Statistics

or “How to find answers to your questions”

Pietro Vischia[1]

[1] CP3 — IRMP, Université catholique de Louvain

UCLouvain
Institut de recherche
en mathématique et physique

LIP Course on Physics at the LHC

A day in the life of a PhD student

Why statistics?

The night before, and the morning

Games, weather

Morning: drawing some histograms

Random variables and their properties

Distributions

After coffee break: Measuring a physical quantity

estimators, maximum likelihood

Early afternoon: finding a new particle

Test of hypotheses

CLs

Significance

Tea time: measuring differential distributions

Unfolding

End of the afternoon: work with difficult final states

Machine Learning

Summary: go home before 18h

Why statistics?

Statistics is all about answering questions

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

Statistics is all about answering questions

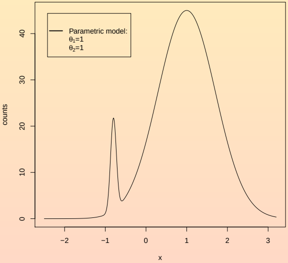
- What is the chance of obtaining a 1 when throwing a six-faced die?
  - We can throw a dice 100 times, and count how many times we obtain 1
- What is the chance of tomorrow being rainy?

Statistics is all about answering questions

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

Where does statistics live

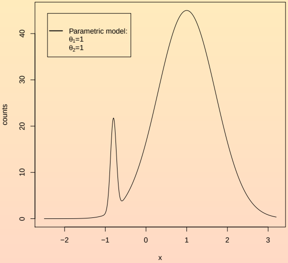
- **Theory**
  - Approximations
  - Free parameters



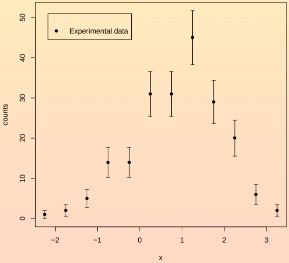
Where does statistics live

UCLouvain
Institut de recherche
en mathématique et physique

- **Theory**
  - Approximations
  - Free parameters

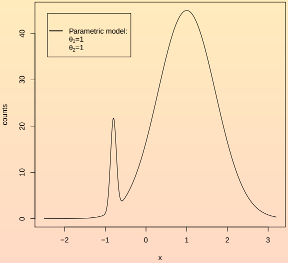


- **Experiment**
  - Random fluctuations
  - Mismeasurements (detector effects, etc)



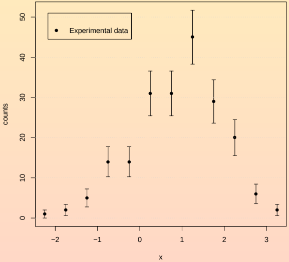
Where does statistics live

- **Theory**
  - Approximations
  - Free parameters



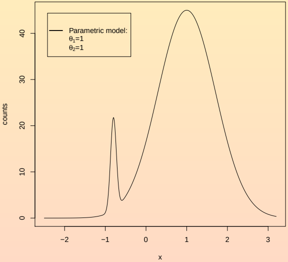
- **Statistics!**

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

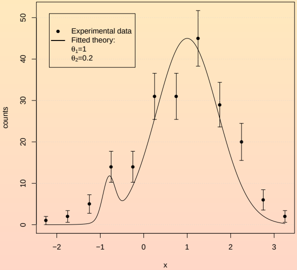


Where does statistics live

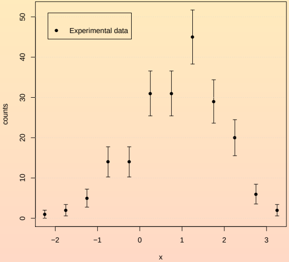
- **Theory**
  - Approximations
  - Free parameters



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



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



Gaming on the night before, walking to work in the morning

What is a “probability”?

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

Andrey Kolmogorov.

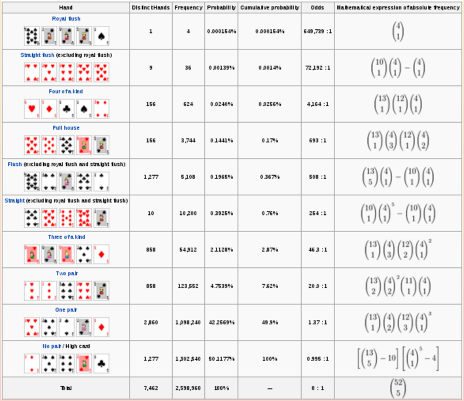
Frequentist probability - 1

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

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

Frequentist probability - 2

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



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

Subjective (Bayesian) probability

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

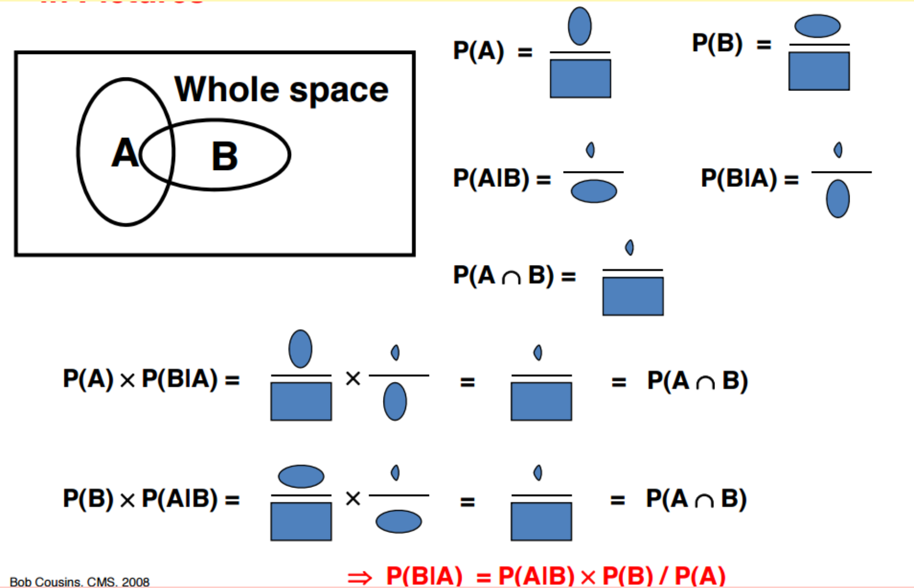
$$P(X) := \frac{\text{The largest amount you are willing to bet}}{\text{The amount you stand to win}} \tag{1}$$

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

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

Conditional probabilities: Bayes theorem

- Probabilities can be combined to obtain more complex expressions



$P(A) =$

$P(B) =$

$P(A|B) =$

$P(B|A) =$

$P(A \cap B) =$

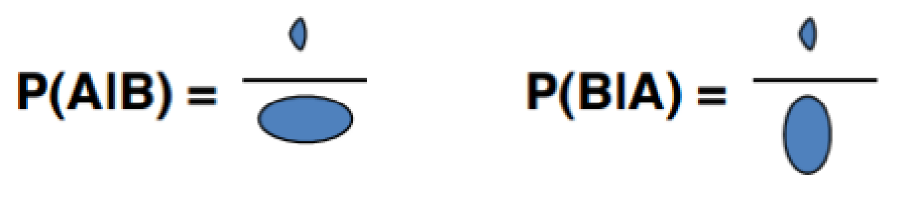
$P(A) \times P(B|A) = \times = = P(A \cap B)$

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

$\Rightarrow P(B|A) = P(A|B) \times P(B) / P(A)$

Bob Cousins, CMS, 2008

A word of advice about conditional probabilities



- Conditional probabilities are not commutative! $P(A|B) \neq P(B|A)$
- Example from Louis Lyons:
  - A: being female
  - B: being pregnant
- The probability for a female to be pregnant, $P(pregnant|female)$, is roughly 3%
- The probability for a pregnant person to be female, $P(female|pregnant)$ is unarguably $>>>>> 3\%$ ☺

A trickier example of conditional probability: the Monty Hall problem

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

A trickier example of conditional probability: the Monty Hall problem

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

A trickier example of conditional probability: the Monty Hall problem

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

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

Bayes Theorem and the Law of Total Probability

UCLouvain
Institut de recherche
en mathématique et physique

- Bayes Theorem (1763):

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

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

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

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

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

A Diagnosis problem

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

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

- We need the incidence of the disease in the population, $P(D)$! It turns out $P(D)$ is a very important to get our answer
  - $P(D) = 0.001$ (very rare disease): then $P(D|+) = 0.0902$, which is fairly small
  - $P(D) = 0.01$ (only a factor 10 more likely): then $P(D|+) = 0.4977$, which is pretty high (and substantially higher than the previous one)

Bayes Theorem and Subjective Probability

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

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

The elements of the Bayes Theorem, in Bayesian Statistics

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

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

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

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

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

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

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

- There is no golden rule for choosing a prior
- Objective Bayesian school: it is necessary to write a golden rule to choose a prior
  - Usually based on an invariance principle
- Consider a theory parameterized with a parameter, e.g. the ratio of vacuum expectation values v in a quantum field theory, $\beta := \frac{v_1}{v_2}$
- Before any experiment, we are Jon Snow about the parameter β: we know nothing
  - We have to choose a very broad prior, or better uniform, in β
- Now we interact with a theoretical physicist, who might have built her theory by using as a parameter of the model the tanged of the ratio, $tan\beta$
  - In a natural way, she will express her pre-experiment ignorance using an uniform prior **in** $tan\beta$.
  - This prior is not constant in β!!!
  - In general, there is no uniquely-defined prior expressing complete ignorance or ambivalence in both parameters (β and $tan\beta$)
- We can build a prior invariant for transformations of the parameter, but this means we have to postulate an invariance principle
  - The prior already deviates from our degree of belief about the parameter (“I know nothing”)

Choosing a prior in Bayesian statistics; in theory... 2/

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

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

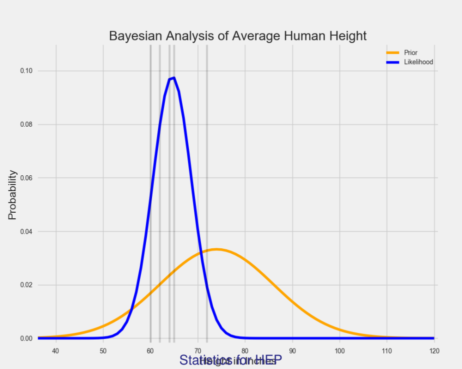
- In particle physics, the typical application of Bayesian statistics is to put an upper limit on a parameter θ
  - Find a value θ_c such that $P(\theta_{true} < \theta_c) = 95\%$
- Typically θ represents the cross section of a physics process, and is proporcional to a variable with a Poisson p.d.f.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

Choosing a prior in Bayesian statistics; in practice... 2/

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

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

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



Short summary on bayesian vs. frequentist

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

Morning: drawing some histograms

Random Variables

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

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

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

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

p.d.f. for many variables

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

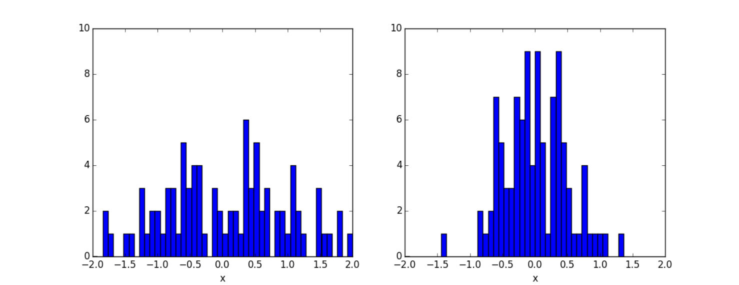
$$f_X(X) := \int f(X, Y) dY \tag{12}$$

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

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

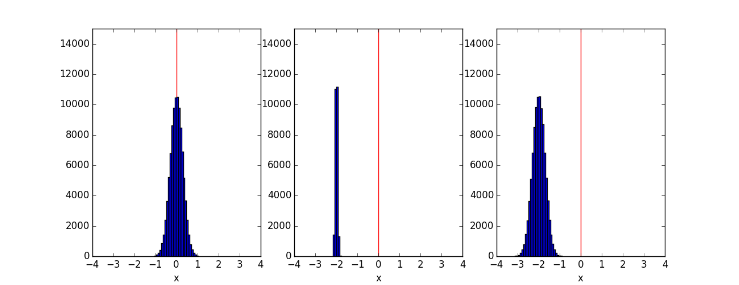
Dispersion and distributions

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



Sources of uncertainty (errors?)

