Statistics for Particle Physics
Theory, methods, and examples

Ricardo Piegaia
Physics Department
Univ Buenos Aires

August 14-16, 2008

What's the big deal about Probability and Statistics?

After all it is just mathematics, isn't it?

What's the big deal about Probability and Statistics?

After all it is just mathematics, isn't it?

Mathematics

- Developped "long ago"

What's the big deal about Probability and Statistics?

After all it is just mathematics, isn't it?

Mathematics

- Developped "long ago"
- Can be easy ...
  - Linear Algebra
  - Elementary Calculus

What's the big deal about Probability and Statistics?

After all it is just mathematics, isn't it?

Mathematics

- Developped "long ago"
- Can be easy ...
  - Linear Algebra
  - Elementary Calculus
- Can be hard ...
  - Semicompact Lie Groups
  - Topological Hausdorff Spaces

What's the big deal about Probability and Statistics?

After all it is just mathematics, isn't it?

Mathematics

- Developped "long ago"
- Can be easy ...
  - Linear Algebra
  - Elementary Calculus
- Can be hard ...
  - Semicompact Lie Groups
  - Topological Hausdorff Spaces
- but it is essentiall "done":
  XIX century, early XX century at the latest...

Same thing when I studied Probability in school:

- To a large extent "closed"
- Fun
- Basically applicable to gaming theory

Same thing when I studied Probability in school:

- To a large extent "closed"
- Fun
- Basically applicable to gaming theory

Statistics was somewhat different (and murkier..)

- Basically simple sampling examples.
- The χ^2 recipe.
- Not at all like the clean Mathematics I was used to.

Same thing when I studied Probability in school:

- To a large extent "closed"
- Fun
- Basically applicable to gaming theory

Statistics was somewhat different (and murkier..)

- Basically simple sampling examples.
- The χ^2 recipe.
- Not at all like the clean Mathematics I was used to.

And then I became an experimental high energy physicists....

Lots of interesting recipes that help solve all kind of useful stuff.

Lots of unsolved problems: Far from closed or done.

Lots of interesting recipes that help solve all kind of useful stuff.

Lots of unsolved problems: Far from closed or done.

Lots of activity of people trying to understand:

- why we do what we do
- how to do it better
- what to do next

Lots of interesting recipes that help solve all kind of useful stuff.

Lots of unsolved problems: Far from closed or done.

Lots of activity of people trying to understand:

- ➩ why we do what we do
- ➩ how to do it better
- ➩ what to do next

In fact most of what I am going to tell in these three lectures comes from papers published by physicists in the last 10 years.

Lots of interesting recipes that help solve all kind of useful stuff.

Lots of unsolved problems: Far from closed or done.

Lots of activity of people trying to understand:

- ➪ why we do what we do
- ➪ how to do it better
- ➪ what to do next

In fact most of what I am going to tell in these three lectures comes from papers published by physicists in the last 10 years.

PHYSICAL REVIEW D VOLUME 57, NUMBER 7 1 APRIL 1998

Unified approach to the classical statistical analysis of small signals

Gary J. Feldman*
Department of Physics, Harvard University, Cambridge, Massachusetts 02138

Robert D. Cousins†
Department of Physics and Astronomy, University of California, Los Angeles, California 90095

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 41, NUMBER 8 AUGUST 2000

The statistical analysis of Gaussian and Poisson signals near physical boundaries

Mark Mandelkern and Jonas Schultz
Department of Physics and Astronomy, University of California, Irvine, California 92697

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 41, NUMBER 8 AUGUST 2000

The statistical analysis of Gaussian and Poisson signals near physical boundaries

Mark Mandelkern and Jonas Schultz
Department of Physics and Astronomy, University of California, Irvine, California 92697

PHYSICAL REVIEW D **67**, 012002 (2003)

Including systematic uncertainties in confidence interval construction for Poisson statistics

J. Conrad, O. Botner, A. Hallgren, and C. Pérez de los Heros
Division of High Energy Physics, Uppsala University, S-75121 Uppsala, Sweden

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 41, NUMBER 8 AUGUST 2000

The statistical analysis of Gaussian and Poisson signals near physical boundaries

Mark Mandelkern and Jonas Schultz
Department of Physics and Astronomy, University of California, Irvine, California 92697

PHYSICAL REVIEW D **67**, 012002 (2003)

Including systematic uncertainties in confidence interval construction for Poisson statistics

