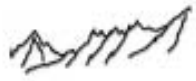
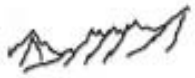


In 2 lessons, it's impossible to cover correctly even the
Fondamentals in Statistics.
Better saying **Introduction to Statistics for Physicists**

IN2P3
Institut National de Physique Nucléaire et de Physique des Particules

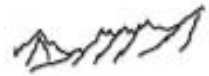
Irfu
cea
saclay



As experimentalists, we need statistics every day

Electronic detector	amplitudes/times → (x,y,z)
Pattern recognition	points → tracks
Particle identification	hypotheses testing
Signal or not signal	hypotheses testing
Extracting results	estimations/fits
Redondancy of measures	combining estimators
Data analysis	Monte-Carlo simulation
	automatic classification (clustering)
	multivariate analyses

How the course is organized



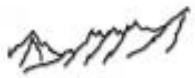
The first lesson (today) will be a quick survey on some main topics:

- Basic concepts
- Point Estimation
- Interval Estimation
- ~~Introduction to Multivariate Discriminants~~

The second lesson (tomorrow) will treat a complete exemple, namely "Observation of a fine structure" through 3 items

- Hypothesis testing
- Classical approach
- Modern approach

I have borrowed many slides from Fred James' course at CERN !
Thank you, Fred !



Classical Statistics have been developed at the end of the 18th century : (GAUSS, BERNOUILLI, BAYES, ...)
During one century (19th) statistics is almost always "Bayesian".

Since then, two major impulses :

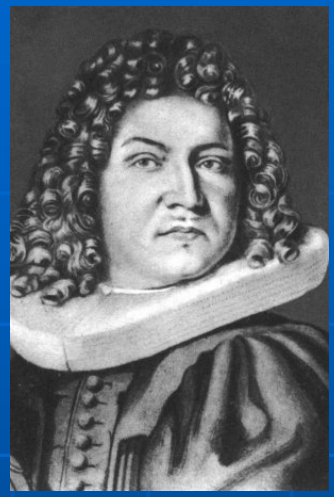
1. begining of the 20th century, the Theory of Probabilities (KOLMOGOROV,...) leads to the frequentist view-point in statistics
   (FISHER, PEARSON(s), NEYMAN, DARMOIS)
2. For the last 50 years or so, due to the computer era (polls)

One can distinguish :

1. Descriptive statistics : any kind of "raw" data
2. Explanatory statistics : estimation, correlations
3. Decision statistics : hypotheses testing
4. Prevision statistics (time is a new parameter) : weather forecast

A portrait gallery

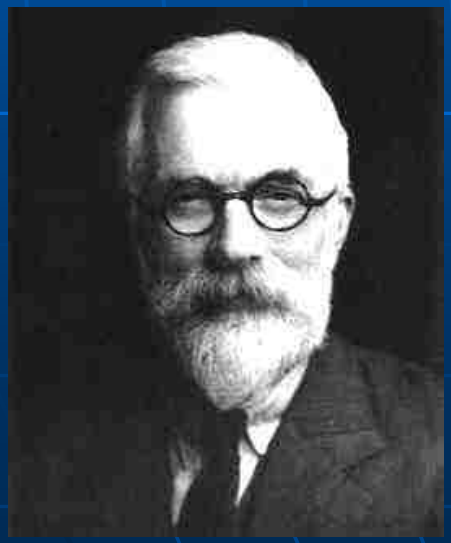
C. F. Gauss



J. Bernoulli

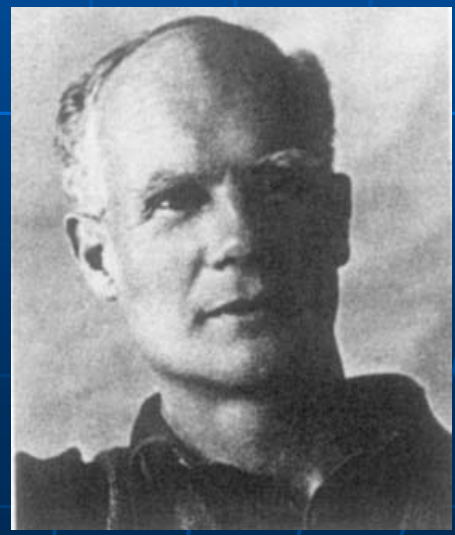
Rev. Bayes

A. Kolmogorov



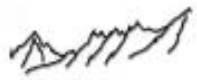
R. A. Fisher

Carl Pearson



Egon Pearson

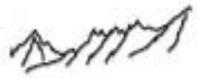
Jerzy Neyman



- The Bible : Kendall et Stuart (Ch. Griffin)
  The advanced theory of statistics, 3 tomes.
- my favorite : Frederick James (World Scientific)
  Statistical Methods in Experimental Physics..
- for the advanced people The PHYSTAT conferences
  `http://www.physics.ox.ac.uk/phystat05/reading.htm`

We will assume that you have sufficient background in Probabilities !
Nevertheless, some remarks may be useful.

Probability



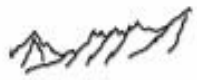
All statistical methods are based on calculations of probability.

- Mathematical probability is an abstract concept which obeys the Kolmogorov axioms.

We will need a specific operational definition. There are two such definitions we can use :

- Frequentist probability is defined as the *limiting frequency* of favourable outcomes in a large number of identical experiments.
- Bayesian probability is defined as a *degree of belief* in a favourable outcome of a single experiment.

Bayesian Probability



For phenomena that are not repeatable, the frequentist approach cannot work. (it's almost always the case if we want a prediction : for example, we want to know the probability that it will rain tomorrow).

Bayesian probability is defined as the degree of belief that the event will happen.
It depends not only on the phenomenon itself, but also on the state of knowledge and beliefs of the observer, and it will in general change with time as the observer gains more knowledge.

We cannot verify if the Bayesian probability is "correct" in terms of counting frequencies.

Bayes' Theorem

Bayes' Theorem says that the probability of both A and B being true simultaneously can be written : $\mathcal{P}(A|B)\mathcal{P}(B) = \mathcal{P}(B|A)\mathcal{P}(A)$, which implies :

$$\mathcal{P}(A|B) = \mathcal{P}(B|A)\mathcal{P}(A)/\mathcal{P}(B)$$

$$p(x|y) = q(y|x)f(x)/g(y)$$

Bayesian statisticians consider parameters as random variables and use the Bayes' theorem also in that case :

$$f(\theta|x) = f(x|\theta)f(\theta)/f(x)$$

where $f(\theta)$ is called the "prior" probability, and $f(\theta|x)$ the "posterior" probability.

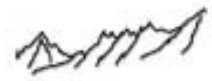
Fundamental Concepts (1)

- The Hypothesis is what we want to test, verify, measure, decide. Examples :
  - H : The standard model is correct
  - H : The tau neutrino is massless

- A Random Variable is data that can take on different values, unpredictable except in probability. $\mathcal{P}(\text{data}|\text{hypothesis})$ is assumed known, provided any "parameters" in the hypothesis are given some values. Example for a POISSON process :

$$\mathcal{P}(N|\mu) = \frac{e^{-\mu}\mu^N}{N!}$$

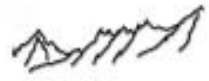
  The possible values of the data (N) are discrete, and μ is the only parameter of the hypothesis.



- Probability Density Function (pdf). When the data are continuous, the probability becomes a pdf, as for the Gaussian

$$\mathcal{P}(x|\mu,\sigma) = \frac{\exp \frac{-(x-\mu)^2}{2\sigma^2}}{\sqrt{2\pi\sigma^2}}$$

where μ and σ are the parameters of the model. Note that μ is the true value of the quantity being measured, while σ is the width of the Gaussian. μ is the parameter of interest, and σ, if unknown, is a nuisance parameter, unknown but unfortunately necessary for the calculation of $\mathcal{P}(\text{data}|\text{hypothesis})$.



The Likelihood function

If in $\mathcal{P}(\text{data}|\text{hypothesis})$, we put in the values of the data observed in the experiment, and consider the resulting function as a function of the (unknown) parameter(s), it becomes

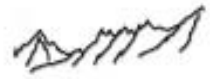
$$\mathcal{P}(\text{data}|\text{hypothesis}) = \mathcal{L}(\text{hypothesis}|\text{data})$$

$\mathcal{L}$ is called the Likelihood function.

R. A. Fisher, who introduced this in 1921, knew that it was not a probability, called it a likelihood.

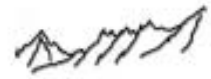
This is THE central concept in statistics

Statistics versus Probabilities



MODEL : a jar containing N marbles (B white and N-B black marbles)

Probabilities	Statistics
DATA : B, N known p =B/N known	p, B, N unknown. One has drawn n marbles and got k white marbles.
QUESTIONS : * Find the probability law of the number of white marbles in n drawings, with or without replacement. * Find the law of the number of drawings necessary for obtaining the first white marble. Expectation of this p.d.f. ...	* Give to p a *reasonable* value point estimation * Find an interval where p has a good chance to be interval estimation * Decide if the true value of p is less then a given value (or between 2 given limits) hypothesis testing
Conclusions : *certains* and in principle with an *arbitrary precision.*	A *certain* answer is impossible The more *precise* the answer, the bigger the prob. that it is *wrong.*



"The probability model, the set of statistical hypotheses, and the data, form a triplet which is the foundation of statistical inference.

Of the many outcomes, each with a specified probability given the hypothesis, which could have occured on the basis of the accepted model, one has occured - the data. What can they reveal about the hypotheses ?"

A. Edwards : Likelihood
Cambridge Univ. Press (1972)

The measuring "errors"(1)

Be $\mathcal{X}$ an observable, X its (unknown) true value. In order to determine X, we make a number of measures x_i pertinent for X, and a computation (estimation) leads to the result x.