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



Expected values of a random variable

- We define the expected value and mathematic expectation

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

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

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

Generalizing expected values to functions of random variables

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

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

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

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

- The variance of X is:

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

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

Let's make it funnier: more variables!

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

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

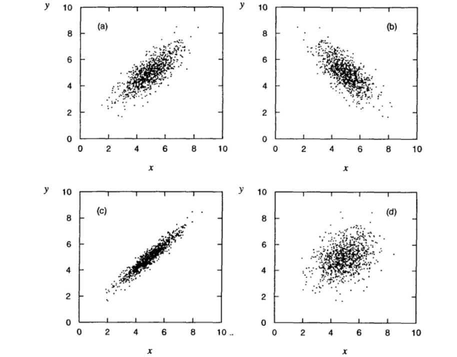


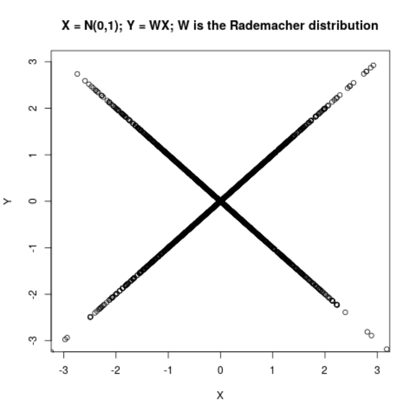
Fig. 1.9 Scatter plots of random variables x and y with (a) a positive correlation, $\rho = 0.75$, (b) a negative correlation, $\rho = -0.75$, (c) $\rho = 0.95$, and (d) $\rho = 0.25$. For all four cases the standard deviations of x and y are $\sigma_x = \sigma_y = 1$.

Mutual information: take it to the next level

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

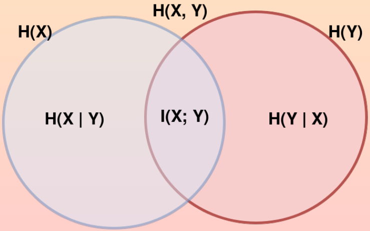
$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y) log\left(\frac{p(x,y)}{p_1(x)p_2(y)}\right)$$

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



- Related to entropy

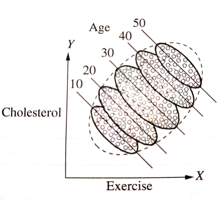
$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

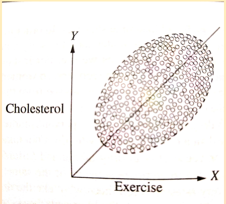


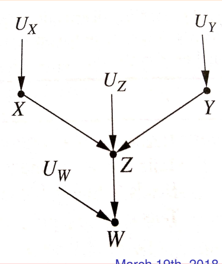
The Simpson paradox: correlation is not causation

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

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



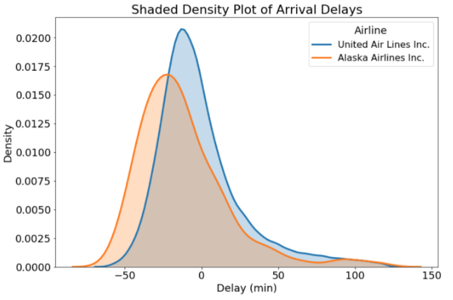


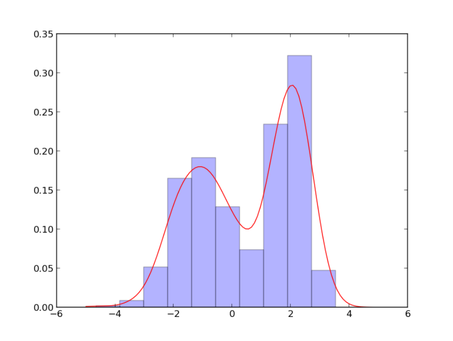


Plots from Pearl, 2016

Distributions... or not?

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

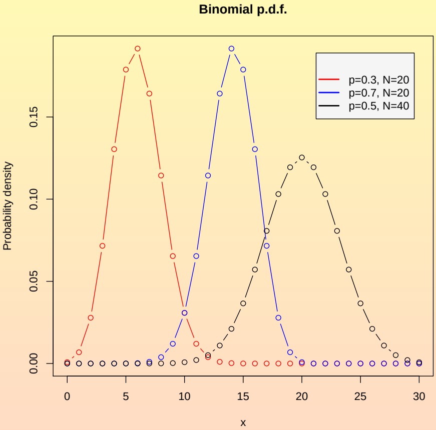




Plots from TheGlowingPython and TowardsDataScience

The Binomial distribution

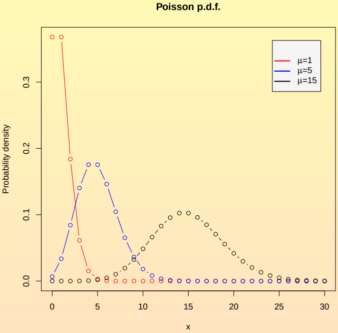
- **Binomial**
  - Discrete variable: r, positive integer $\leq N$
  - Parameters:
    - N, positive integer
    - p, $0 \leq p \leq 1$
  - Probability function: $P(r) = \binom{N}{r} p^r (1-p)^{N-r}$, $r = 0, 1, \ldots, N$
  - $E(r) = Np$, $V(r) = Np(1-p)$
  - Usage: probability of finding exactly r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



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

The Poisson distribution

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



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

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

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

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

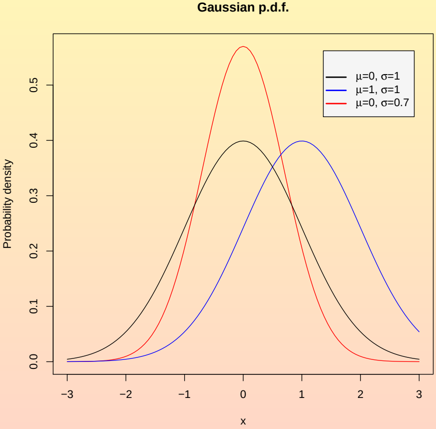
$$P(X) = \frac{0.6^1 e^{-0.6}}{1!} = 0.32929$$

The Gaussian distribution

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

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

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



The χ^2 distribution

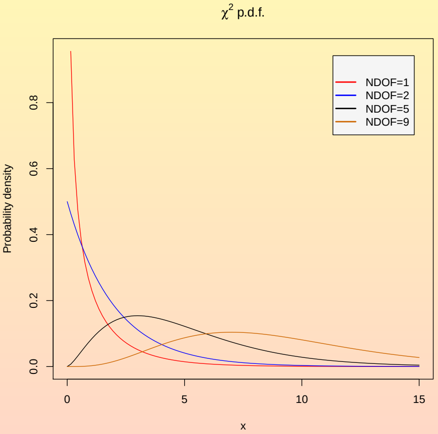
- Parameter: integer $N > 0$ degrees of freedom
- Continuous variable $X \in \mathcal{R}$
- p.d.f., expected value, variance

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

$$E[r] = N$$

$$V(r) = 2N$$

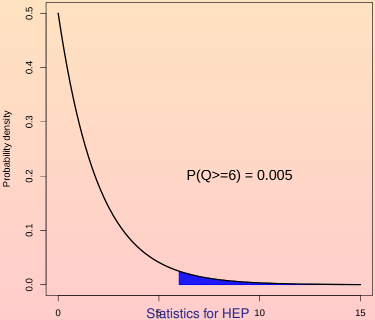
- It describes the distribution of the sum of the squares of a random variable, $\sum_{i=1}^{N} X_i^2$



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

The χ^2 distribution: why degrees of freedom?

- Sample randomly from a Gaussian p.d.f., obtaining X_1 y X_2
- $Q = X_1^2 + X_2^2$ (or in general $Q = \sum_{i=1}^N X_i^2$) is itself a random variable
  - What is $P(Q \geq 6)$? Just integrate the $\chi^2(N = 2)$ distribution from 6 to ∞
- Depends only on N!
  - If we sample 12 times from a Gaussian and compute $Q = \sum_{i=1}^{12} X_i^2$, then $Q \sim \chi^2(N = 12)$
- Theorem: if $Z_1, ..., Z_N$ is a sequence of normal random variables, the sum $V = \sum_{i=1}^N Z_i^2$ is distributed as a $\chi^2(N)$
  - The sum of squares is closely linked to the variance $E[(X - \mu)^2] = E[X^2] - \mu^2$ from Eq. 18
- The χ^2 distribution is useful for goodness-of-fit tests that check how much two distributions diverge point-by-point
- It is also the large-sample limit of many distributions (useful to simplify them to a single parameter)



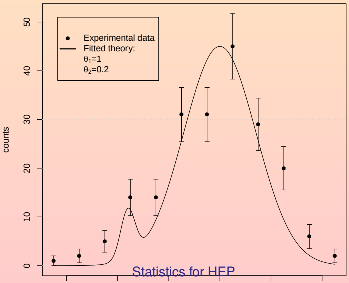
The χ^2 distribution: goodness-of-fit tests 1/

- Consider a set of M measurements $\{(X_i, Y_i)\}$
  - Suppose Y_i are affected by a random error representable by a gaussian with variance σ_i
- Consider a function $g(X)$ with predictive capacity, i.e. such that for each i we have $g(X_i) \sim Y_i$
- Pearson's χ^2 function related to the difference between the prediction and the experimental measurement in each point

$$\chi^2_P := \sum_{i=1}^{M} \left[\frac{Y_i - g(X_i)}{\sigma_i} \right]^2 \tag{19}$$

- Neyman's χ^2 is a similar expression under some assumptions
  - If the gaussian error on the measurements is constant, it can be factorized
  - If Y_i represent event counts $Y_i = n_i$, then the errors can be approximated with $\sigma_i \propto \sqrt{n_i}$

$$\chi^2_N := \sum_{i=1}^{M} \frac{\Big(n_i - g(X_i)\Big)^2}{n_i} \tag{20}$$



The χ^2 distribution: goodness-of-fit tests 2/

- If $g(X_i) \sim Y_i$ (i.e. $g(X)$ reasonably predicts the data), then each term of the sum is approximately 1
- Consider a function of $\chi^2_{N,P}$ and of the number of measurements M
  - $E[f(\chi^2_{N,P}, M)] = M$
  - The function is analytically a χ^2:

$$f(\chi^2, M) = \frac{2^{-\frac{M}{2}}}{\Gamma\left(\frac{N}{2}\right)} \chi^{N-2} e^{-\frac{\chi^2}{2}} \tag{21}$$

  - The cumulative of f is

$$1 - cum(f) = P(\chi^2 > \chi^2_{obs} | g(x) \text{ is the correct model}) \tag{22}$$

- Comparing χ^2 with the number of degrees of freedom M, we therefore have a criterion to test for goodness-of-fit
  - For a given M, the p.d.f. is known ($\chi^2(M)$) and the observed value can be computed and compared with it
  - Null hypothesis: there is no difference between prediction and observation (i.e. g fits well the data)
  - Alternative hypothesis: there is a significant difference between prediction and observation
  - Under the null, the sum of squares is distributed as a $\chi^2(M)$
  - p-values can be calculated by integration of the χ^2 distribution

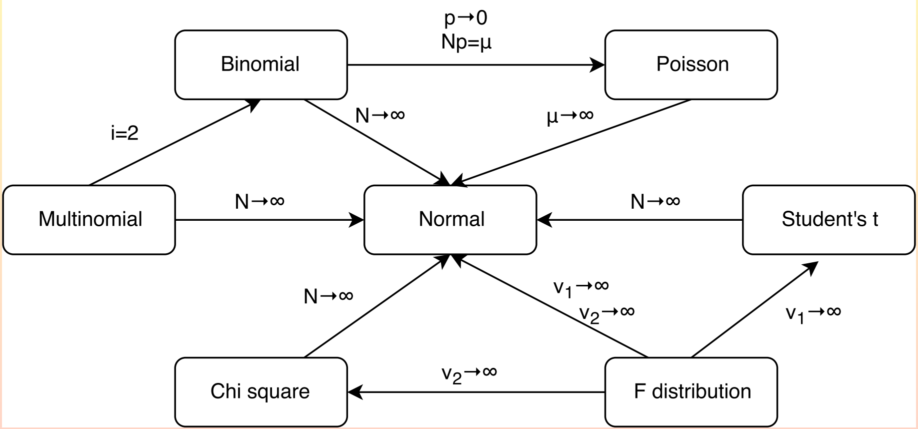
$$\begin{aligned} \frac{\chi^2}{M} &\sim 1 \Rightarrow \text{g(X) approximates well the data} \\ \frac{\chi^2}{M} &>> 1 \Rightarrow \text{poor model (increases } \chi^2\text{), or statistically improbable fluctuation} \\ \frac{\chi^2}{M} &<< 1 \Rightarrow \text{overestimated } \sigma_i\text{, or fraudulent data, or statistically improbable fluctuation} \end{aligned} \tag{23}$$