J. Conrad, O. Botner, A. Hallgren, and C. Pérez de los Heros
Division of High Energy Physics, Uppsala University, S-75121 Uppsala, Sweden

ELSEVIER Nuclear Instruments and Methods in Physics Research A 551 (2005) 493–503
RESEARCH Section A
www.elsevier.com/locate/nima

Limits and confidence intervals in the presence of nuisance parameters

Wolfgang A. Rolke[a,*], Angel M. López[b], Jan Conrad[c]

ELSEVIER

Nuclear Instruments and Methods in Physics Research A 570 (2007) 159–164

RESEARCH
Section A

www.elsevier.com/locate/nima

Statistical errors in Monte Carlo estimates of systematic errors ☆

Byron P. Roe*

ELSEVIER Nuclear Instruments and Methods in Physics Research A 570 (2007) 159–164
RESEARCH Section A
www.elsevier.com/locate/nima

Statistical errors in Monte Carlo estimates of systematic errors ☆

Byron P. Roe*

Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment
Article in Press, Uncorrected Proof - Note to users

Abstract | Article | Figures/Tables | References | Purchase PDF (1050 K)

doi:10.1016/j.nima.2008.07.086 Cite or Link Using DOI

Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process

Robert D. Cousins[a], James T. Linnemann[b], and Jordan Tucker[a],

[a]Department of Physics and Astronomy, University of California, Los Angeles, CA 90095, USA [b]Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48840, USA

Probability: Relation to Statistics

Statistics is to a large extent the inverse problem of Probability

Probability:
Know parameters that describe theory $\Rightarrow$ predict probability of result

Statistics:
Know result $\Rightarrow$ extract information on the parameters and/or the theory

Probability:

b-tagging efficiency is 97% $\Rightarrow$

$P\,(\text{tag } 65 \leq n \leq 72 \text{ out of } N = 75\ b\text{-jets}) = 39.165\%$

Probability:

b-tagging efficiency is 97% $\Rightarrow$

$P\,(\text{tag } 65 \leq n \leq 72 \text{ out of } N = 75\ b\text{-jets}) = 39.165\%$

Statistics:

b-tagging algorithm selects 73 out of 75 b-jets.

What can we say about the algorithm efficiency?

Probability:

b-tagging efficiency is 97% $\Rightarrow$

$P\,(\text{tag } 65 \leq n \leq 72 \text{ out of } N = 75 \; b\text{-jets}) = 39.165\%$

Statistics:

b-tagging algorithm selects 73 out of 75 *b*-jets.

What can we say about the algorithm efficiency?

Well, we can say it's in [91.8,99.5] with 90% CL

Probability:

b-tagging efficiency is 97% $\Rightarrow$

$P\,(\text{tag } 65 \leq n \leq 72 \text{ out of } N = 75\ b\text{-jets}) = 39.165\%$

Statistics:

b-tagging algorithm selects 73 out of 75 b-jets.

What can we say about the algorithm efficiency?

Well, we can say it's in [91.8,99.5] with 90% CL

or in [93.9,99.1] with 68% CL

Probability:

b-tagging efficiency is 97% $\Rightarrow$

$P\,(\text{tag } 65 \leq n \leq 72 \text{ out of } N = 75\ b\text{-jets}) = 39.165\%$

Statistics:

b-tagging algorithm selects 73 out of 75 *b*-jets.

What can we say about the algorithm efficiency?

Well, we can say it's in [91.8,99.5] with 90% CL

or in [93.9,99.1] with 68% CL, that is $\varepsilon = 97.3^{+1.8}_{-3.4}$

Binomial

Something very common in HEP:
An experiment of probability p is repeated N times.

Binomial

Something very common in HEP:
An experiment of probability p is repeated N times.

$$Binom\,(k\,|\,N,p) = \binom{N}{k} p^k (1-p)^{N-k}, \quad \sigma(k) = \sqrt{\mathrm{Var}(k)} = \sqrt{N\,p\,(1-p)}$$

Binomial

Something very common in HEP:
An experiment of probability p is repeated N times.

$$Binom\,(k\,|\,N, p) = \binom{N}{k} p^k (1-p)^{N-k}, \quad \sigma(k) = \sqrt{\mathrm{Var}(k)} = \sqrt{N\,p\,(1-p)}$$

Examples:

- Coin tossing!

