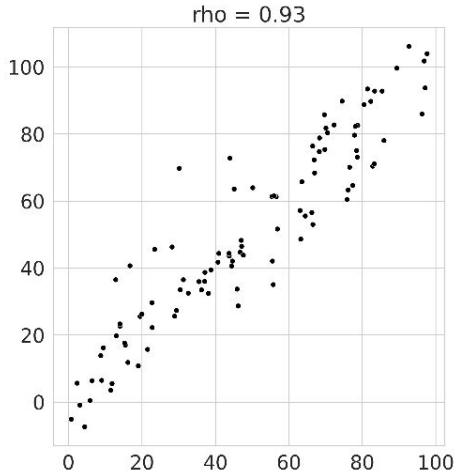
Covariance:

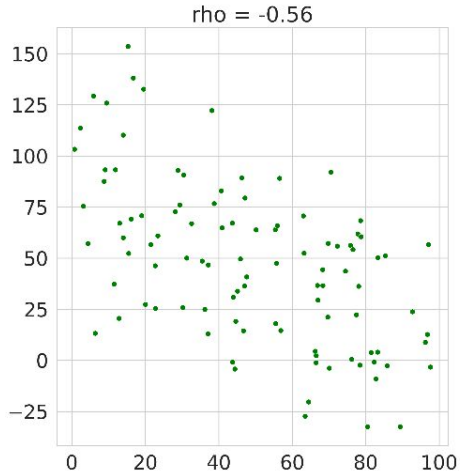
$$\text{cov}[x, y] = E[xy] - \mu_x \mu_y = E[(x - \mu_x)(y - \mu_y)]$$

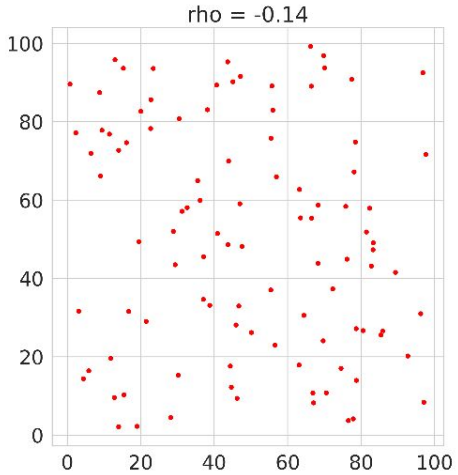
Correlation coefficient:

$$\rho_{xy} = \frac{\text{cov}[x, y]}{\sigma_x \sigma_y}$$









If we have several random variables....

covariance matrix

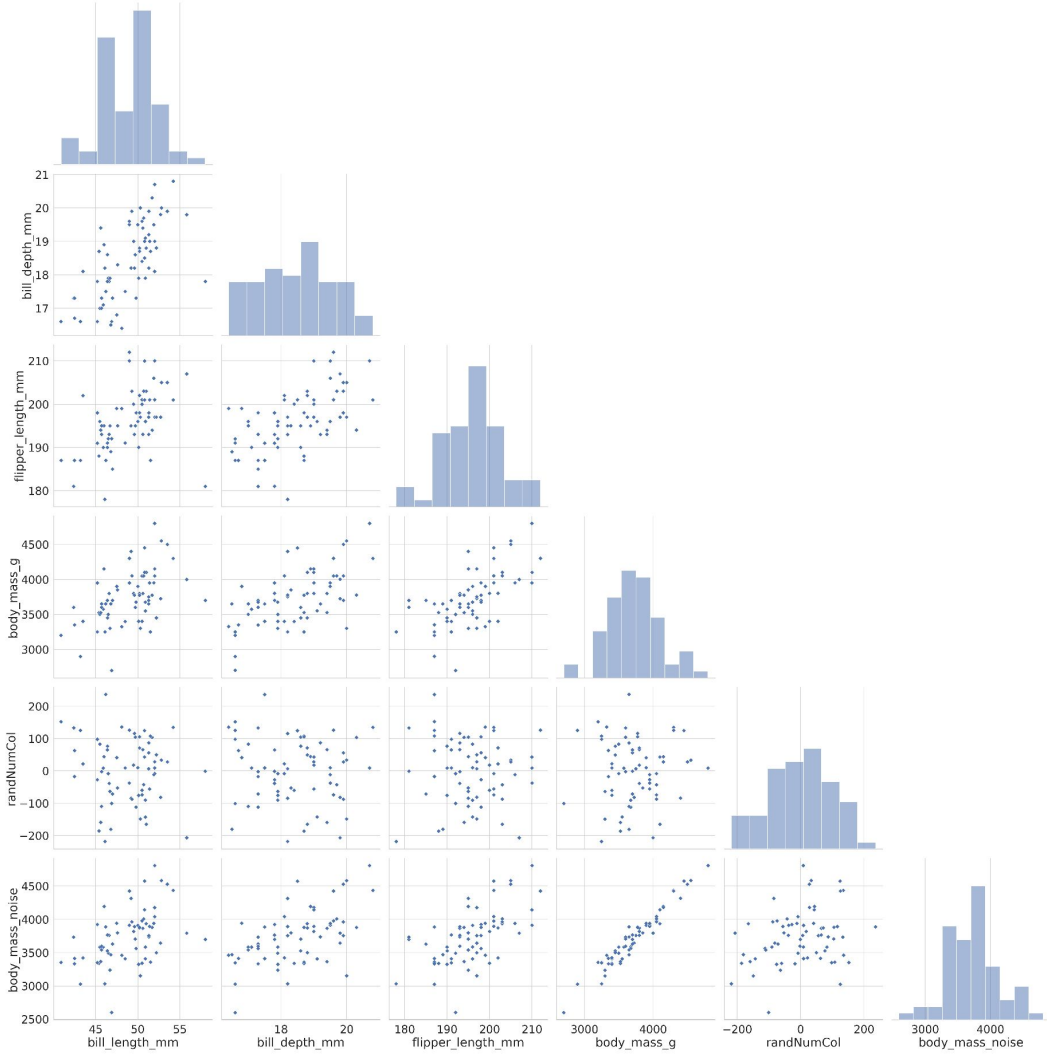
$$V_{ij} = \mathrm{cov}[x_i, x_j] = \rho_{ij}\sigma_i\sigma_j$$

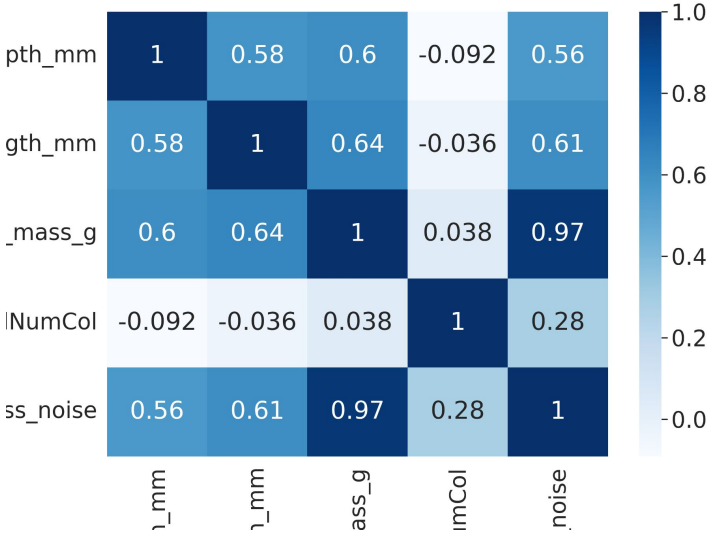
$$V = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & & \ddots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \rho_{n2}\sigma_n\sigma_2 & \cdots & \sigma_n^2 \end{pmatrix}$$

Correlation

$$\rho_{ij} = \frac{\mathrm{cov}[x_i, x_j]}{\sigma_i\sigma_j}$$

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix}$$


bill_depth_mm
flipper_length_mm
body_mass_g
randNumCol
body_mass_noise
bill_length_mm
bill_depth_mm
flipper_length_mm
body_mass_g
randNumCol
body_mass_noise


pth_mm
gth_mm
_mass_g
lNumCol
ss_noise
1
0.58
0.6
-0.092
0.56
0.58
1
0.64
-0.036
0.61
0.6
0.64
1
0.038
0.97
-0.092
-0.036
0.038
1
0.28
0.56
0.61
0.97
0.28
1
1.0
0.8
0.6
0.4
0.2
0.0

Statistics underground

PART II

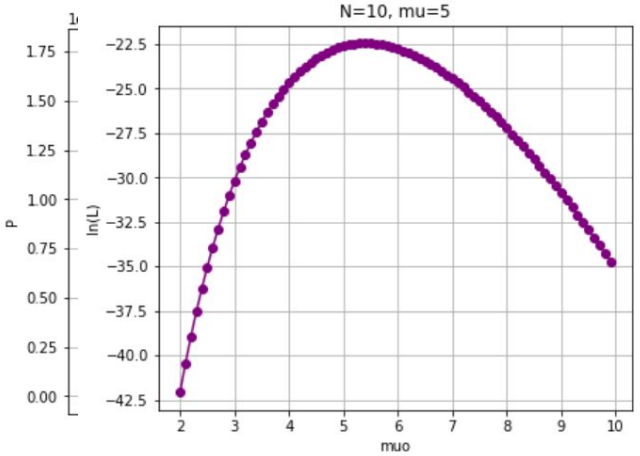
The INFN School on Underground Physics

The likelihood function

Probability of data given the parameter

Data value(s) {x1,x2,x3,...xn} are drawn from some f(x;θ):

⇒ Their joint pdf will be: $f(x_1;\theta)\cdot f(x_2;\theta)\cdots f(x_n;\theta)$ $P(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$



For example: 10 poisson distributed values around 5:

E.g. [9 5 8 4 1 5 5 4 8 5]

P(5)= $\prod_{i=1}^{10}$Poiss(x_i;5)=Poiss(9;5)·Poiss(5;5)·Poiss(8;5)·Poiss(4;5)...= 0.036·0.175·0.065·0.175… =1.52e-10

P(4)= $\prod_{i=1}^{10}$Poiss(x_i;4)=Poiss(9;4)·Poiss(5;4)·Poiss(8;4)·Poiss(4;4)...=0.013·0.156·0.029·0.195… =1.96e-11

P(6)= $\prod_{i=1}^{10}$Poiss(x_i;6)=1.3e-10

$$L(x_i \mid \theta) = \prod_{i=1}^{n} f(x_i;\theta) \qquad \ln L(x_i \mid \theta) = \sum_{i=1}^{N} \ln f(x_i;\theta) \qquad -\ln L(x_i \mid \theta) = -\sum_{i=1}^{N} \ln f(x_i;\theta) \qquad \frac{\partial \ln(L)}{\partial \theta}\Big|_{\theta=\hat{\theta}} = 0$$

Binomial	Branching ratio	Chi-square	Goodness-of-fit
Multinomial	Histogram with fixed N	Cauchy	Mass of resonance
Poisson	Number of events found	Landau	Ionization energy loss
Uniform	Monte Carlo method	Beta	Prior pdf for efficiency
Exponential	Decay time	Gamma	Sum of exponential variables
Gaussian	Measurement error	Student's t	Resolution function with adjustable tails

A Bernouli trial

Bernoulli trial - is an experiment where s trials are made of an event, with an independent probability p for success, and q=1-p for failure, in any given trial.

⇒ Each trial has two possible outcomes success/fail

⇒ The probability p of success is constant for each trial.

⇒ The probability q=1-p for failure is constant as well

⇒ Each trial is independent.

Binomial distribution

The **discrete** probability distribution to obtaining exactly n successes out of N Bernoulli trials. Each trial is true with probability p, and false with q=1-p

⇒ n is the random variable is n

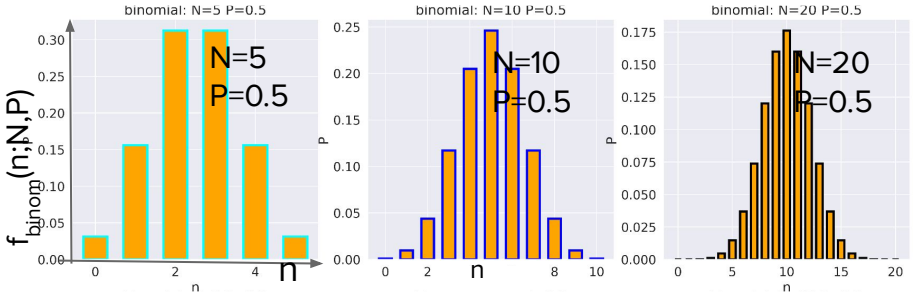
⇒ N and P are the parametres

$$f(n;N,p) = \binom{N}{n} p^n q^{N-n} = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n},$$

Number of permutations getting n out of N x Probability per one permutation

$$E[n] = \sum_{n=0}^{N} n f(n;N,p) = Np$$

$$V[n] = E[n^2] - (E[n])^2 = Np(1-p)$$



Example, Bayesian coin flip:

The likelihood:probability of observing the data, given H. H is the binomial distribution so:

$$P(X \mid H) = \frac{(h+t)!}{h!t!} p^h (1-p)^t$$

Prior: We will start with a fair coin

$$P(H \mid X) = \frac{P(X \mid H) \cdot P(H)}{P(X)}$$

- x=event of getting h heads and t tails
- p=probability for head

The Posterior!

Marginalization - probability of observing the data summed over all hypotheses (values of p)

$$P(X) = \int_0^1 dp \cdot P(H = p) \cdot \frac{(h+t)!}{h!t!} p^h (1-p)^t$$

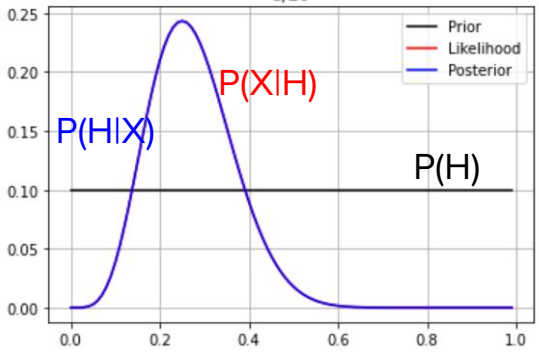
$$P(H \mid X) \propto P(X \mid H) \cdot P(H)$$

Example, Bayesian coin flip

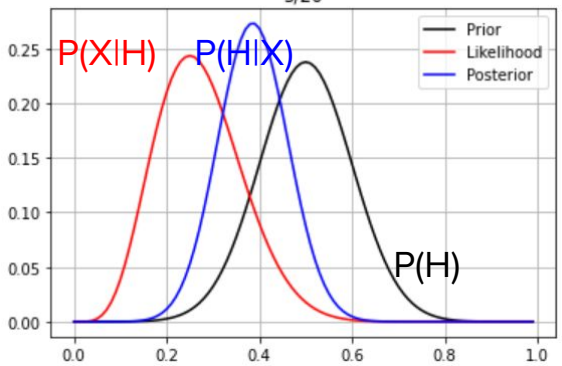
What if I threw the coin 20 times and got 5 heads…

Fair coin

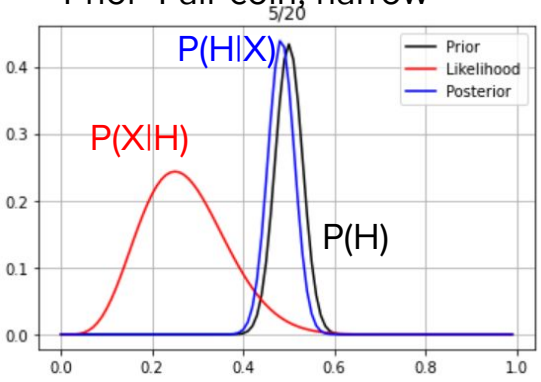
Prior =Flat Prior



PriorFair coin, wide



Prior=Fair coin, narrow



$$P(H \mid X) \propto P(X \mid H) \cdot P(H) = \frac{(h+t)!}{h!t!} p^h (1-p)^t \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(p-0.5)^2}{2\sigma^2}}$$

E.g. alternative priors:

$$P(H) = 0.9 \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(p-0.3)^2}{2\sigma^2}} + 0.1 \times 1 \qquad P(H) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(p-0.3)^2}{2\sigma^2}}$$

Poisson distribution

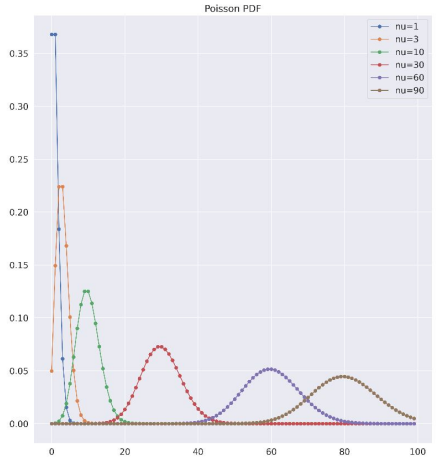
Binomial f(n;N,p) = $= \frac{N!}{n!\,(N-n)!}\, p^n\,(1-p)^{N-n},$

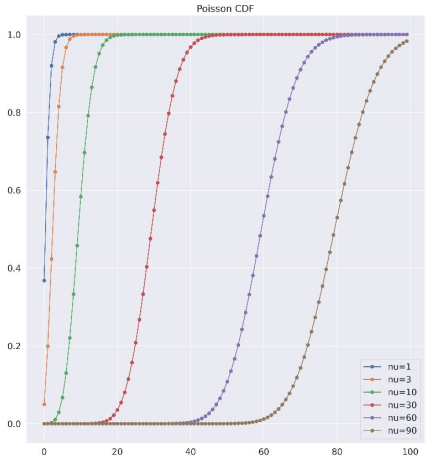
With

N➔∞ and p➔0 and use ν=Np=finite

$$f(n;\nu) = \frac{\nu^n}{n!}e^{-\nu} \quad (n \geq 0)$$

$$E[n] = \nu\,, \quad V[n] = \nu\,.$$





Discrete distribution that describes the probability of getting exactly N events in a given time, if they appear independently and randomly in a constant rate.

An example of a poisson process

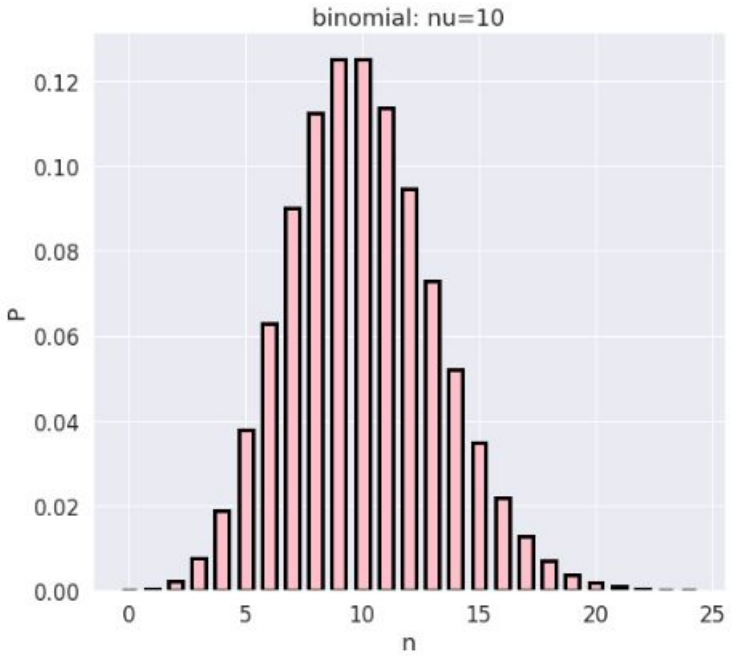
Assume that I get emails at rate of 10 messages per hour, as a poisson process

What is the probability of me getting exactly 10 emails in the next hour?

$$P(X = 10) = f_X(10) = \frac{e^{-10}\lambda^{10}}{10!} = 0.1251$$

What is the probability of me getting no emails in the next hour?

$$P(X = 0) = f_X(0) = \frac{e^{-10}\lambda^{0}}{0!} = e^{-10} = 4.54 * 10^{-5}$$



Radioactive sources : Binomial or Poissonian?

Is it binomial?

- Two outcomes: Nucleus either decays or doesn't
- Events are independent of each other.
- The mean rate (events per time period) is constant.

$$\Gamma = \frac{\ln 2}{\tau_{1/2}}$$

- Two events cannot occur at the same time.

Alas! Needs to know N ! $N = \frac{A}{\Gamma}$

Is it poissonian?

Yes, if:

- Decayed nuclei can "decay again" , so N is constant (which is nonsense)

  Or

- Number of nuclei is "large" compared to the decay rate and measuring time, so N is ~Constant.

Radioactive sources : Binomial or Poissonian?

Measuring the decays over 1 hour, for 1000 sources

Cs-137 $\tau_{1/2} \sim 30\ years$ 1uCi (3.7e4 Bq).

$$\Gamma = \frac{\ln 2}{\tau_{1/2}} = \frac{\ln 2}{30\ years} = 7.3 \cdot 10^{-10}\ \frac{1}{sec}$$

$N_0 = \frac{A}{\Gamma} = 51 \times 10^{12}$

In 1 hour: $\Delta n = 130 \times 10^6 << N_0$

Probability of a nucleon to decay over 1 second

Rb-82 $\tau_{1/2} \sim 75\ sec$ 1mCi (3.7e7 Bq).

$$\Gamma = \frac{\ln 2}{\tau_{1/2}} = \frac{\ln 2}{75\ sec} = 9 \cdot 10^{-3}\ \frac{1}{sec}$$

$N_0 = \frac{A}{\Gamma} = 4 \times 10^9$

In 1 hour: $\Delta n = 30 \times 10^9 > N_0 (!!!!!)$

Don't

1 hour ~ 50 half lives

PMT example 1
Trigger rate on a single PMT

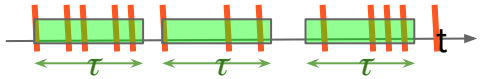
A PMT triggers at rate Γ Hz. Each trigger is τ seconds long. All triggers arriving within the time window are counted as n hits.