Definitions

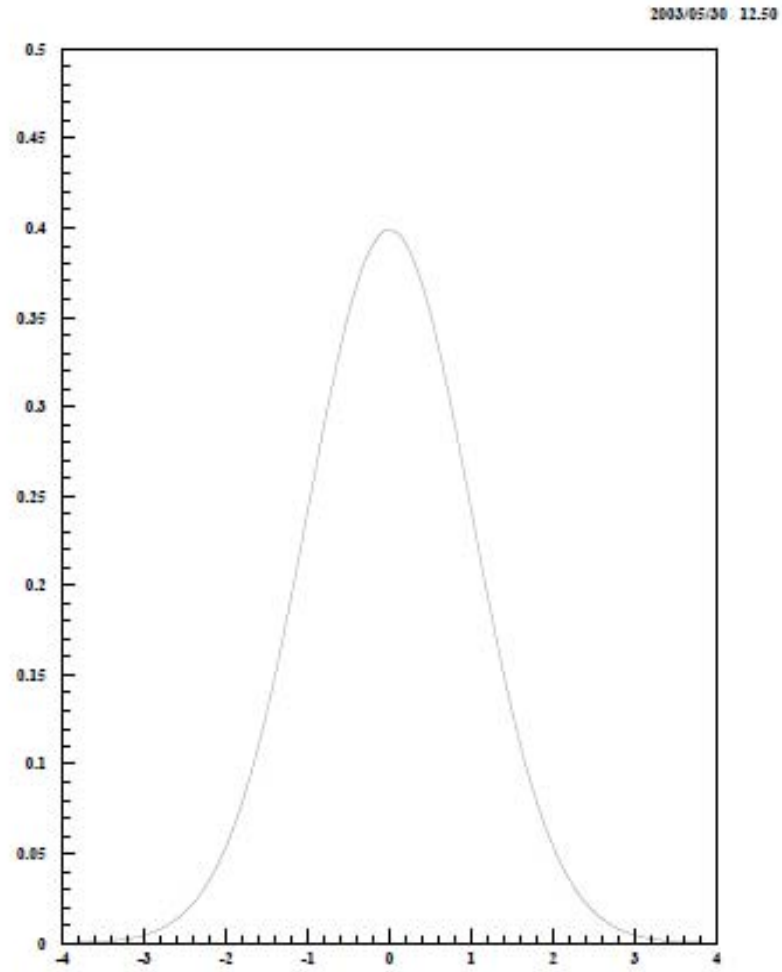
- $\epsilon = X - x$ is the error on X
- $\epsilon_s = E[\epsilon]$ is the systematic error on X
- $\epsilon_a = \epsilon - \epsilon_s$ is the statistical (or accidental) error on X.

With these definitions, x, ϵ and ϵ_a are random var., but ϵ_s is NOT.

Statistical error

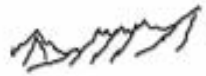
According to the Law of Large Numbers, $\epsilon_a = E[x] - x = \lim_{n\to\infty} \overline{x}_n - x$. The dispersion of ϵ_a is thus the dispersion of the x_i, measures done in the same conditions by the same instrument. It is called the dispersion of the instrument

The measuring "errors"(2)



The ϵ_a p.d.f. is usually taken as a GAUSSIAN.
The smaller the dispersion, the more trusty the apparatus.
2σ is called the precision of the apparatus.
Dispersion (or precision) are easy to estimate, by the usual estimator

$$S_a^2 = \frac{1}{n-1}\sum (x_i - \overline{x})^2$$



Systematic error

Recall : by its definition, $\epsilon_s = E[\epsilon]$, it's NOT a random variable. It is linked to the measuring instrument: the smaller it is, the more accurate the apparatus. It appears more like a bias, contrarily to the statistical error, which looks like the root of a variance.

In order to estimate a systematic error, one should use (gedanken experiment) many measuring devices of the same type, in order to measure the same quantity X. On should moreover admit that all devices are fabricated in such a way that their systematic uncertainty has a zero mean. Thus ϵ_s becomes a random variable ! (somehow bayesian)

So we have k instruments, and a large number of measures for each device. We have thus k means $\overline{x}_j$, where j goes from 1 to k. For each instrument, $\overline{x}_j \xrightarrow{\mathcal{P}} X - \epsilon_j^s$. Consequently, (since $E[\epsilon_s] = 0$), $\frac{1}{k}\sum \epsilon_j^s \xrightarrow{\mathcal{P}} 0$ and then $\hat{x} = \frac{1}{k}\sum \overline{x}_j \xrightarrow{\mathcal{P}} X$. The ϵ_j^s estimator is then $\hat{x} - \overline{x}_j$ and the variance of ϵ_s can be estimated as $\hat{\sigma_s^2} = \frac{1}{k-1}\sum(\overline{x}_j - \hat{x})^2$

Note : ϵ_s and ϵ_a are independent in probability.

Systematics come in 3 types

P. Sinervo (2003)

Type 1 : « The Good »

Can be constrained by other measurements (sideband/auxiliary) ; can then be treated as statistical uncertainties

* scale with luminosity

Type 2 : « The Bad »

Arise from model assumptions in the measurement or from poorly understood features in data or analysis technique.

* eg : « shape » systematics

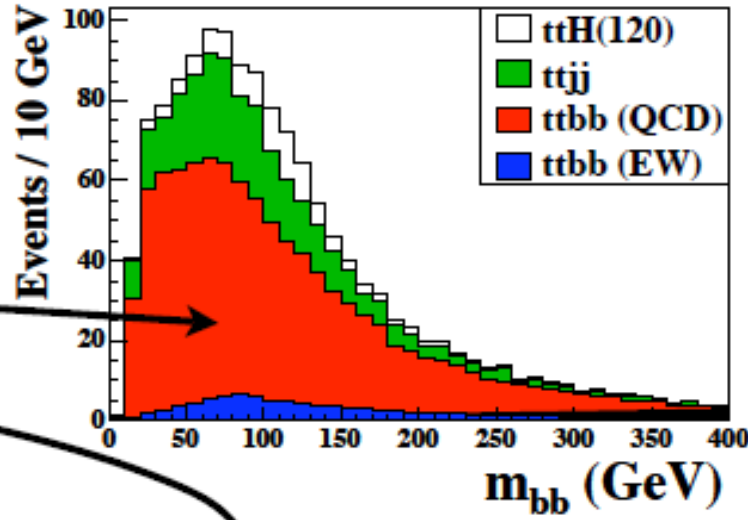
Type 3 : « The Ugly »

Arise from uncertainties in underlying theoretical paradigm used to make the inference (somewhat philosophical)

Type 2 systematics

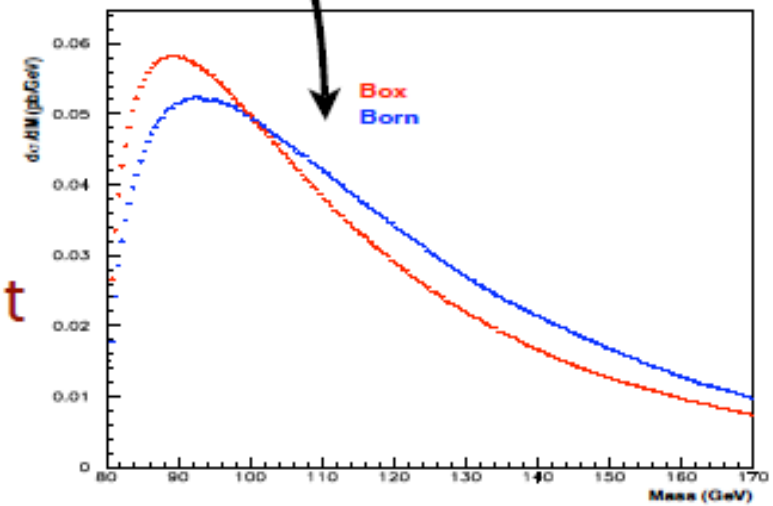
Type II systematics generally due to uncertainty in shape of background

- this uncertainty is limiting factor in $ttH(H \to bb)$ analysis
- also relevant for $H \to \gamma\gamma$



A huge amount of effort goes into identifying other measurements that can be used to estimate or constrain the background

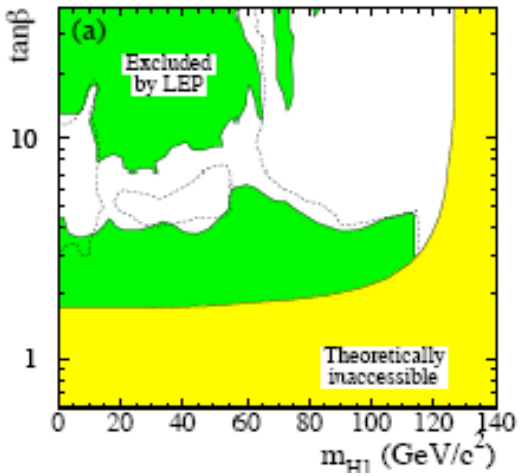
- control samples are an important tool for experimentalists



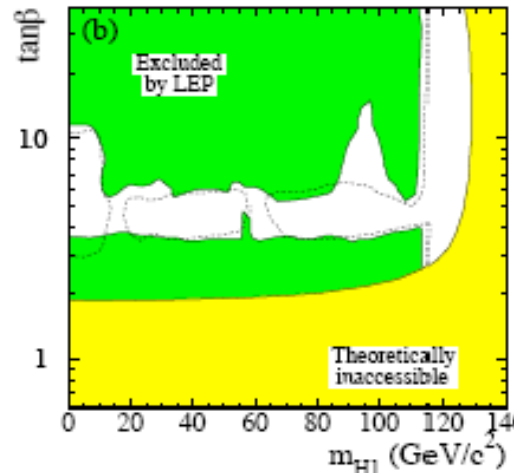
Type 3 systematics

Two theoretical tools used to exclude regions of CPX Higgs scenario using the same measurement & statistical techniques

CPH calculation



FeynHiggs calculation

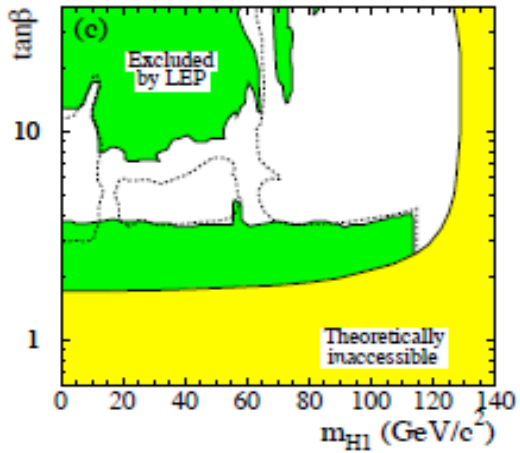


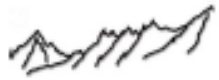
Do we want to weight these plots with a Bayesian prior,

– or –

Do we want to only exclude in the region where they both exclude?

CPH .OR. FeynHiggs





So much for the introduction

Estimation (flying over)

Point Estimation

The goal of Point Estimation is
to find the function of the data X which gives
the "best" estimate (measurement) of the parameter μ.

We assume, as always, $P(\text{data}|\text{hypothesis}) = P(X|\mu)$ known.

What we mean by the "best" estimate depends very much on whether we will use a frequentist or Bayesian method.

Historically, the Bayesian was the first method, so we start there.

Point Estimation - Bayesian

For parameter estimation, we can rewrite Bayes' Theorem:

$$P(\text{hyp}|\text{data}) = \frac{P(\text{data}|\text{hyp})P(\text{hyp})}{P(\text{data})}$$

and if the hypothesis concerns the value of μ:

$$P(\mu|\text{data}) = \frac{P(\text{data}|\mu)P(\mu)}{P(\text{data})}$$

which is a **probability density function** in the unknown μ.
Since it is a **pdf**, it must be normalized:
$\int_\Omega P(\mu|\text{data}) = 1$ which determines $P(\text{data})$.