Binomial

Something very common in HEP:
An experiment of probability p is repeated N times.

$$Binom\,(k\,|\,N,p) = \binom{N}{k} p^k (1-p)^{N-k}, \quad \sigma(k) = \sqrt{\mathrm{Var}(k)} = \sqrt{N\,p\,(1-p)}$$

Examples:

- Coin tossing!
- Efficiencies (detector, method, selection)

Binomial

Something very common in HEP:
An experiment of probability p is repeated N times.

$$Binom\,(k\,|\,N,p) = \binom{N}{k} p^k (1-p)^{N-k}, \quad \sigma(k) = \sqrt{\mathrm{Var}(k)} = \sqrt{N\,p\,(1-p)}$$

Examples:

- Coin tossing!
- Efficiencies (detector, method, selection)
- Branching Ratios

Binomial

Something very common in HEP:
An experiment of probability p is repeated N times.

$$Binom\,(k\,|\,N,p) = \binom{N}{k} p^k (1-p)^{N-k}, \quad \sigma(k) = \sqrt{\mathrm{Var}(k)} = \sqrt{N\,p\,(1-p)}$$

Examples:

- Coin tossing!
- Efficiencies (detector, method, selection)
- Branching Ratios
- Asymmetries

Poisson

Limit of binomial when $N \to \infty$ and $p \to 0$ with $N \cdot p = \mu$ finite

$$Poiss\,(k \mid \mu) = \frac{e^{-\mu}\mu^k}{k!} \qquad \sigma(k) = \sqrt{\mu}$$

LOTS of examples.

Any counting observable in colliders.

Poisson

Limit of binomial when $N \to \infty$ and $p \to 0$ with $N \cdot p = \mu$ finite

$$Poiss\,(k\,|\,\mu) = \frac{e^{-\mu}\mu^k}{k!} \qquad \sigma(k) = \sqrt{\mu}$$

LOTS of examples.

Any counting observable in colliders.

For instance, in LHC:

$$N = 1.30 \times 10^{+22} \qquad (p\text{-}p \text{ crossings per bunch})$$

$$p = 1.93 \times 10^{-21} \qquad (\text{production of a minbias event})$$

$$\mu = N \cdot p = 25 \qquad (\text{av. minbias per bunch crossing})$$

Actually the number of p per bunch is Poissonian itself because there is a tiny probability that a proton ends up in a bunch out of a huge number of starting protons.

Actually the number of p per bunch is Poissonian itself because there is a tiny probability that a proton ends up in a bunch out of a huge number of starting protons.

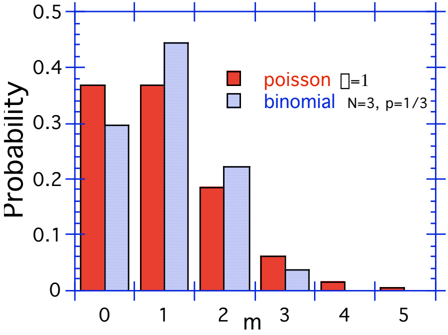
But at least somewhere we start with a true binomial experiment with (large) fixed N:

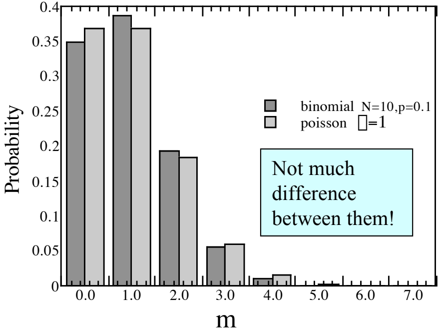
Actually the number of p per bunch is Poissonian itself because there is a tiny probability that a proton ends up in a bunch out of a huge number of starting protons.

But at least somewhere we start with a true binomial experiment with (large) fixed N:

The bottle where it *all* starts ...

But don't need to go up to $N = 10^{24}$!
At $N = 30$ Poisson and Binomial already equivalent.