The χ^2 distribution: goodness-of-fit tests 3/

- $\chi^2(M)$ tends to a Normal distribution for $M \to \infty$
  - Slow convergence
  - It is generally not a good idea to substitute a χ^2 distribution with a Gaussian
- The goodness of fit seen so far is valid only if the model (the function $g(X)$) is fixed
- Sometimes the model has k free parameters that were not given and that have been fit to the data
- Then the observed value of χ^2 must be compared with $\chi^2(N')$, with $N' = N - k$ degrees of freedom
  - $N' = N - k$ are called reduced degrees of freedom
  - This however works only if the model is linear in the parameters
  - If the model is not linear in the parameters, when comparing χ^2_{obs} with $\chi^2(N-k)$ then the p-values will be deceptively small!
- Variant of the χ^2 for small datasets: the G-test
  - $g = 2\sum O_{ij} ln(O_{ij}/E_{ij})$
  - It responds better when the number of events is low (Petersen 2012)

Some relationships among distributions

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



After the coffee break: measuring a physical quantity

Information, the Fisher way

- The information of a set of observations should increase with the number of observations
  - Double the data should result in double the information if the data are independent
- Information should be conditional on what we want to learn from the experiment
  - Data which are irrelevant to our hypothesis should carry zero information relative to our hypothesis
- Information should be related to precision
  - The greatest the information carried by the data, the better the precision of our result

Likelihood and Fisher Information

- The narrowness of the likelihood can be estimated by looking at its curvature
- The curvature is the second derivative with respect to the parameter of interest
- A very narrow (peaked) likelihood is characterized by a very large and positive $-\frac{\partial^2 lnL}{\partial\theta^2}$
- The second derivative of the likelihood is linked to the Fisher Information

$$I(\theta) = -E\left[\frac{\partial^2 lnL}{\partial\theta^2}\right] = E\left[\left(\frac{\partial lnL}{\partial\theta}\right)^2\right] \tag{24}$$

- A very narrow likelihood will provide much information about θ_{true}
  - The posterior probability will be more localized than the prior (in the regimen in which the likelihood function dominates the product $L(\vec{x};\vec{\theta}) \times \pi$)
  - The Fisher Information will be large
- A very broad likelihood will not carry much information, and in fact the computed Fisher Information will turn out to be small

Fisher Information and Jeffreys priors

- When changing variable, the change of parameterization must not result in a change of the information
  - The information is a property of the data only, through the likelihood—that summarizes them completely (likelihood principle)
- Search for a parametrization $\theta'(\theta)$ in which the Fisher Information is constant
- Compute the prior as a function of the new variable

$$\begin{aligned}
\pi(\theta) = \pi(\theta')\left|\frac{d\theta'}{d\theta}\right| &\propto \sqrt{E\left[\left(\frac{\partial lnN}{\partial\theta'}\right)^2\right]\left|\frac{\partial\theta'}{\partial\theta}\right|} \\
&= \sqrt{E\left[\left(\frac{\partial lnL}{\partial\theta'}\frac{\partial\theta'}{\partial\theta}\right)^2\right]} \\
&= \sqrt{E\left[\left(\frac{\partial lnL}{\partial\theta}\right)^2\right]} \\
&= \sqrt{I(\theta)}
\end{aligned} \tag{25}$$

- For any θ, $\pi(\theta) = \sqrt{I(\theta)}$; with this choice, the information is constant under changes of variable
- Such priors are called Jeffreys priors, and assume different forms depending on the type of parametrization
  - Location parameters: uniform prior
  - Scale parameters: prior $\propto \frac{1}{\theta}$
  - Poisson processes: prior $\propto \frac{1}{\sqrt{\theta}}$

Sufficient statistic and data reduction

- A test statistic is a function of the data (a quantity derived from the data sample)
- A statistic $T = T(\boldsymbol{X})$ is sufficient for θ if the density function $f(\boldsymbol{X}|T)$ is independent of θ
  - If T is a sufficient statistic for θ, then also any strictly monotonic $g(T)$ is sufficient for θ
- The statistic T carries as much information about θ as the original data $\boldsymbol{X}$
  - No other function can give any further information about θ
  - Same inference from data $\boldsymbol{X}$ with model E and from sufficient statistic $T(\boldsymbol{X})$ with model E'
- Example: data $1, 2, 3, 4, 5$; sample mean (estimate of population mean) $\hat{x} = \frac{1+2+3+4+5}{5} = 3$
  - Imagine we don't have the data; we only know that the sample mean is 3
  - Since the sample mean is 3, we also estimate the population mean to be 3
  - Knowing the data (the set $1, 2, 3, 4, 5$) or knowing only the sample mean does not improve our estimate for the population mean
- Data can be reduced; we only need to store a sufficient statistic
  - Binomial test in coin toss
  - Record heads and tails, with their order: *HTTHHHTHHTTTHTHTH*
  - Recording only the number of heads (no tails, no order) gives exactly the same information
  - Storage needs are reduced

The Likelihood Principle

- Common enunciation: given a set of observed data $\vec{x}$, the likelihood function $L(\vec{x};\theta)$ contains all the information relevant to the measurement of θ
  - The likelihood function is seen as a function of θ, for a fixed set (a particular realization) of observed data $\vec{x}$
  - As we have seen, the likelihood is used to define the information contained in a sample
- Bayesian statistics normally complies, frequentist statistics usually does not, because a frequentist has to consider the hypothetical set of data that might have been obtained.
- This on one side implies that a frequentist always needs multiple sets of observations (simulations of the day of tomorrow, or counting the past frequency of la abuela con dolor a la espalda)
- On the other side a Bayesian would say “Probably tomorrow will rain”, a frequentist “the sentence -tomorrow it will rain- is probably true”

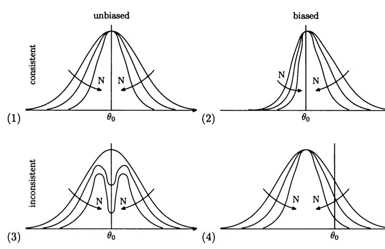
Estimators

- Set $\vec{x} = (x_1, ..., x_N)$ of N statistically independent observations x_i, sampled from a p.d.f. $f(x)$.
- Mean and width of $f(x)$ (or some parameter of it: $f(x;\vec{\theta})$, with $\vec{\theta} = (\theta_1, ..., \theta_M)$ unknown)
  - In case of a linear p.d.f., the vector of parameters would be $\vec{\theta} = (intercept, slope)$
- We call estimator a function of the observed data $\vec{x}$ which returns numerical values $\hat{\vec{\theta}}$ for the vector $\vec{\theta}$.

- $\hat{\vec{\theta}}$ is (asymptotically) consistent if it converges to $\vec{\theta}_{true}$ for large N:

$$\lim_{N\to\infty} \hat{\vec{\theta}} = \vec{\theta}_{true}$$

- $\hat{\vec{\theta}}$ is unbiased if its bias is zero, $\vec{b} = 0$
  - Bias of $\hat{\vec{\theta}}$: $\vec{b} := E[\hat{\vec{\theta}}] - \vec{\theta}_{true}$
  - If bias is known, can redefine $\hat{\vec{\theta}}' = \hat{\vec{\theta}} - \vec{b}$, resulting in $\vec{b}' = 0$.



Plot from James, 2nd ed.

- $\hat{\vec{\theta}}$ is efficient if its variance $V[\hat{\vec{\theta}}]$ is the smallest possible
- An estimator is robust when it is insensitive to small deviations from the underlying distribution (p.d.f.) assumed (ideally, one would want distribution-free estimates, without assumptions on the underlying p.d.f.)

The Maximum Likelihood Method 1/

- Let $\vec{x} = (x_1, ..., x_N)$ be a set of N statistically independent observations x_i, sampled from a p.d.f. $f(x; \vec{\theta})$ depending on a vector of parameters
- Under independence of the observations, the likelihood function factorizes to the individual p.d.f. s

$$L(\vec{x}; \vec{\theta}) = \prod_{i=1}^{N} f(x_i, \vec{\theta}) \tag{26}$$

- The maximum-likelihood estimator is the $\vec{\theta}_{ML}$ which maximizes the joint likelihood

$$\vec{\theta}_{ML} := argmax_{\theta}\left(L(\vec{x}, \vec{\theta})\right) \tag{27}$$

  - The maximum must be global
- Numerically, it's usually easier to minimize

$$-lnL(\vec{x}; \vec{\theta}) = -\sum_{i=1}^{N} lnf(x_i, \vec{\theta}) \tag{28}$$

  - Easier working with sums than with products
  - Easier minimizing than maximizing
- If the minimum is far from the range of permitted values for $\vec{\theta}$, then the minimization can be performed by finding solutions to

$$-\frac{lnL(\vec{x}; \vec{\theta})}{\partial\theta_j} = 0 \tag{29}$$

  - It is assumed that the p.d.f. s are correctly normalized, i.e. that $\int f(\vec{x}; \vec{\theta})dx = 1$ (→ integral does not depend on $\vec{\theta}$)

The Maximum Likelihood Method 2/

- Solutions to the likelihood minimization are found via numerical methods such as MINOS
  - Fred James' Minuit: https://root.cern.ch/root/htmldoc/guides/minuit2/Minuit2.html
- $\vec{\theta}_{ML}$ is an estimator → let's study its properties!
  1. **Consistent:** $\lim_{N\to\infty}\vec{\theta}_{ML} = \vec{\theta}_{true}$;
  2. **Unbiased:** only asymptotically. $\vec{b} \propto \frac{1}{N}$, so $\vec{b} = 0$ only for $N \to \infty$;
  3. **Efficient:** $V[\vec{\theta}_{ML}] = \frac{1}{I(\theta)}$
  4. **Invariant:** for change of variables $\psi = g(\theta)$; $\hat{\psi}_{ML} = g(\vec{\theta}_{ML})$
- $\vec{\theta}_{ML}$ is only asymptotically unbiased, and therefore it does not always represent the best trade-off between bias and variance
- Remember that in frequentist statistics $L(\vec{x};\vec{\theta})$ is not a p.d.f. . In Bayesian statistics, the posterior probability is a p.d.f.:

$$P(\vec{\theta}|\vec{x}) = \frac{L(\vec{x}|\vec{\theta})\pi(\vec{\theta})}{\int L(\vec{x}|\vec{\theta})\pi(\vec{\theta})d\vec{\theta}} \tag{30}$$

  - Note that if the prior is uniform, $\pi(\vec{\theta}) = k$, then the MLE is also the maximum of the posterior probability, $\vec{\theta}_{ML} = maxP(\vec{\theta}|\vec{x})$.

Nuclear Decay with Maximum Likelihood Method 1/

- A nuclear decay with half-life τ is described by the p.d.f., expected value, and variance

$$\begin{aligned} f(t;\tau) &= \frac{1}{\tau}e^{-\frac{t}{\tau}} \\ E[f] &= \tau \\ V[f] &= \tau^2 \end{aligned} \tag{31}$$

- Sampling N independent measurements t_i from the same p.d.f. results in a set of measurements identically distributed
- The joint p.d.f. can be factorized

$$f(t_1, ...t_N;\tau) = \prod_i f(t_i;\tau) \tag{32}$$

- For a particular set of N measurements t_i, the p.d.f. can be written as a function of τ only, $L(\tau) := f(t_i;\tau)$
- The logarithm of the likelihood, $lnL(\tau) = \sum \left(ln\frac{1}{\tau} - \frac{t_i}{\tau} \right)$, can be maximized analytically

$$\frac{\partial lnL(\tau)}{\partial \tau} = \sum_i \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) \equiv 0 \tag{33}$$

Nuclear Decay with Maximum Likelihood Method 1/

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N} \sum_i t_i \tag{34}$$

- It's the simple arithmetical mean of the individual measurements!
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0 \tag{35}$$

- The variance interestingly decreases when N increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N} \sum_i t_i\right] = \frac{1}{N^2} \sum_i V[t_i] = \frac{\tau^2}{N} \tag{36}$$

- The MLE is not the only estimator we can think of

	Consistente	Insesgado	Eficiente
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1+\ldots+t_N}{N}$	✓	✓	✓
$\hat{\tau} = \frac{t_1+\ldots+t_N}{N-1}$	✓	✗	✗
$\hat{\tau} = t_i$	✗	✓	✗

Table: Propiedades de diferentes estimadores de la vida media de un decaimiento nuclear.