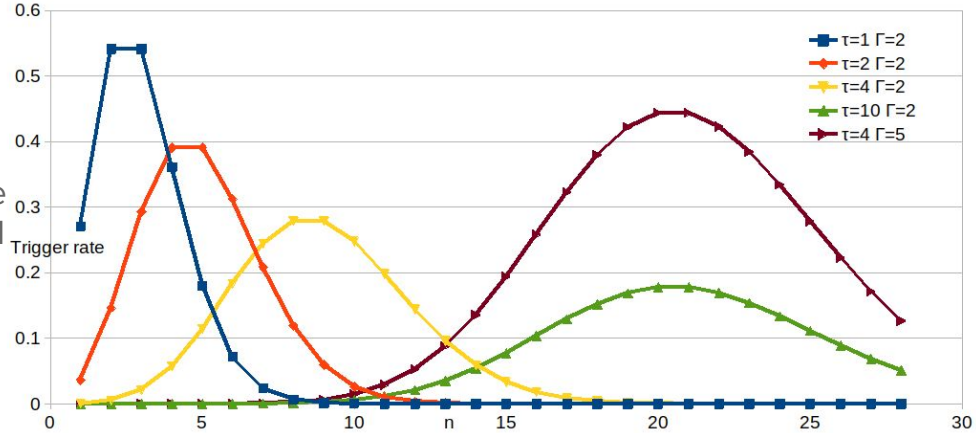
The expected number of hits in the window τ is :

$$\mu_1 = \Gamma\tau$$

The rate of it triggering n times in this time window is the rate of a single trigger times the probability of getting n-1 triggers following:

$$\begin{aligned}\Gamma_n &= \Gamma \cdot poiss(n-1 \mid \mu) \\ &= \Gamma \cdot e^{-\mu_1} \cdot \frac{\mu_1^{n-1}}{(n-1)!} \\ &= \Gamma \cdot e^{-\Gamma\tau} \cdot \frac{(\Gamma\tau)^{n-1}}{(n-1)!}\end{aligned}$$





- The longer the time window, the slower the trigger rate
- The longer the time window, the larger n
- The faster Γ the trigger rate

PMT example 2
Rate of n hits over m PMTs

In this scenario we are counting hits. If a PMT got 2 hits in the same trigger window, it will be counted as “2”

A PMT triggers at rate $\boldsymbol{\Gamma}$ Hz.

The rate of a single hit over all m PMTs is : $\Gamma_m^1 = m \cdot \Gamma$

The expected number of hits in the time window is: $\mu = \Gamma_m^1 \cdot \tau = m \cdot \Gamma \cdot \tau$

The rate in which total of n hits (not PMTs) are observed over all PMTs is:

$$\begin{aligned}\Gamma_n^m &= \Gamma_1^m \cdot poiss(n-1 \,|\, \mu) \\ &= m \cdot \Gamma \cdot e^{-\mu} \frac{\mu^{n-1}}{(n-1)!} \\ &= m \cdot \Gamma \cdot e^{-m\Gamma\tau} \frac{(m\Gamma\tau)^{n-1}}{(n-1)!}\end{aligned}$$

This scenario (using m PMTs with rate $\boldsymbol{\Gamma}$) is **identical** to the previous (using 1 PMT with rate m$\boldsymbol{\Gamma}$)

So this is like example 1, but with a trigger rate m times higher.

PMT example 3
Rate of m PMTs over n PMTs

In this scenario we are counting PMTs... If a PMT got 2 hits, it will be counted as “1”....

A PMT triggers at rate Γ Hz.

The probability that a PMT didn’t trigger is: $p_0 = poiss(0 \mid \mu) = e^{-\mu}$

The probability of **exactly** m PMTs triggering is the probability that m triggered, and (n-m) did not:

$$p_n = \frac{m!}{(m-n)! \cdot n!}(1-p_0)^n \cdot p_0^{m-n}$$

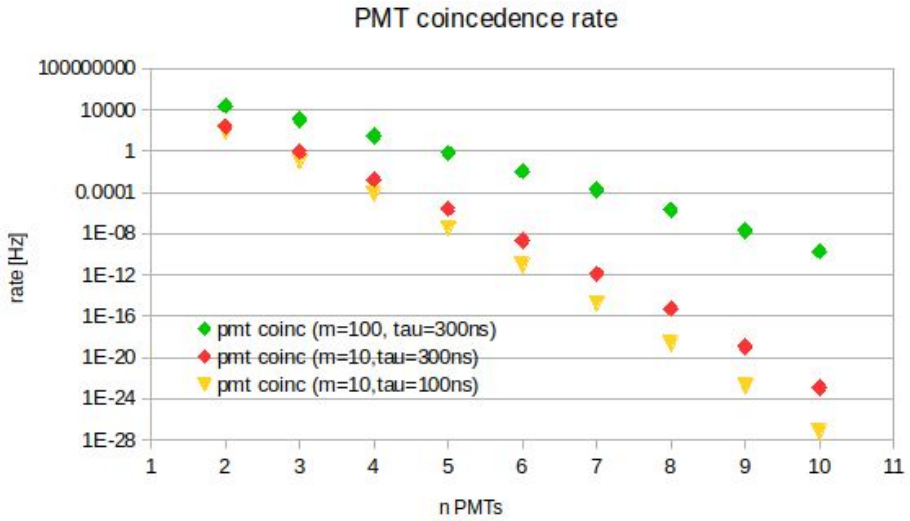
So...using $\Gamma_m^1 = m \cdot \Gamma$ which is the rate of a single hit over m PMTs is

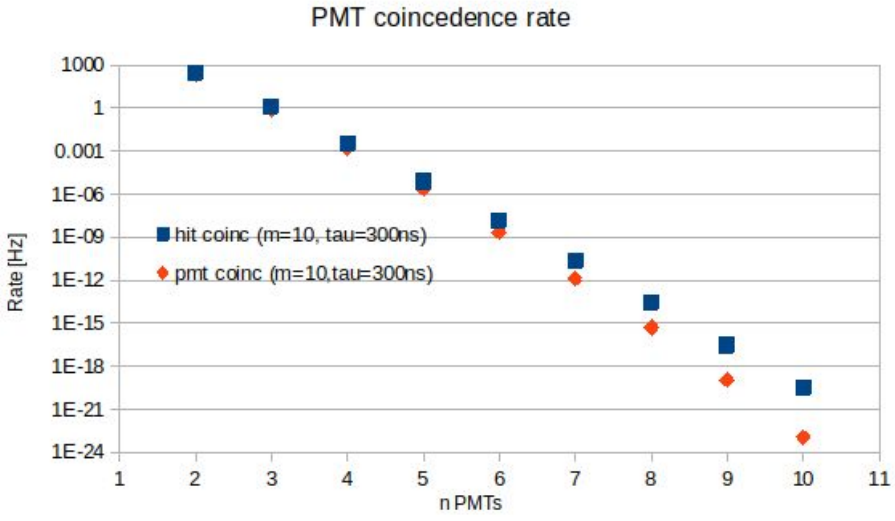
$$\Gamma_n^m = \Gamma_1^m \cdot p_{n-1} = m \cdot \Gamma \cdot \frac{(m-1)!}{(m-n)! \cdot (n-1)!} \cdot (1-p_0)^{n-1} \cdot p_0^{m-n} =$$

$$= m \cdot \Gamma \cdot \frac{(m-1)!}{(m-n)! \cdot (n-1)!}\left(e^{+\Gamma\tau} - 1\right)^{n-1} \cdot e^{-\Gamma\tau(m-1)}$$

How coincedence lowers the trigger ratre

Γ=3 kHz





Time between triggers in a Poisson process

In a poisson process events occurs on average at rate λ per unit time

In average there will be λt occurrences per time t.

$$p(x;\lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}$$

The probability of observing no events in time t is

$$P(0) = e^{-\lambda t}$$

This is the cumulative distribution, we will differentiate by t and get the PDF:

$$f(t) = \lambda e^{-\lambda t}$$

Time difference is distributed exponentially

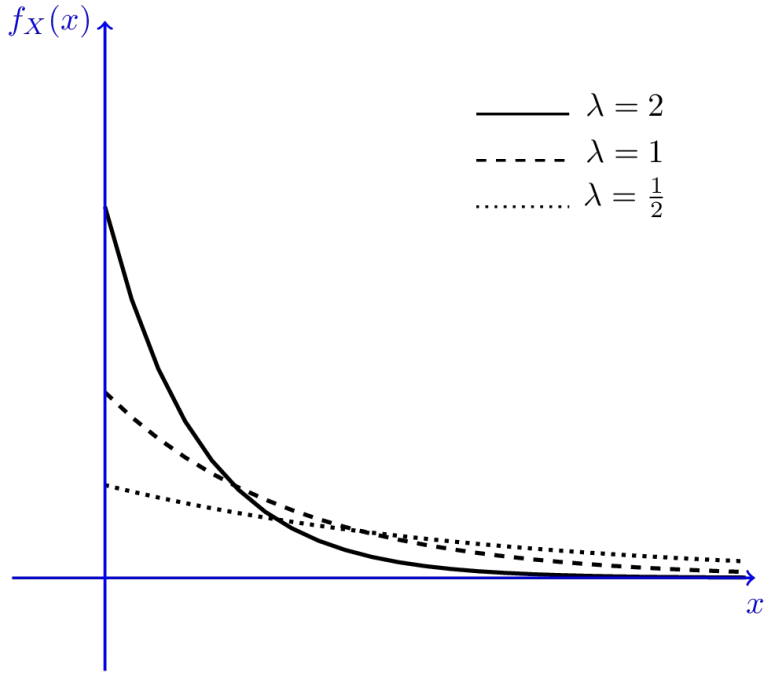
The exponential distribution

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

$$F(t) = \begin{cases} 1 - e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

$$E(x) = \frac{1}{\lambda}$$

$$V(x) = \frac{1}{\lambda^2}$$

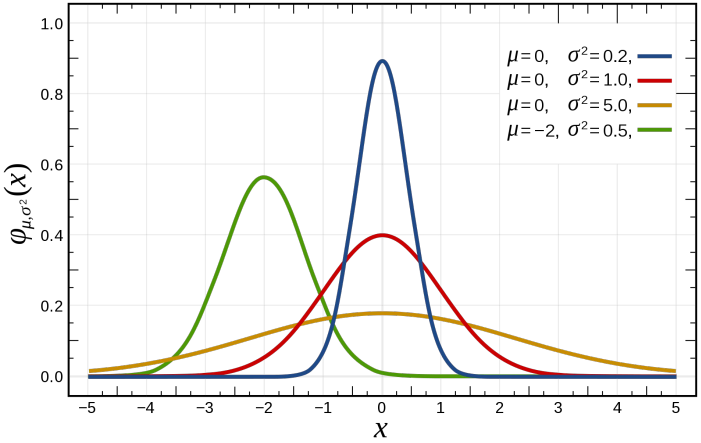


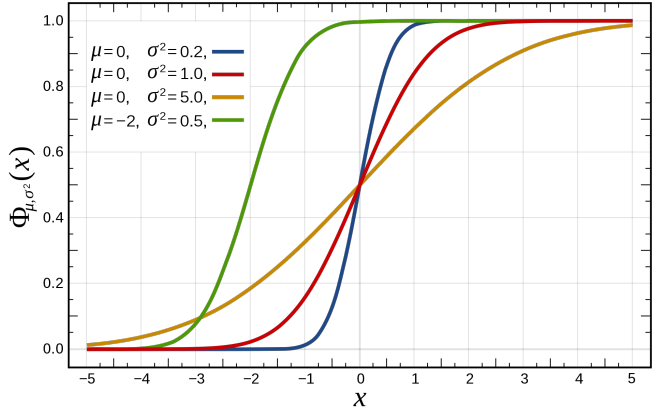
Memoryless - The past has no bearing on its future behaviour ➔ Waiting time paradox"

$$P(x > s + t \,|\, x > s) = P(x > t) = e^{-\lambda t}$$

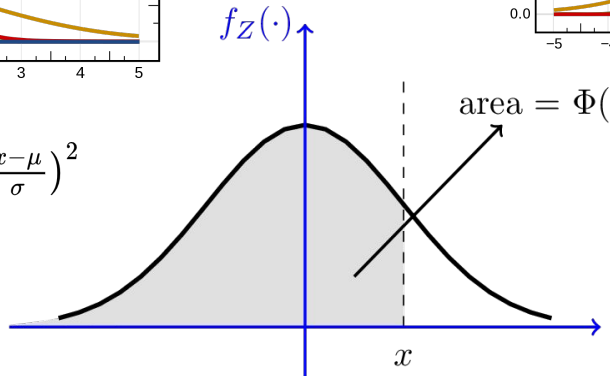
Gaussian distribution

For large λ, Poisson ➔ Gaussian with μ=λ and σ=sqrt(λ)





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

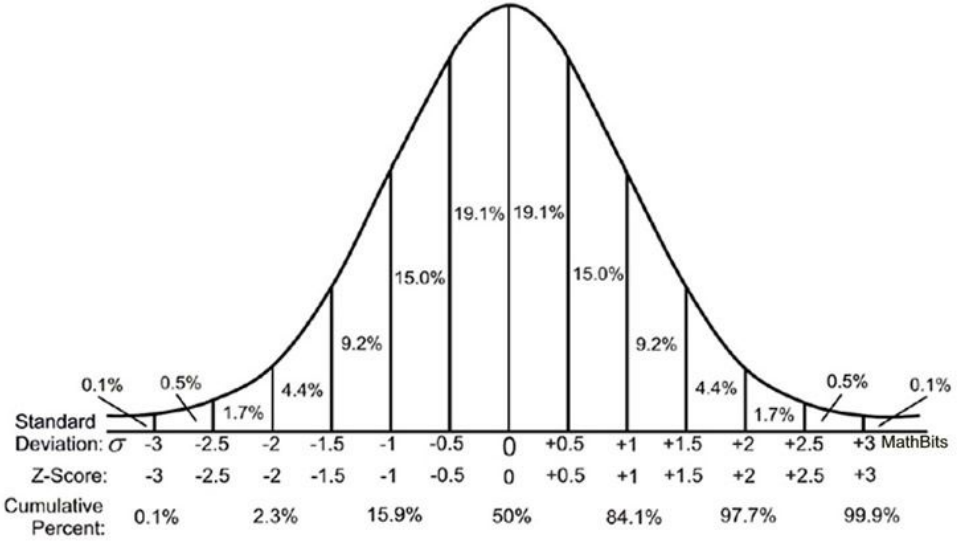


The Z score

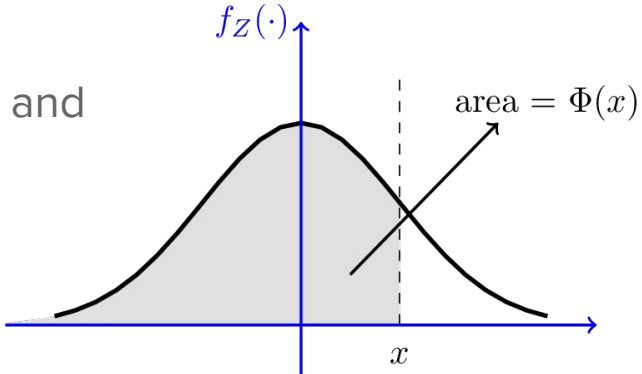
$$Z = \frac{x - \mu}{\sigma}$$

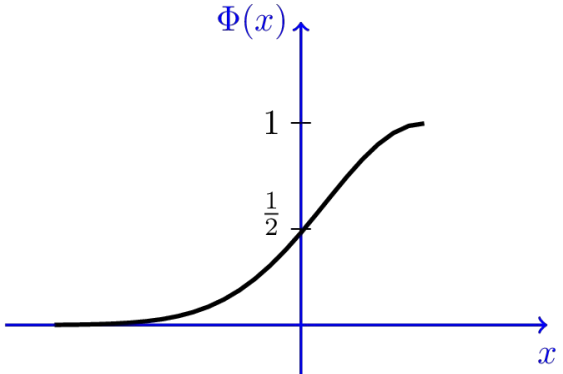
And on the cumulative distribution:

- ϕ(z)= P(Z< z)
- 1-ϕ(z)= P(Z> z)



Random variable can be standardized so its mean is 0, and standard deviation=1 so:





The Z table

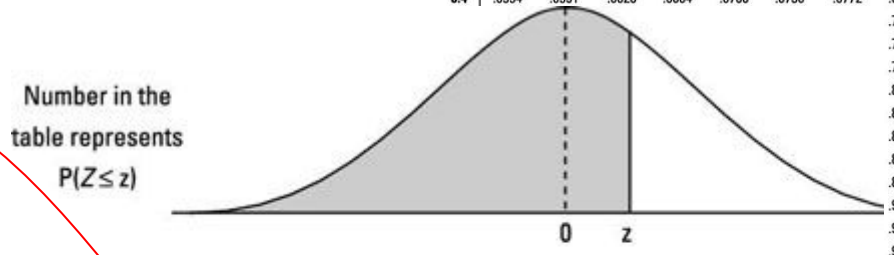
⇒ P(z<0) =0.5 = ϕ(0.5)

⇒ P(z<0.44)=ϕ(0.44)=0.67

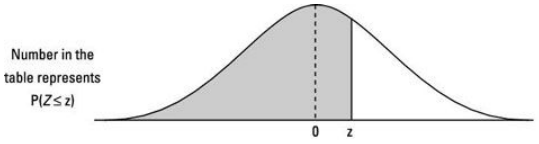
⇒ p(z>0.44)=1-ϕ(0.44)= 1-0.67=0.33

⇒ p(z<-0.44)=ϕ(-0.44)

=-(0.67-0.5)+0.5=0.33



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794



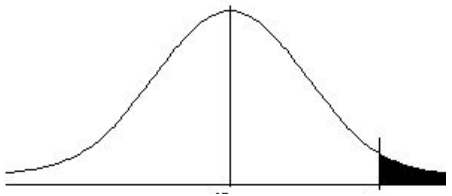
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
								.7157	.7190	.7224
								.7486	.7517	.7549
								.7794	.7823	.7852
								.8078	.8106	.8133
								.8340	.8365	.8389
								.8577	.8599	.8621
								.8790	.8810	.8830
								.8980	.8997	.9015
								.9147	.9162	.9177
								.9292	.9306	.9319
								.9418	.9429	.9441
								.9525	.9535	.9545
								.9616	.9625	.9633
								.9693	.9699	.9706
								.9756	.9761	.9767
								.9808	.9812	.9817
								.9850	.9854	.9857
								.9884	.9887	.9890
								.9911	.9913	.9916
								.9932	.9934	.9936
								.9949	.9951	.9952
								.9962	.9963	.9964
								.9972	.9973	.9974
								.9979	.9980	.9981
								.9985	.9986	.9986
								.9989	.9990	.9990
								.9992	.9993	.9993
								.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

Example

A company manufactures Al foils with a mean thickness of 100µm and a standard deviation of 10µm. What is the approximate probability that a foil will be more than 120µm thick?

$$Z = \frac{120 - 100}{10} = 2$$

$$\begin{aligned} p(z > 2) &= 1 - \phi(2) = \\ &= 1 - 0.977 \\ &= 0.023 \\ &= 2.3\% \end{aligned}$$



What is the probability a sample of 25 foils will have an average thickness of more than 95µm?

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

$$\begin{aligned} \phi\left(\frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right) &= \phi\left(\frac{95 - 100}{2}\right) \\ &= \phi(-2.5) \\ &= 1 - \phi(2.5) \\ &= 1 - 0.9338 \\ &= 0.0062 \end{aligned}$$

Central limit theorem (CLT)

The sum of n independent random variables xi, each taken from a distribution with finite variance σ_i^2 and Mean value μ_i

Then the sum y :

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

Has an expectation value of:

$$E[y] = \sum_{i=1}^{n} \mu_i$$

Has a variance of:

$$V[y] = \sum_{i=1}^{n} \sigma_i^2$$

When n➔∞, the random variable X becomes a gaussian

In the case of repeated measurements:

All μ_i have the same value " μ" and:

E[y]=n μ

All σ_i have the same value "σ" and:

V[y]=nσ2

And the averages are:

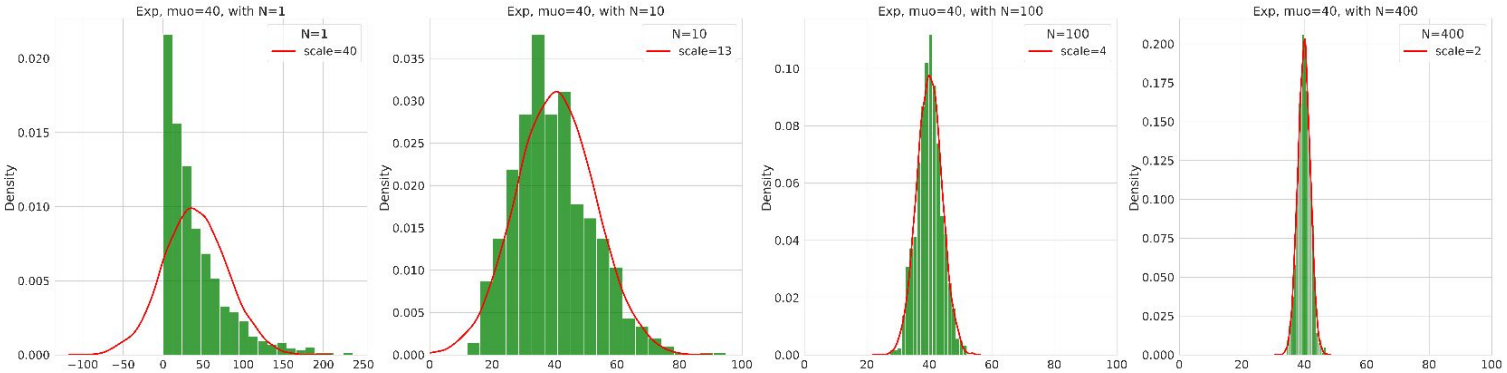
$$E[\bar{y}] = \frac{y}{n}$$

$$V[\bar{y}] = Var\left(\frac{y}{n}\right) = \frac{1}{n^2} Var(y) = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}$$

$$\sigma[\bar{y}] = \frac{\sigma}{\sqrt{n}}$$

Central limit theorem (CLT)

The sample mean will approximately be normally distributed for large sample sizes, regardless of the distribution from which we are sampling.