Binomial $\rightarrow$ Poisson, in addition to “large” N, requires

$$\frac{10}{N} \lesssim p \lesssim 1 - \frac{10}{N}$$

And we basically always forget about binomial errors,
unless p gets very close to 0 or 1:

Binomial $\rightarrow$ Poisson, in addition to "large" N, requires

$$\frac{10}{N} \lesssim p \lesssim 1 - \frac{10}{N}$$

And we basically always forget about binomial errors, unless p gets very close to 0 or 1:

Example

75 events out of 75 pass a given cut $\quad \Rightarrow \quad \varepsilon = 100\%$

Binomial $\rightarrow$ Poisson, in addition to "large" N, requires

$$\frac{10}{N} \lesssim p \lesssim 1 - \frac{10}{N}$$

And we basically always forget about binomial errors,
unless p gets very close to 0 or 1:

Example

75 events out of 75 pass a given cut $\quad \Rightarrow \quad \varepsilon = 100\%$

With what error?

$$\varepsilon = 1 \quad \text{in} \quad \sigma = \sqrt{N\varepsilon\,(1-\varepsilon)} \quad \text{yields} \quad \sigma = 0$$

Binomial $\longrightarrow$ Poisson, in addition to “large” N, requires

$$\frac{10}{N} \lesssim p \lesssim 1 - \frac{10}{N}$$

And we basically always forget about binomial errors, unless p gets very close to 0 or 1:

Example

75 events out of 75 pass a given cut $\quad \Rightarrow \quad \varepsilon = 100\%$

With what error?

$$\varepsilon = 1 \quad \text{in} \quad \sigma = \sqrt{N\varepsilon(1-\varepsilon)} \quad \text{yields} \quad \sigma = 0$$

In this case the result is [0.976, 1.0] @68% CL

Binomial $\rightarrow$ Poisson, in addition to "large" N, requires

$$\frac{10}{N} \lesssim p \lesssim 1 - \frac{10}{N}$$

And we basically always forget about binomial errors,
unless p gets very close to 0 or 1:

Example

75 events out of 75 pass a given cut $\Rightarrow$ $\varepsilon = 100\%$

With what error?

$$\varepsilon = 1 \quad \text{in} \quad \sigma = \sqrt{N\varepsilon(1-\varepsilon)} \quad \text{yields} \quad \sigma = 0$$

In this case the result is [0.976, 1.0] @68% CL

- Need Confidence Intervals,
- A recipe for taking them into account in fits,
- No χ^2 fit, but maximum likelihood...

Multinomial

Generalization of the Binomial distribution.

N_{T} repetitions of an experiment with n possible outcomes.

Most important example: Histogram with n bins and N_{T} total entries

$$Mult\,(\boldsymbol{k}\mid N_{\mathrm{T}},\boldsymbol{p}) = \frac{N_{\mathrm{T}}!}{k_1!k_2!\cdots k_n!}\,p_1^{k_1}p_2^{k_2}\cdots p_n^{k_n}, \quad \sigma(k_i) = \sqrt{N_{\mathrm{T}}\,p_i\,(1-p_i)}$$

k_i is the number of events on the i-th bin, $\displaystyle\sum_{i=1}^{n} k_i = N_{\mathrm{T}}$.

p_i is the probability for an event to fall on the i-th bin, $\displaystyle\sum_{i=1}^{n} p_i = 1$.

Composition of Binomial and Poisson

A Binomial experiment: $Binom\,(k \,|\, N, p)$

but N itself a Poisson variable: $Poiss\,(N \,|\, \mu)$

$$\Longrightarrow \quad k \quad \text{is} \quad Poiss\,(k \,|\, \mu\, p)$$

Example:

The number k of $t\bar{t}$ triggered on a sample N is $Binom\,(n \,|\, N, \varepsilon)$

The number N of $t\bar{t}$ pairs during Run2a is $Poiss\,(N \,|\, \sigma \mathcal{L})$

$$\Longrightarrow \quad k \quad \text{is} \quad Poiss\,(n \,|\, \varepsilon \sigma \mathcal{L})$$

Composition of Multinomial and Poisson

A multinomial experiment, $Mult\,(k_i \,|\, N, p_i)$,

where N itself is a Poisson variable $Poiss\,(N \,|\, \mu)$.

$\Longrightarrow \quad k_i$ are n independent Poisson variables

$$k_i \quad \text{are} \quad Poiss\,(k_i \,|\, \mu\, p_i) \quad \Rightarrow \quad \sigma(k_i) = \sqrt{\mathrm{E}(k_i)}$$

➩ The number of entries in each bin of an histogram is Poisson.