Bayesian Point Estimates 2

Assigning names to the different factors, we get:

$$\text{Posterior pdf}(\mu) = \frac{\mathcal{L}(\mu) \times \text{Prior pdf}(\mu)}{\text{normalization factor}}$$

The Prior pdf represents your belief about μ before you do any experiments. If you already have some experimental knowledge about μ (for example from a previous experiment), you can use the posterior pdf from the previous expt. as the prior for the new one. But this implies that somewhere in the beginning there was a prior which contained no experimental evidence, just belief.

This very first prior can be thought of as a kind of phase space, or density of possible states of nature. But there is no law of nature that tells us what this density is.

In the true Bayesian spirit, the posterior density represents all our knowledge and belief about μ, so there is no need to process this pdf any further, but since we usually want a point estimate (and an interval estimate), we take another step.

Bayesian Point Estimates 2.1

Given the Bayesian posterior density for μ,
the Bayesian point estimate $\hat{\mu}$ is usually taken as
the value of μ for which the Posterior pdf takes on its maximum value.

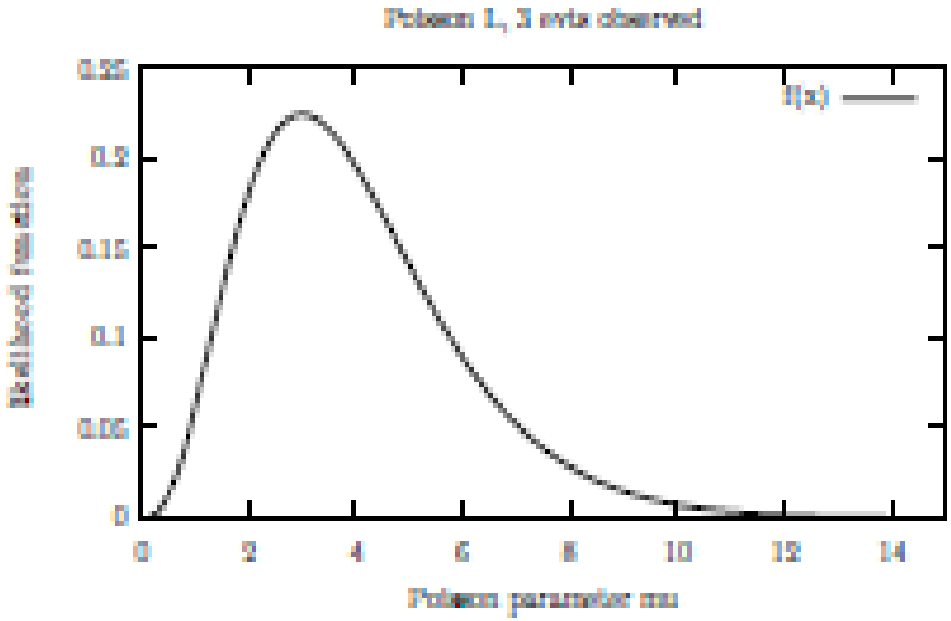
This is sometimes called incorrectly the most probable value.
The correct terminology is
the highest posterior density, or HPD point.

If the Prior probability density is taken as a uniform density,
then the maximum of the Posterior density will occur
at the maximum of the likelihood $\mathcal{L}(\mu)$.

Bayes Point Estimates 3

Example: In a Poisson process, we observe 3 events.

$$\mathcal{L}(\mu) = P(3|\mu) = \frac{e^{-\mu}\mu^3}{3!}$$



If the prior $P(\mu)$ is flat, the peak in the pdf occurs at $\mu = 3$.

Bayesian Point Estimates 4

Generalizing from 3 to n, we see that this method gives the expected result:

with n events observed, $\hat{\mu} = n$

However, in order to get this result we had to do two suspicious things:

- We used a flat prior: That means our prior belief of the true value of μ, integrated between **any two finite values** is zero. That is hardly possible.
- Our point estimate was the mode (position of the peak) of the pdf, which is not invariant under change of variables. That means that if we had chosen instead to estimate μ^2, we would not have obtained $\hat{\mu}^2$.

We will now consider alternatives to both these choices.

Point Estimation – Bayesian alternatives

Another possible Bayesian estimate of μ would be to use the Posterior Expectation $E(\mu)$. With a uniform prior on the Poisson parameter μ, when N events are observed, this gives $E(\mu) = N+1$.

Since $E(N) = \mu$, one might prefer to see $E(\mu) = N$

In fact, you get $E(\mu) = N$ if you use the prior pdf $P(\mu) = 1/\mu$.

The $1/\mu$ prior also has other advantages:

- It is a proper (normalizable) prior, unlike the uniform.
- It could actually represent someone's prior belief, since it goes to zero at $\mu = \infty$, and it produces a uniform density on a log scale.
- It is a Jeffreys Prior, proposed by physicist H. Jeffreys as being "objective" because it is scale-invariant.

Point Estimation – Bayesian alternatives

Unfortunately, the $1/\mu$ prior doesn't work, because:

- If you observe $N = 0$, $P(\mu \,|\, 0)$ is a delta-function at $\mu = 0$.
- When there is background to the Poisson process, none of the integrals converge, and all the point estimates come out $= 0$.

We will consider the problem of background in more detail later (under interval estimation).

To summarize the most common Bayesian choices:

1. The priors:
   - Uniform: Good properties, but cannot be belief.
   - Jeffreys $1/\mu$: Better in theory, not in practice.
2. The point estimate:
   - HPD gives good results, but violently non-invariant.
   - Expectation $E(\mu)$ is possible, also not invariant.
   - Median (50th percentile) of posterior is invariant, but is not used much.

Point Estimation – Bayesian Summary

Assuming that you like the Bayesian definition of Probability (degree of belief), Bayesian point estimation is a coherent methodology which provides a reasonable way to estimate parameters. But it involves two arbitrary choices:

- The Prior pdf.
- The Mapping from Posterior pdf to Point estimate.

In practice, both these problems become less important as the amount of data increases, so that

- the data dominates the prior and
- the Posterior pdf tends toward a Gaussian.

However, in this limit,
almost any statistical method would give the same result.

Point Estimation - from Bayesian to Frequentist

Up to the early 1900's, the only statistical theory was Bayesian.

In fact, frequentist methods were already being used:

Linear least-squares fitting of data had been in use for many years, and in 1900, Karl Pearson published the Chi-square test to be treated later under *goodness-of-fit*.
Karl Pearson also had a famous son: Egon Pearson
and he founded a famous journal: Biometrika.

Another biologist, R. A. Fisher, was one of several people looking for a statistical theory that would not require as input prior belief. He succeeded in making a frequentist theory of point estimation, but was unable to produce an acceptable theory of interval estimation.

Point Estimation - Frequentist

An Estimator $\mathcal{E}_\theta$ is a function of the data X which can be used to estimate (measure) the unknown parameter θ.

$$\hat{\theta} = \mathcal{E}_\theta(X)$$

The goal:
Find that function $\mathcal{E}_\theta$ which gives estimates $\hat{\theta}$
closest to the true value of θ.

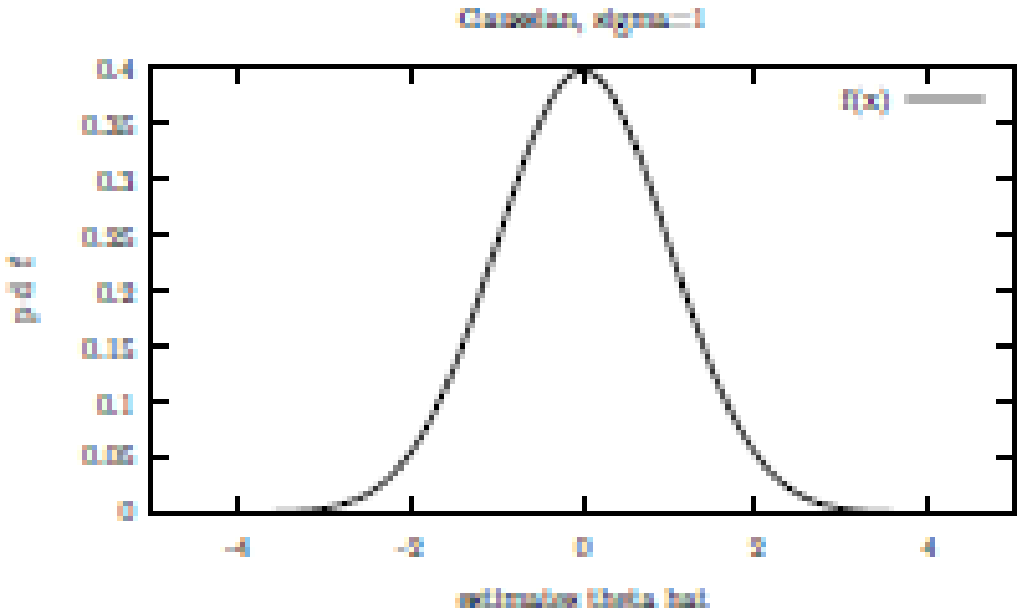
As usual, we know $P(X|\theta)$
and because the estimate is a function of the data,
we also know the distribution of $\hat{\theta}$, for any given value of θ:

$$P(\hat{\theta}|\theta) = \int_X \mathcal{E}_\theta(X) P(X|\theta) dX$$

.

Point Estimation - Frequentist Estimates

For our trial estimator $\mathcal{E}_\theta$, assuming $\theta = 0$,
the distribution of estimates $\hat{\theta}$ might look something like this:



Now we can see whether this estimator has the desired properties.
Is it (1) consistent, (2) unbiased, (3) efficient, and (4) robust?

Frequentist Point Estimation - Consistency

Let $\mathcal{E}_\theta$ be an estimator producing estimates $\hat{\theta}_n$, where n is the number of observations entering into the estimate.

Given any $\varepsilon > 0$ and any $\eta > 0$, $\mathcal{E}_\theta$ is a consistent estimator of θ if an N exists such that

$$P(|\hat{\theta}_n - \theta_0| > \varepsilon) < \eta$$

for all $n > N$, where θ_0 is the assumed true value.

This says that $\hat{\theta}_n$ converges (in probability) to the true value of θ as n increases.

Frequentist Point Estimation - Bias

We define the bias b of the estimate $\hat{\theta}$ as the difference between the expectation of $\hat{\theta}$ and the assumed true value θ_0,

$$b_N(\hat{\theta}) = E(\hat{\theta}) - \theta_0 = E(\hat{\theta} - \theta_0)\,.$$

Thus, an estimator is unbiased if, for all N and θ_0,

$$b_N(\hat{\theta}) = 0$$

or

$$E(\hat{\theta}) = \theta_0\,.$$

Frequentist Point Estimation - Efficiency

Among those estimators that are consistent and unbiased, we clearly want the one whose estimates have the smallest spread around the true value, that is, estimators with a small variance.

We define the efficiency of an estimator in terms of the variance of its estimates $V(\hat{\theta})$:

$$\text{Efficiency} = \frac{V_{\text{min}}}{V(\hat{\theta})}$$

where V_{min} is the smallest variance of any estimator.

The above definition is possible because, as we shall see, V_{min} is given by the Cramér-Rao lower bound.