Bonus exercise : DIY - 👆👆👆

A few words on Monte Carlo

Numerically generating “random” data sets, usually on a computer

What we need for simulation?

1. Model(s)
2. Random number (but not too random)
3. Sampler

⇒ Using random numbers? Make sure to seed correctly...

How to generate an arbitrary PDF from a flat distribution?

Numerous methods exist.

I will briefly review three:

- Simulate the physical process
- Inverse transform method
- Rejection method

Monte Carlo sampling by “physicsal processes”

Example: Poisson process example

The goal: Write a Poissonian sampler, using only flat random number generator. Go into the “core” - simulate “decays”

⇒ The probability is given as input Γ.

⇒ Loop over time in some stamps.

⇒ Draw a flat distributed random number,

- If it is greater than the given time constant do nothing.
- If it is lighter than the given time constant count an event.

⇒ Choose your units and scale carefully.

MC sampling using Inverse Transformation

Example: generating an exponential distribution from flat

$$f(z)=\frac{1}{\tau}\exp(-z/\tau)$$

$$F(x)=\int_0^x \frac{dz}{\tau}\exp(-z/\tau)=\int_0^{x/\tau} dy\, e^{-y}=1-\exp(-x/\tau)$$

This method is efficient, fast, accurate

But not always feasible…

1. If Y is a random number between 0-1, we'll set:

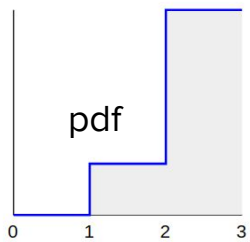
   $$Y = F(x) = 1 - e^{-x/\tau}$$

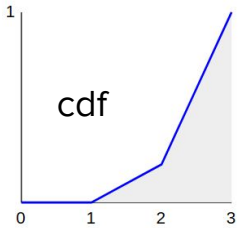
2. And solve for x:

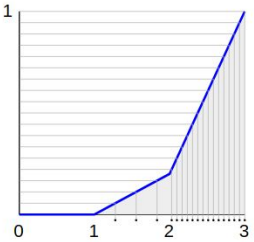
   $$x=-\tau \ln(1-Y)$$

3. But since both Y and 1-Y are uniform we can also use:

   $$x=-\tau \ln Y$$







Now, if we put "flat Y" we will get exponential x with the correct coverage

Illustrations from https://programming.guide/generate-random-value-with-distribution.html

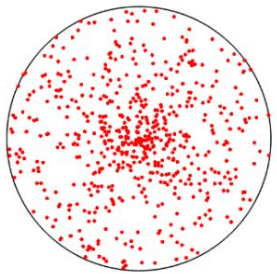
MC sampling using rejection method

Example: unified sampling from polar coordinates

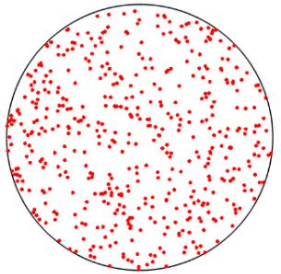
Draw random numbers inside a circle with radius R starting from flat distributions and the following techniques:

1. Rejection method: draw random x and y in [−R,R] Keep only points inside the circle.
2. Using variable transformation (random r in [0,R] a random θ in [0,2π won't work]

Another useful and powerful sampling method is MCMC (Markov chain Monte Carlo) - especially important in bayesian inference when a lot of sampling is needed to account for a wide hypothesis space.



Incorrect: points cluster around the center.



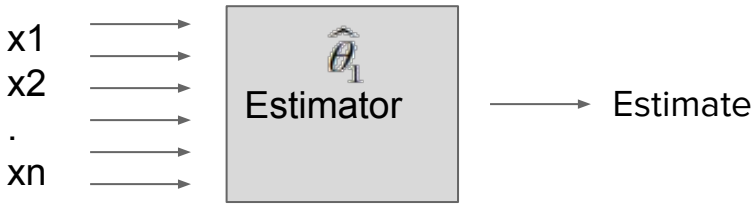
Correct: points are evenly spread out.

Part 3: Estimation

Estimation

Data value(s) (random variables) {x1,x2,x3,...xn} are drawn from some probability density function f(x;θ).

- The PDF is characterized by parameter(s) θ
- The estimator for θ, will provide an estimate for the parameter θ.
- The data values will be different each time. The estimator will remain the same but the estimate will change.

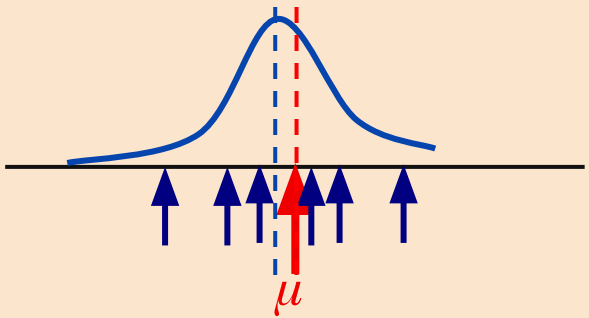


A model describes data with a **mean μ**
$\Rightarrow$ μ is a fixed unknown parameter

We estimate the mean, by calculating the **average** ("mean of the set") of our data:

$$\bar{x_1} = \frac{\sum_1^n x_i}{n}$$

A different set of measurements will give us a different average...etc...



Estimators properties

- **Consistent**

$$\lim_{n\to\infty} P\left(|\hat{\theta} - \theta| > \epsilon\right) = 0$$ (for any ε>0)

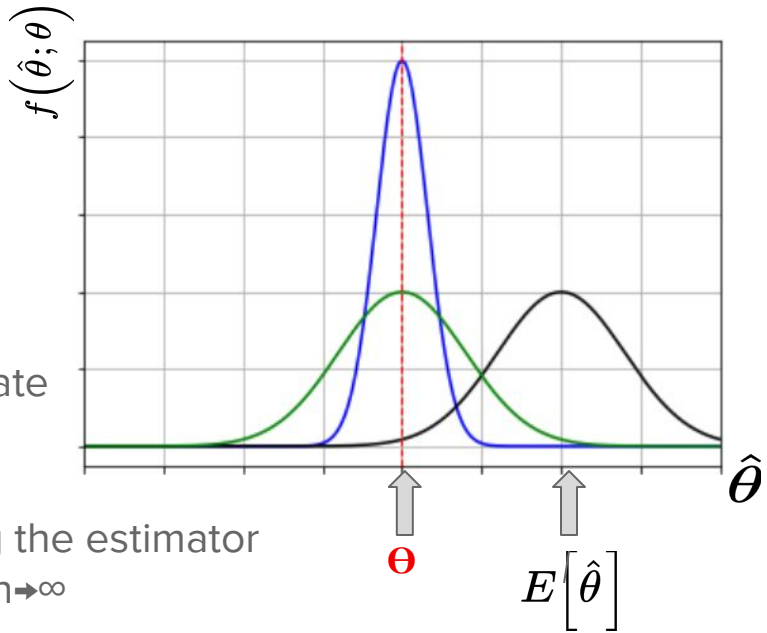
- **None-biased**

The actual value = expectation value of the estimate

$$b = E\left[\hat{\theta}\right] - \theta$$

⇒ If bias is found it can be corrected for, by tuning the estimator

⇒ For a consistent estimator a bias will vanish as n➔∞



- **Efficient**

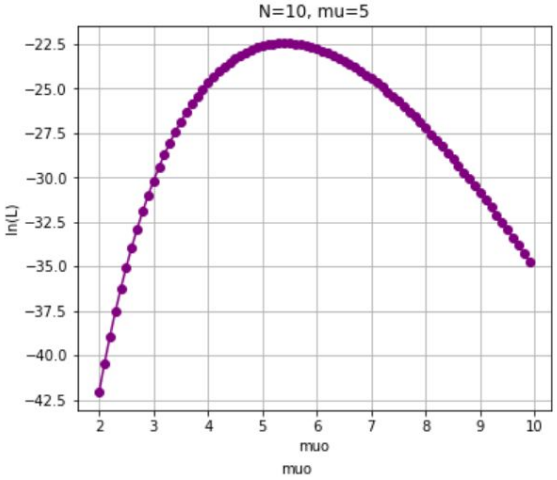
Minimum variance. An estimator is said to be efficient if its variance is at a minimum value called “Minimum Variance Bound”

$$\mathrm{var}(\hat{\theta} \mid \theta) = \mathbb{E}[\left(\hat{\theta} - \mathbb{E}[\hat{\theta} \mid \theta]\right)^2 \mid \theta]$$

Also exist: robustness, cimplicity...

The likelihood function

Probability of data given a parameter (model)



Data value(s) {x1,x2,x3,...xn} are drawn from some f(x;θ):

⇒ Their joint pdf will be: f(x_1;θ)·f(x_2;θ)...·f(x_n;θ) $P(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$

For example: 10 poisson distributed values around 5:

E.g. [9 5 8 4 1 5 5 4 8 5]

P(5)= $\prod_{i=1}^{10}$Poiss(x_i;5)=Poiss(9;5)·Poiss(5;5)·Poiss(8;5)·Poiss(4;5)...= 0.036·0.175·0.065·0.175… =1.52e-10

P(4)= $\prod_{i=1}^{10}$Poiss(x_i;4)=Poiss(9;4)·Poiss(5;4)·Poiss(8;4)·Poiss(4;4)...=0.013·0.156·0.029·0.195… =1.96e-11

P(6)= $\prod_{i=1}^{10}$Poiss(x_i;6)=1.3e-10

$$L(x_i \mid \theta) = \prod_{i=1}^{n} f(x_i;\theta) \qquad \ln L(x_i \mid \theta) = \sum_{i=1}^{N} \ln f(x_i;\theta) \qquad -\ln L(x_i \mid \theta) = -\sum_{i=1}^{N} \ln f(x_i;\theta) \qquad \frac{\partial \ln(L)}{\partial \theta}\Big|_{\theta=\hat{\theta}} = 0$$

Maximum Likelihood Estimator

Maximum likelihood (ML) estimator for θ

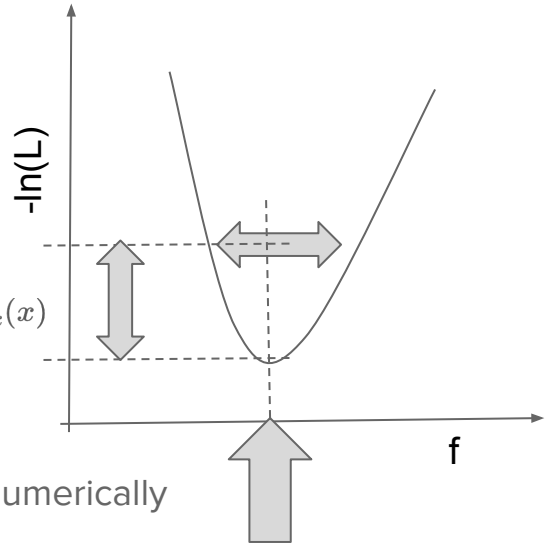
$$-\ln L(x_i \mid \theta) = -\sum_{i=1}^{N} \ln f(x_i; \theta)$$

The game plan:

1. Take N measurement of random variable x
2. Hypothesize a model e.g.
3. Write the log likelihood

$$-\ln L(x \mid f) = -\sum_{i=1}^{N} \ln \left(P(x \mid f)\right)$$

$$P(x) = f \cdot P_{signal}(x) + (1-f) \cdot P_{bck}(x)$$



4. Minimize -ln(L) w.r.t. Your parameter (f) - analytically if possible. Or numerically

The most probable value of the likelihood is the maximum likelihood estimator

The spread around the minimum is usually the measure of the accuracy

Maximum Likelihood Estimator

- Maximum likelihood estimators are usually consistent
- Maximum likelihood estimators are usually biased, but it gets better as N➔infinity
- In the asymptotic limit, the estimator is efficient. *

$$V\left(\hat{\theta}\right) \geq \frac{-1}{\frac{d^2 \ln L}{d\theta^2}} \quad \Longrightarrow \quad \sigma^2_{\hat{\theta}} = V\left(\hat{\theta}\right) = \frac{-1}{\frac{d^2 \ln L}{d\theta^2}}$$

- Maximum likelihood estimators are invariant under parameter transformation

* *This statement is based on a term called "Minimum Variant Bound" (or MVB). It says there is a "best case estimator" which gives smallest RMS value when averaged over thousands of expereriments.*

$$V\left(\hat{\theta}\right) \geq \frac{-\left(1 + \frac{db}{d\theta}\right)^2}{\frac{d^2 \ln L}{d\theta^2}}$$

What about the uncertainty?

4 options

⇒ **Option 1:** In some cases, this can be calculated analytically a'la $V\left[\hat{\theta}\right] = E\left[\hat{\theta}^2\right] - \left(E\left[\hat{\theta}\right]\right)^2$

⇒ **Option 2:** Monte Carlo: Simulate "many" experiments with similar sample size, collect the expectation values in an histogram, and estimate the variance.

Asymptotic normality - ML estimators, for large sample limit, the distribution will be apx gaussian.

Gentle practical points

- What values to use ("truth" or "expectation")?
- Simulate all identical sample sizes (e.g. N=50) or do N=Poisson(50)?
- What happens when parameter is near its limit (i.e. positive only).

If we take a taylor series expansion about the ML estimator θ

What about the uncertainty? (cont)

4 options

⇒ **Option 3:** "Information inequality" or "Rao Cramer Frechet inequality" (Holds not only for ML estimators). For efficient, unbiased estimators

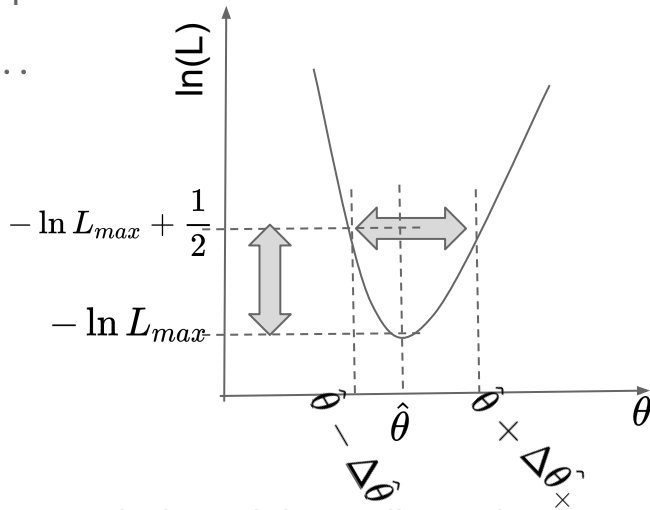
$$\hat{\sigma}^2{}_{\theta=\theta_0} = \left(-\frac{\partial^2 \ln L}{\partial \theta^2} \right)^{-1}_{\theta=\theta_0}$$

⇒ **Option 4:** The graphical method: If we take a taylor series expansion about the ML estimator θ

$$\ln L(\theta) = \underbrace{\ln L\left(\hat{\theta}\right)}_{=\ln L_{max}} + \underbrace{\frac{\partial \ln L}{\partial \theta}\bigg|_{\theta=\hat{\theta}} \left(\theta - \hat{\theta}\right)}_{=0} + \underbrace{\frac{1}{2!}\frac{\partial^2 \ln L}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}} \left(\theta - \hat{\theta}\right)^2}_{= -\frac{1}{2}\frac{1}{\hat{\sigma}^2_{\hat{\theta}}}\left(\theta - \hat{\theta}\right)^2} + \ldots$$

$$\ln L\left(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}\right) = \ln L_{\max} - \frac{1}{2}\frac{\left(\hat{\sigma}_{\hat{\theta}}\right)^2}{\hat{\sigma}_{\hat{\theta}}{}^2} = \ln L_{\max()} - \frac{1}{2}$$

$$\ln L\left(\hat{\theta} \pm 2\hat{\sigma}_{\hat{\theta}}\right) = \ln L_{\max} - \frac{1}{2}\frac{\left(2\hat{\sigma}_{\hat{\theta}}\right)^2}{\hat{\sigma}_{\hat{\theta}}{}^2} = \ln L_{\max()} - 2$$



$\Delta\theta+$ and $\Delta\theta+$ are not necessarily equal. With good statistics this will become parabola and they will equal.

For example...

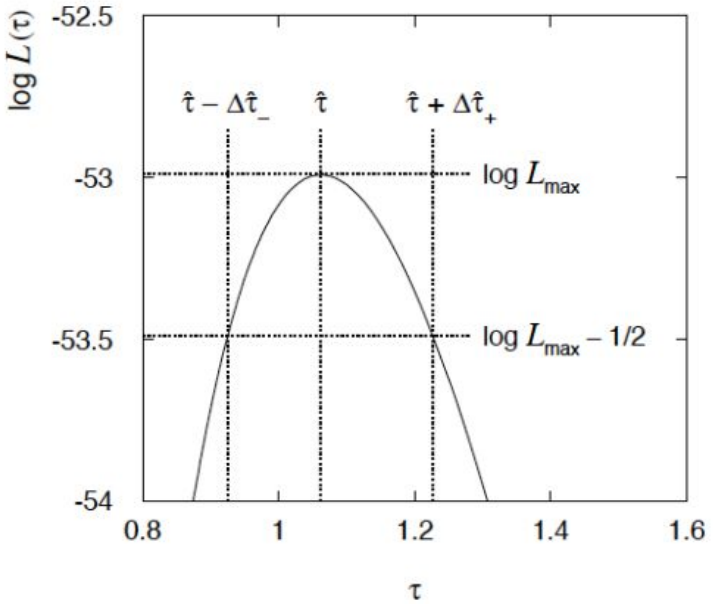
Not quite parabolic ln L since finite sample size (n = 50).

$$\hat{\tau} = 1.062$$

$$\Delta\hat{\tau}_- = 0.137$$

$$\Delta\hat{\tau}_+ = 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$$



Unbinned extended likelihood

So far we looked at “shapes” only: $L\left(\vec{\theta}\right) = \prod_{i=1}^{n} f\left(x_i; \vec{\theta}\right)$

We can include also the expected number of events by including a Poisson term

$$L\left(N; \vec{\theta}\right) = Poiss(n \mid N) \prod_{i=1}^{n} f\left(x_i \mid \vec{\theta}\right) = \frac{N^n e^{-N}}{n!} \prod_{i=1}^{n} f\left(x_i \mid \vec{\theta}\right) = \frac{e^{-N}}{n!} \prod_{i=1}^{n} N \cdot f\left(x_i \mid \vec{\theta}\right)$$

Let’s log it:

$$\ln(L) = -N - \ln(n!) + \sum_{i=1}^{n} \ln N + \sum_{i=1}^{n} \ln f\left(\vec{x_i} \mid \theta\right) = -N + n \cdot \ln N + \sum_{i=1}^{n} \ln f\left(\vec{x_i} \mid \theta\right)$$

To find the minimum of -ln(L):

$$-\frac{\partial \ln L}{\partial N} = 1 - \frac{n}{N} = 0 \implies n = N \qquad -\frac{\partial \ln L}{\partial \theta_1} = 0\,, \; -\frac{\partial \ln L}{\partial \theta_2} = 0$$

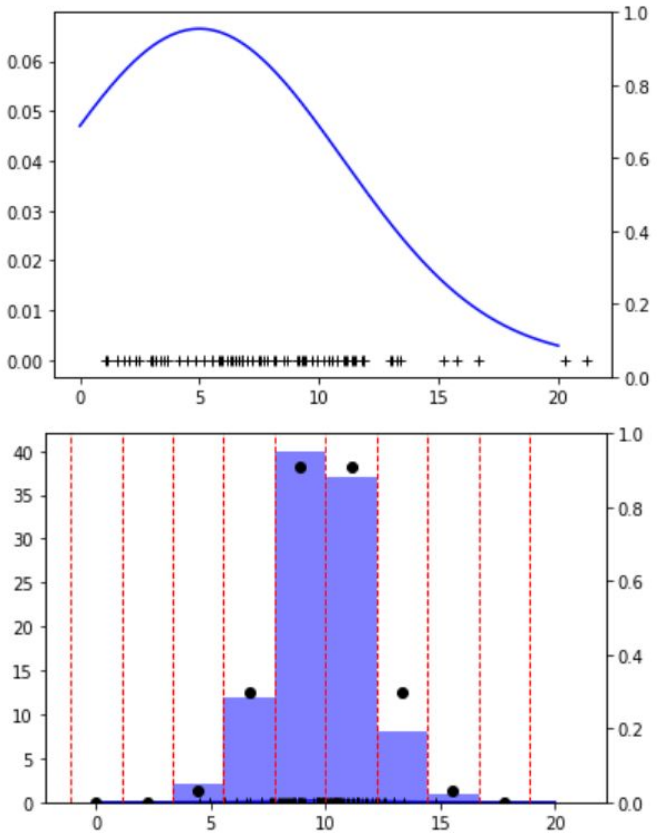
Unbinned extended likelihood

Same idea, only binned...