Joint Distribution of Poisson variables

Joint probability of two Poisson $\{x, y\}$, is the product of single Poisson $z = x + y$ times a Binomial for observing x events in z trials.

$$Poiss\,(x\,|\,\mu) \times Poiss\,(y\,|\,\nu) =$$

$$= \frac{e^{-\mu}\,\mu^x}{x!} \times \frac{e^{-\nu}\,\nu^y}{y!}$$

$$= \frac{e^{-\mu}\,\mu^x}{x!} \times \frac{e^{-\nu}\,\nu^{z-x}}{(z-x)!}$$

$$= \frac{e^{-(\mu+\nu)}\,(\mu+\nu)^z}{z!} \times \frac{z!}{x!\,(z-x)!}\left(\frac{\mu}{\mu+\nu}\right)^x \left(1-\frac{\mu}{\mu+\nu}\right)^{z-x}$$

$$= Poiss\,(z\,|\,\mu+\nu) \quad \times \quad Binom\,(x\,|\,z, \frac{\mu}{\mu+\nu})$$

Application to test of Poisson ratios

$$\text{Measure} \begin{cases} n_\mathrm{P} \text{ in the peak region} & n_\mathrm{P} \sim \mathit{Poiss}(s+b) \\ n_\mathrm{C} \text{ in control region (“sidebands”)} & n_\mathrm{C} \sim \mathit{Poiss}(\tau b) \end{cases}$$

with τ the ratio of expected backgrounds in control and peak region

Application to test of Poisson ratios

$$\text{Measure} \begin{cases} n_{\mathrm{P}} \text{ in the peak region} & n_{\mathrm{P}} \sim Poiss(s+b) \\ n_{\mathrm{C}} \text{ in control region (“sidebands”)} & n_{\mathrm{C}} \sim Poiss(\tau b) \end{cases}$$

with τ the ratio of expected backgrounds in control and peak region

Suppose we want to test $H_0 : s = 0$ via $n_{\mathrm{C}}/n_{\mathrm{P}} \approx \tau$.

Application to test of Poisson ratios

Measure $\begin{cases} n_{\mathrm{P}} \text{ in the peak region} \\ n_{\mathrm{C}} \text{ in control region ("sidebands")} \end{cases}$ $\quad \begin{array}{l} n_{\mathrm{P}} \sim \mathit{Poiss}\,(s+b) \\ n_{\mathrm{C}} \sim \mathit{Poiss}\,(\tau b) \end{array}$

with τ the ratio of expected backgrounds in control and peak region

Suppose we want to test $H_0 : s = 0$ via $n_{\mathrm{C}}/n_{\mathrm{P}} \approx \tau$.

But if $s = 0$, $\quad n_{\mathrm{P}} \sim \mathit{Poiss}\,(b)$ and $n_{\mathrm{C}} \sim \mathit{Poiss}\,(\tau b)$,

or $\quad n_{\mathrm{P}} + n_{\mathrm{C}} \sim \mathit{Poiss}\,(b + \tau b)$ and $n_{\mathrm{p}} \sim \mathit{Binom}\,(n_{\mathrm{P}} + n_{\mathrm{C}}, \frac{1}{1+\tau})$

Application to test of Poisson ratios

Measure $\begin{cases} n_{\mathrm{P}} \text{ in the peak region} & n_{\mathrm{P}} \sim \mathit{Poiss}(s+b) \\ n_{\mathrm{C}} \text{ in control region ("sidebands")} & n_{\mathrm{C}} \sim \mathit{Poiss}(\tau b) \end{cases}$

with τ the ratio of expected backgrounds in control and peak region

Suppose we want to test $H_0 : s = 0$ via $n_{\mathrm{C}}/n_{\mathrm{P}} \approx \tau$.

But if $s = 0$, $n_{\mathrm{P}} \sim \mathit{Poiss}(b)$ and $n_{\mathrm{C}} \sim \mathit{Poiss}(\tau b)$,

or $n_{\mathrm{P}} + n_{\mathrm{C}} \sim \mathit{Poiss}(b + \tau b)$ and $n_{\mathrm{p}} \sim \mathit{Binom}(n_{\mathrm{P}} + n_{\mathrm{C}}, \frac{1}{1+\tau})$

The fraction of measured events that are in the "peak" region, $n_{\mathrm{P}}/(n_{\mathrm{P}} + n_{\mathrm{C}})$, is a Binomial variable that measures $\frac{1}{1+\tau}$

Application to test of Poisson ratios

Measure $\begin{cases} n_{\mathrm{P}} \text{ in the peak region} & n_{\mathrm{P}} \sim \mathit{Poiss}\,(s+b) \\ n_{\mathrm{C}} \text{ in control region (“sidebands”)} & n_{\mathrm{C}} \sim \mathit{Poiss}\,(\tau b) \end{cases}$

with τ the ratio of expected backgrounds in control and peak region

Suppose we want to test $H_0 : s = 0$ via $n_{\mathrm{C}}/n_{\mathrm{P}} \approx \tau$.