Point Estimation - Fisher Information

Let the pdf of the data X be denoted by f or by L:

$$P(\text{data}|\text{hypothesis}) = f(X|\theta) = L(X|\theta)$$

depending on whether we are primarily interested in the dependence on X or θ.

The amount of information given by an observation X about the parameter θ is defined by the following expression (if it exists)

$$\begin{aligned} I_X(\theta) &= E\left[\left(\frac{\partial \ \ln \ L(X|\theta)}{\partial\theta}\right)^2\right] \\ &= \int_{\Omega_\theta} \left(\frac{\partial \ \ln \ L(X|\theta)}{\partial\theta}\right)^2 L(X|\theta)dX\,. \end{aligned}$$

Point Estimation - Fisher Information cont.

If $\boldsymbol{\theta}$ has k dimensions, the definition becomes

$$\begin{aligned}\left[\mathcal{I}_X(\boldsymbol{\theta})\right]_{ij} &= E\left[\frac{\partial \ln L(X|\boldsymbol{\theta})}{\partial\theta_i}\cdot\frac{\partial \ln L(X|\boldsymbol{\theta})}{\partial\theta_j}\right] \\ &= \int_{\Omega_0}\left[\frac{\partial \ln L(X|\boldsymbol{\theta})}{\partial\theta_i}\cdot\frac{\partial \ln L(X|\boldsymbol{\theta})}{\partial\theta_j}\right]L(X|\boldsymbol{\theta})dX\,.\end{aligned}$$

Thus, in general, $\mathcal{I}_X(\boldsymbol{\theta})$ is a $k \times k$ matrix. Assuming certain regularity conditions, the same matrix can be expressed as the expectation of the second derivative matrix see next slide:

$$\left[\mathcal{I}_X(\boldsymbol{\theta})\right]_{ij} = -E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln L(X|\boldsymbol{\theta})\right]\,.$$

Point Estimation - Fisher Information cont.

So the Fisher information in the sample X about the parameter(s) θ is

$$\left[\mathcal{I}_X(\theta) \right]_{ij} = -E\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln L(X|\theta) \right].$$

It can be seen that $\mathcal{I}_X(\theta)$ has the additive property: If I_N is the information in N events, then $I_N(\theta) = N I_1(\theta)$.

We will also see that information about θ is related to the minimum variance possible for an estimator of *theta*.

But first we introduce the concept of Sufficient Statistics

Point Estimation - Sufficiency

Any function of the data is called a statistic.

A sufficient statistic for θ is a function of the data that contains all the information about θ.

A statistic $T(X)$ is sufficient for θ if the conditional density function for X given T, $f(X|T)$ is independent of θ.

Sufficient statistics are clearly important for data reduction.

The Darmois Theorem says that a number of sufficient statistics independent of N can exist only if $f(X|\theta)$ belongs to the exponential family

$$f(X|\theta) = \exp[\alpha(X)a(\theta) + \beta(X) + c(\theta)] .$$

Point Estimation - Cramér-Rao Inequality

Let the estimator $\hat{\theta}$ have the sampling distribution $q(\hat{\theta}|\theta)$. The bias is a function of the true value θ

$$b = E(\hat{\theta}) - \theta = \int \hat{\theta}(\mathbf{X}) f(\mathbf{X}|\theta) d\mathbf{X} \quad - \quad \theta \,.$$

Then the variance of the sampling distribution,

$$V(\hat{\theta}) = \int [\hat{\theta} - E(\hat{\theta})]^2 q(\hat{\theta}|\theta) d\hat{\theta} \,,$$

is related to the information by the Cramér–Rao inequality:

$$V(\hat{\theta}) \geq \frac{[1 + (db/d\theta)]^2}{I_\theta} = \frac{[1 + (db/d\theta)]^2}{E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]} \,.$$

Point Estimation - Robustness

An estimator is said perfectly robust if its p.d.f. $g(\hat{\theta})$ does not depend on the model ($\mathcal{P}(x|\theta)$). None of the usual estimators is perfectly robust. Thus, robustness is more a qualitative property : a robust estimator is such that $g(\hat{\theta})$ is weakly sensitive to (small) variations of the probablity model. This property is particularly useful when the chosen analytical form of $\mathcal{P}$ is not well known, for example due to some tails in the distribution.

Example : the trimmed mean and the Winsorized mean are more robust than the usual sample mean in order to estimate the center of a symmetrical distribution.

Point Estimation - The Usual Estimators

The most common general-purpose estimators are:

- The method of moments is based on approximating $f(X|\theta)$ by its first few moments. It is surprisingly efficient for an approximate method, but will not be treated here.
- Maximum likelihood is the most important method, mostly because it can be shown to be **asymptotically efficient**.
- Least squares is **asymptotically efficient** for data that is already grouped into bins or points, and is generally easier to apply than M.L.

Point Estimation - Maximum Likelihood

The likelihood of a set of N independent observations $\mathbf{X}$ is

$$L(\mathbf{X}|\theta) = \prod_{i=1}^{N} f(X_i, \theta)\,,$$

where $f(X, \theta)$ is the p.d.f. of any observation X.
The maximum likelihood estimate of the parameter θ is that value $\hat{\theta}$ for which $L(\mathbf{X}|\theta)$ has its maximum, given the particular observations $\mathbf{X}$.
Note that maximizing $\ln L$ or L gives the same result.
The likelihood equation is

$$\frac{\partial}{\partial\theta} \sum_{i=1}^{N} \ln f(X_i, \theta) = \frac{\partial}{\partial\theta} \ln L(\mathbf{X}, \theta) = 0\,.$$

since that is the analytic way to find the maximum, but in practice we will usually find the maximum numerically.

Asymptotic Properties of Maximum Likelihood

Asymptotically (for very large data samples), the M. L. estimator has optimal properties:

- It is consistent.
- It is efficient, the variance $V(\hat{\theta})$ being given by the Cramer–Rao lower bound

$$V(\hat{\theta}) \underset{N\to\infty}{\longrightarrow} \left\{ E\left[\left(\frac{\partial \ln L}{\partial \theta} \right)^2 \right] \right\}^{-1}.$$

- The estimates $\hat{\theta}$ are Normally distributed.
- Since it is consistent, it is asymptotically unbiased.

Asymptotic Properties of Maximum Likelihood 2

If the range of the data is independent of the parameters θ, then the variance $V(\hat{\theta})$ may be estimated by

$$\hat{V}(\hat{\theta}) = \left\{ \left(-\frac{\partial^2 \ln L}{\partial \theta^2} \right) \Big|_{\theta=\hat{\theta}} \right\}^{-1}.$$

The estimate $\sqrt{N}(\hat{\theta} - \theta)$ is distributed as $N[0, I_1^{-1}(\theta)]$.
(Estimates are asymptotically Gaussian-distributed.)

Finite Sample Properties of Maximum Likelihood

- For finite samples, M.L. estimates are efficient only when there exist sufficient statistics for the parameter(s) being evaluated, and that can be shown only for the exponential family, consistent with the Darmois Theorem.
- Although the estimates are in general biased, they have a more important property, invariance, which is incompatible with unbiasedness because the definition of bias is not invariant.

Point Estimation - Least Squares

Consider a set of observations $Y_1, \ldots, Y_N$ from a distribution with expectations $E(Y_i, \boldsymbol{\theta})$ and covariance matrix $\underset{\sim}{V}$. The $\boldsymbol{\theta}$ are unknown parameters and the $E(Y_i, \boldsymbol{\theta})$ and $V_{ij}(\boldsymbol{\theta})$ are known functions of $\boldsymbol{\theta}$.

In the method of least squares the estimates of the θ_k are those values $\hat{\theta}_k$ which minimize

$$
\begin{aligned}
Q^2 &= \sum_{i=1}^{N} \sum_{j=1}^{N} [Y_i - E(Y_i, \boldsymbol{\theta})] (\underset{\sim}{V}^{-1})_{ij} [Y_j - E(Y_j, \boldsymbol{\theta})] \\
&= [\mathbf{Y} - E(\mathbf{Y}, \boldsymbol{\theta})]^{\mathrm{T}} \, \underset{\sim}{V}^{-1} \, [\mathbf{Y} - E(\mathbf{Y}, \boldsymbol{\theta})] \, .
\end{aligned}
$$

Point Estimation - Least Squares 2

When the observations Y_i are independent, it follows that they are uncorrelated, and the covariance matrix is diagonal, with elements

$$V_{ii} = \sigma_i^2(\boldsymbol{\theta}) .$$

The covariance form then simplifies to the familiar sum of squares

$$Q^2 = \sum_{i=1}^{N} \frac{[Y_i - E(Y_i, \boldsymbol{\theta})]^2}{\sigma_i^2(\boldsymbol{\theta})}$$

The $\hat{\boldsymbol{\theta}}$ are found by solving the Normal equations

$$\partial Q^2 / \partial \boldsymbol{\theta} = 0 ,$$

Point Estimation - Linear Least Squares

The method of linear least squares is applicable when the variances σ_i^2 are independent of the r parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$, and the expectations $E(Y_i, \boldsymbol{\theta})$ are linear in the θ_j's,

$$E(Y_i, \boldsymbol{\theta}) = \sum_{j=1}^{r} a_{ij}\theta_j\,, \quad i = 1, \ldots, N$$

or in matrix notation

$$E(\mathbf{Y}, \boldsymbol{\theta}) = \mathcal{A}\,\boldsymbol{\theta}$$

The elements a_{ij} of the *design matrix* $\mathcal{A}$ are given by a model.

In the linear case, the solution of the Normal equations is

$$\widehat{\boldsymbol{\theta}} = (\mathcal{A}^{\mathrm{T}}\mathcal{V}^{-1}\mathcal{A})^{-1}\,\mathcal{A}^{\mathrm{T}}\mathcal{V}^{-1}\,\mathbf{Y}\,.$$

Point Estimation - Linear Least Squares

Since the linear least squares solution is found by matrix inversion and multiplication (no minimization needed), one often solves the non-linear problem by linearization, setting:

$$a_{ij} = \frac{\partial E(Y_i, \theta)}{\partial \theta_j}$$

Example of linear least squares: fitting a curve to a polynomial.

$$Y_i = Y(X_i) = \theta_0 + \theta_1 X_i + \theta_2 X_i^2 + \theta_3 X_i^3$$

is clearly of the linear form. To find the matrix $\mathcal{A}$ one only needs to evaluate the $(j-1)^{\text{th}}$ power of X_i.
Solving the Normal equations $\partial Q^2/\partial \boldsymbol{\theta} = 0$, we find:

$$\tilde{\boldsymbol{\theta}} = (\mathcal{A}^{\mathrm{T}} \mathcal{V}^{-1} \mathcal{A})^{-1} \mathcal{A}^{\mathrm{T}} \mathcal{V}^{-1} \mathbf{Y}.$$

which is exact and unique as long as $\mathcal{A}^{\mathrm{T}} \mathcal{V}^{-1} \mathcal{A}$ is non-singular.