- Switching from n events, to N bins so:

  $n=n_1+n_2+n_3+...+n_N$

- The number of expected events in each bin will be integrated for the data, and for the models



Some other useful extensions to the likelihood

Model combination $f(x_i \mid \bar{\theta}) = \sum_{components} \theta_j f_j(x)$ where $\sum_{components} \theta_i = 1$ e..g:

$$L = L(x_i \mid N, \mu_s, \sigma_s, \mu_b, \sigma_b, \alpha_s, \alpha_b, \alpha_r, \Gamma) = Poiss(n \mid N) \cdot \prod_{i=1}^{n} (\alpha_s gaus(\mu_s, \sigma_s) + \alpha_b gaus(\mu_b, \sigma_b) + \overbrace{1 - \alpha_s - \alpha_b}^{=} \alpha_r \exp(\Gamma))$$

Additional data sets

For example include auxiliary calibration data, constraining some of the used parameters. e.g:

$$L = L(x_i, x_j \mid N, \bar{\theta}, N_{cal}, \bar{\theta_{cal}}, \bar{\theta^{\star}}) = \left[Poiss(n \mid N) \cdot \prod_{i=1}^{n} f_{data}(x_i \mid \theta, \bar{\theta^{\star}}) \right] \cdot \left[Poiss(n_{cal} \mid N_{cal}) \cdot \prod_{j=1}^{n_{cal}} f_{cal}(x_j \mid \bar{\theta_{cal}}, \bar{\theta^{\star}}) \right]$$

Parameter constraints

Add distributions with (e.g) gaussian constraint to the likelihood function. E.g:

$$gaus(\mu_s \mid \mu_{model}, \sigma)$$

😊 Mix & Match

Bad news! More parameters means...

No analytical solution

For each parameter we set the derivative to zero. The equations become more and more complex.

Chances to solve analytically are low.

Solution: program it.

However: Minimizers only partly useful.

Larger uncertainties

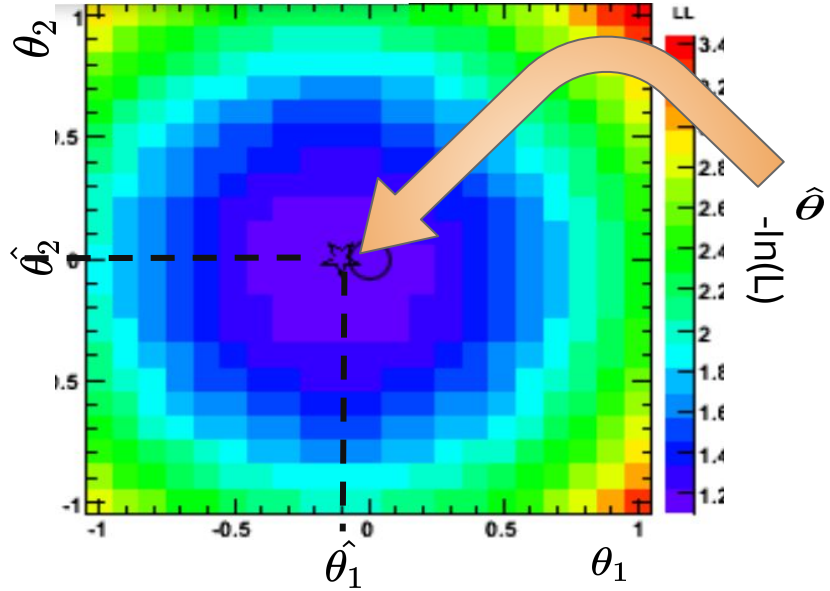
The likelihood curve will become wider and wider

(don’t believe me? Simulate it and check)

$$L\left(\underbrace{x_i, x_j;}_{\text{Data points}} \underbrace{N, \alpha_s, \alpha_b, \bar{\theta}, \bar{\theta}_{aux}, \bar{\theta}_{constra}}_{\text{parameters}} \ldots\right)$$

(list of (s1,s2))

Use the data-set(s) and find the minimum of this



$$L\left(x_i, x_j; N, \alpha_s, \alpha_b, \bar{\theta}, \bar{\theta}_{aux}, \bar{\theta}_{constra} \dots\right)$$

"μ"
Parameter
of interest*

"θ"
Nuisance parameters

profile likelihood

* To be more precise, the parameter of interest is not N_{signal}, but rather σ... but we will discuss this issue later

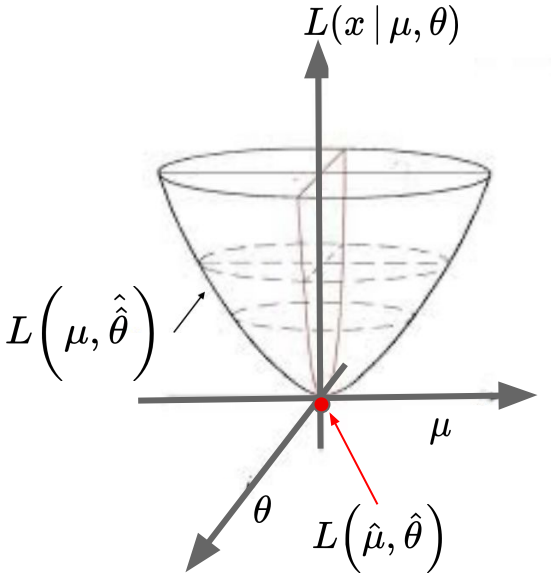
Profile likelihood

μ - Parameter of interest
θ - Nuisance parameter
x - data

$$\lambda(\bar{x} \mid \mu) = \frac{L\left(\bar{x} \mid \mu, \hat{\hat{\theta}}\right)}{L\left(\bar{x} \mid \hat{\mu}, \hat{\theta}\right)}$$

Value of nuisance parameter that maximizes L for a specific μ

Maximized value of L. (unconditional)

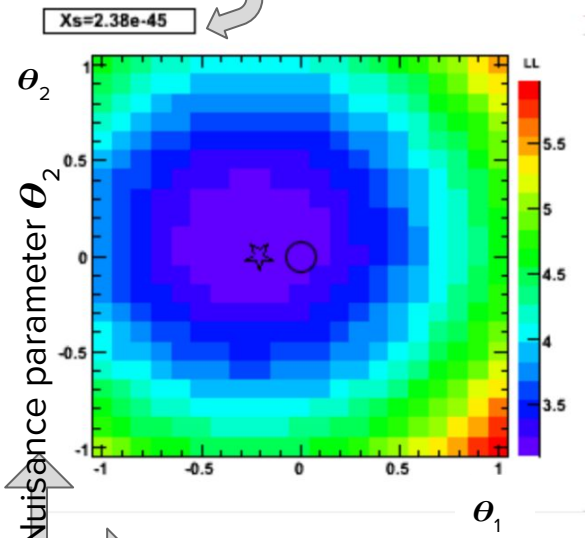


Profiling in action

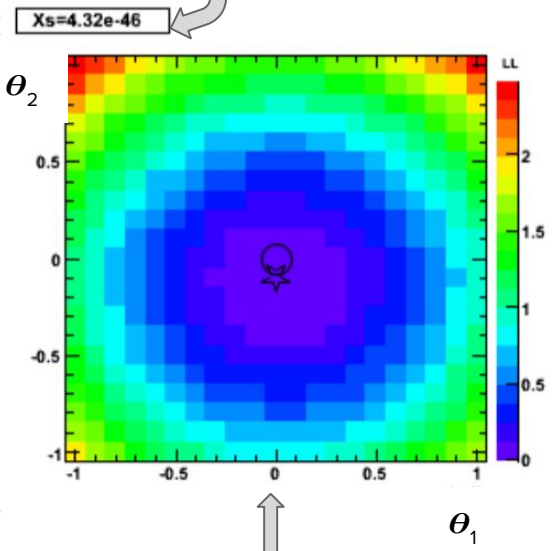
☆ = $\left(\hat{\theta}_1, \hat{\theta}_2\right)$ Minimum for a specific μ

◯ = (0,0) Nominal

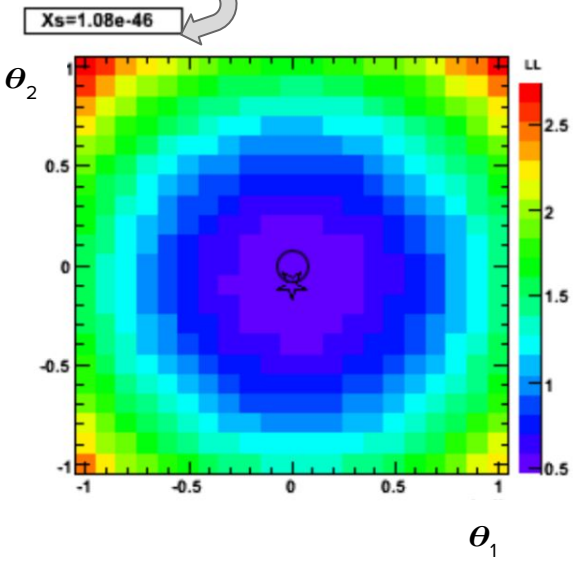
Param of interest = val1



Param of interest = val2



Param of interest = val2



Nuisance parameter θ_2

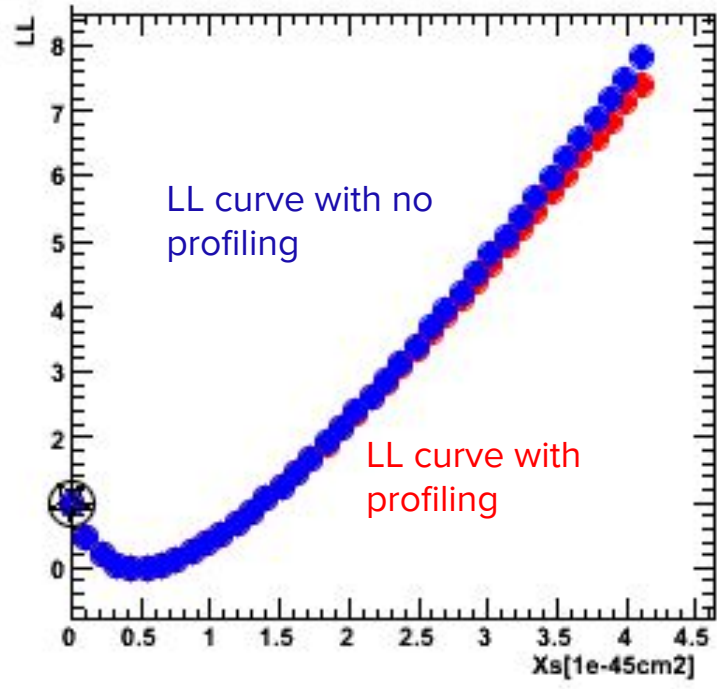
Nuisance parameter θ_1
In σ units around nominal

Nominal value

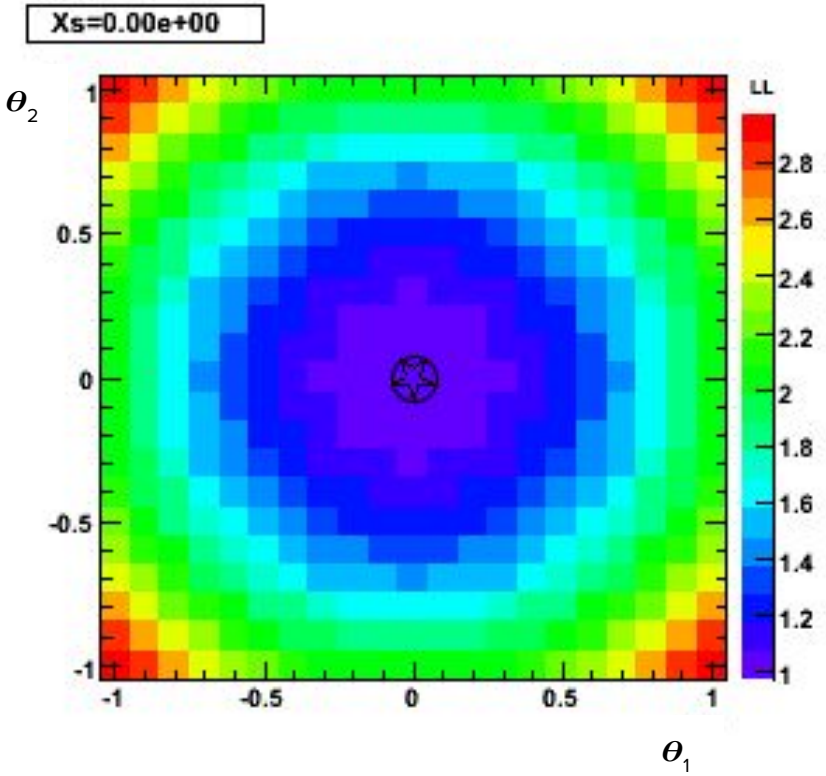
Profiling in live action

☆ = $\left(\hat{\theta}_1, \hat{\theta}_2\right)$ Minimum for a specific μ

◯ = (0,0) Nominal



Param of interest



Nuisance parameters

Profile likelihood
Wilks' theorem

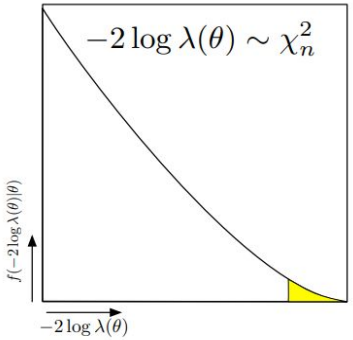
μ - Parameter of interest
θ - Nuisance parameter

$$\lambda(\bar{x} \mid \mu) = \frac{L\left(\bar{x} \mid \mu, \hat{\hat{\theta}}\right)}{L\left(\bar{x} \mid \hat{\mu}, \hat{\theta}\right)}$$

The distribution of q evaluated at μ approaches a chi-square pdf asymptotically

$$q_\mu = -2 \ln \lambda(\bar{x} \mid \mu)$$

$$f(q_\mu(\mu)) = f(-2 \ln \lambda(\mu \mid \mu))$$



What does it mean?

- We pick a model
- We generate MC experiments using this model
- For each experiment we check the value of the test statistics under the same assumption it was generated with, and add this value to a histogram
- We will get a chi2 distribution

Why is this important

- For our "real" measurement, we can now estimate q for a given model.
- From the value of q we can estimate probabilities assuming that the model we tested is correct

$$P_{stb} = \int_{q_r}^{\infty} f(q_r|r)\,dq \overset{r=r'}{=} 1 - \phi(\sqrt{q_r}) = 0.1 \rightsquigarrow \boxed{\phi(\sqrt{q_r}) = 0.9 = 1-\alpha}$$

90%: $\sqrt{q_r} = \phi^{-1}(0.9) = 1.28$

TMath: NormQuantile(0.9)

95%: $\sqrt{q_r}$ $\phi^{-1}(0.95) = 1.64$

To estimate the sensitivity we use $\hat{\mu} = \mu'$ for the median, and $\hat{\mu} = \mu' \pm N\sigma$ for the bands

* with No CLs correction :

like before:

$$P_{stb} = \alpha = 1 - \phi(\sqrt{q_r}) = 1 - \phi\left(\frac{\mu - \hat{\mu}}{\sigma}\right)$$

$$q_r = \left(\frac{\mu - \hat{\mu}}{\sigma}\right)^2 \quad r > \hat{r}$$

$$1 - \alpha = \phi\left(\frac{\mu - \hat{\mu}}{\sigma}\right)$$

$$\phi^{-1}(1-\alpha) = \frac{\mu - \hat{\mu}}{\sigma}$$

$$\mu = \hat{\mu} + \sigma \cdot \phi^{-1}(1-\alpha)$$

For median we will replace : $\hat{\mu} = \mu'$

$$\mu_{median} = \mu' + \sigma \cdot \phi^{-1}(1-\alpha)$$

For band we replace: $\hat{\mu} = \mu' + N\sigma$

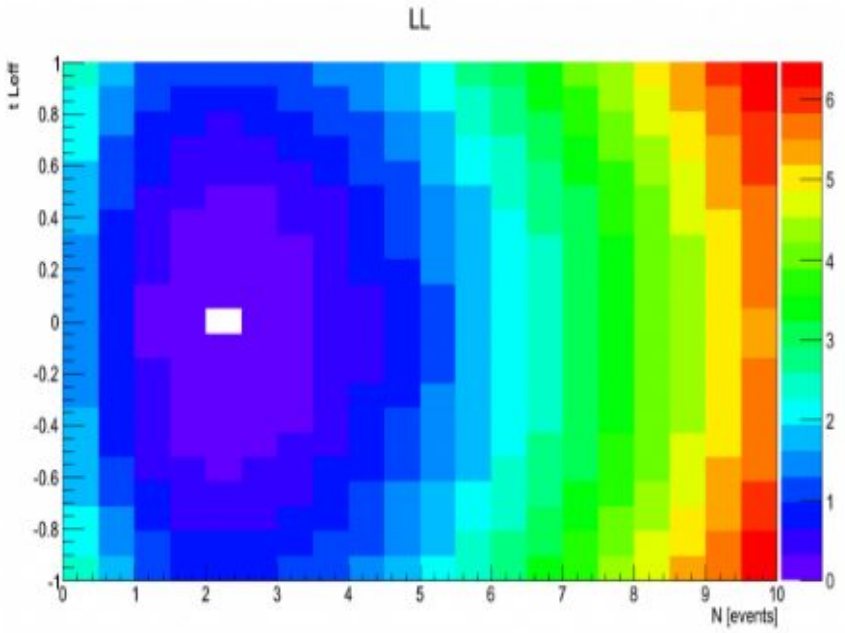
$$\boxed{\mu_\alpha = \mu' + \sigma\left(N + \phi^{-1}(1-\alpha)\right)} \longrightarrow$$

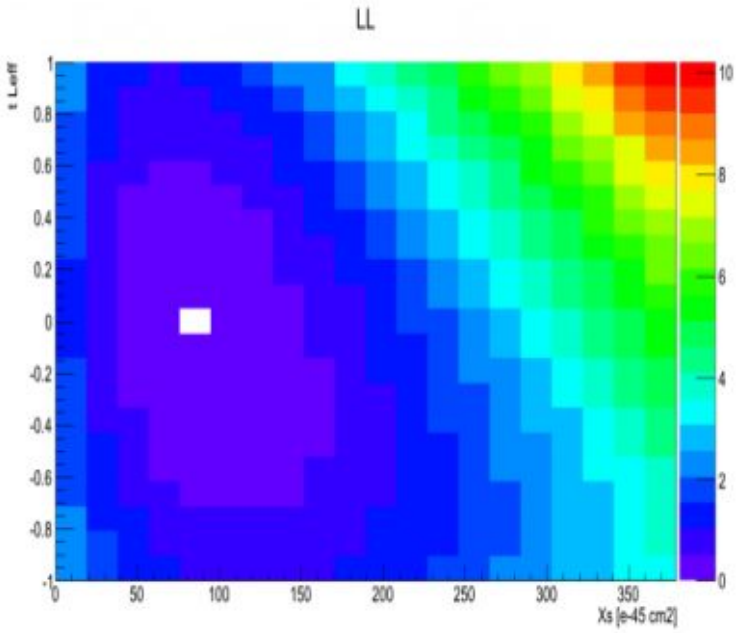
No CLs

for $\alpha = 0.1$ (90% cl)

$n = -2 \rightarrow \mu_{-2} = (-0.78)\,\sigma$

$n = -1 \rightarrow \mu_{-1} = (0.28)\,\sigma$

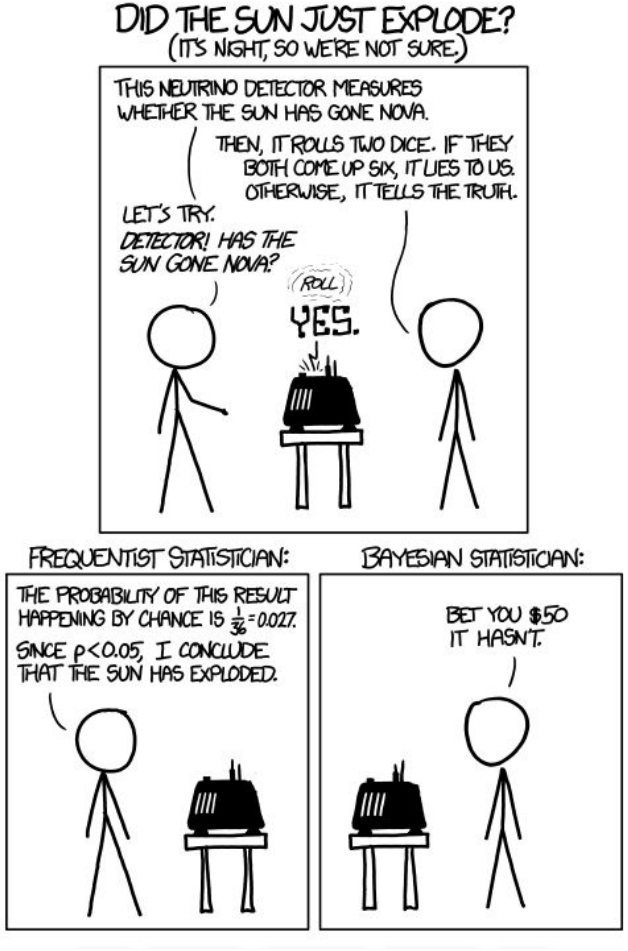




Similar information

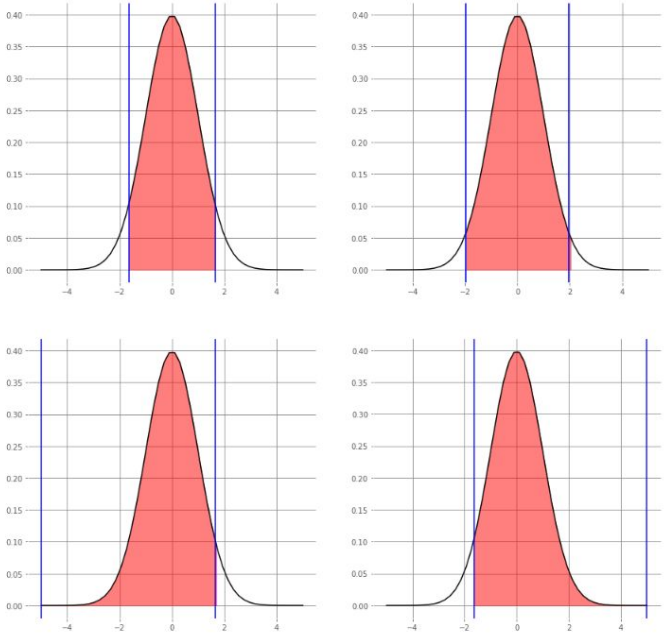
Comic relief

https://xkcd.com/1132/



Part 4: Inference

Significant intervals



Experiment ➔ Random variable ➔ Estimator ➔ Result

How confident are we about this result?