Bias-variance tradeoff and optimal variance 1/

- We usually want to optimize both bias $\vec{b}$ and variance $V[\hat{\vec{\theta}}]$
- While we can optimize each one separately, optimizing them simultaneously leads to none being optimally optimized, in genreal
  - Optimal solutions in two dimensions are often suboptimal with respect to the optimization of just one of the two properties
- The variance is linked to the width of the likelihood function, which naturally leads to linking it to the curvature of $L(\vec{x};\vec{\theta})$ near the maximum
- However, the curvature of $L(\vec{x};\vec{\theta})$ near the maximum is linked to the Fisher information, as we have seen
- Information is therefore a limiting factor for the variance (no data set contains infinite information, variance cannot collapse to zero)
- Variance of an estimator satisfies the Rao-Cramér-Frechet (RCF) bound

$$V[\hat{\theta}] \geq \frac{1}{\hat{\theta}} \tag{37}$$

Information Inequality – 1

- Rao-Cramer-Frechet (RCF) bound
  $V[\hat{\theta}] \geq \frac{(1+\partial b/\partial\theta)^2}{-E\left[\partial^2 lnL/\partial\theta^2\right]}$
  - In multiple dimensions, this is linked with the Fisher Information Matrix:
    $I_{ij} = E\left[\partial^2 lnL/\partial\theta_i\partial\theta_j\right]$
- Approximations
  - Neglect the bias ($b = 0$)
  - Inequality is an approximate equality (true for large data samples)
- $V[\hat{\theta}] \simeq \frac{1}{-E\left[\partial^2 lnL/\partial\theta^2\right]}$
- Estimate of the variance of the estimate of the parameter!
- $\hat{V}[\hat{\theta}] \simeq \frac{1}{-E\left[\partial^2 lnL/\partial\theta^2\right]|_{\theta=th\hat{e}ta}}$

Bias-variance tradeoff and optimal variance 2/

- For multidimensional parmaeters, we can build the information matrix with elements:

$$
\begin{aligned}
I_{jk}(\vec{\theta}) &= -E\Big[\sum_{i}^{N} \frac{\partial^2 lnf(x_i;\vec{\theta})}{\partial\theta_k \partial\theta_k}\Big] \\
&= N\int \frac{1}{f}\frac{\partial f}{\partial\theta_j}\frac{\partial f}{\partial\theta_k}dx
\end{aligned}
\tag{38}
$$

- (the last equality is due to the integration interval not being dependent on $\vec{\theta}$)

Estimating variance non-analytically

- We have calculated the variance of the MLE in the simple case of the nuclear decay
- Analytic calculation of the variance is not always possible
- Write the variance approximately as:

$$V[\hat{\theta}] \geq \frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{-E\left[\frac{\partial^2 lnL}{\partial \theta^2}\right]} \tag{39}$$

- This expression is valid for any estimator, but if applied to the MLE then we can note $\vec{\theta}_{ML}$ is efficient and asymptotically unbiased
- Therefore, when $N \to \infty$ then $b = 0$ and the variance approximate to the RCF bound, and $\geq$ becomes $\simeq$:

$$V[\vec{\theta}_{ML}] \simeq \frac{1}{-E\left[\frac{\partial^2 lnL}{\partial \theta^2}\right]\Big|_{\theta=\vec{\theta}_{ML}}} \tag{40}$$

How to extract an interval from the likelihood function 1/

- For a Gaussian p.d.f., $f(x;\vec{\theta}) = N(\mu, \sigma)$, the likelihood can be written as:

$$L(\vec{x};\vec{\theta}) = ln\Big[- \frac{(\vec{x} - \vec{\theta})^2}{2\sigma^2} \Big] \qquad (41)$$

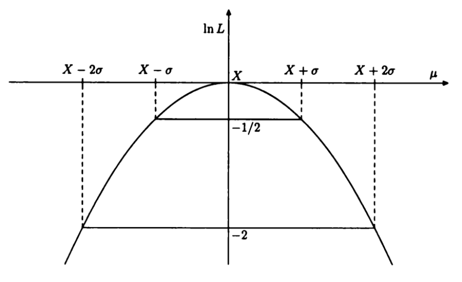
- Moving away from the maximum of $L(\vec{x};\vec{\theta})$ by one unit of σ, the likelihood assumes the value $\frac{1}{2}$, and the area enclosed in $[\vec{\theta} - \sigma, \vec{\theta} + \sigma]$ will be—because of the properties of the Normal distribution—equal to 68.3%.

How to extract an interval from the likelihood function 2/

- We can therefore write

$$\begin{aligned} P\left((\vec{x}-\vec{\theta})^2 \leq \sigma)\right) &= 68.3\% \\ P(-\sigma \leq \vec{x}-\vec{\theta} \leq \sigma) &= 68.3\% \\ P(\vec{x}-\sigma \leq \vec{\theta} \leq \vec{x}+\sigma) &= 68.3\% \end{aligned} \tag{42}$$

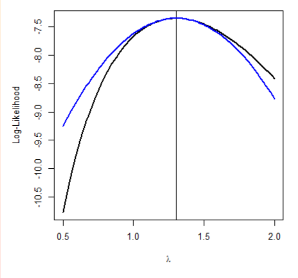
- Taking into account that it is important to keep in mind that probability is a property of sets, in frequentist statistics
  - Confidence interval: interval with a fixed probability content
- This process for computing a confidence interval is exact for a Gaussian p.d.f.
  - Pathological cases reviewed later on (confidence belts and Neyman construction)
- Practical prescription:
  - Point estimate by computing the Maximum Likelihood Estimate
  - Confidence interval by taking the range delimited by the crossings of the likelihood function with $\frac{1}{2}$ (for 68.3% probability content, or 2 for 95% probability content—2σ, etc)

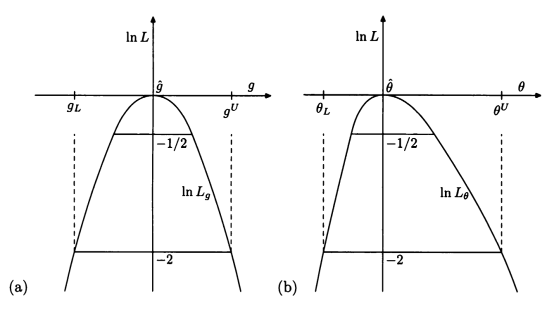


Plot from James, 2nd ed.

How to extract an interval from the likelihood function 3/

- MLE is invariant for monotonic transformations of θ
  - This applies not only to the maximum of the likelihood, but to all relative values
  - The likelihood ratio is therefore an invariant quantity (we'll use it for hypothesis testing)
  - Can transform the likelihood such that $log(L(\vec{x};\vec{\theta}))$ is parabolic, but not necessary (MINOS/Minuit)
- When the p.d.f. is not normal, either assume it is, and use symmetric intervals from Gaussian tails...
  - This yields symmetric approximate intervals
  - The approximation is often good even for small amounts of data
- ...or use asymmetric intervals by just looking at the crossing of the $log(L(\vec{x};\vec{\theta}))$ values
  - Naturally-arising asymmetrical intervals
  - No gaussian approximation
- In any case (even asymmetric intervals) still based on asymptotic expansion
  - Method is exact only to $\mathcal{O}(\frac{1}{N})$





Plot from James, 2nd ed.

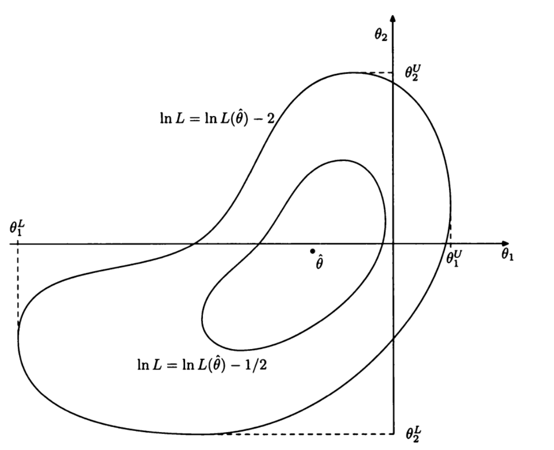
And in many dimensions...

- Construct $log\mathcal{L}$ contours and determine confidence intervals by MINOS
- Elliptical contours correspond to gaussian Likelihoods
  - The closer to MLE, the more elliptical the contours, even in non-linear problems
  - All models are linear in a sufficiently small region
- Nonlinear regions not problematic (no parabolic transformation of $log\mathcal{L}$ needed)
  - MINOS accounts for non-linearities by following the likelihood contour

- Confidence intervals for each parameter

$$\max_{\theta_j, j\neq i} log\mathcal{L}(\theta) = log\mathcal{L}(\hat{\theta}) - \lambda$$

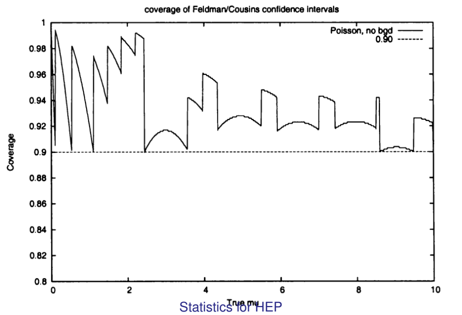
- $\lambda = \frac{Z^2_{1-\beta}}{2}$
  - $\lambda = 1/2$ for $\beta = 0.683$ ("1σ")
  - $\lambda = 2$ for $\beta = 0.955$ ("2σ")



Plot from James, 2nd ed.

What if I have systematic uncertainties?

- Parametrize them into the likelihood function; conventional separation of parameters in two classes
  - the Parameter(s) of Interest (POI), often representing σ/σ_{SM} and denoted as μ (*signal strength*)
  - the parameters representing uncertainties, *nuisance parameters* θ
- H_0: $\mu = 0$ (Standard Model only, no Higgs)
- H_1: $\mu = 1$ (Standard Model + Standard Model Higgs)
- Find the maximum likelihood estimates (MLEs) $\hat{\mu}$, $\hat{\theta}$
- Find the conditional MLE $\hat{\hat{\theta}}(\mu)$, i.e. the value of θ maximizing the likelihood function for each fixed value of μ
- Write the test statistics as $\lambda(\mu) = \frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})}$
  - Independent on the nuisance parameters (profiled, i.e. their MLE has been taken as a function of each value of μ)
  - Can even freeze them one by one to extract their contribution to the total uncertainty
- Asymptotically, $\lambda(\mu) \sim \chi^2$ (Wilks Theorem, under some regularity conditions)



How to extract an interval from the likelihood function 4/

- Theorem: for any p.d.f. $f(x|\vec{\theta})$, in the large numbers limit $N \to \infty$, the likelihood can always be approximated with a gaussian:

$$L(\vec{x};\vec{\theta}) \propto_{N\to\infty} e^{-\frac{1}{2}(\vec{\theta}-\vec{\theta}_{ML})^{T}H(\vec{\theta}-\vec{\theta}_{ML})} \tag{43}$$

- where H is the information matrix $I(\vec{\theta})$.
- Under these conditions, $V[\vec{\theta}_{ML}] \to \frac{1}{I(\vec{\theta}_{ML})}$, and the intervals can be computed as:

$$\Delta lnL := lnL(\theta') - lnL_{max} = -\frac{1}{2} \tag{44}$$

- The resulting interval has in general a larger probability content than the one for a gaussian p.d.f., but the approximation grows better when N increases
  - The interval overcovers the true value $\vec{\theta}_{true}$

How to extract an interval from the likelihood function 5/

- $\vec{\theta}_{true}$ is therefore stimated as $\hat{\theta} = \vec{\theta}_{ML} \pm \sigma$. This is another situation in which frequentist and Bayesian statistics differ in the interpretation of the numerical result
- Frequentist: $\vec{\theta}_{true}$ is fixed
  - "if I repeat the experiment many times, computing each time a confidence interval around $\vec{\theta}_{ML}$, on average 68.3% of those intervals will contain $\vec{\theta}_{true}$"
  - Coverage: "the interval covers the true value with 68.3% probability"
  - Direct consequence of the probability being a property of data sets
- Bayesian: $\vec{\theta}_{true}$ is not fixed
  - "the true value $\vec{\theta}_{true}$ will be in the range $[\vec{\theta}_{ML} - \sigma, \vec{\theta}_{ML} + \sigma]$ with a probabilty of 68.3%"
  - This corresponds to giving a value for the posterior probability of the parameter $\vec{\theta}_{true}$