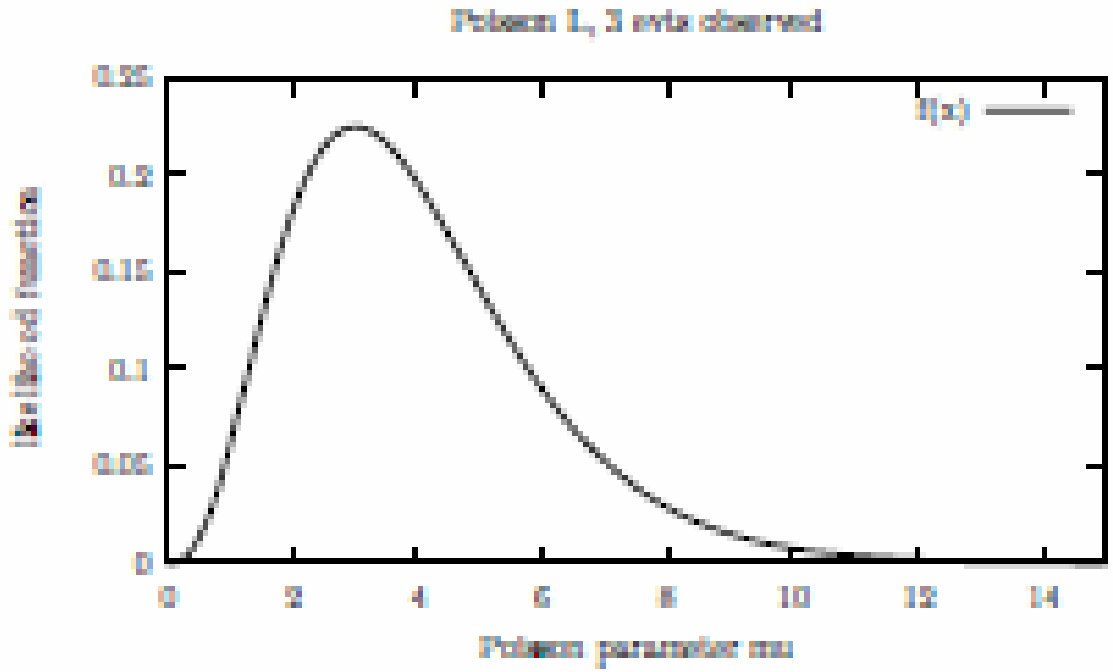
Point Estimation - Least Squares

The asymptotic properties of least squares are the same as for maximum likelihood, and in fact the two methods are often identical. When they are different, it is believed that M.L. generally approaches the asymptotic limit faster than L.S.
The biggest difference is largely practical.
If the data are already grouped into bins or points, L.S. is more convenient and there is no advantage in using M.L.

Point Estimation: Example: Poisson data

Example: In a Poisson process, we observe 3 events.

$$\mathcal{L}(\mu) = P(3|\mu) = \frac{e^{-\mu}\mu^3}{3!}$$



The peak in the likelihood occurs at $\mu = 3$.

Generalizing from 3 to n, we get the expected result:

with n events observed, $\hat{\mu} = n$

Point Estimation - Example: Weighted Average

Suppose we have Normally-distributed observations X_i of a quantity μ, each X_i being distributed with standard deviation σ_i:

$$f(X_i|\mu) = N(\mu, \sigma_i^2) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(X_i-\mu)^2}{\sigma_i^2}\right].$$

We wish to use this data to estimate μ. The likelihood function is the product of the $f(X_i|\mu)$, and its logarithm is:

$$\ln \mathcal{L}(\mu) = k - \sum_i \frac{1}{2}\frac{(X_i-\mu)^2}{\sigma_i^2}$$

where k is a constant. It is clear that in this case, maximizing the log likelihood is equivalent to minimizing χ^2. In both cases, the solution is the familiar weighted average:

$$\hat{\mu} = \frac{\sum_i \frac{X_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}$$

Interval Estimation

The goal of interval estimation is to find an interval which will contain the true value of the parameter with a given probability.

The meaning of this probability, and hence the meaning of the interval, will of course be very different for the Bayesian and frequentist methods.

In both methods, the interval with the required probability content will not generally be unique. Then one must find the best interval with the specified probability content.

Interval Estimation

We may distinguish four different theories of Interval Estimation:

1. Bayesian Theory is based on Bayes' Theorem, and requires only a straightforward extension of the Bayesian Theory of Point Estimation. However, it will cause us to look more carefully at the problem of Priors.
2. Frequentist Normal Theory, is an asymptotic theory valid when estimates are approximately Normally distributed, which is nearly always the case. MOST BOOKS AND COURSES PRESENT ONLY THIS THEORY.
3. Exact Frequentist Theory was developed by Jerzy Neyman with the help of Pearson's son Egon and a few others around 1930.
4. Likelihood-based Methods, intermediate between 2. and 3., are what you will probably use most of the time. (You can get these intervals easily with Minuit.)

Interval Estimation - Bayesian

Recall that in the Bayesian method of parameter estimation, all the knowledge about the parameter(s) is summarized in the posterior pdf $P(\theta|data)$.

To find an interval (θ_1, θ_2) which contains probability β, one simply has to find two points such that

$$\int_{\theta_1}^{\theta_2} P(\theta|X)\, d\theta \;=\; \beta.$$

where β is usually chosen either 0.683 for one-standard-deviation intervals, or 0.900 for safer intervals. This is the degree of belief that the true value of θ lies within the interval. The Bayesian interval with probability β is called a credible interval to distinguish it from its frequentist equivalent, the confidence interval.

Interval Estimation - Bayesian

Since the credible interval of content β is not unique, we can impose an additional condition, which is usually taken to be one of:

- Accept into the interval the points of highest posterior density (H.P.D.). [This interval is not invariant under change of variable $\theta \to \theta'$.]
- A central interval, such that the integral in each tail is $= (1 - \beta)/2$. Central intervals are invariant, but do not produce one-sided intervals (upper limits) in cases where they are obviously appropriate.
- A one-sided interval, usually an upper limit, when there is reason to believe that θ is near one end of the allowed region. One-sided intervals are invariant.

Bayesian Intervals – The Physical Region

One of the most attractive features of the Bayesian method:
Since the Prior is always zero in the non-physical region,
the entire credible interval is necessarily in the allowed region.

But the negative side of this property is:
A measurement near the edge of the physical region will always be biased toward the interior of the physical region.

This is to be expected, since the credible interval represents belief,
but it means that we lose the information about what comes from the actual measurement and what comes from the prior.

Interval Estimation - Frequentist

The Problem: Given β, find the optimal range $[\theta_a, \theta_b]$ in θ-space such that:

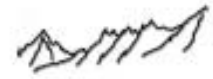
$$P(\theta_a \leq \theta_{\text{true}} \leq \theta_b) = \beta\,.$$

The interval (θ_a, θ_b) is then called a confidence interval. A method which yields intervals (θ_a, θ_b) satisfying the above is said to possess the property of coverage.

Formally, if an interval does not possess the property of coverage, it is not a confidence interval, although we will consider sometimes approximate confidence intervals, which have only approximate coverage.

Overcoverage occurs when $P > \beta$.
Undercoverage occurs when $P < \beta$.

Normal Theory Interval Estimation



Suppose we are sampling X from the Gaussian $\mathcal{N}(\mu, \sigma^2)$, with μ parameter (unknown) and σ known. Thus $\overline{x}$ follows the Gaussian $\mathcal{N}(\mu, \sigma^2/n)$ and $z = (\overline{x} - \mu)/\sigma/\sqrt{n}$ follows the reduced Gaussian $\mathcal{N}(0, 1)$. So, if we choose (ex.) $\beta = 0.95$, we know that $\mathcal{P}(|z| < 1.96) = 0.95 = \beta$. In other words

$$\overline{x} - 1.96\sigma/\sqrt{n} < \mu < \overline{x} + 1.96\sigma/\sqrt{n}$$

with a 95% probability.

This has been possible because we found a statistic (z) the pdf of which does not depend on the parameter. z is a function of μ, not its pdf. We thus have been able to reverse the inequalities as

$$\beta = \mathcal{P}(a < z < b) = \mathcal{P}(m_1 < \mu < m_2)$$

This is possible only for very few situations;

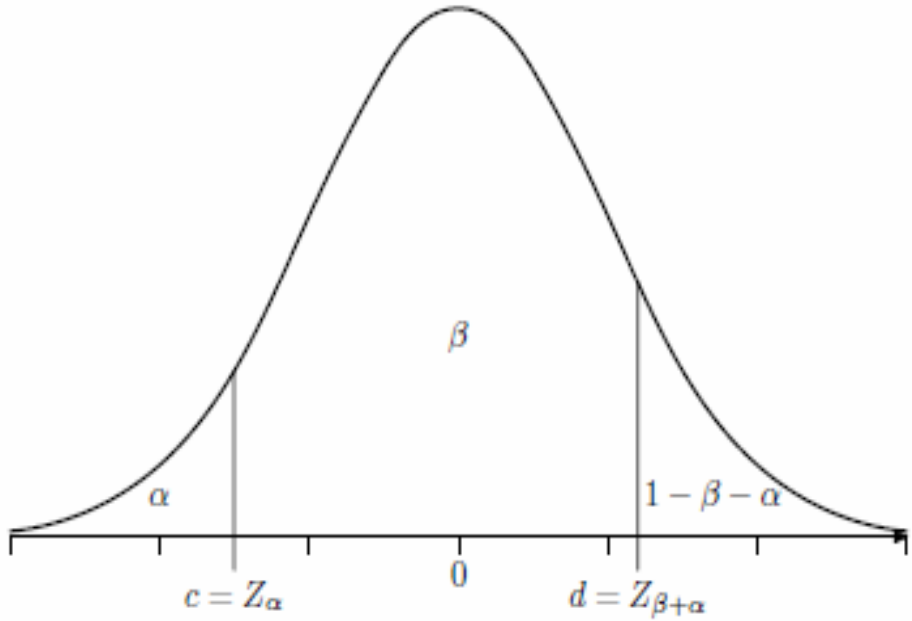
- Gaussian law (all cases)
- Poisson law
- Binomial law
- Asymptoticaly through the Central Limit Theorem

Normal Theory Interval Estimation

Given a random variable X with p.d.f. $f(X)$ and cumulative distribution $F(X)$, the α-*point* X_α is defined by

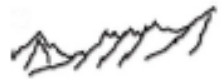
$$\int_{-\infty}^{X_\alpha} f(X)dX = F(X_\alpha) = \alpha\,.$$

In terms of α-points, the interval $[c, d]$ is obviously $[Z_\alpha, Z_{\alpha+\beta}]$.



$N(0, 1)$ with regions of probability content α, β, and $1 - \beta - \alpha$. c is the α-point and d the $(\alpha + \beta)$-point.

Normal Theory Interval Estimation



Clearly for a given value of β, there are many possible intervals, corresponding to different values of α. The most usual choice is $\alpha = (1-\beta)/2$, which gives the central interval, symmetric about zero.