How close is the “real” theta to our theta estimation?

- 1-α is called confidence level (0-1 or 0-100%)
- [a,b] is the confidence interval

the probability that the “true” value of parameter θ is in the interval [a,b] is greater than 1-α

$$\hat{m} \pm \sigma_m \Longrightarrow [\hat{m} - \sigma_m, \hat{m} + \sigma_m]$$
$$\hat{\Gamma} \pm \sigma_\Gamma \Longrightarrow \left[\hat{\Gamma} - \sigma_\Gamma, \hat{\Gamma} + \sigma_\Gamma\right]$$
$$\hat{\tau} \pm \sigma_\tau \Longrightarrow [\hat{\tau} - \sigma_\tau, \hat{m} + \sigma_\tau]$$

For a given α i we can choose different regions

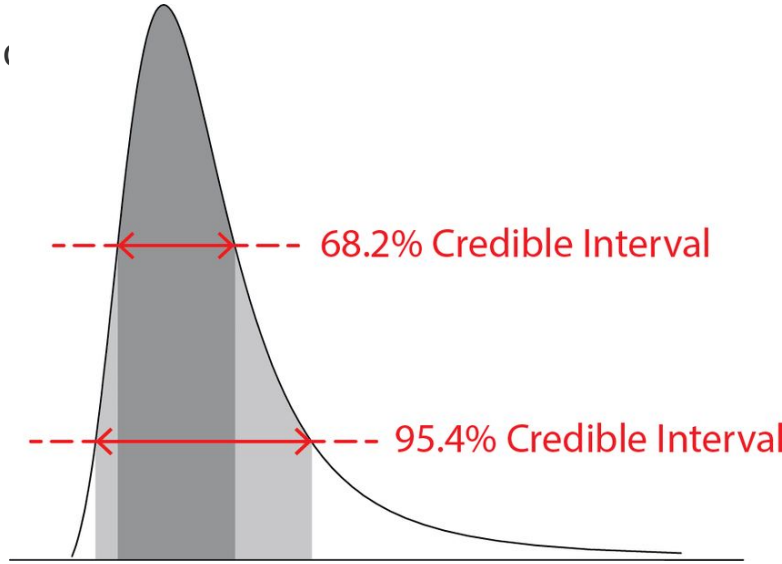
Bayesian credible intervals

$$p(\mu_t|x) = \frac{\mathcal{L}(x|\mu_t)\ p(\mu_t)}{p(x)}$$

A Bayesian interval [μ1, μ2] with confidence level α is

$$\int_{\mu_1}^{\mu_2} p(\mu_t|x)\ d\mu_t = \alpha$$

Easily obtained from the posterior pdf



Frequentist confidence intervals

What does it mean?

Experiment ➔ Random variable ➔ Estimator ➔ Result

$$\hat{m} \pm \sigma_m \Longrightarrow [\hat{m} - \sigma_m, \hat{m} + \sigma_m]$$

$$\hat{\Gamma} \pm \sigma_\Gamma \Longrightarrow \left[\hat{\Gamma} - \sigma_\Gamma, \hat{\Gamma} + \sigma_\Gamma\right]$$

$$\hat{\tau} \pm \sigma_\tau \Longrightarrow [\hat{\tau} - \sigma_\tau, \hat{m} + \sigma_\tau]$$

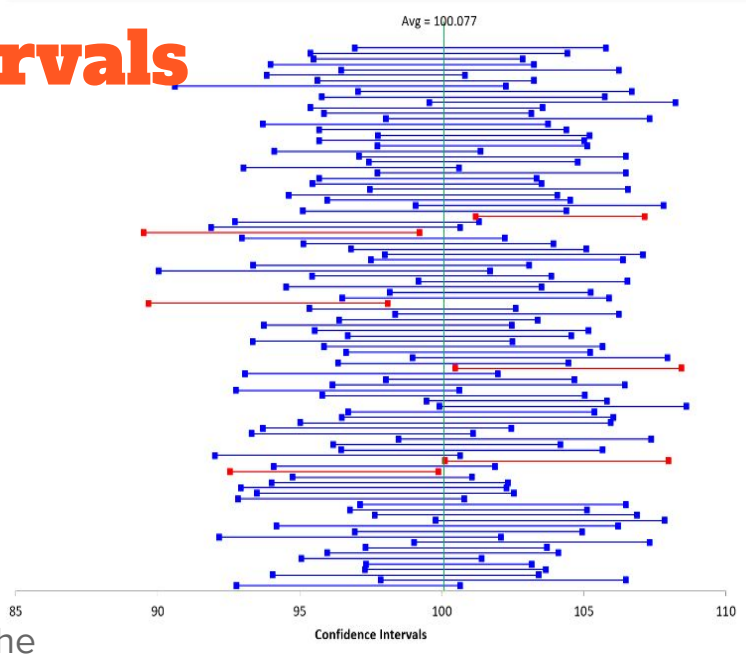
What does it mean?

Repeating our experiment many times....

✅ 68% of the resulting τ̂± σ intervals include the true value **τ** of the parameter

❌ In 68% of the experiments the true value is the τ̂± σ range

❌ There is 68% probability that the true value is in the τ̂± σ range



100 simulated experiments with 5% confidence intervals.

Frequentist Confidence interval

A confidence interval [μ1, μ2] is a member of a set, such that the set has the property that $P(\mu \in [\mu_1, \mu_2]) = \alpha$

⇒ Ensemble of experiments with a fixed unknown μ

⇒ μ1 and μ2 depends on the measured x

⇒Intervals will contain the unknown true μ in fraction α of experiments

⇒ If it holds for every allowed μ the intervals cover μ with α confidence (“correct coverage”)

⇒ If for any value of μ for which P($\mu \in [\mu 1, \mu 2]$) < α then the intervals undercover this μ

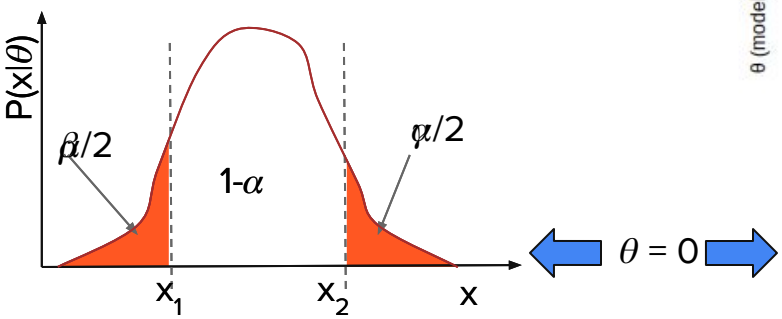
⇒ If for any value of μ for which P($\mu \in [\mu 1, \mu 2]$) > α then the intervals overcover this μ

⇒ Conservative intervals - only overcover. The price : loss of power in rejecting false hypothesis

- Frequentist: [μ1, μ2] contains the fixed, unknown μt in a fraction α of hypothetical experiments
- Bayesian: the degree of belief that μt is in [μ1, μ2] is α
- These views can correspond, but they don’t have to

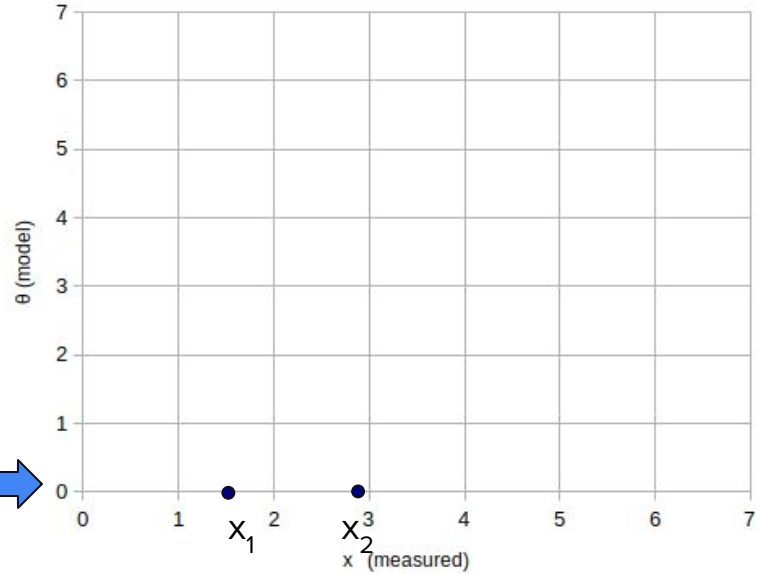
Neyman construction (of "Neyman Belt")

Method of constructing confidence intervals with the desired level of coverage



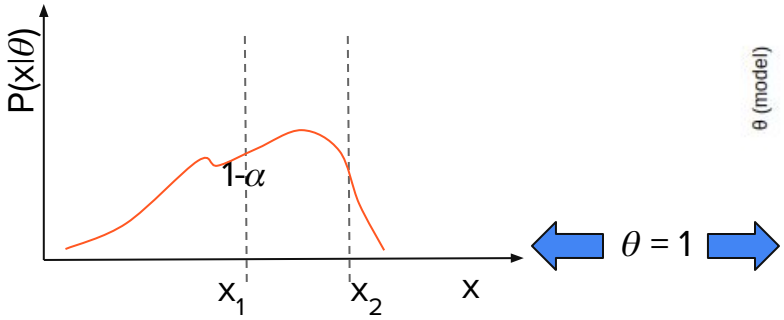
$$P(x < x_1 \,|\, \theta) + P(x > x_2 \,|\, \theta) = \alpha$$

$\theta = 0$

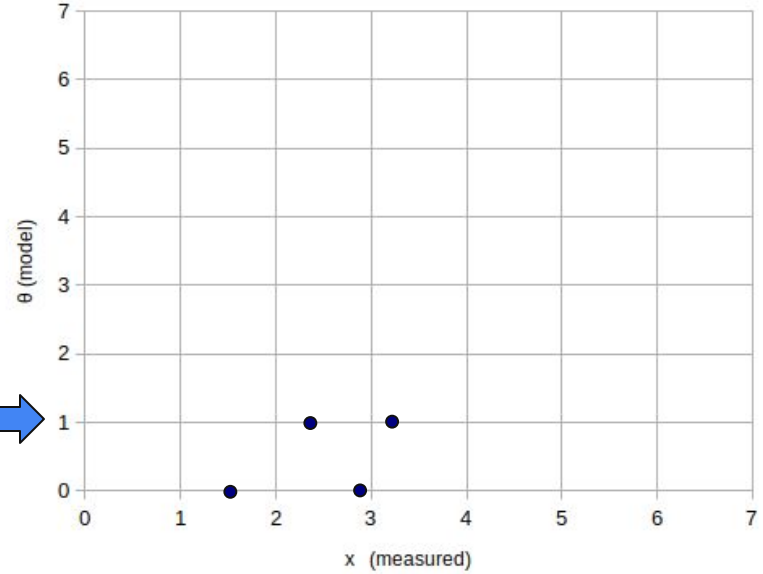


Neyman construction (of “Neyman Belt”)

Method of constructing confidence intervals with the desired level of coverage



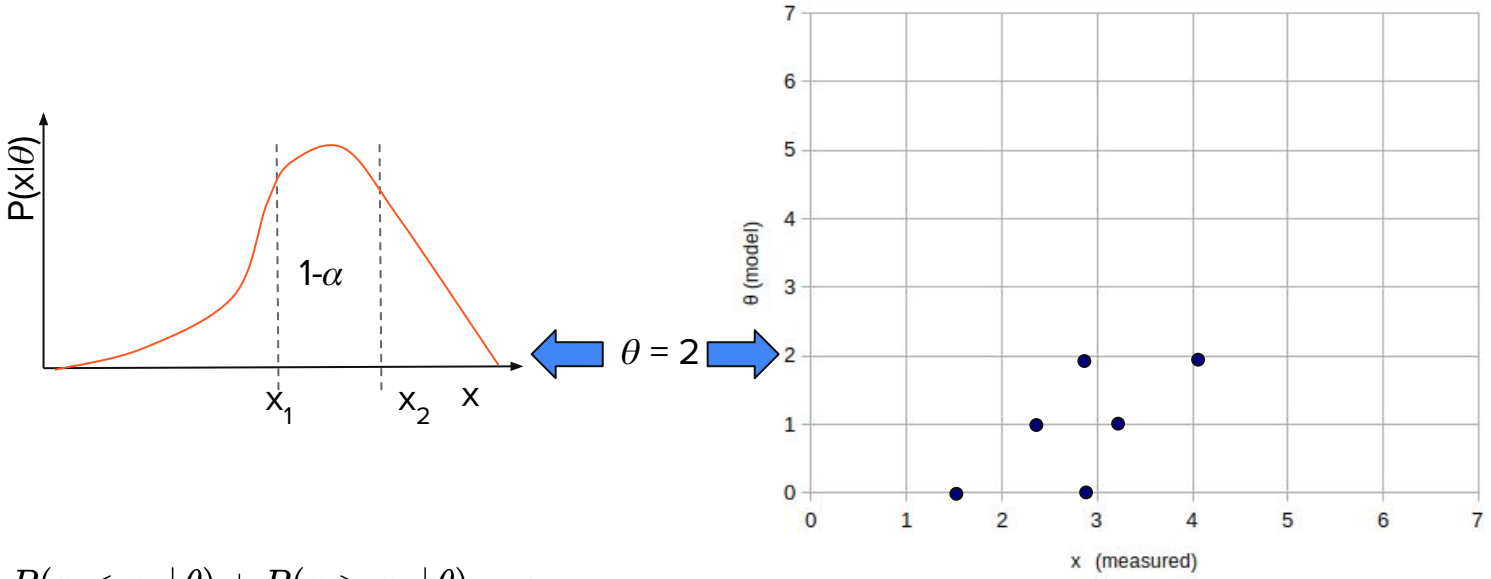
$\theta = 1$



$$P(x < x_1 \,|\, \theta) + P(x > x_2 \,|\, \theta) = \alpha$$

Neyman construction (of "Neyman Belt")

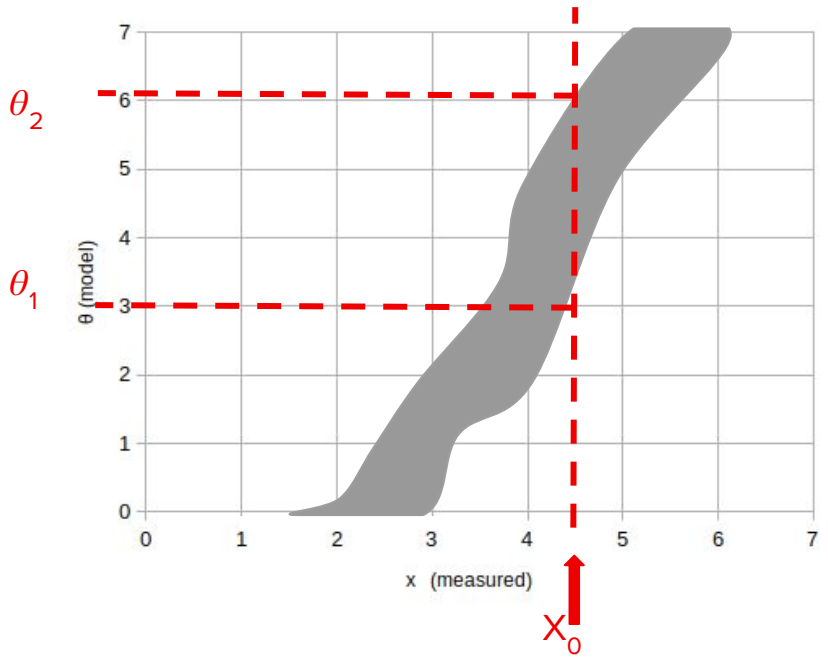
Method of constructing confidence intervals with the desired level of coverage



$$P(x < x_1 \,|\, \theta) + P(x > x_2 \,|\, \theta) = \alpha$$

Neyman construction (of “Neyman Belt”)

Method of constructing confidence intervals with the desired level of coverage



Example for Neyman construction

Constructing a 90% two sided interval for a normal gaussian distributed random variable:

⇒ α=1-0.9=0.1. We'll take 0.05 from each side.

⇒ For μ=0 the relevant PDF is normal:

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

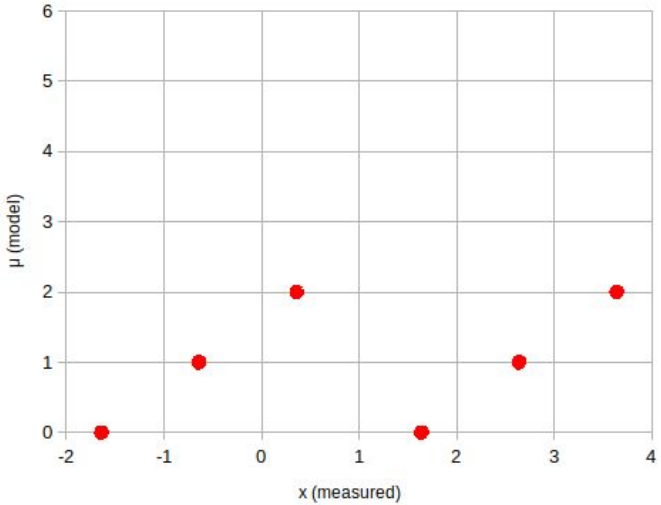
- X2=Z_0.95 = scipy.stats.norm.ppf(.95)= 1.64
- X1=Z_0.05= - 1.64

⇒ For μ=1 the relevant PDF shifts by 1:

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

X1=-0.64, X2=2.64

etc...



Example for Neyman construction

Constructing a 90% two sided interval for a normal gaussian distributed random variable with know width, and unknown mean $\mu \geq 0$:

⇒ α=1-0.9=0.1. We'll take 0.05 from each side.

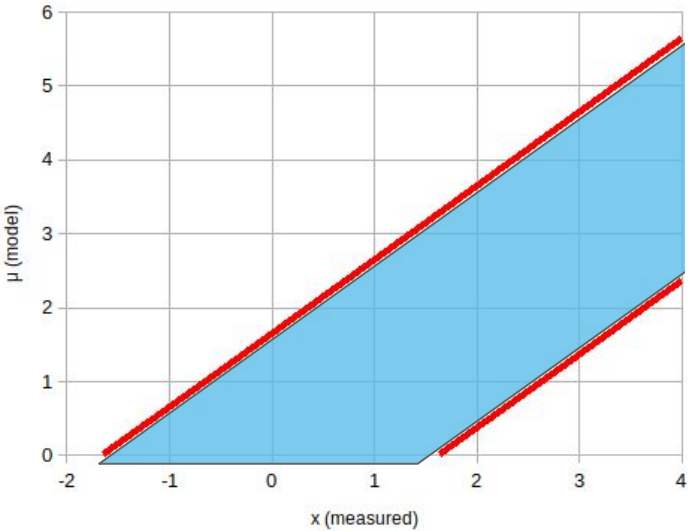
⇒ For μ=0 the relevant PDF is normal: $f(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}$

- X2(Z=0.95) = scipy.stats.norm.ppf(.95)= 1.64
- X1(Z=0.05)= - 1.64

⇒ For μ=1 the relevant PDF shifts by 1: $f(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-1}{\sigma}\right)^2}$

X1=-0.64, X2=2.64

etc...



Example for Neyman construction

Constructing a 90% upper limit interval for a normal gaussian distributed random variable:

⇒ α=1-0.9=0.1. We'll take 0.1 from each side.