The Central Limit Theorem

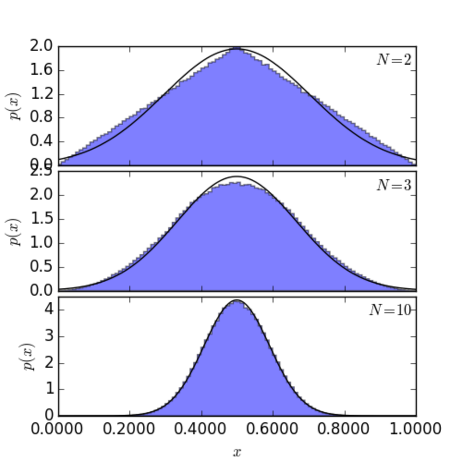
- The convergence of the likelihood $L(\vec{x};\vec{\theta})$ to a gaussian is a direct consequence of the central limit theorem
- Take a set of measurements $\vec{x} = (x_i, ..., x_N)$ affected by experimental errors that results in uncertainties $\sigma_1, ..., \sigma_N$ (not necessarily equal among each other)
- In the limit of a large number of events, $M \to \infty$, the random variable built summing M measurements is gaussian-distributed:

$$Q := \sum_{j=1}^{M} x_j \sim N\Big(\sum_{j=1}^{M} x_j, \sum_{j=1}^{M} \sigma_j^2\Big), \qquad \forall f(x,\vec{\theta}) \tag{45}$$

- The demonstration runs by expanding in series the characteristic function $y_i = \frac{x_j - \mu_j}{\sqrt{\sigma_j}}$
- The theorem is valid for any p.d.f. $f(x,\vec{\theta})$ that is reasonably peaked around its expected value.
  - If the p.d.f. has large tails, the bigger contributions from values sampled from the tails will have a large weight in the sum, and the distribution of Q will have non-gaussian tails
  - The consequence is an alteration of the probability of having sums Q outside of the gaussian

Asymptoticity of the Central limit theorem

- The condition $M \to \infty$ is reasonably valid if the sum is of many small contributions, and M does not need to be very large



Combination of measurements

- Measure N times the same quantity: values x_i and uncertainties σ_i. MLE and variance are:

$$\begin{aligned} \hat{x}_{ML} &= \frac{\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}} \\ \frac{1}{\hat{\sigma}_x^2} &= \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \end{aligned} \tag{46}$$

- The MLE is obtained when each measurement is weighted by its own variance
  - This is because the variance is essentially an estimate of how much information lies in each measurement
- This works if the p.d.f. is known
  - Compare this method with an alternative one that does not assume knowledge of the p.d.f.
  - The second method will be the only one applicable to cases in which the p.d.f. is unknown

Combination of measurements: alternative method 1/

- Take a set of measures sampled from an unknown p.d.f. $f(\vec{x}, \vec{\theta})$
- Compute the expected value and variance of a combination of such measurements described by a function $g(\vec{x})$.
- The expected value and variance of x_i are elementary:

$$\mu = E[x] V_{ij} = E[x_i x_j] - \mu_i \mu_j \tag{47}$$

- If we want to extract the p.d.f. of $g(\vec{x})$, we would normally use the jacobian of the transformation of f to g, but in this case we assumed $f(\vec{x})$ is <u>unknown</u>.

Combination of measurements: alternative method 2/

- We don't know f, but we can still write an expansion in series for it:

$$g(\vec{x}) \simeq g(\vec{\mu}) + \sum_{i=1}^{N} \left(\frac{\partial g}{\partial x_i} \right) \Big|_{x=\mu} (x_i - \mu_i) \tag{48}$$

- We can compute the expected value and variance of g by using the expansion:

$$\begin{aligned} E\big[g(\vec{x})\big] &\simeq g(\mu), \qquad (E[x_i - \mu_i] = 0) \\ \sigma_g^2 &= \sum_{ij=1}^{N} \left[\frac{\partial g}{\partial x_i} \frac{\partial g}{\partial x_j} \right] \Big|_{\vec{x}=\vec{\mu}} V_{ij} \end{aligned} \tag{49}$$

- The variances are propagated to g by means of their jacobian!
- For a sum of measurements, $y = g(\vec{x}) = x_1 + x_2$, the variance of y is $\sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2V_{12}$, which is reduced to the sum of squares for independent measurements

Combination of measurements: example 1/

- Let's compare the two ways of combining measurements, and check the role of the Fisher Information
- Let's estimate the number of married people, N_M, in a given country
  - We have data corresponding to a census that permits us to estimate separately the number of married men N_{MM} and the number of married women N_{MW}:

$$\begin{aligned} N_{HC} &= 10.0 \pm 0.5\ M \\ N_{MC} &= 8 \pm 3\ M \end{aligned} \tag{50}$$

- Evidently, the number of married people is $N_M = N_{MM} + N_{MW}$, and we can apply Eq. 49
  - $N_M = 10.0 + 8 \pm \sqrt{3^2 + 0.5^2}\ M = 18 \pm 3\ M$, corresponding to a precision of $\frac{\sigma_{N_M}}{N_M} \sim 17\%$.

Combination of measurements: example 2/

- Imagine the country is somehow incivil, and the marriage can be only between a woman and a man
- We can use this additional information to note that in this case the two estimates N_{MM} and N_{MW} are independent estimates of the same physical quantity $\frac{N_M}{2}$
- We can therefore use Eq. 46 to compute $\frac{N_M}{2}$ and multiply the result by 2, obtaining

$$N_M = 20 \pm 1\ M \tag{51}$$

- This estimate corresponds to a precision of only 5%!!!
- The dramatic improvement in the precision of the measurement, from 17% to 5%, is a direct consequence of having used additional information under the form of a relationship (constraint) between the two available measurements.
- A good physicist exploits as many constraints as possible in order to improve the precision of a measurement
  - Sometimes the contraints are arbitrary or correspond to special cases
  - Is is very important to explicitly mention any constraint used to derive a measurement, when quoting the result.

Early afternoon: finding a new particle

What is an hypothesis...

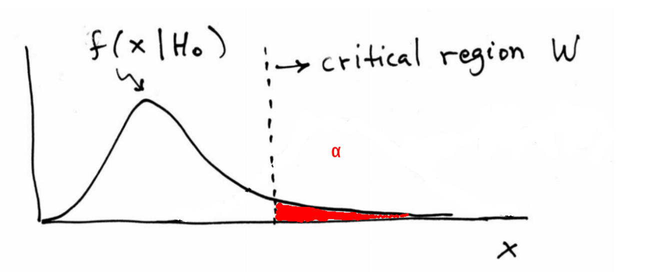
- Is our hypothesis compatible with the experimental data? By how much?
- Hypothesis: a complete rule that defines probabilities for data.
  - An hypothesis is simple if it is completely specified (or if each of its parameters is fixed to a single value)
  - An hypothesis is complex if it consists in fact in a family of hypotheses parameterized by one or more parameters
- “Classical” hypothesis testing is based on frequentist statistics
  - An hypothesis—as we do for a parameter $\vec{\theta}_{true}$—is either true or false. We might improperly say that $P(H)$ can only be either 0 or 1
  - The concept of probability is defined only for a set of data $\vec{x}$
- We take into account probabilities for data, $P(\vec{x}|H)$
  - For a fixed hypotesis, often we write $P(\vec{x};H)$, skipping over the fact that it is a conditional probability
  - The size of the vector $\vec{x}$ can be large or just 1, and the data can be either continuos or discrete.

...and how do we test it?

- The hypothesis can depend on a parameter
  - Technically, it consists in a family of hypotheses scanned by the parameter
  - We use the parameter as a proxy for the hypothesis, $P(\vec{x};\theta) := P(\vec{x};H(\theta))$.
- We are working in frequentist statistics, so there is no $P(H)$ enabling conversion from $P(\vec{x}|\theta)$ to $P(\theta|\vec{x})$.
- Statistical test
  - A statistical test is a proposition concerning the compatibility of H with the available data.
  - A binary test has only two possible outcomes: either accept or reject the hypothesis

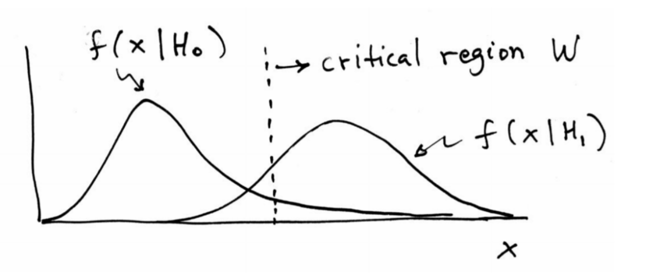
Testing the world as we know it...

- Suppose we want to test an hypothesis H_0
- H_0 is normally the hypothesis that we assume true in absence of further evidence
- Let $\mathbf{X}$ be a function of the observations (called "*test statistic*")
- Let W be the space of all possible values of $\mathbf{X}$, and divide it into
  - A critical region w: observations X falling into w are regarded as suggesting that H_0 is NOT true
  - A region of acceptance $W - w$
- The size of the critical region is adjusted to obtain a desired *level of significance* α
  - Also called *size of the test*
  - $P(X \in w|H_0) = \alpha$
  - α is the probability of rejecting H_0 when H_0 is actually true
- Once $\mathcal{W}$ is defined, given an observed value $\vec{x}_{obs}$ in the space of data, we define the test by saying that we reject the hypothesis H_0 if $\vec{x}_{obs} \in W$.
- If $\vec{x}_{obs}$ is inside the critical region, then H_0 is rejected; in the other case, H_0 is accepted
  - In this context, accepting H_0 does not mean demonstrating its truth, but simply not rejecting it
- Choosing a small α is equivalente to giving a priori preference to H_0!!!



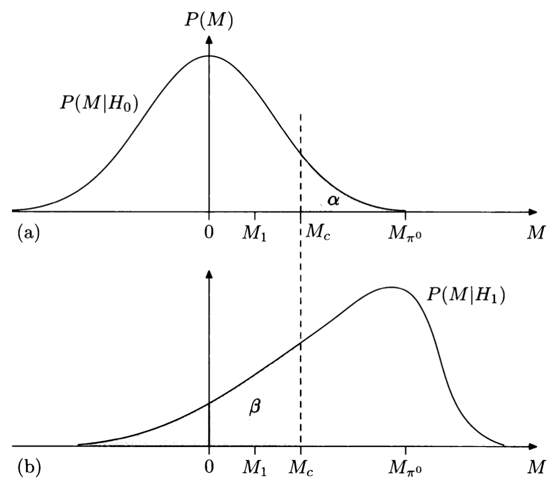
...while introducing some spice in it

- The definition of $\mathcal{W}$ depends only on its area α, without any other condition
  - Any other area of area α can be defined as critical region, independently on how it is placed with respect to $\vec{x}_{obs}$
  - In particular, for an infinite number of choices of $\mathcal{W}$, the point $\vec{x}_{obs}$—which beforehand was situated outside of $\mathcal{W}$—is now included inside the critical region
  - In this condition, the result of the test switches from accept H_0 to reject H_0
- To remove or at least reduce this arbitrariness in the choice of $\mathcal{W}$, we introduce the alternative hypothesis, H_1
- The idea is to choose the critical region so that the probability of a point $\vec{x}$ being inside $\mathcal{W}$ be α under H_0, and that it is as large as possible under H_1



A small example

- H_0: $pp \rightarrow pp$ elastic scattering
- H_1: $pp \rightarrow pp\pi^0$
- Compute the missing mass M (as total rest energy of unseen particles)
- Under H_0, $M = 0$
- Under H_1, $M = 135\ MeV$

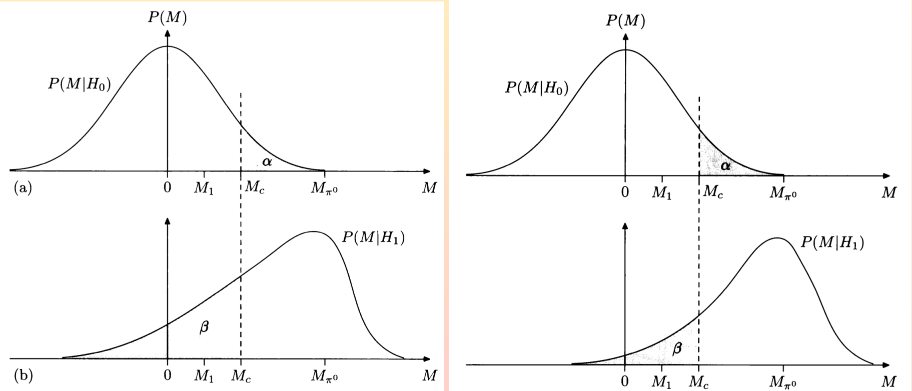


	Choose H_0	Choose H_1
H_0 is true	$1-\alpha$	α (Type I error)
H_1 is true	β (Type II error)	$1-\beta$

Plot from James, 2nd ed.