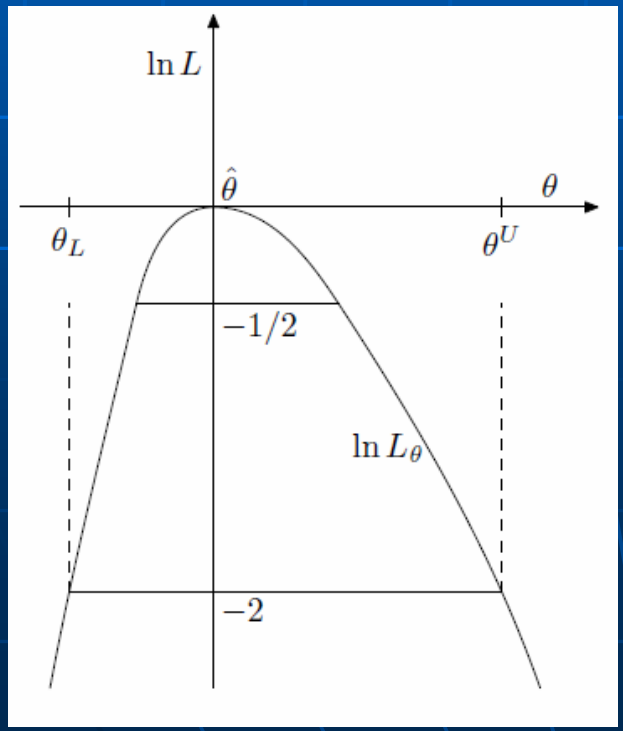
Example: Central Intervals for N(0,1).

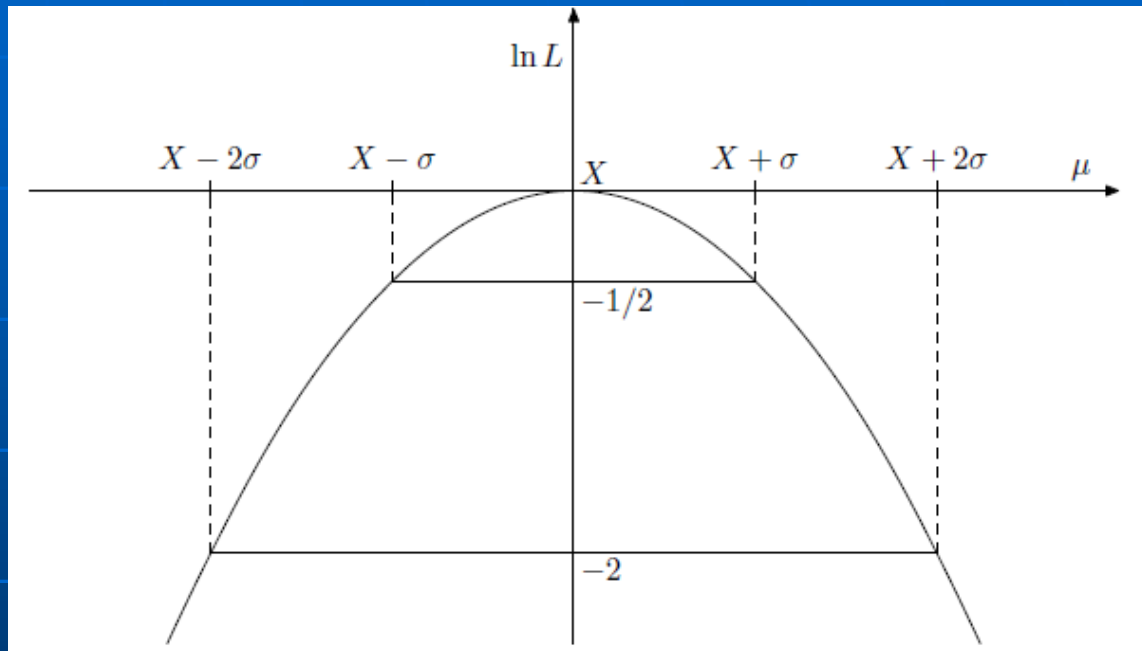
$\beta = (1-\alpha)/2$	Z_α	$Z_{\alpha+\beta}$
0.6827	-1.00	1.00
0.9000	-1.65	1.65
0.9500	-1.96	1.96
0.9545	-2.00	2.00
0.9900	-2.58	2.58
0.9973	-3.00	3.00

Likelihood-based Confidence Intervals

If X is Gaussian, its log-Likelihood function is a parabola and finding a confidence interval is easy.

This property can be used in all cases asymptotically since a ML estimator is asymptotically gaussian.

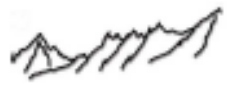




Log-likelihood function for Gaussian X, distributed $N(\mu, \sigma^2)$.

In case of a finite sample, we have no more a parabola. One can give a dissymmetric interval (with respect to the point est).

Normal Theory Intervals in Many Variables



In more than one dimension, the confidence interval becomes a confidence region, and the Normal pdf becomes:

$$f(\mathbf{t}|\theta) = \frac{1}{(2\pi)^{N/2}|\underset{\sim}{V}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{t}-\theta)^T \underset{\sim}{V}^{-1}(\mathbf{t}-\theta)\right].$$

It follows from the Normality of the $\mathbf{t}$ that the covariance form

$$Q(\mathbf{t},\theta) = (\mathbf{t}-\theta)^T \underset{\sim}{V}^{-1}(\mathbf{t}-\theta)$$

has a $\chi^2(N)$ distribution. This means that the distribution of Q is independent of θ, and we have

$$P[Q(\mathbf{t},\theta) \leq K_\beta^2] = \beta$$

where K_β^2 is the β-point of the $\chi^2(N)$ distribution.

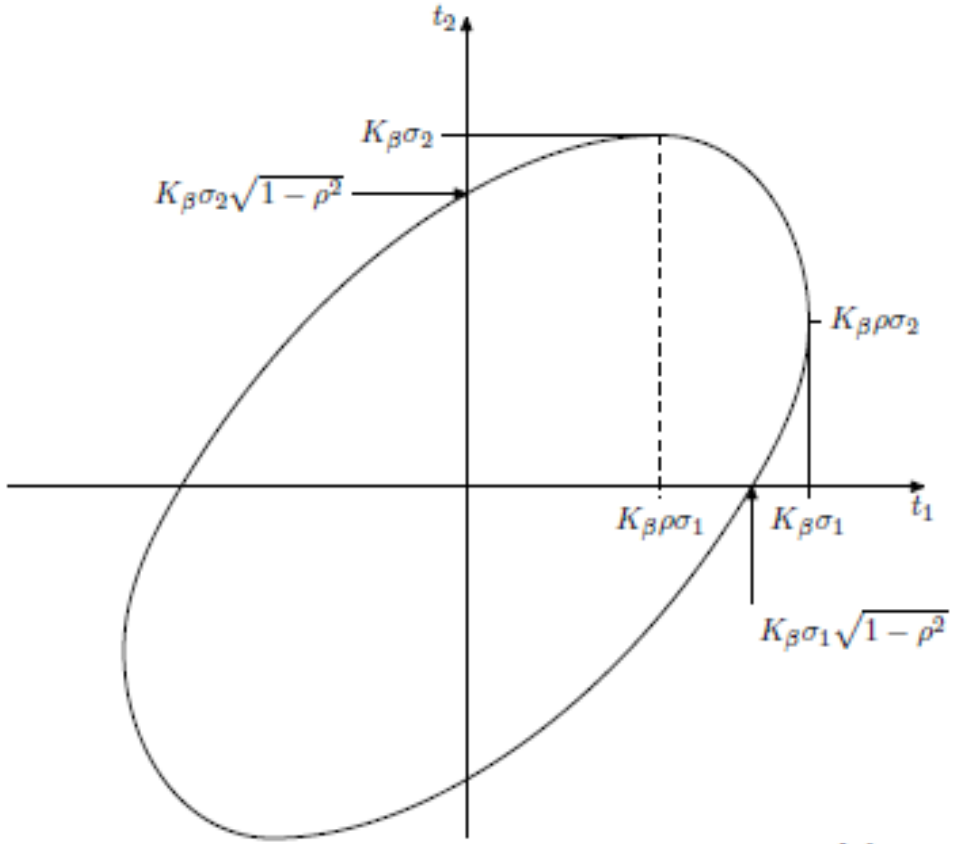
The region in $\mathbf{t}$-space defined by $Q(\mathbf{t},\theta) \leq K_\beta^2$
is a hyperellipsoid of constant probability density for the Normal pdf, a region with probability content β.

Normal Theory Intervals in Two Variables

For two Normally-distributed variables with covariance matrix

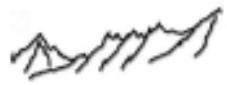
$$\underset{\sim}{V} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

The elliptical confidence region will look like this:



Confidence region of probability content β for covariance matrix $\underset{\sim}{V}$. Shown here is the case $\rho = 0.5$. If $\rho = 0$, the axes of the ellipse are horizontal and vertical. If $\rho = 1$, the ellipse degenerates to a diagonal line.

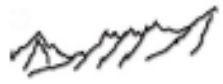
Exact Frequentist Intervals



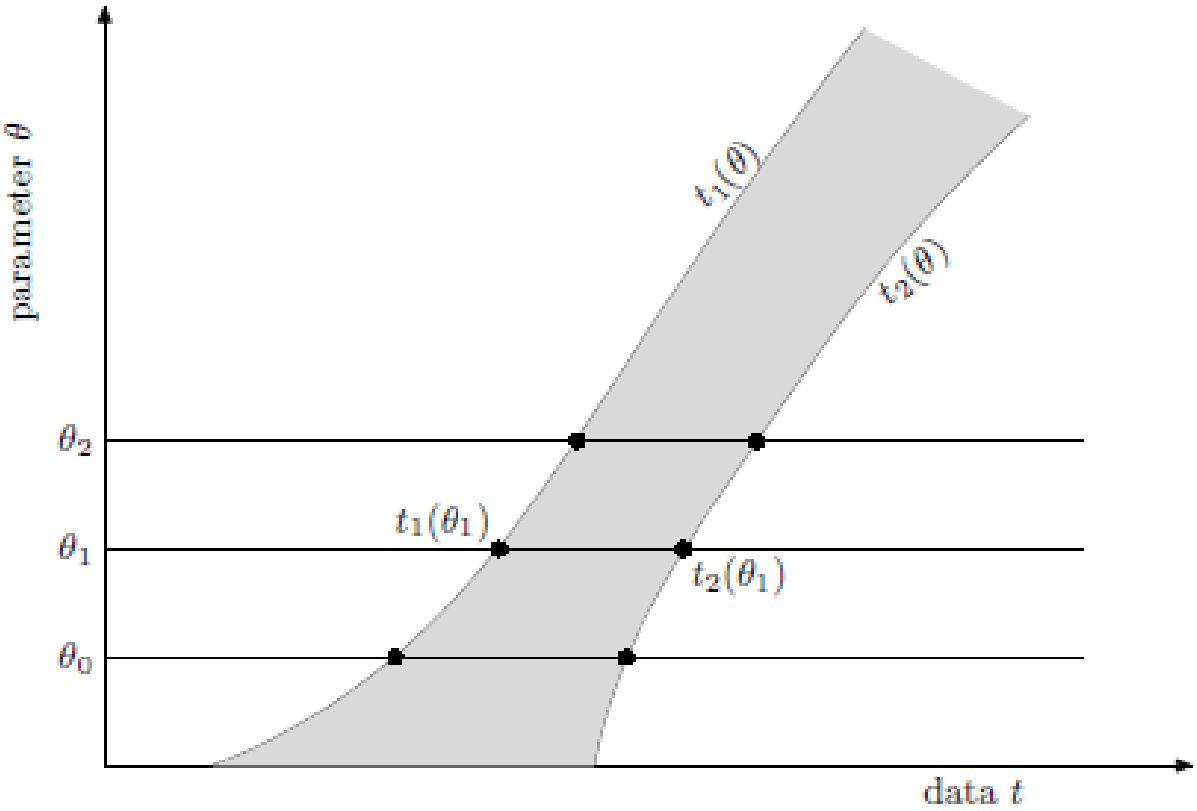
The first important step in finding an exact theory was to work in the right space: $P(\text{data}|\text{hypothesis})$, with one axis (or set of axes) for data, and another for hypotheses.