⇒ For μ=0 the relevant PDF is normal:

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

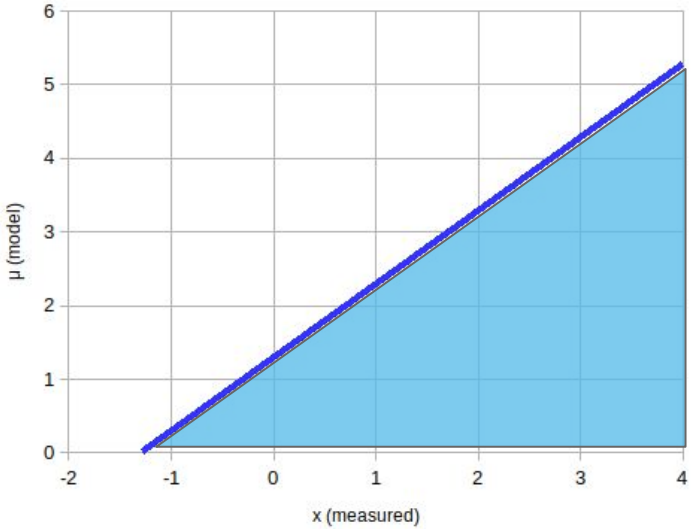
- X1(Z=0.1) = scipy.stats.norm.ppf(.1)= -1.28

⇒ For μ=1 the relevant PDF shifts by 1:

X1=-0.28,

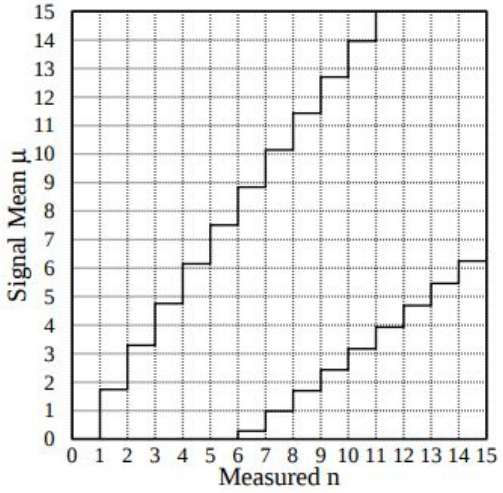
$$f(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-1}{\sigma}\right)^2}$$

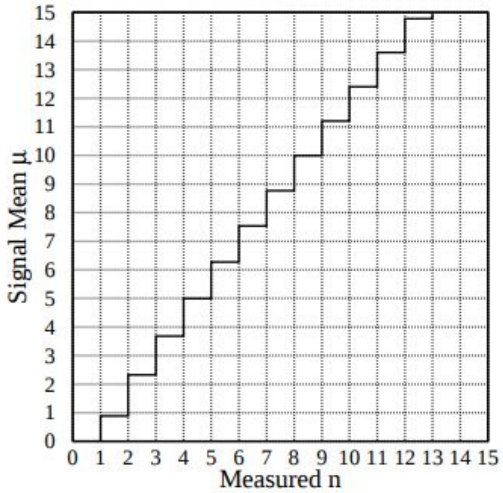
etc...



Neyman construction: Try it yourself

Construct Neyman belt for poisson with constant background of Nb=3

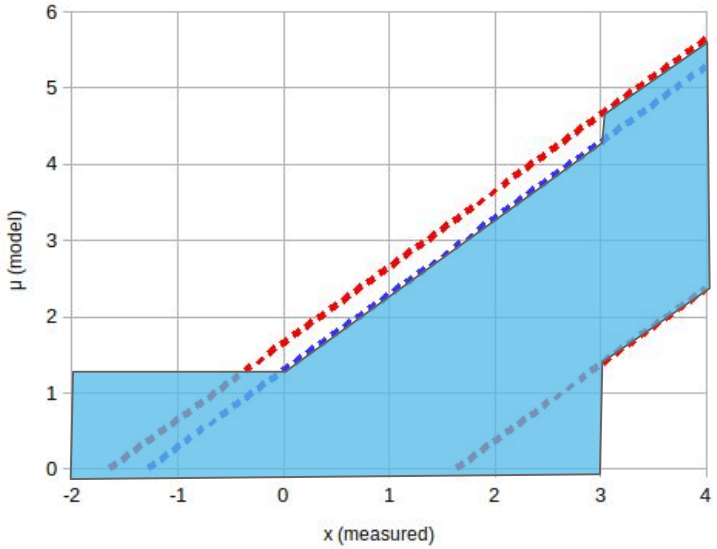




Flip flopping

- If more than 3 sigma: Discovery!
- If between 0 and 3 sigma: Limit!
- If less than 0: Set it to 0!

For μ=2⇒ X1=2-1.28 , x2=2+1.64



Feldman-Cousins
Likelihood ratio ordering principle

Based on the likelihood ratio:

$$R(x) = \frac{P(x|\mu)}{P(x|\mu_{\text{best}})}$$

μ_{best} is the (physically allowed) value that maximizes p(x|μ) for that specific x .

For fixed μ, add values of x to the interval from higher to lower R until the desired probability content is realized.

$$P(x|\mu_{\text{best}}) = \begin{cases} 1/\sqrt{2\pi}, & x \geq 0 \\ \exp(-x^2/2)/\sqrt{2\pi}, & x < 0. \end{cases}$$

We then compute R in analogy to Eq. 4.1, using Eqs. 3.1 and 4.2:

$$R(x) = \frac{P(x|\mu)}{P(x|\mu_{\text{best}})} = \begin{cases} \exp(-(x-\mu)^2/2), & x \geq 0 \\ \exp(x\mu - \mu^2/2), & x < 0. \end{cases}$$

For an only positive gaussian gaussian:

if x≥0 then $\hat{\mu}$=x ⇒ $R = \frac{e^{-(x-\mu)^2/2}}{1}$

If x<0 then $\hat{\mu}$=0⇒ $R = \frac{e^{-(x-\mu)^2/2}}{e^{-(x)^2/2}}$

Type I and Type II errors

Type I

Reject the null when it is actually true/

The probability for this is the “significance level”

(” P-Value”, “alpha”).

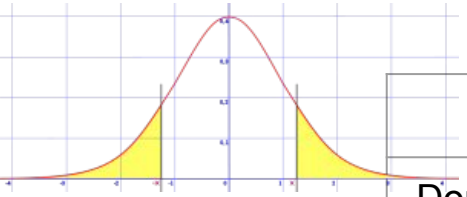
We can choose it to be as small as we wish

Type II

Fail to reject the null when it is wrong/

If the probability of this to happen is beta,

Then “1-beta” is called “the power”



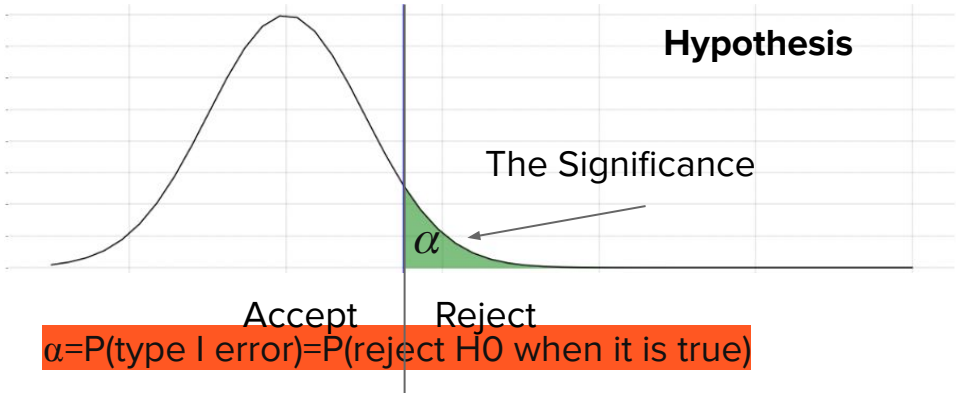
	H0 is True	H0 is False
Don’t reject H0	✅	Type II error β
Reject H0	Type I error α	✅

Significance and power

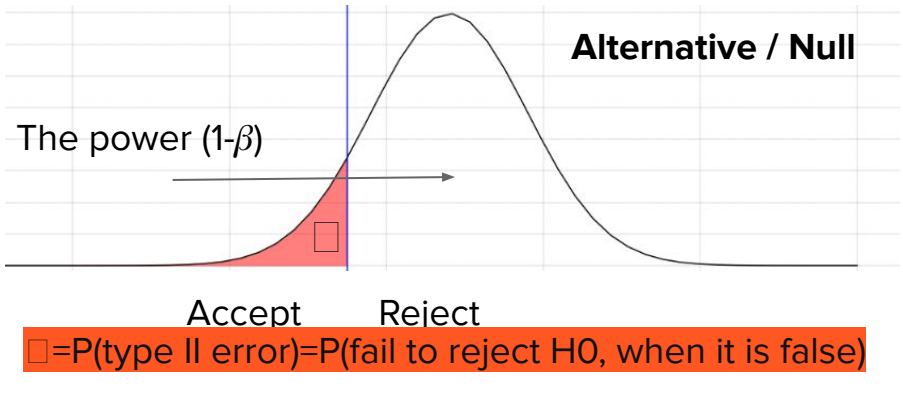
- Significance - The probability to reject H when it is true
- Confidence level = (1-α)* 100%
- Type I and Type II are related - when one increase, the other decrease

Neyman-Pearson Lemma - The likelihood-ratio test statistics is the most powerful test for a given significance level (alpha) . Any other test will have less power

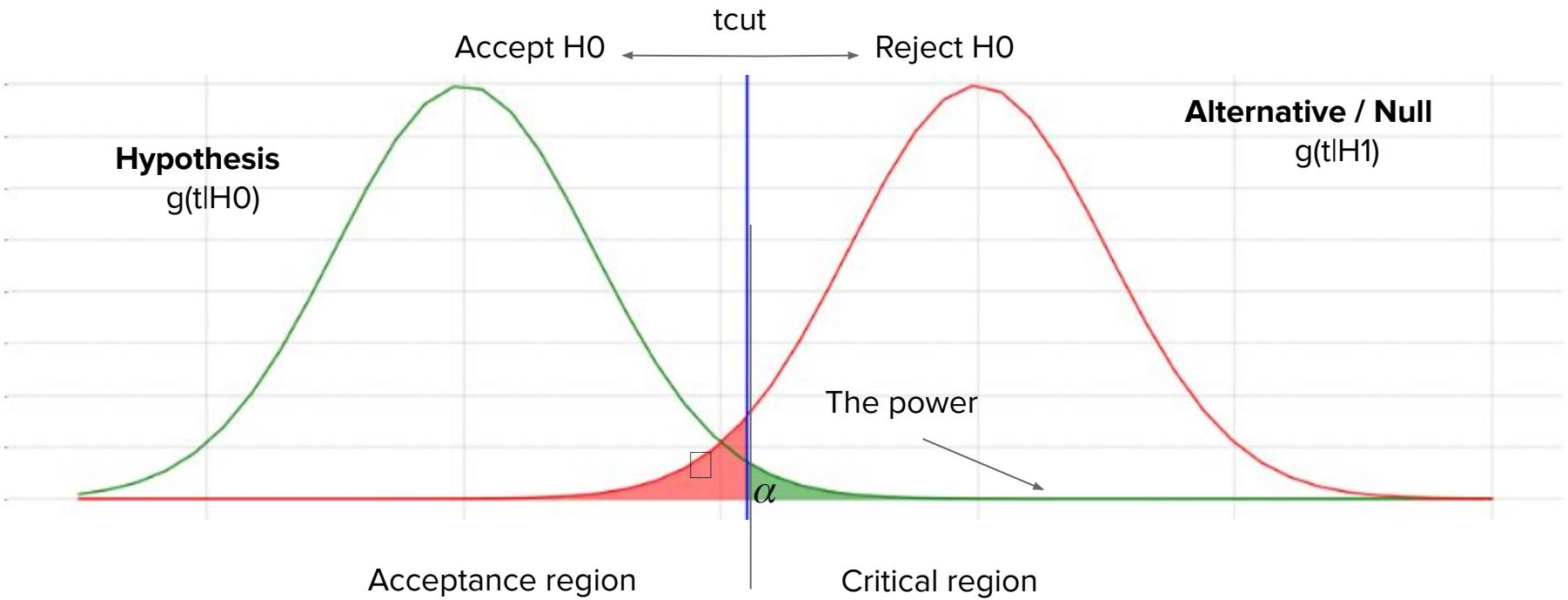
$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

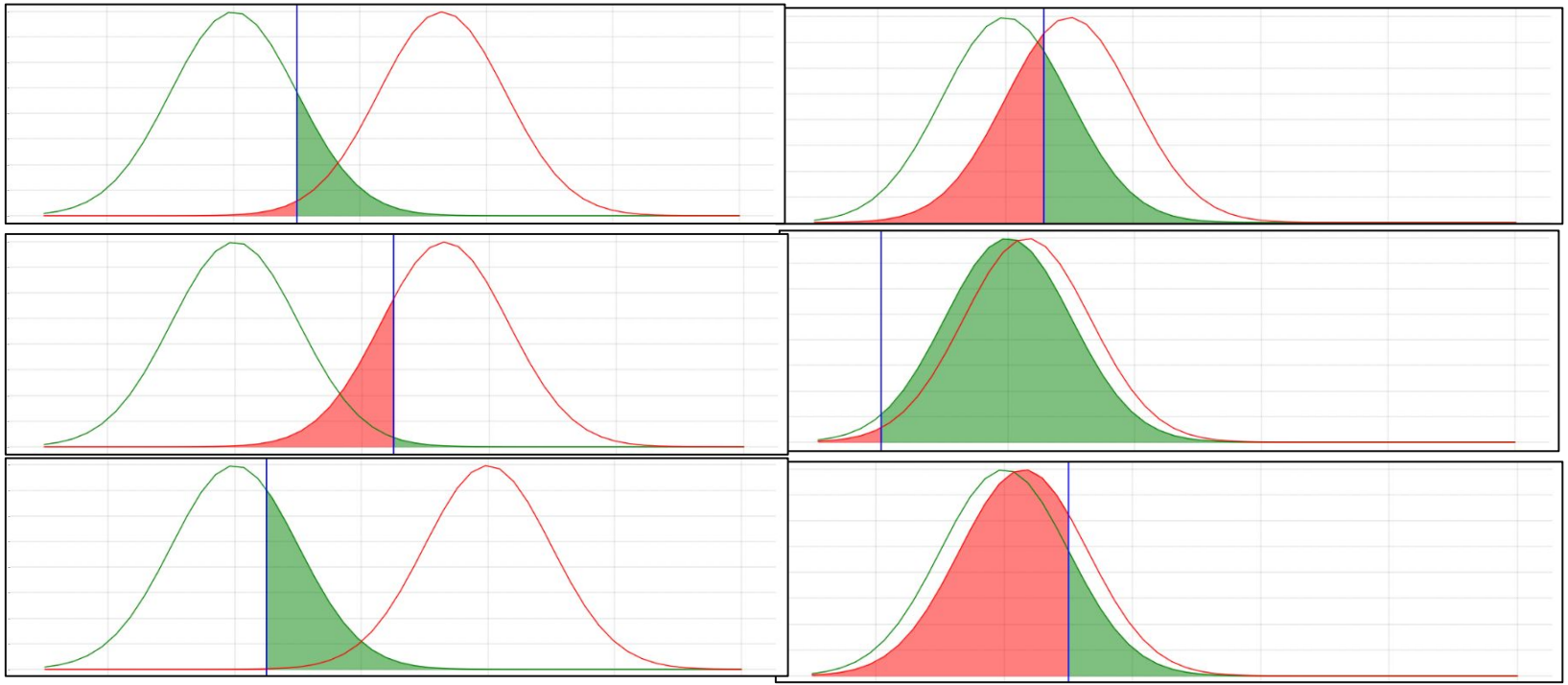


α=P(type I error)=P(reject H0 when it is true)



□=P(type II error)=P(fail to reject H0, when it is false)


tcut
Accept H0
Reject H0
Hypothesis
g(t|H0)
Alternative / Null
g(t|H1)
The power
α
Acceptance region
Critical region



P-value

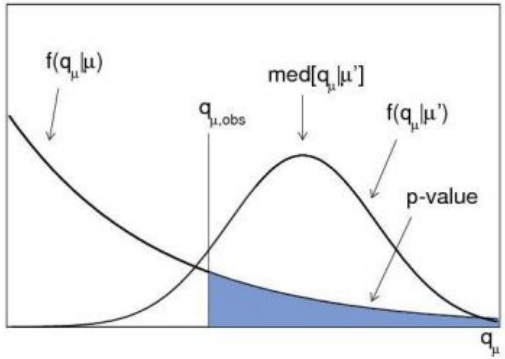
- P-value = Probability that a test statistic will take on a value that is at least as extreme as the observed value when the null hypothesis H0 is true

⇒ If P-value > α , fail to reject H0 at significance level α;

⇒ If P-value < α , reject H0 at significance level α.

- Equivalently use significance, Z, defined as equivalent number of sigmas for a Gaussian fluctuation in one direction:

$$Z = \Phi^{-1}(1 - p)$$



Asymptotic formulae for likelihood-based tests of new physics

Glen Cowan[1], Kyle Cranmer[2], Eilam Gross[3], Ofer Vitells[3]

[1] Physics Department, Royal Holloway, University of London, Egham, TW20 0EX, U.K.
[2] Physics Department, New York University, New York, NY 10003, U.S.A.
[3] Weizmann Institute of Science, Rehovot 76100, Israel