Basic hypothesis testing – 4

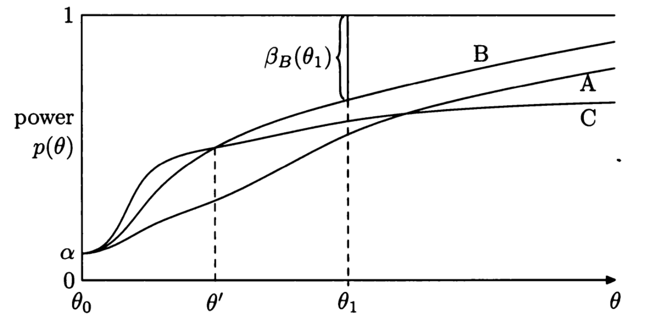
- The usefulness of the test depends on how well it discriminates against the alternative hypothesis
- The measure of usefulness is the *power of the test*
  - $P(X \in w|H_1) = 1 - \beta$
  - Power $(1 - \beta)$ is the probabiliity of X falling into the critical region if H_1 is true
  - $P(X \in W - w|H_1) = \beta$
  - β is the probability that X will fall into the acceptance region if H_1 is true
- NOTE: some authors use β where we use $1 - \beta$. Pay attention, and live with it.



Plots from James, 2nd ed.

Comparing tests

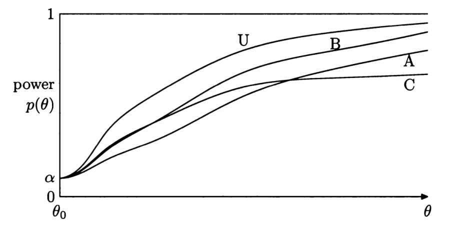
- For parametric (families of) hypotheses, the power depends on the parameter
  - $H_0 : \theta = \theta_0$
  - $H_1 : \theta = \theta_1$
  - Power: $p(\theta_1) = 1 - \beta$
- Generalize for all possible alternative hypotheses: $p(\theta) = 1 - \beta(\theta)$
  - For the null, $p(\theta_0) = 1 - \beta(\theta_0) = \alpha$

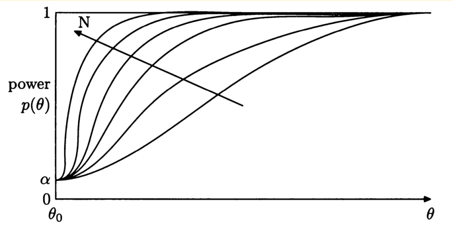


Plot from James, 2nd ed.

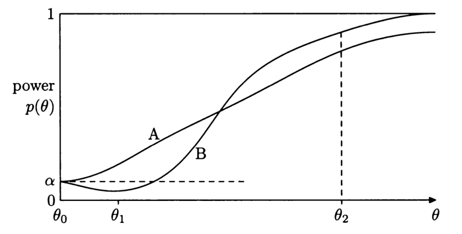
Properties of tests

- More powerful test: a test which at least as powerful as any other test for a given θ
- Uniformly more powerful test: a test which is the more powerful test for any value of θ
  - A less powerful test might be preferrable if more robust than the UMP[1]
- If we increase the number of observations, it makes sense to require consistency
  - The more observations we add, the more the test distinguishes between the two hypotheses
  - Power function tends to a step function for $N \to \infty$





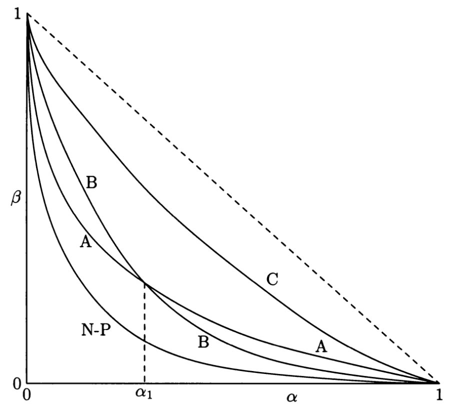
- Biased test: $argmin(p(\theta)) \neq \theta_0$
- More likely to accept H_0 when it is false than when it is true
- Big no-no for θ_0 vs $\theta_1]$
- Still useful (larger power) for θ_0 vs θ_2



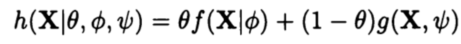
Plot from James, 2nd ed.

[1] Robust: a test with low sensitivity to unimportant changes of the null hypothesis

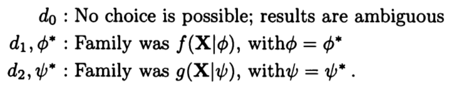
Play with Type I (α) and Type II (β) errors freely



- Comparing only based on the power curve is asymmetric w.r.t. α
- For each value of $\alpha = p(\theta_0)$, compute $\beta = p(\theta_1)$, and draw the curve
  - Unbiased tests fall under the line $1 - \beta = \alpha$
  - Curves closer to the axes are better tests
- Ultimately, though, choose based on the cost function of a wrong decision
  - Bayesian decision theory



$$h(\mathbf{X}|\theta,\phi,\psi) = \theta f(\mathbf{X}|\phi) + (1-\theta) g(\mathbf{X},\psi)$$



d_0 : No choice is possible; results are ambiguous
d_1, ϕ^* : Family was $f(\mathbf{X}|\phi)$, with$\phi = \phi^*$
d_2, ψ^* : Family was $g(\mathbf{X}|\psi)$, with$\psi = \psi^*$.

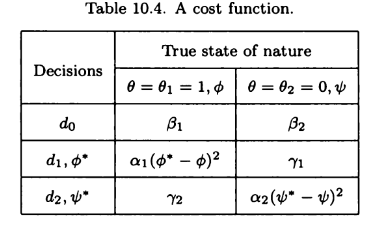


Table 10.4. A cost function.

Decisions	True state of nature	
	$\theta = \theta_1 = 1, \phi$	$\theta = \theta_2 = 0, \psi$
d_0	β_1	β_2
d_1, ϕ^*	$\alpha_1(\phi^* - \phi)^2$	γ_1
d_2, ψ^*	γ_2	$\alpha_2(\psi^* - \psi)^2$

Plot from James, 2nd ed.

Find the most powerful test

- Testing simple hypotheses H_0 vs H_1, find the best critical region
- Maximize power curve $1 - \beta = \int_{w_\alpha} f(\mathbf{X}|\theta_1) d\mathbf{X}$, given $\alpha = \int_{w_\alpha} f(\mathbf{X}|\theta_0) d\mathbf{X}$
- The best critical region w_α consists in the region satisfying the likelihood ratio equation

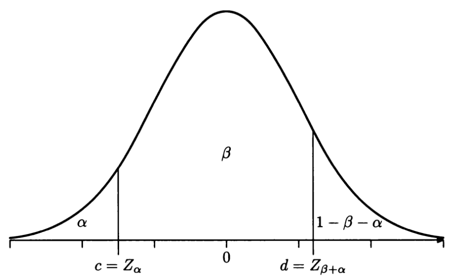
$$\ell(\mathbf{X}, \theta_0, \theta_1) := \frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)} \geq c_\alpha$$

- The criterion, called Neyman-Pearson test is therefore
  - If $\ell(\mathbf{X}, \theta_0, \theta_1) > c_\alpha$ then choose H_1
  - If $\ell(\mathbf{X}, \theta_0, \theta_1) \leq c_\alpha$ then choose H_0
- The likelihood ratio must be calculable for any $\mathbf{X}$
  - The hypotheses must therefore be completely specified simple hypotheses
  - For complex hypotheses, ℓ is not necessarily optimal

Confidence intervals!

- Confidence interval for θ with probability content β
  - The range $\theta_a < \theta < \theta_b$ containing the true value θ_0 with probability β
  - The physicists sometimes improperly say the uncertainty on the parameter θ
- Given a p.d.f., the probability content is $\beta = P(a \leq X \leq b) = \int_a^b f(X|\theta)dX$
- If θ is unknown (as is usually the case), use auxiliary variable $Z = Z(X,\theta)$ with p.d.f. $g(Z)$ independent of θ
- If Z can be found, then the problem is to estimate interval $P(\theta_a \leq \theta_0 \leq \theta_b) = \beta$
  - Confidence interval
  - A method yielding an interval satisfying this property has coverage

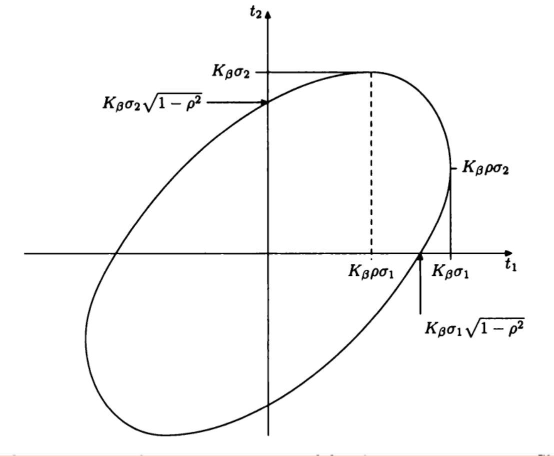
- Example: if $f(X|\theta) = N(\mu,\sigma^2)$ with unknown μ, σ, choose $Z = \frac{X-\mu}{\sigma}$
- Find $[c,d]$ in $\beta = P(c \leq Z \leq d) = \Phi(d) - \Phi(c)$ by finding $[Z_\alpha, Z_{\alpha+\beta}]$
- Infinite interval choices: here central interval $\alpha = \frac{1-\beta}{2}$



Plot from James, 2nd ed.

Confidence intervals in many dimensions

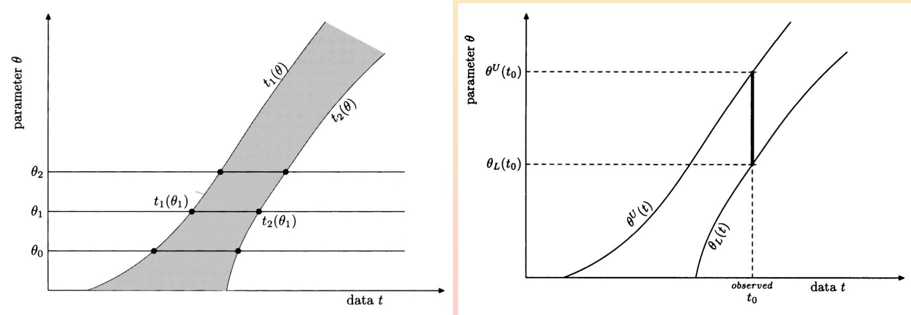
- Generalization to multidimensional $\boldsymbol{\theta}$ is immediate
- Probability statement concerns the whole $\boldsymbol{\theta}$, not the individual θ_i
- Shape of the ellipsoid governed by the correlation coefficient (or the mutual information) between the parameters
- Arbitrariety in the choice of the interval is still present



Plot from James, 2nd ed.

Confidence belts: the Neyman construction

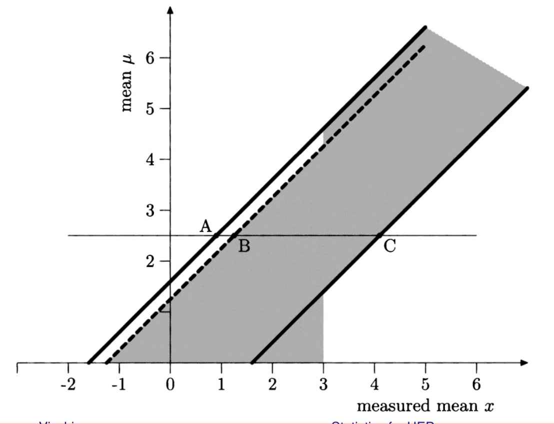
- Unique solutions to finding confidence intervals are infinite
  - Central intervals, lower limits, upper limits, etc
- Let's suppose we have chosen a way
- Build horizontally: for each (hypothetical) value of θ, determine $t_1(\theta)$, $t_2(\theta)$ such that $\int_t 1^t 2P(t|\theta)dt = \beta$
- Read vertically: from the observed value t_0, determine $[\theta_L, \theta^U]$ by intersection
  - The resulting interval might be disconnected in severely non-linear cases
- Probability content statements to be seen in a frequentist way
  - Repeating many times the experiment, the fraction of $[\theta_L, \theta^U]$ containing θ_0 is β



Plot from James, 2nd ed.