Trying to plot "true values" and "measured values" on the same axis is not a good approach, since we know that $P(\text{data}|\text{hypothesis})$ transforms differently as a function of data or as a function of the hypothesis.

The Neyman Construction



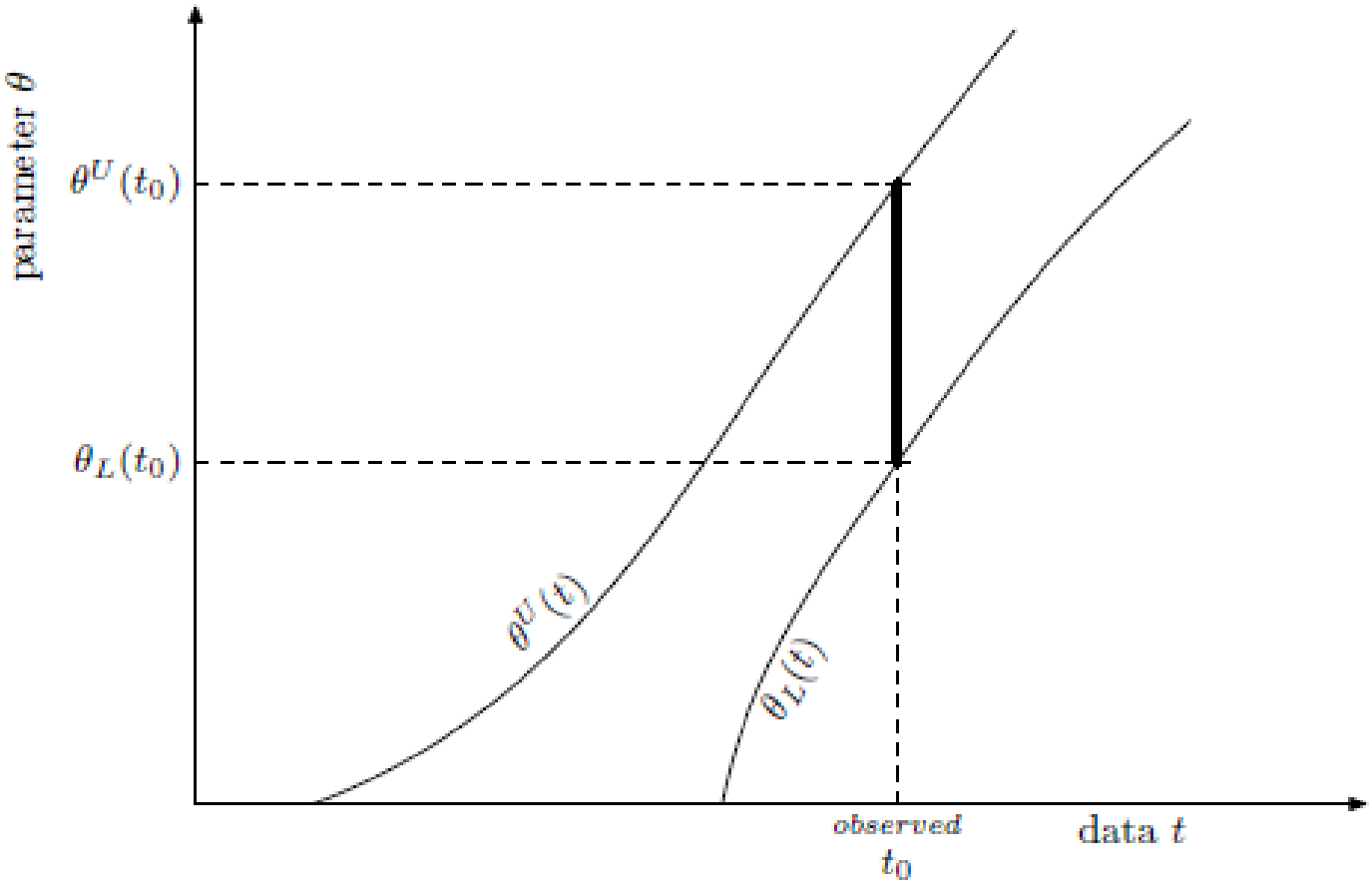
The confidence belt is constructed horizontally in the space of $P(t|\theta)$.



$t_1(\theta)$ and $t_2(\theta)$ are such that: $P(t_1 < \text{data} < t_2) = \beta$ where β is usually chosen to be 0.683 or 0.900.

The Neyman Construction 2

The two curves of $t(\theta)$ are re-labelled as $\theta(t)$, and the confidence limit is read vertically.

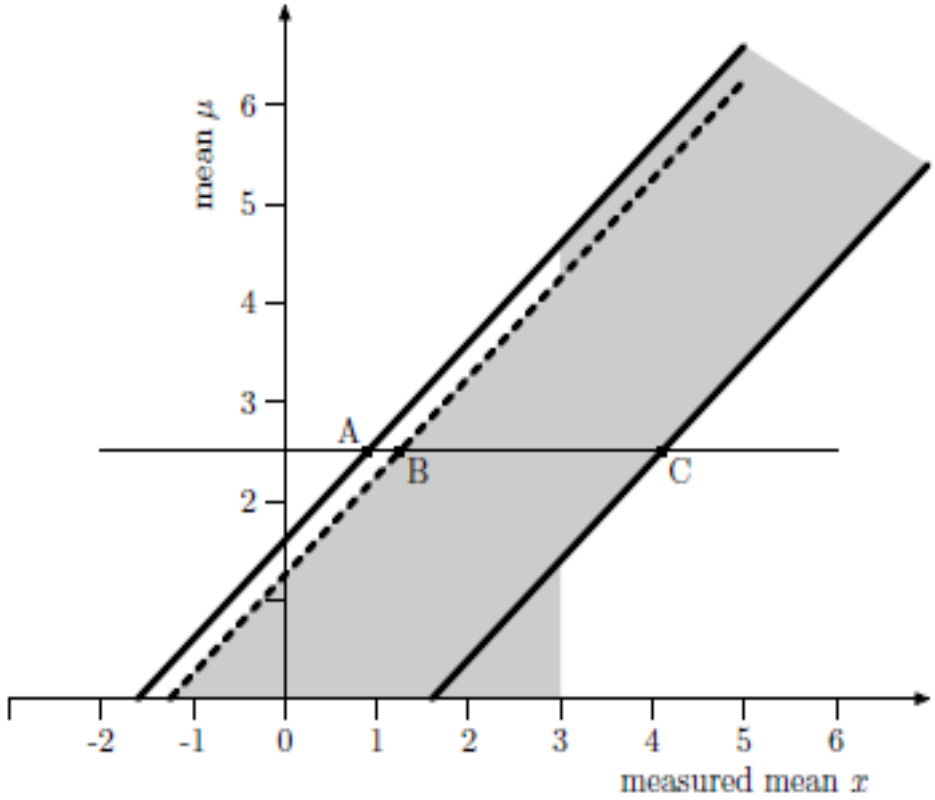


For observed data t_0, the confidence interval is (θ^U, θ_L)

Upper limits, Flip-flopping and Empty Intervals

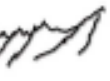
When the parameter cannot be negative but is very close to zero, one often quotes an Upper limit rather than a two-sided interval.

empty intervals down here $\longrightarrow$



Flip-flopping for a Gaussian measurement. The solid lines delimit the central 90% confidence belt, the dashed line the 90% upper limit, and the shaded area the effective confidence belt resulting from choosing between the two after seeing the data. This effective belt undercovers for $1.2 < \mu < 4.3$, for example at $\mu = 2.5$ where the intervals AC and $B\infty$ each contain 90% probability but BC contains only 85%.

The Unified Approach (Feldman-Cousins)



The elegant way to solve all the problems (flip-flopping and empty intervals) would be to find an ordering principle which automatically gives intervals with the desired properties.