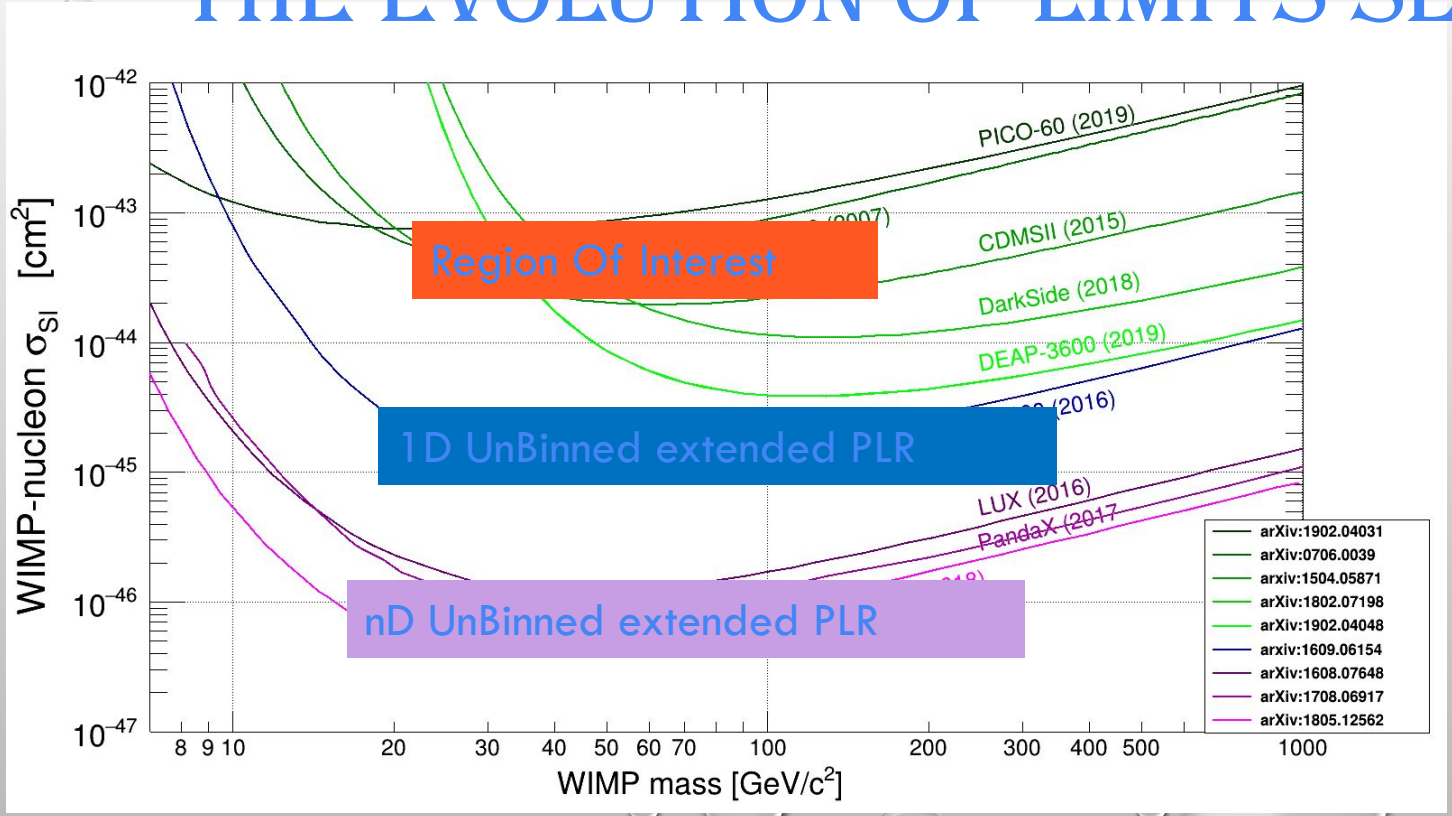
THE EVOLUTION OF LIMITS SETTING

Region Of Interest

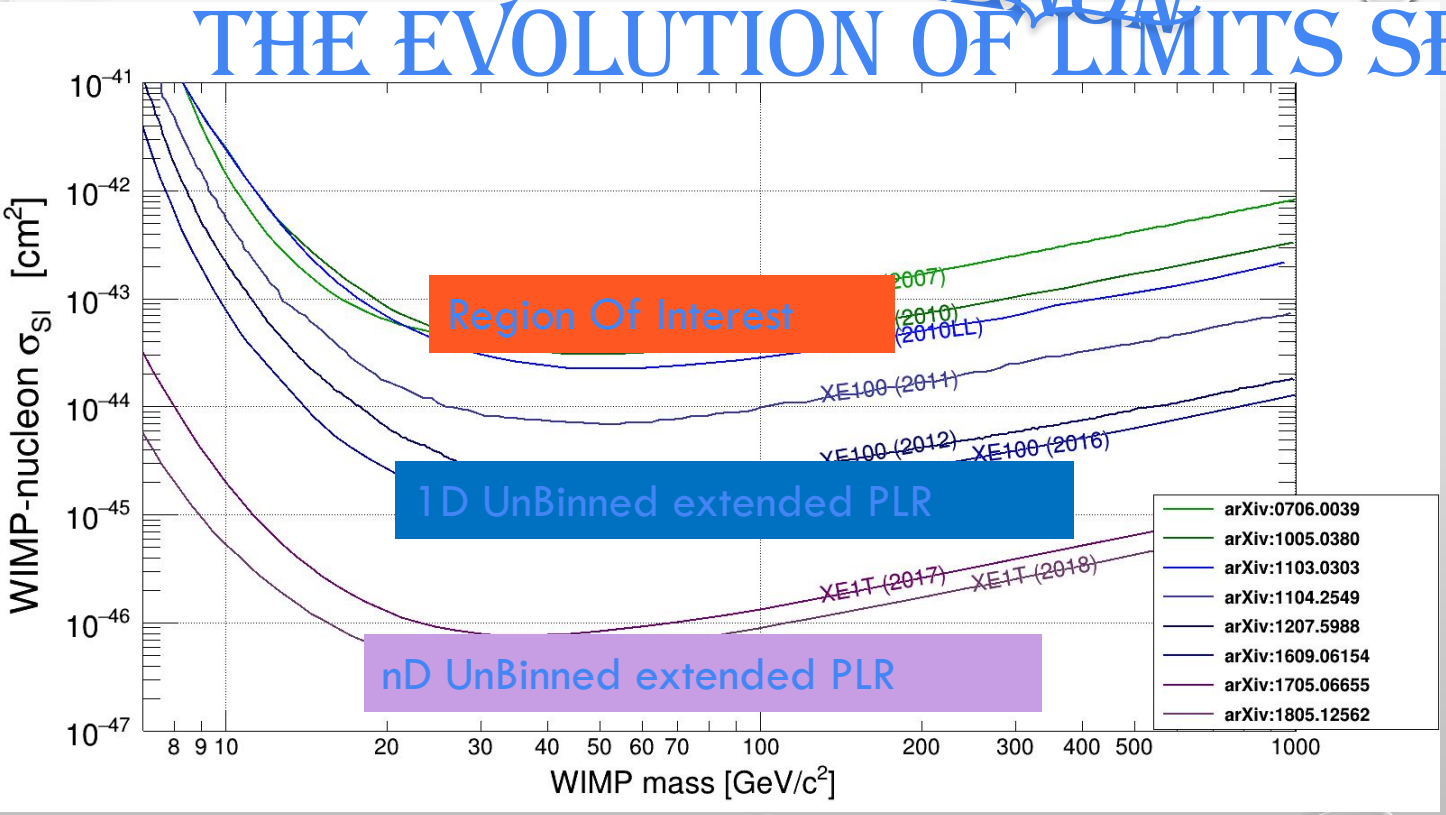
1D UnBinned extended PLR

nD UnBinned extended PLR

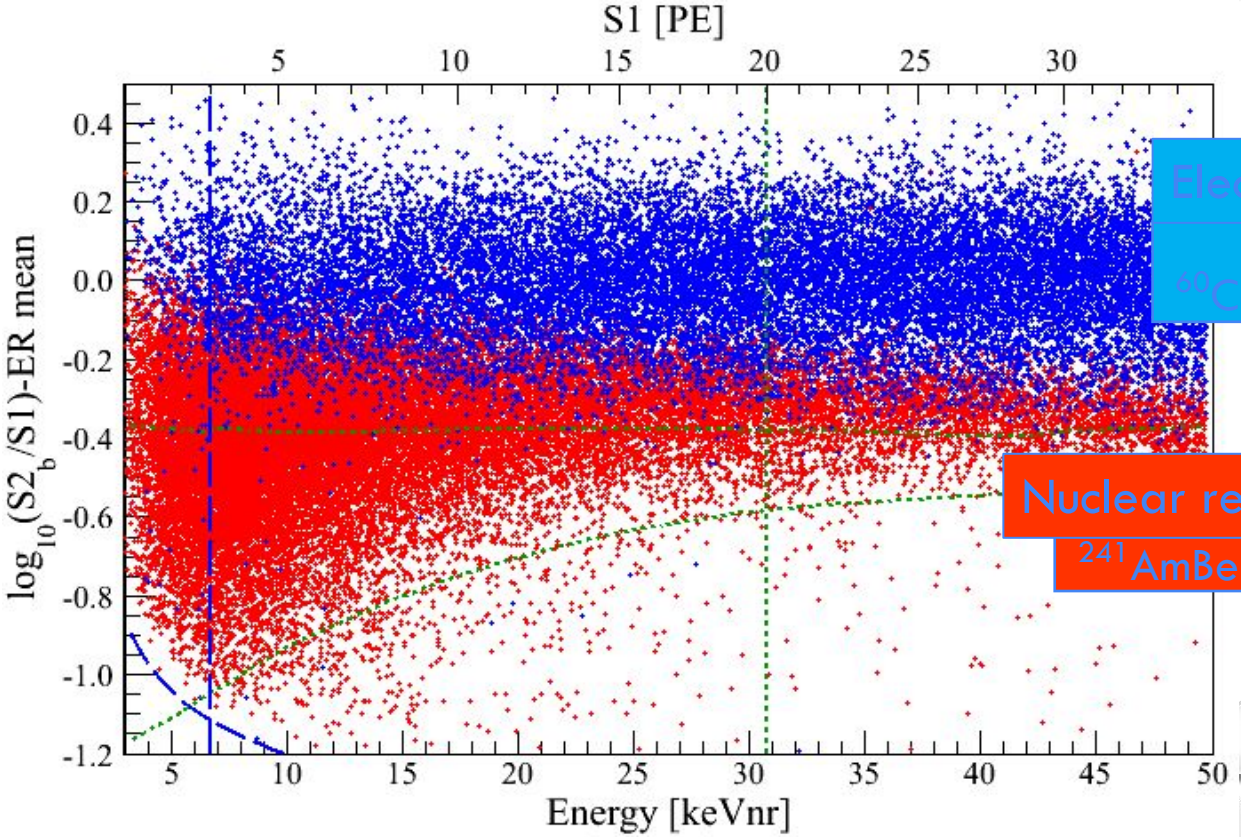
THE EVOLUTION OF LIMITS SETTING



THE EVOLUTION OF LIMITS SETTING



DETECTION PRINCIPLE DISCRIMINATION VARIABLES



Electronic recoil

^{60}Co and ^{232}Th sources

Nuclear recoil

^{241}AmBe

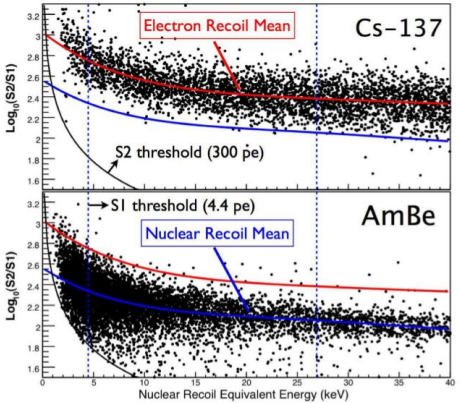
Region Of Interest

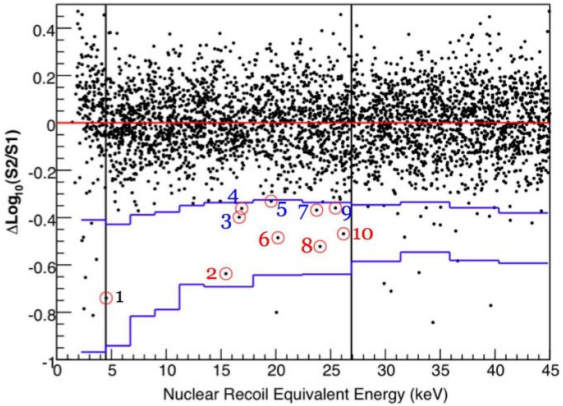
PRL 100, 021303 (2008) PHYSICAL REVIEW LETTERS

First Results from the XENON10 Dark Matter Experiment at the Gran Sasso National Laboratory

J. Angle,[1,2] E. Aprile,[3,*] F. Arneodo,[4] L. Baudis,[2] A. Bernstein,[5] A. Bolozdynya,[6] P. Brusov,[6] L. C. C. E. Dahl,[6,8] L. DeViveiros,[9] A. D. Ferella,[2,4] L. M. P. Fernandes,[7] S. Fiorucci,[9] R. J. Gaitskell,[9] K. R. Gomez,[10] R. Hasty,[11] L. Kastens,[11] J. Kwong,[6,8] J. A. M. Lopes,[7] N. Madden,[5] A. Manalaysay,[1,2] D. N. McKinsey,[11] M. E. Monzani,[3] K. Ni,[11] U. Oberlack,[10] J. Orboeck,[2] G. Plante,[3] R. Santorelli,[3] J. M P. Shagin,[10] T. Shutt,[6] P. Sorensen,[9] S. Schulte,[2] C. Winant,[5] and M. Yamashita[3]

XENON10 **2005-2007**





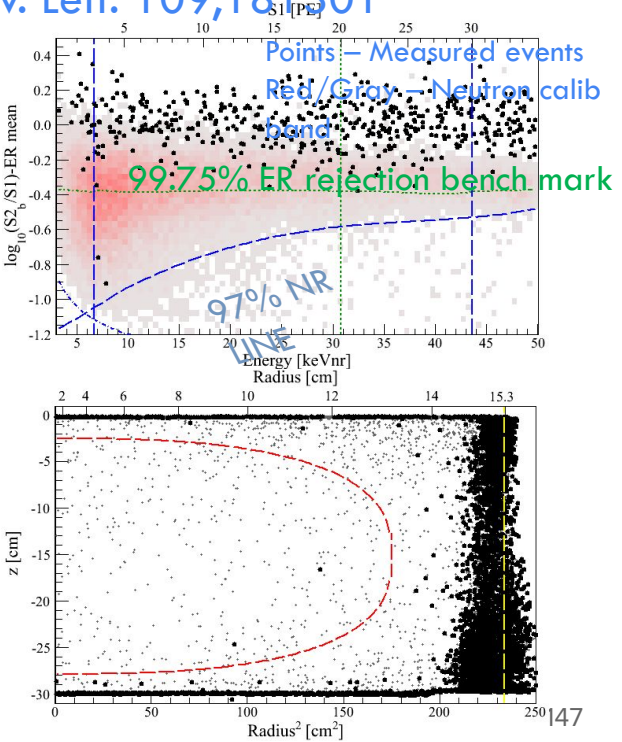
"However, the uncertainty of the estimated number of leakage events for each energy bin in the analysis of the WIMP search data is currently limited by available calibration statistics. Based on the analysis of multiple scatter events, no neutron induced recoil event is expected in the single scatter WIMP-search data set. To set conservative limits on WIMP-nucleon spin-independent cross section, we consider all ten observed events, with no background subtraction. "

What do we learn from a TPC event?

Illustrated by XENON100 2011/2012 data set
225 Live days

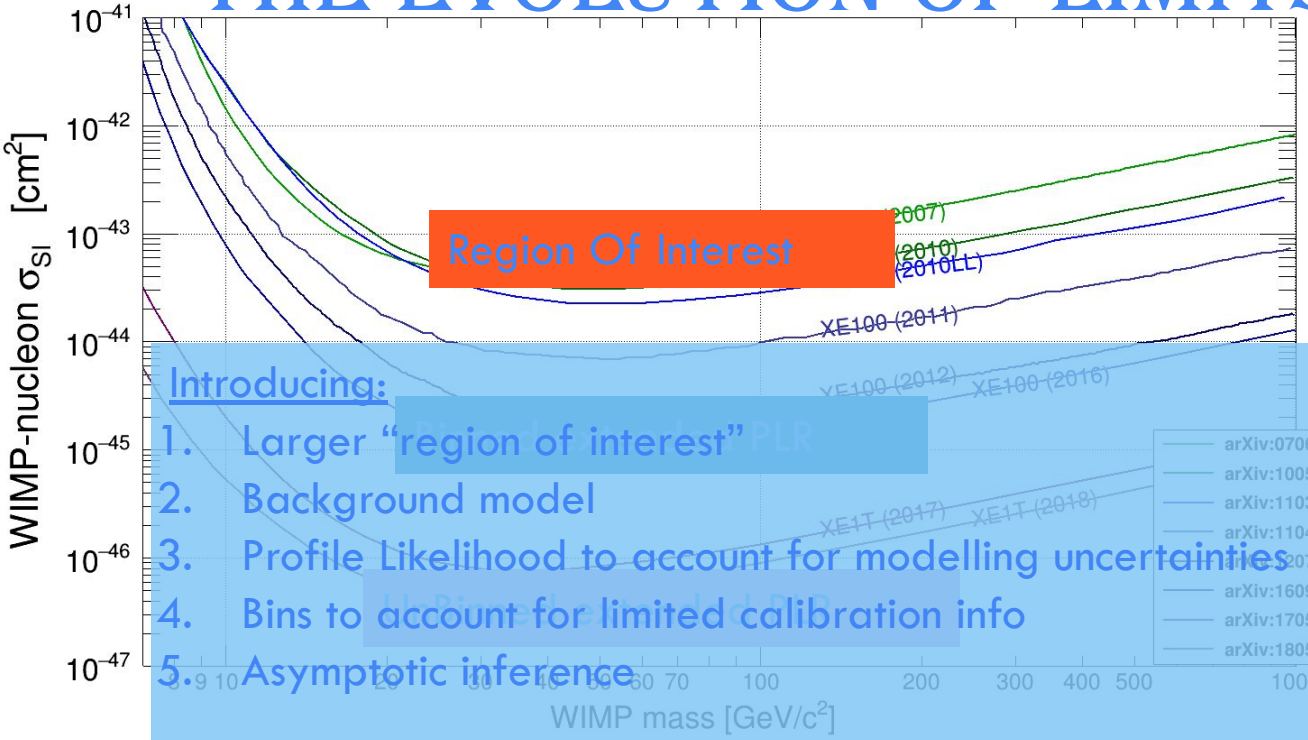
- s1,s2:
  - Energy scale
  - Discrimination: ER vs. Nr (s1/s2)
- Vertex reconstruction
  - Fiducialization
  - Single vs. Multiple scatters
- Waveforms
- Event epoch time
- Slow control (detector stability)

Phys. Rev. Lett. 109,181301 (2012)



Bold – below 99.75 ER line (likely to be NR)
dots – above 99.75 ER line (likely to be ER)

THE EVOLUTION OF LIMITS SETTING

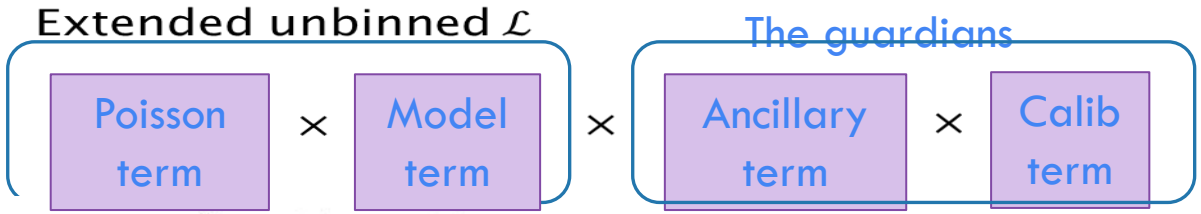


Introducing:

1. Larger "region of interest"
2. Background model
3. Profile Likelihood to account for modelling uncertainties
4. Bins to account for limited calibration info
5. Asymptotic inference

11.7 days analysis limit improved by × ~2

THE LIKELIHOOD FUNCTION



$$\mathscr{L}_{ub} = \text{Poiss}(N|N_s+N_b)\prod_{i=1}^{N}\frac{N_s f_s(x_i)+N_b f_b(x_i)}{N_s+N_b}.$$

$N_s\ (\sigma;\bar{\theta}_s,\bar{\theta}_g)$

$N_b^j\ (\bar{\theta}_b,\bar{\theta}_g)$

n_{data}

$f_s\ (\bar{x}|\bar{\theta}_s,\ \bar{\theta}_g)$

$f_b^j(\bar{x}|\bar{\theta}_b,\ \bar{\theta}_g)$

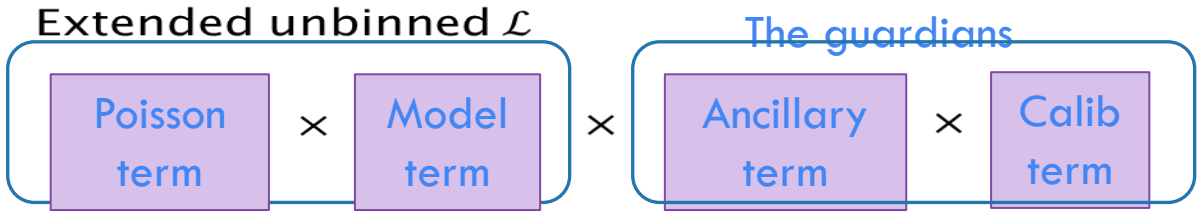
e.g.

$\text{Gaus}\left(\hat{N}_b^j \middle| N_b^j, \sigma^j\right)$

$\text{Gaus}\left(\hat{\theta} \middle| \theta, \sigma_\theta\right)$

- Long list of observables x: S_1, S_2, (R,z,θ), t
- Long list of parameters: $\bar{\theta}_s, \bar{\theta}_g, \bar{\theta}_b$
  Some are correlated, some are not…

THE LIKELIHOOD FUNCTION



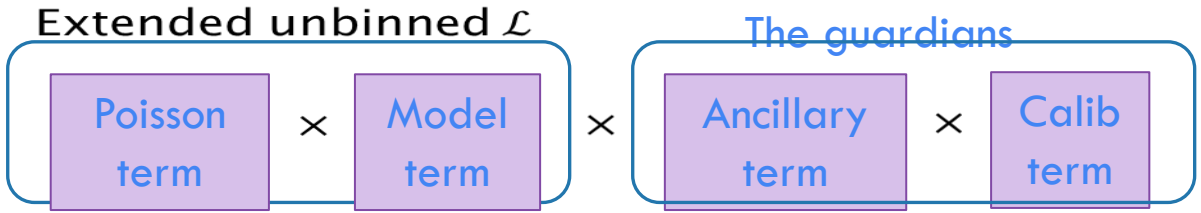
- Long list of observables x: S_1, S_2, (R,z,θ), t
- Long list of parameters: $\bar{\theta}_s, \bar{\theta}_g, \bar{\theta}_b$

Three choices:

1. Ignore –
   That's easy to implement
2. Binned model –
   Bins in discrimination space ("bands")
   Spatial bins (r,z,q)
   Temporal bins (E variations, background conditions, "runs")
3. Unbinned Model
   Higher dimensions for f_s, f_b
   Add nuisance parameters

$$\mathcal{L} = \mathcal{L}^{I} \times \mathcal{L}^{II} \times \mathcal{L}^{III}$$

THE LIKELIHOOD FUNCTION



Some (hopefully) good reasons to take it slowly:

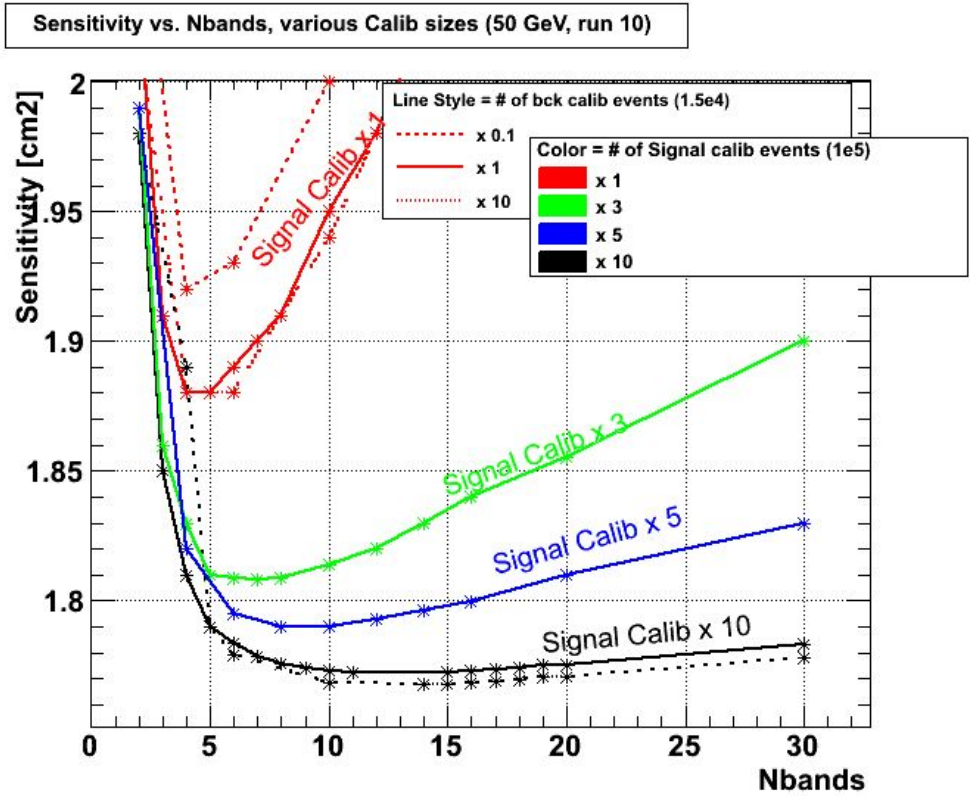
- Limited knowledge – risk of under/over coverage
  - Limited calibration
  - Lack of model
  - Always risk of mis-modeling

Use the power of the guardians wisely!

- Not needed
  - The additional information / resolution is not needed
- Save on resources
  - Modeling and minimizing. Asymptoticness (checks or bypass)

    Required cpus, diskspace, people, nerves, sanity

How many bins to use?