Upper limits for non-negative parameters

- Gaussian measurement (variance 1) of a non-negative parameter $\mu \sim 0$ (physical bound)
- Individual prescriptions are self-consistent
  - 90% central limit (solid lines)
  - 90% upper limit (single dashed line)
- Other choices are problematic (flip-flopping): never choose after seeing the data!
  - “quote upper limit if x_{obs} is less than 3σ from zero, and central limit above” (shaded)
  - Coverage not guaranteed anymore (see e.g. $\mu = 2.5$)
- Unphysical values and empty intervals: choose 90% central interval, measure $x_{obs} = -2.0$
  - Don’t extrapolate to an unphysical interval for the true value of μ!
  - The interval is simply empty, i.e. does not contain any allowed value of μ
  - The method still has coverage (90% of other hypothetical intervals would cover the true value)

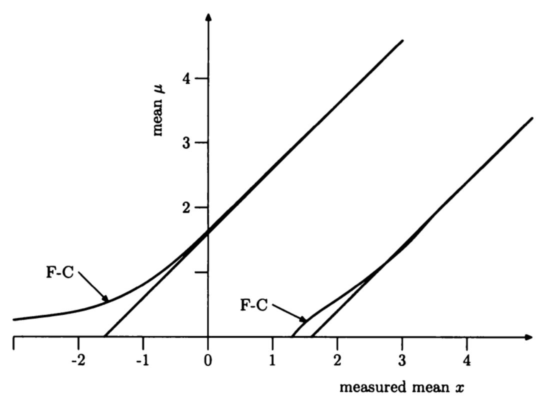


Plot from James, 2nd ed.

Unphysical values: Feldman-Cousins

- The Neyman construction results in guaranteed coverage, but choice still free on how to fill probability content
  - Different ordering principles are possible (e.g. central/upper/lower limits)
- Unified approach for determining interval for $\mu = \mu_0$: the likelihood ratio ordering principle
  - Include in order by largest $\ell(x) = \frac{P(x|\mu_0)}{P(x|\hat{\mu})}$
  - $\hat{\mu}$ value of μ which maximizes $P(x|\mu)$ within the physical region
  - $\hat{\mu}$ remains equal to zero for $\mu < 1.65$, yielding deviation w.r.t. central intervals

- Minimizes Type II error (likelihood ratio for simple test is the most powerful test)
- Solves the problem of empty intervals
- Avoids flip-flopping in choosing an ordering prescription



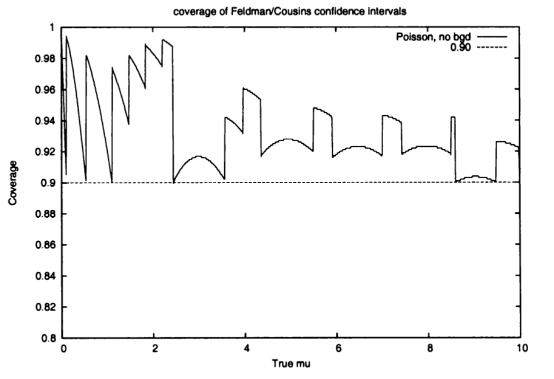
Plot from James, 2nd ed.

Feldman-Cousins in HEP

- The most typical HEP application of F-C is confidence belts for the mean of a Poisson distribution
- Discreteness of the problem affects coverage
- When performing the Neyman construction, will add discrete elements of probability
- The exact probability content won't be achieved, must accept overcoverage

$$\int_{x_1}^{x_2} f(x|\theta)dx = \beta \qquad \rightarrow \qquad \sum_{i=L}^{U} P(x_i|\theta) \geq \beta$$

- Overcoverage larger for small values of μ (but less than other methods)



Plot from James, 2nd ed.

Bayesian intervals

- Often numerically identical to frequentist confidence intervals
  - Particularly in the large sample limit
- Interpretation is different: credible intervals
- Posterior density summarizes the complete knowledge about θ

$$\pi(\theta|\boldsymbol{X}) = \frac{\prod_{i=1}^{N} f(X_i, \theta)\pi(\theta)}{\int \prod_{i=1}^{N} f(X_i, \theta)\pi(\theta)d\theta}$$

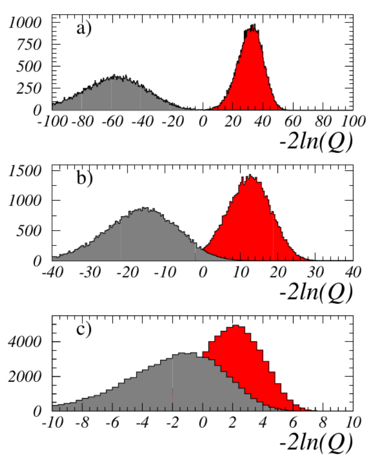
- An interval $[\theta_L, \theta^U]$ with content β defined by $\int_{\theta_L}^{\theta^U} \pi(\theta|\boldsymbol{X})d\theta = \beta$
- Bayesian statement! $P(\theta_L < \theta < \theta^U = \beta$
  - Again, non unique
- Issues with empty intervals don't arise, though, because the prior takes care of defining the physical region in a natural way!
  - But this implies that central intervals cannot be seamlessly converted into upper limits
  - Need the notion of shortest interval
  - Issue of the metric (present in frequentist statistic) solved because here the preferred metric is defined by the prior

Beyond coverage: CLs

- Goal: seamless transition between exclusion, observation, discovery (historically for the Higgs)
  - Exclude Higgs as strongly as possible in its absence (in a region where we would be sensitive to its presence)
  - Confirm its existence as strongly as possible in its presence (in a region where we are sensitive to its presence)
  - Maintain Type I and Type II errors below specified (small) levels
- Identify observables, and a suitable test statistic Q
- Define rules for exclusion/discovery, i.e. ranges of values of Q leading to various conclusions
  - Specify the significance of the statement, in form of confidence level (CL)
- Confidence limit: value of a parameter (mass, xsec) excluded at a given confidence level CL
  - A confidence limit is an upper(lower) limit if the exclusion confidence is greater(less) than the specified CL for all values of the parameter below(above) the confidence limit
- The resulting intervals are neither frequentist nor bayesian!

Get your confidence levels right

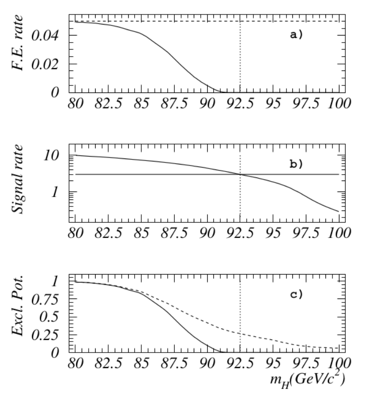
- Find a monotonic Q for increasing signal-like experiments (e.g. likelihood ratio)
- $CL_{s+b} = P_{s+b}(Q \leq Q_{obs})$
  - Small values imply poor compatibility with $S + B$ hypothesis, favouring B-only
- $CL_b = P_b(Q \leq Q_{obs})$
  - Large (close to 1) values imply poor compatibility with B-only, favouring $S + B$
- What to do when the estimated parameter is unphysical?
  - The same issue solved by Feldman-Cousins
  - If there is also underfluctuation of backgrounds, it's possible to exclude even zero events at 95$CL!
  - It would be a statement about future experiments
  - Not enough information to make statements about the signal
- Normalize the $S + B$ confidence level to the B-only confidence level!



Plot from Read, CERN-open-2000-205

Avoid issues at low signal rates

- $CL_s := \frac{CL_{s+b}}{CL_b}$
- Exclude the signal hypothesis at confidence level CL if $1 - CL_s \leq CL$
- Ratio of confidences is not a confidence
  - The hypotetical false exclusion rate is generally less than the nominal $1 - CL$ rate
  - CL_s and the actual false exclusion rate grow more different the more $S + B$ and B p.d.f. become similar
- CL_s increases coverage, i.e. the range of parameters that can be exclude is reduced
  - It is more conservative
  - Approximation of the confidence in the signal hypothesis that might be obtained if there was no background
- Avoids the issue of CL_{s+b} with experiments with the same small expected signal
  - With different backgrounds, the experiment with the larger background might have a better expected performance

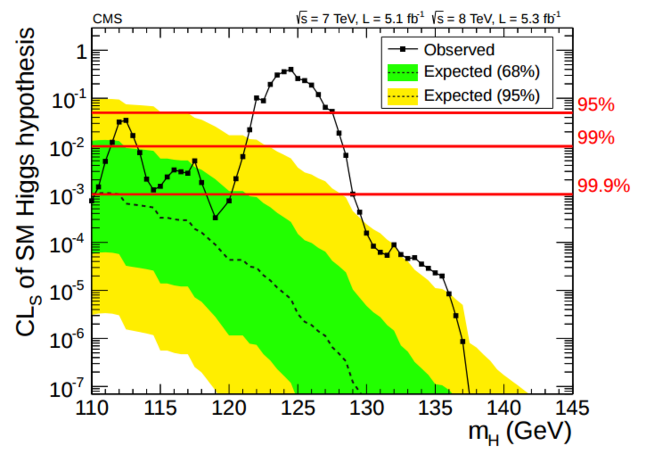


Dashed: CL_{s+b}
Solid: CL_s
$S < 3$: exclusion for a B-free search $\equiv 0$

Plot from Read, CERN-open-2000-205

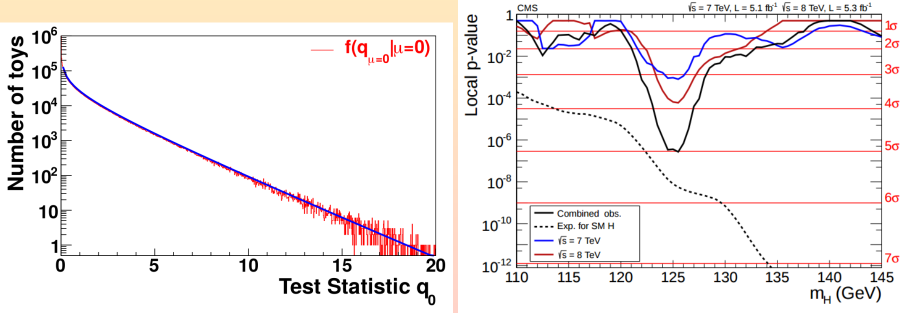
A practical example: Higgs discovery - 1

- Apply the CL_s method to each Higgs mass point
- Green/yellow bands indicate the $\pm 1\sigma$ and $\pm 2\sigma$ intervals for the expected values under B-only hypothesis



Quantifying excesses

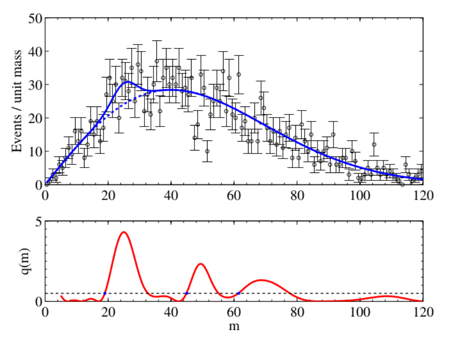
- Quantify the presence of the signal by using the background-only p-value
  - Probability that the background fluctuates yielding and excess as large or larger of the observed one
- For the mass of a resonance, $q_0 = -2log\frac{\mathcal{L}(data|0,\hat{\theta}_0)}{\mathcal{L}(data|\hat{\mu},\hat{\theta})}$, with $\hat{\mu} \geq 0$
  - Interested only in upwards fluctuation, accumulate downwards one to zero
- Use pseudo-data to generate background-only Poisson counts and nuisance parameters θ_0^{obs}
  - Use distribution to evaluate tail probability $p_0 = P(q_0 \leq q_0^{obs})$
  - Convert to one-sided Gaussian tail areas by inverting $p = \frac{1}{2}P_{\chi_1^2}(Z^2)$