But if $s = 0$, $n_{\mathrm{P}} \sim \mathit{Poiss}\,(b)$ and $n_{\mathrm{C}} \sim \mathit{Poiss}\,(\tau b)$,

or $n_{\mathrm{P}} + n_{\mathrm{C}} \sim \mathit{Poiss}\,(b + \tau b)$ and $n_{\mathrm{p}} \sim \mathit{Binom}\,(n_{\mathrm{P}} + n_{\mathrm{C}}, \frac{1}{1+\tau})$

The fraction of measured events that are in the “peak” region, $n_{\mathrm{P}}/(n_{\mathrm{P}} + n_{\mathrm{C}})$, is a Binomial variable that measures $\frac{1}{1+\tau}$

➪ Test on ratio of Poisson variables is test on a Binomial.

Story of a rediscovery ...

This standard method was elucidated for Botanics (testing clover seed for dodder) by

Przyborowski and Wilenski, Biometrika 31 (1940) 313

and generalized for Zoology (studying salmon fry migration)

Chapman, Ann. Inst. Stat. Math. (Tokyo) 4 (1952) 45

The same result was obtained in the HEP community by

F. James, M. Roos, Nucl. Phys. B 172 (1980) 475.
"Errors on Ratios of Small Numbers of Events"

and in the GRA community

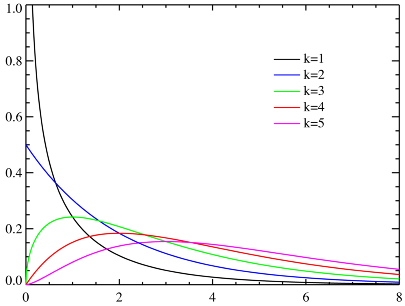
N. Gehrels, Astrophysical Journal, 303 (1986) 336
"Confidence limits for small numbers of events in astrophysical data"

Chi-square

$y \equiv x^2$:

If $x \in (-\infty, \infty)$ is $x \sim N(0,1)$

then $y \in [0, \infty)$ is $y \sim \chi^2(1)$.



For n independent $x_i \sim N(0,1)$: $\quad y \equiv \sum_i^n x_i^2 \;\Rightarrow\; y \sim \chi^2(n)$.

The exponent in the n-dim multinormal

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{V}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathbf{T}} \boldsymbol{V}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right]$$

is itself a $\chi^2(n)$ random variable.

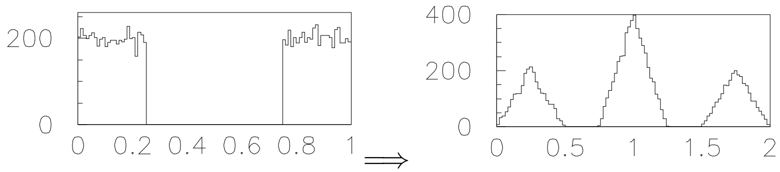
Central limit theorem

Given 2 random variables x_1 and x_2, its sum $y = x_1 + x_2$

will be a new random variable with a different distribution

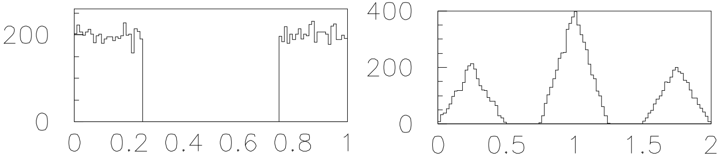
Example: the sum of two flat distributions is the triangular distribution.