XENON'S 1ST LIKELIHOOD FUNCTION

Parameter of interest:

Ns – total number of signal events

Nuisance parameters:

Nb – background events

ε_s^i, ε_b^i – distribution along bands of sig/bck

tLeff – deviation of Leff from median

$$\mathcal{L} = \prod_j^K \mathrm{Poiss}(n^j | \epsilon_s^j N_s + \epsilon_b^j N_b)$$

Poisson on data, per band.

$$\times \prod_{i=1}^{n^j} \frac{\epsilon_s^j N_s f_s(S1_i) + \epsilon_b^j N_b f_b(S1_i)}{\epsilon_s^j N_s + \epsilon_b^j N_b}$$

Distribution of events in each band

$$\times \prod_j^K \mathrm{Poiss}(m_s^j | \epsilon_s^j M_s) \times \prod_j^K \mathrm{Poiss}(m_b^j | \epsilon_b^j M_b)$$

Poisson on calibration data

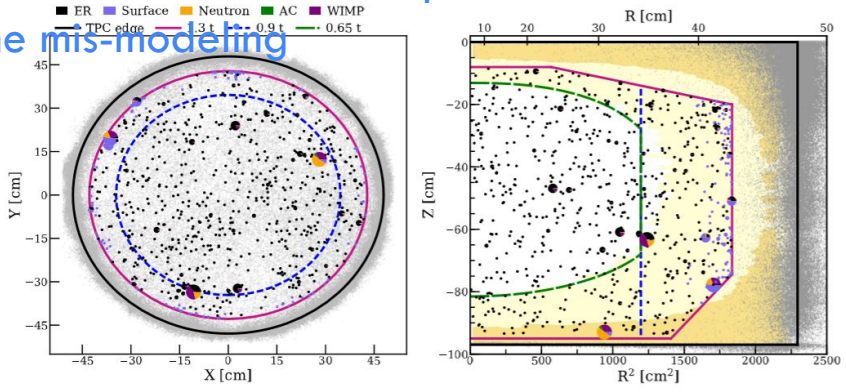
$$\times \exp(-(t - t_{\mathrm{obs}})^2/2)$$

Leff penalty

XENON1T - PLR RESULTS- 1TON-YEAR 2018 ANALYSIS

Introducing…

- PDFs in higher dimensions (s1,s2,r). No bands in s2.
- Larger volume used
- 4 independent background models constraint by calibration and simulation
- More nuisance parameters
- More complete interaction model
- More sophisticated background model with some a-priori fits
- Safeguard to account for some mis-modeling



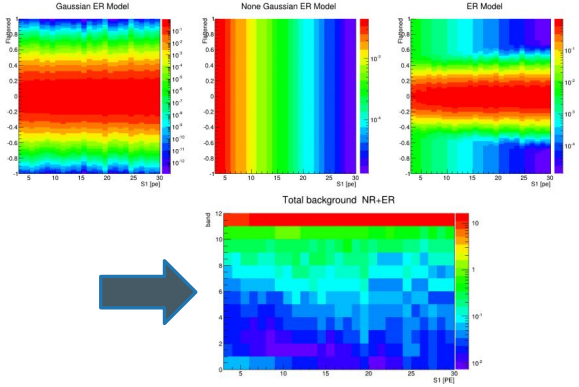
SOME THOUGHT ON SIGNAL MODEL

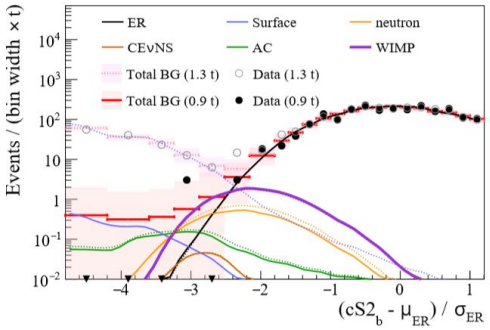
- Signal model sets: f_s and $n_s(\sigma)$ and (ε_s)
- Don't forget our parameter of interest is σ (not n_s)
- Energy scale: pe□ □kev_{nr}
- Nuisance parameters in astrophysical model, interaction model, detector response
- No calibration sample available

  (calibration data can be used to constraint parameters)
- Need to artificially incorporate spatial and temporal detector instabilities

SOME THOUGHT ON BACKGROUND MODEL

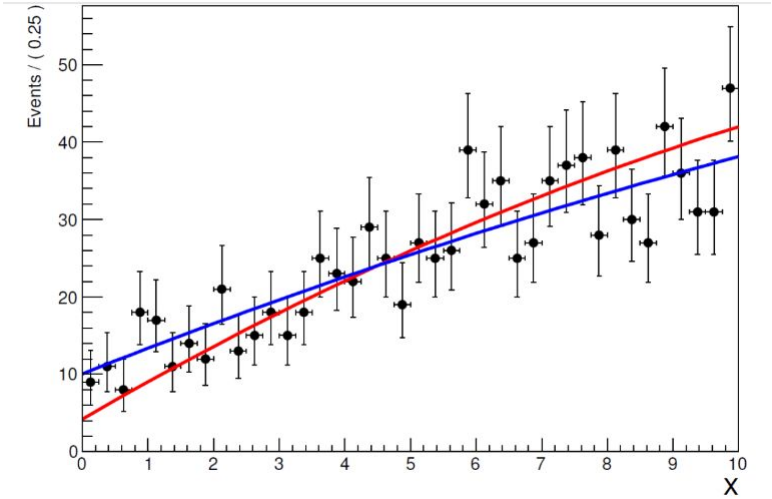
- Background model sets: f_b (ε_b) and sometimes N_b
- Several components of background: Fractions can be “frozen” or be nuisance
- Shape and magnitudes modeling
- Calibration samples may exist – statistic decreases with #variables
- Is our background model accurate “enough”?





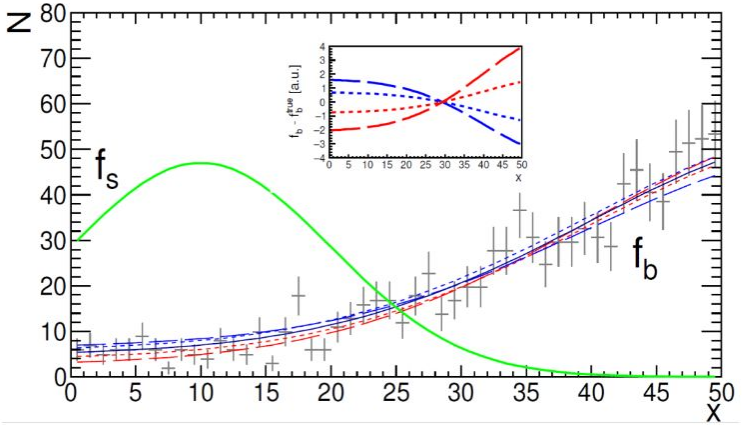
THE CURSE OF BACKGROUND MISMODELLING THE PROBLEM

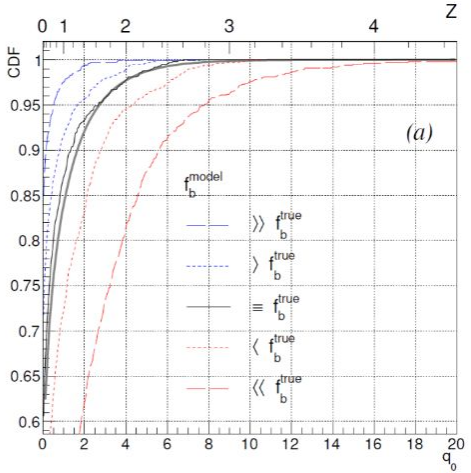
- Too many parameters
- Hidden parameters
- Partial underlying model
- ....Mistakes...



Might lead to enhanced false discovery rate or overly constrained limits

THE CURSE OF BACKGROUND MISMODELLING THE PROBLEM





Arxiv:1610.02643

THE CURSE OF BACKGROUND MISMODELLING
THE PROBLEM

Arxiv:1610.02643

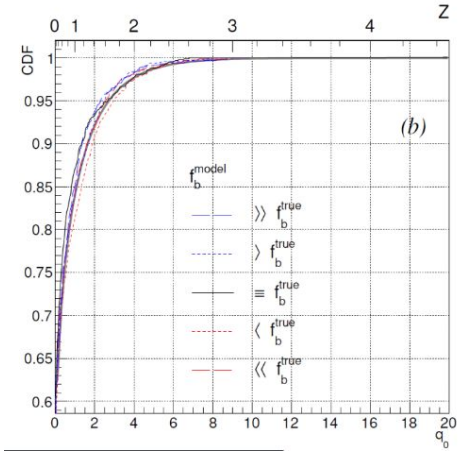
- Use the benchmark model
- Do not add extra nuisance parameters

$$f_b(x) \rightarrow (1-\varepsilon)f_b(x) + \varepsilon f_s(x)$$

$$L_{overall} = Poiss(N|N_s+N_b)\prod\frac{N_s f_s(x_i) + N_b(1-\varepsilon)f_b(x_i) + N_b\varepsilon f_s(x_i)}{N_s+N_b} \times L_{cal}(\varepsilon)$$

$$L_{cal}(\varepsilon) = \prod\prod(1-\varepsilon)\, f_b(x_i) + \varepsilon f_s(x_i)$$

- Works for limits and discoveries
- Safeguards background components that are based on calibration
- We found out that a similar technique used for cross checks in the LHC, “spurious signal”



WHERE IT HURTS..

EXAMPLE 1:“THE CURSE OF MISMODELLING”

The “safeguard” can provide some protection for models constructed based on calibration samples.

Nuisance parameters can be added, but

– Require some model assumption

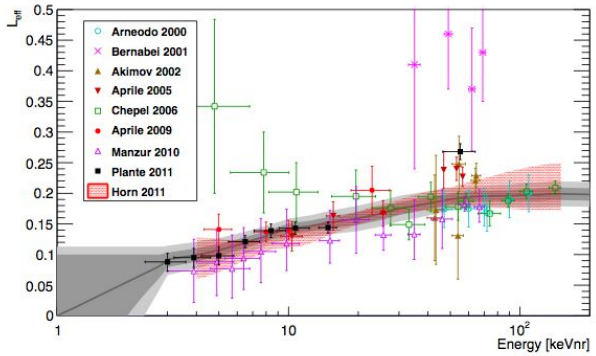
- Complicates analysis – heavier, slower

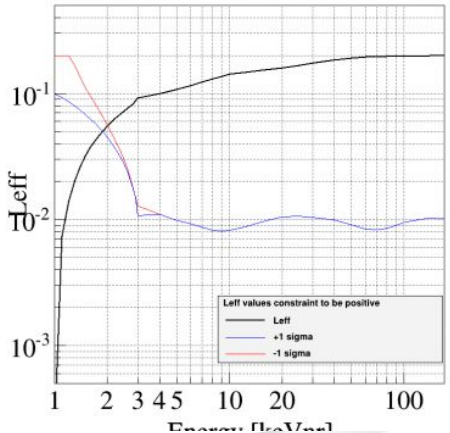
It is not enough

WHERE IT HURTS..

EXAMPLE 2: "THE CURSE OF THE UN-MODELLED"

- Include nuisance parameters without an underlying model
- Non physical regions
- Non symmetric nuisance uncertainties





Phys. Rev. C 84, 045805 (2011)

WHERE IT HURTS..

EXAMPLE 3: "THE BLESSING OF ASYMPTOTICNESS"

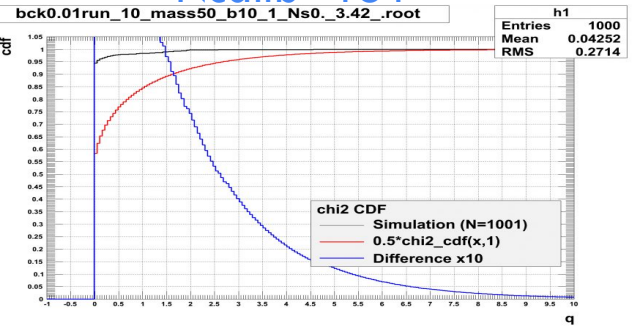
(Or "we ❤ wilks & arxiv1007.1727")

Low bg

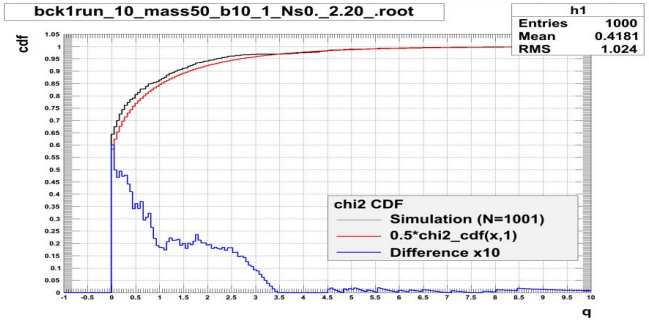
$$q_\sigma = \begin{cases} -2\ln\lambda(\sigma) & \hat{\sigma} < \sigma \\ 0 & \hat{\sigma} > \sigma \end{cases}$$

$$p_s = \int_{q_\sigma^{\mathrm{obs}}}^{\infty} f(q_\sigma|H_\sigma)\,\mathrm{d}q_\sigma.$$

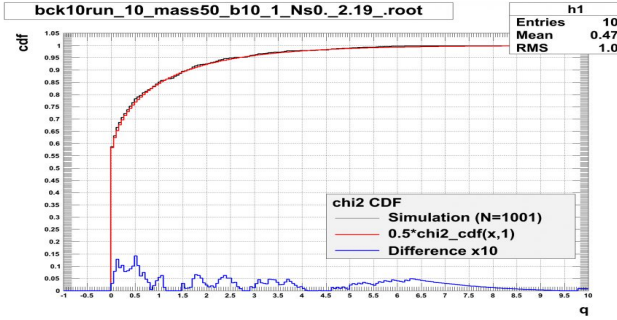
Ncalib=151



Ncalib=15128



Ncalib=151280



Need to verify asymptoticness and run MC if broken

WHERE IT HURTS..

EXAMPLE 4: "THE CURSE OF MULTIPLE DIMENSIONS"

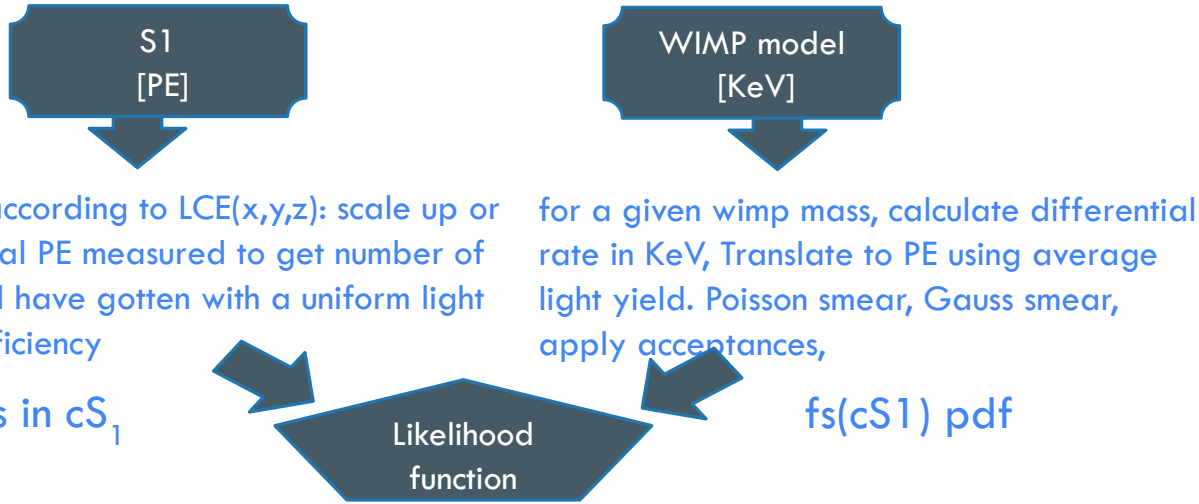
Generating multidimensional (s1,s2,r,z...) pdf maps for "many" nuisance parameters variations

- Algorithm:
  - Prepare a model bank ahead of time
  - Or build the necessary model during minimization
  - (Possibly with smart book keeping and archiving)
- Nuisance parameter resolution
  - How large a step in modeling
  - Interpolate?
- Verifying asymptoticness or doing mc instead becomes painful
- Also: complicated codes

$$\lambda(\sigma) = \frac{\max_{\sigma \text{ fixed}} \mathscr{L}(\sigma; \mathcal{L}_{\text{eff}}, v_{\text{esc}}, N_b, \epsilon_s, \epsilon_b)}{\max \mathscr{L}(\sigma, \mathcal{L}_{\text{eff}}, v_{\text{esc}}, N_b, \epsilon_s, \epsilon_b)} \equiv \frac{\mathscr{L}\left(\sigma, \hat{\hat{\mathcal{L}}}_{\text{eff}}, \hat{\hat{v}}_{\text{esc}}, \hat{\hat{N}}_b, \hat{\hat{\epsilon}}_s, \hat{\hat{\epsilon}}_b\right)}{\mathscr{L}\left(\hat{\sigma}, \hat{\mathcal{L}}_{\text{eff}}, \hat{v}_{\text{esc}}, \hat{N}_b, \hat{\epsilon}_s, \hat{\epsilon}_b\right)}.$$

WHERE IT HURTS..

EXAMPLE 5: "THE CURSE OF HANDWAVING"



Loop on all events in each band. For each event, use its cS1 to check how likely it is to come from the signal pdf, or background pdf.

Problem: cS1 is not physical. Low PE cut, Poisson smearing should be done on s1!

WHERE IT HURTS...

EXAMPLE 6: "THE CURSE OF DIVERSITY"

e.g. Over coverage:

- Power constraint
- Cls (Roughly 90%CL□95%CL)
- Ce la vie

$$p'_s = \frac{p_s}{1 - p_b}$$

where

$$1 - p_b = \int_{q_\sigma^{\mathrm{obs}}}^{\infty} f(q_\sigma | H_0)\, \mathrm{d}q_\sigma$$

WHERE IT HURTS...

EXAMPLE 7: "THE CURSE OF PAGE LIMIT"

- Many details to the models, inference method…
- Information in papers is limited. Very often summarized to: "…as was done in [xx]."
- Would be nice to see more detailed likelihood functions…
- Would be nice to see more likelihood curves…
- Many consistency checks, verifications to be made . usually n acknowledged.
- Follow up papers become more popular, but

  …cannot make everyone happy….

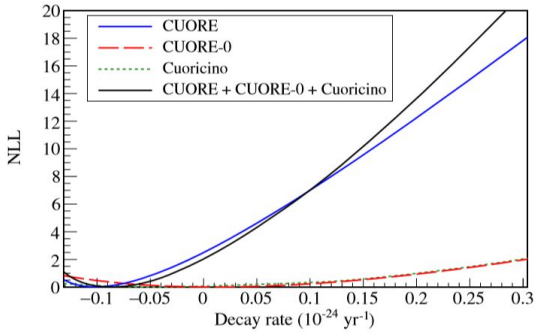
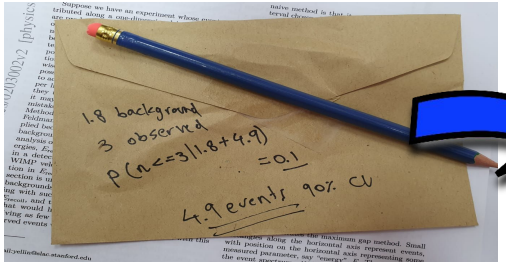
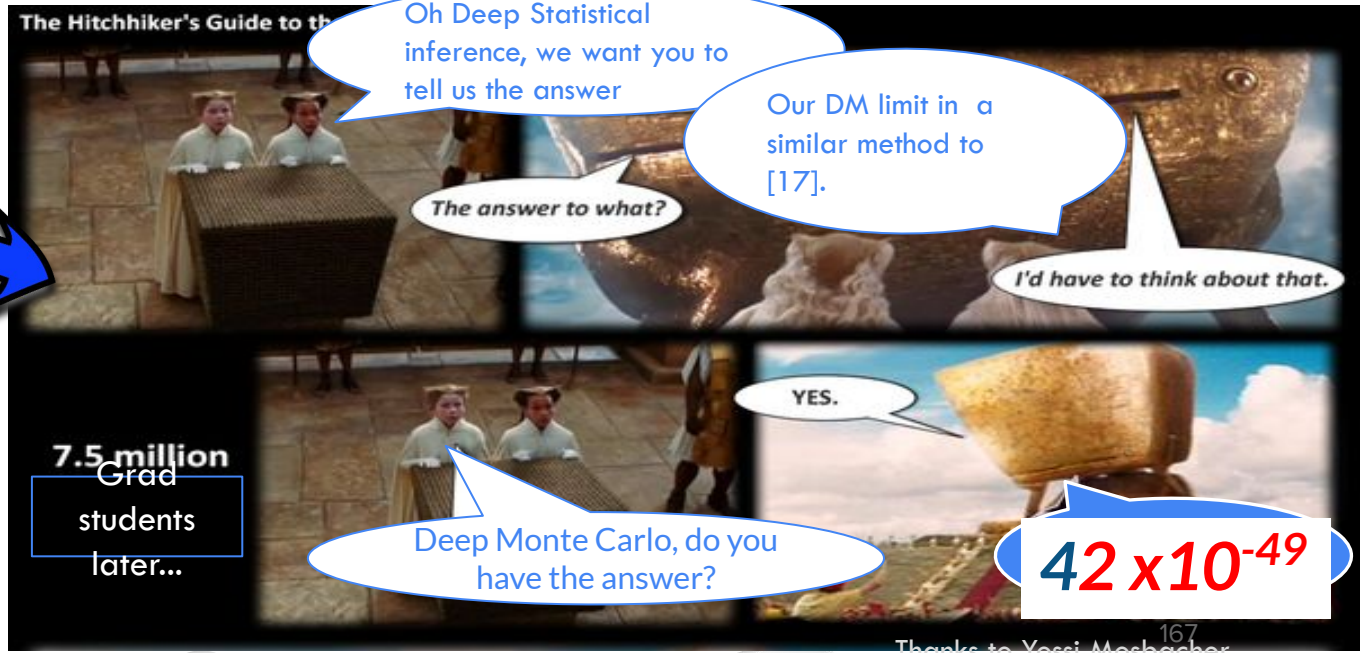


FIG. 4. Profile negative-log-likelihood curves for CUORE, CUORE-0, Cuoricino, and their combination.

GRAPHICAL SUMMARY

Back then....





Thanks to Yossi Mosbacher