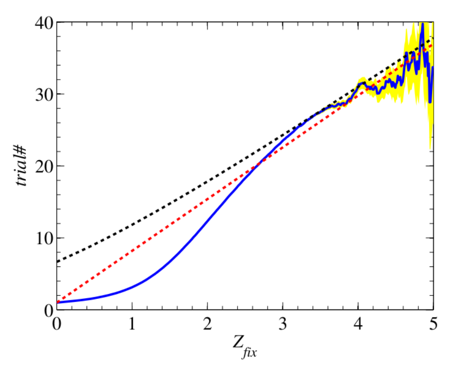


Plots from ATL-PHYS-PUB-2011-011 and from Higgs discovery

The Look-elsewhere effect

- Searching for a resonance X of arbitrary mass
  - $H_0 =$ no resonance, the mass of the resonance is not defined (Standard Model)
  - $H_1 = H(M \neq 0)$, but there are infinte possible values of M
- Wilks theorem not valid anymore, no unique test statistic encompassing every possible H_1
- Quantify the compatibility of an observation with the B-only hypothesis
  - $q_0(\hat{m_X}) = \max_{m_X} q_0(m_X)$
  - Write a global p-value as $p_b^{global} := P(q_0(\hat{m_X}) > u) \leq \langle N_u \rangle + \frac{1}{2} P_{\chi^2_1}(u)$
  - u fixed confidence level
  - Crossings computable using pseudo-data (toys)
  - Ratio of global and local p-value: trial factor
  - Asymptoticly linear in the number of search regions and in the fixed significance level





Plot from Gross-Vitells, 10.1140/epjc/s10052-010-1470-8

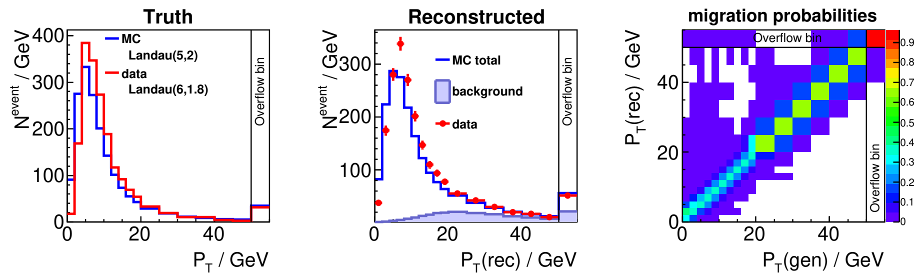
Tea time: measuring differential distributions

Unfolding: the problem

- Unfolding it's about how to invert a matrix that should not be inverted

$$\mathcal{L} = (\mathbf{y} - \mathbf{A}\mathbf{x})^T \mathbf{V}_{\mathbf{yy}} (\mathbf{y} - \mathbf{A}\mathbf{x}),$$

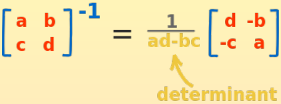
- Observations y, to be transformed in the theory space into x
  - Model the detector as a response matrix
  - Invert the response to convert experimental data to theory space distributions
  - Usually to compare with models in the theory space
- The best solution is to fold any new theory and make comparisons in the experimental data space



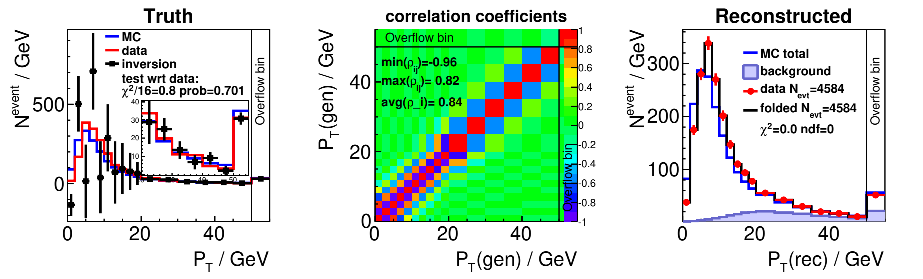
Plot from ArXiv:1611.01927

Unfolding: naïve solutions

- Bin-by-bin correction factors $\hat{x}_i = (y_i - b_i)\frac{N_i^{\text{gen}}}{N_i^{\text{rec}}}$; disfavoured
  - Heavy biases due to the underlying MC truth
  - Yields the wrong normalization for the unfolded distribution
- Invert the response matrix $\hat{\boldsymbol{x}} = \boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{b})$
  - Only for square matrices, but always unbiased
  - Oscillation patterns (small determinants in matrix inversion)
  - Patterns also seen as large negative $\rho_{ij} \sim -1$ near diagonal
  - Result is correct within uncertainty envelope given by $\boldsymbol{V_{xx}}$



$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \underbrace{\frac{1}{ad-bc}}_{\text{determinant}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$



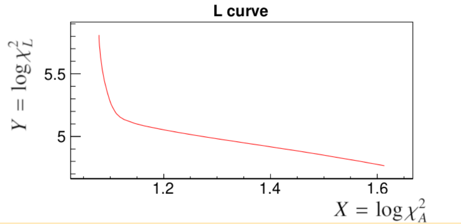
Cartoon from https://www.mathsisfun.com/algebra/matrix-inverse.html, plots from ArXiv:1611.01927

Unfolding: regularization 1/

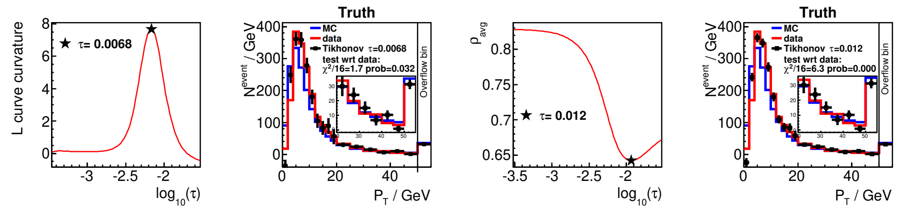
$$\chi^2_{\text{TUnfold}} = \chi^2_A + \tau^2 \chi^2_L$$

$$\chi^2_A = (\boldsymbol{A}\hat{\boldsymbol{x}} + \boldsymbol{b} - \boldsymbol{y})^{\mathsf{T}} (\boldsymbol{V_{yy}})^{-1} (\boldsymbol{A}\hat{\boldsymbol{x}} + \boldsymbol{b} - \boldsymbol{y})$$

$$\chi^2_L = (\hat{\boldsymbol{x}} - \boldsymbol{x_B})^{\mathsf{T}} \boldsymbol{L}^{\mathsf{T}} \boldsymbol{L} (\hat{\boldsymbol{x}} - \boldsymbol{x_B})$$



- Choose τ corresponding to maximum curvature of L-curve
- Or minimize the global $\rho_{\text{avg}} = \frac{1}{M_x} \sum_{j=1}^{M_x} \rho_j$
    - Often results in stronger regularization than L-curve

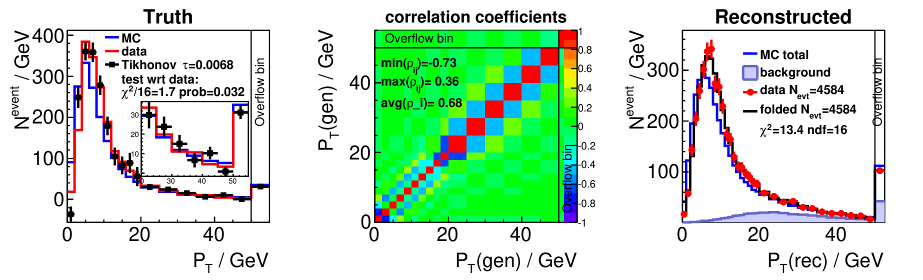


Plots from ArXiv:1611.01927

Unfolding: regularization 2/

$$
\begin{aligned}
\mathcal{L}(\mathbf{x},\lambda) &= \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3,\\
\mathcal{L}_1 &= (\mathbf{y} - \mathbf{A}\mathbf{x})^T \mathbf{V}_{\mathbf{yy}} (\mathbf{y} - \mathbf{A}\mathbf{x}),\\
\mathcal{L}_2 &= \tau^2 (\mathbf{x} - f_b \mathbf{x_0})^T (\mathbf{L}^T \mathbf{L}) (\mathbf{x} - f_b \mathbf{x_0}),\\
\mathcal{L}_3 &= \lambda (Y - \mathbf{e}^T \mathbf{x}),\\
Y &= \sum_i y_i,\\
e_j &= \sum_i A_{ij}.
\end{aligned}
$$

- $\mathbf{y}$: observed yields
- $\mathbf{A}$: response matrix
- $\mathbf{x}$: the unfolded result
- $\mathcal{L}_1$: least-squares minimization ($V_{ij} = e_{ij}/e_{ii}e_{jj}$ correlation coefficients)
- $\mathcal{L}_2$: regularization with strength τ
- Bias vector $f_b \mathbf{x_0}$: reference with respect to which large deviations are suppressed
- $\mathcal{L}_3$; area constraint (bind unfolded normalization to the total yields in folded space)



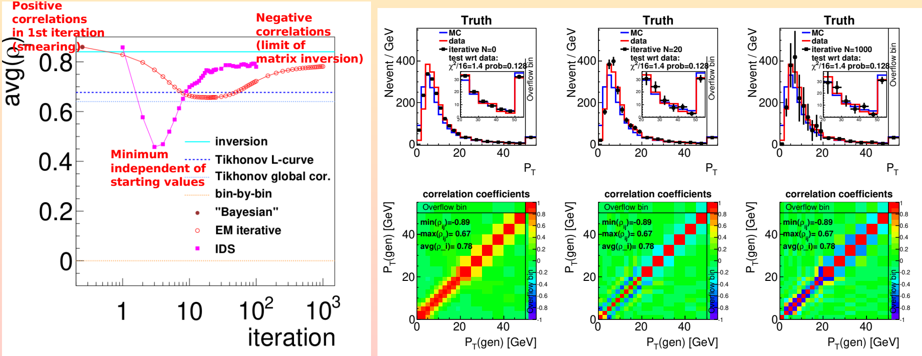
Plots from ArXiv:1611.01927

Unfolding: Iterative Unfolding

- Iterative improvement over the result of a previous iteration;

$$x_j^{(n+1)} = x_j^{(n)} \sum_{i=1}^{M} \frac{A_{ij}}{\epsilon_j} \frac{y_i}{\sum_{k=1}^{N} A_{ik} x_k^{(n)} + b_i}$$

  - It converges (slowly, $N_{iter} \sim N_{bins}^2$) to the MLE of the likelihood for independent Poisson-distributed y_i
  - Not necessarily unbiased for correlated data (does not make use of covariance of input data $\boldsymbol{V_{yy}}$)
- In HEP most people don't iterate until convergence
  - Fixed N_{iter} is often used; the dependence on starting values provides regularization
- Intrinsically frequentist method
  - for $N_{iter} \to \infty$ converges to matrix inversion, if all $\hat{x}_j$ from matrix inversion are positive
  - $N_{iter} = 0$ sometimes called improperly "Bayesian" unfolding (the author, D'Agostini, is Bayesian)
- Don't use software defaults!!! (e.g. some software has $N_{iter} = 4$)
  - Minimizing the global ρ is a good objective criterion, but there are others (Akaike information, etc)



Plots from ArXiv:1611.01927

End of the afternoon: work with difficult final states

Machine learning

- Machine learning is a generalization of fitting functions
- The basics you got today are more important for a small course
- I preferred going more in detail about the basics of point and interval estimation and hypothesis tests
- Leaving Machine Learning for another time ☺

Summary: go home before 18h[2]

Have a healthy 8h/day work schedule
Don’t work outside those hours
Have long nights of sleep
It’s very important!

[2] Except during this Course ☺

Non-exhaustive list of references

- Frederick James: Statistical Methods in Experimental Physics - 2nd Edition, World Scientific
- Glen Cowan: Statistical Data Analysis - Oxford Science Publications
- Louis Lyons: Statistics for Nuclear And Particle Physicists - Cambridge University Press
- Louis Lyons: A Practical Guide to Data Analysis for Physical Science Students - Cambridge University Press
- Annis?, Stuard, Ord, Arnold: Kendall's Advanced Theory Of Statistics I and II
- R.J.Barlow: A Guide to the Use of Statistical Methods in the Physical Sciences - Wiley
- Kyle Cranmer: Lessons at HCP Summer School 2015
- Kyle Cranmer: Practical Statistics for the LHC - http://arxiv.org/abs/1503.07622
- Harrison Prosper: Practical Statistics for LHC Physicists - CERN Academic Training Lectures, 2015 https://indico.cern.ch/category/72/

THANKS FOR THE ATTENTION!

Backup