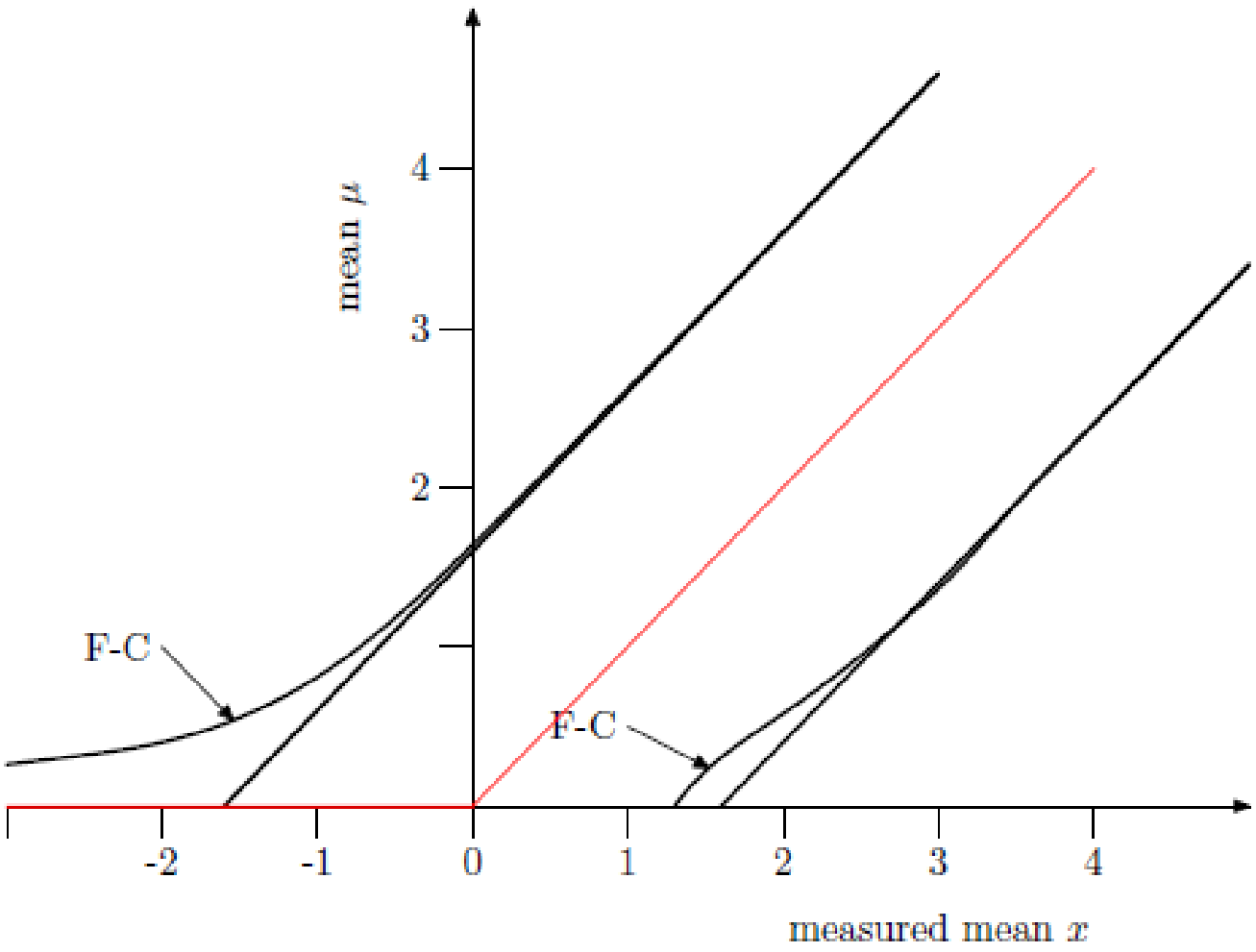
Inspired by an important result in hypothesis testing
which we will see in the next chapter,
Feldman and Cousins proposed the
likelihood ratio ordering principle:
see Feldman and Cousins, *Unified Approach ...*
Phys. Rev. D 57 (1998) 3873

When determining the interval for $\mu = \mu_0$, include the elements of probability $P(x|\mu_0)$ which have the largest values of the likelihood ratio

$$R(x) = \frac{P(x|\mu_0)}{P(x|\hat{\mu})},$$

where $\hat{\mu}$ is the value of μ for which the likelihood $P(x|\mu)$ is maximized within the physical region.

The Unified Approach (Feldman-Cousins)



Belts of 90% confidence for a Gaussian measurement showing the effect of using different ordering principles. The Feldman-Cousins belt is labelled "F-C", and the straight lines give central intervals. The red line is the M.L. solution

Backup slides

Combining Bayesian Intervals

In the Bayesian system, both point estimates and interval estimates may be highly biased, so it would seem impossible to combine the estimates from different experiments to produce a "world average", and indeed it is.

However, the Bayesian framework offers an elegant way to combine results from several experiments by extending Bayes' Rule:

$$\text{Posterior pdf}(\mu) = \frac{\mathcal{L}_1(\mu) \times \mathcal{L}_2(\mu) \times \mathcal{L}_3(\mu) \times \text{Prior pdf}(\mu)}{\text{normalization factor}}$$

where $\mathcal{L}_i(\mu)$ is the likelihood function from the i^{th} experiment.

To get the Posterior, you may use as many likelihoods as you want, but you must use one and only one Prior.

Background in Poisson Processes

We may distinguish different cases:

1. The background expectation is exactly known.
   - Observe 10 events.
     Expect 3 bgd.
   - Observe 0 events.
     Expect 3 bgd.
2. The background expectation is measured with some uncertainty.
   ("side-bands", or "signal off".)
   example: $b = 3.1 \pm 1.2$
   In this case, b is a nuisance parameter.

Bayesian Intervals: Poisson with Estimated Background

In the Bayesian framework, everything has its probability distribution, including of course nuisance parameters. The distribution of background is some pdf $P(b)$.

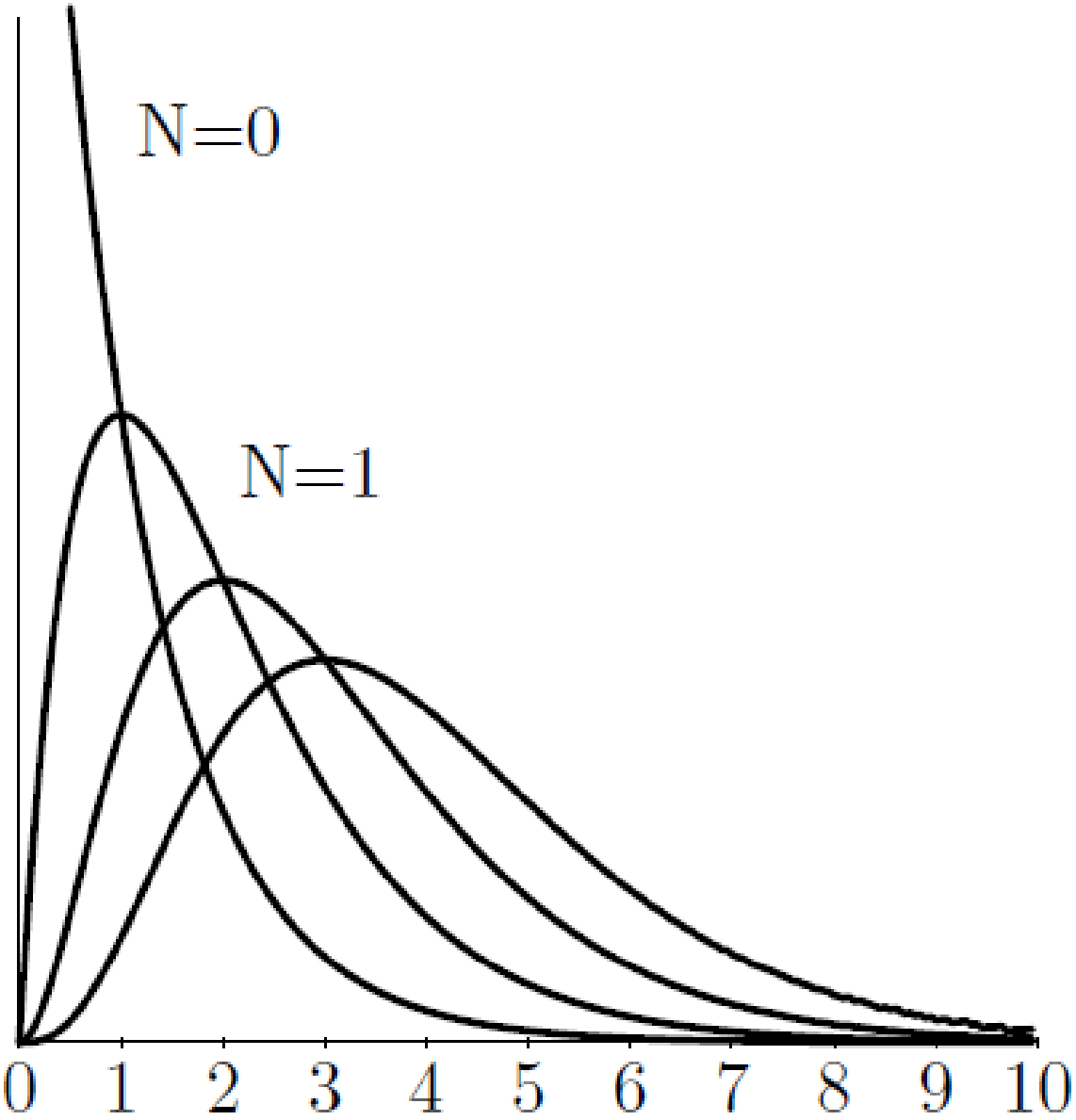
Therefore, in calculating the posterior pdf,
one simply integrates over all nuisance parameters:

$$P(\mu|\text{data}) = \int_b \frac{P(\text{data}|\mu)P(\mu)}{P(\text{data})} P(b)db$$

This may be very heavy numerically, but it is conceptually easy.

Bayesian Intervals for Poisson, Uniform Prior

Bayesian Posterior for Poisson, $N_{obs} = 0, 1, 2, 3,$ Uniform Prior



Bayesian Intervals: Poisson with Known Background

Bayesian 90% Upper Limits (Uniform Prior)

observed =	0	1	2	3
background = 0.0	2.30	3.89	5.32	6.68
0.5	2.30	3.50	4.83	6.17
1.0	2.30	3.26	4.44	5.71
2.0	2.30	3.00	3.87	4.92
3.0	2.30	2.83	3.52	4.37

The uniform prior gives very reasonable upper limits for Poisson observations, with or without background.

However, the Uniform Prior $U(x)$ cannot represent belief, because

$$\int_a^b U(\mu)\, d\mu \;=\; 0 \qquad \text{for all finite a,b}$$

Bayesian Intervals: Non-Uniform Priors

So let us try the famous Jeffreys Priors.

Jeffreys Priors were derived in order to be
invariant under certain coordinate transformations.
The $1/\mu$ Jeffreys Prior is scale-invariant.
It could represent belief, since it goes to zero at infinity.
We have used it earlier in Bayesian Point Estimation.

Bayesian 90% Upper Limits ($1/\mu$ Jeffreys Prior)

observed =	0	1	2	3
background = 0.0	0.00	2.30	3.89	5.32
0.5	0.00	0.00	0.00	0.00
1.0	0.00	0.00	0.00	0.00
2.0	0.00	0.00	0.00	0.00
3.0	0.00	0.00	0.00	0.00

Bayesian Intervals with Jeffreys Priors

Can Jeffreys Priors be saved?

For parameters μ, $\quad 0 \leq \mu \leq \infty$, there is another Jeffreys Prior,

$$P(\mu) = 1/\sqrt{\mu}$$

which minimizes the Fisher information contained in the prior.
Unfortunately, this very good idea also doesn't work.
The divergences remain.

The prior that gives the desired Poisson intervals in the presence of background is

$$P(\mu) = 1/\sqrt{\mu + b}$$

where b is the expected background. This means that the prior for μ depends on b, which is completely crazy.

Frequentist Upper Limits for Poisson data

Naive Frequentist 90% Upper Limits for Poisson with Background

observed =	0	1	2	3
background = 0.0	2.30	3.89	5.32	6.68
0.5	1.80	3.39	4.82	6.18
1.0	1.30	2.89	4.32	5.58
2.0	0.30	1.89	3.32	4.68
3.0	-0.70	0.89	2.32	3.68

Feldman-Cousins 90% Upper Limits for Poisson with Background

observed =	0	1	2	3
background = 0.0	2.44			
0.5	1.94	3.86		
1.0	1.61	3.36	4.91	
2.0	1.26	2.53	3.91	5.42
3.0	1.08	1.88	3.04	4.42