Example 2:



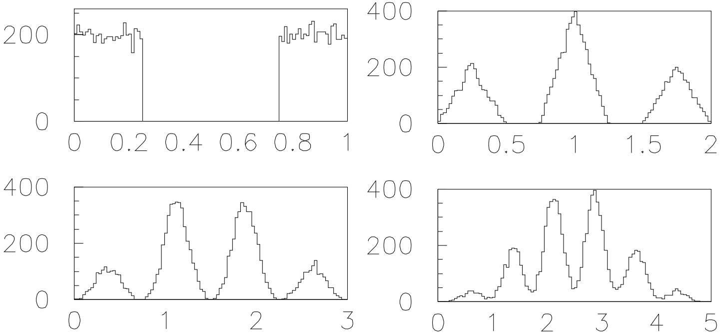
Sum of n independent random x_i, with $\mathrm{E}(x_i) = \mu_i$ and $\mathrm{Var}(x_i) = \sigma_i^2$.

tends to a $N(\mu, \sigma)$, with $\mu = \sum_i^n \mu_i$ and $\sigma^2 = \sum_i^n \sigma_i^2$



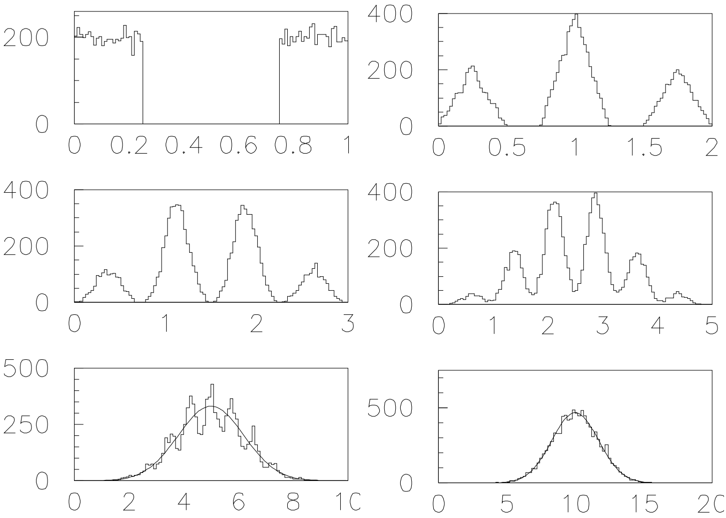
Sum of n independent random x_i, with $\mathrm{E}(x_i) = \mu_i$ and $\mathrm{Var}(x_i) = \sigma_i^2$.

tends to a $N(\mu, \sigma)$, with $\mu = \sum_i^n \mu_i$ and $\sigma^2 = \sum_i^n \sigma_i^2$



Sum of n independent random x_i, with $\mathrm{E}(x_i) = \mu_i$ and $\mathrm{Var}(x_i) = \sigma_i^2$.

tends to a $N(\mu, \sigma)$, with $\mu = \sum_i^n \mu_i$ and $\sigma^2 = \sum_i^n \sigma_i^2$



Central limit theorem: Special cases

- Sum of Binomials with equal p is Binomial:

  $$Binom\,(n_1, p) + Binom\,(n_2, p) = Binom\,(n_1 + n_2, p)$$

  $$\Longrightarrow \quad Binom\,(n, p) \rightarrow \mathrm{N}\,(np, \sqrt{np(1-p)}\,) \text{ for large } n$$

- Sum of Poissonians is Poisson:

  $$Poiss\,(\mu_1) + Poiss\,(\mu_2) = Poiss\,(\mu_1 + \mu_2)$$

  $$\Longrightarrow \quad Poiss\,(\mu) \rightarrow \mathrm{N}\,(\mu, \sqrt{\mu}\,) \text{ for large } \mu.$$

- Sum of Chi-squares is Chi-square:

  $$\chi^2(n_1) + \chi^2(n_2) = \chi^2(n_1 + n_2)$$

  $$\Longrightarrow \quad \chi^2(n) \rightarrow \mathrm{N}\,(n, \sqrt{2n}\,) \text{ for large } n.$$

The typical analysis we face is composed of roughly four steps

Physics language	*Statisticians terminology*
"Best fit" of parameters	Point estimation
Errors on the parameters	Confidence region (at given C.L.)
Judging quality of the fit	Goodness-of-fit testing
Compare to theory	Hypothesis testing (at significance level)

Point Estimation

A random variable depends on a parameter θ: $f(x \mid \theta)$

By measuring a sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$

we want to infer the value of θ.

An estimator $\hat{\theta}$ of the parameter θ

- is a random variable,
- function of the sample $\boldsymbol{x}$: $\hat{\theta} = \hat{\theta}\,(x_1, \ldots, x_n)$
- that can have the following properties: Consistency, Bias, Efficiency, Sufficiency, Robustness

Consistency (for an infinite sample):

$$\lim_{n\to\infty} \hat{\theta} = \theta$$

.

Bias

Bias is defined for a finite sample: $b \equiv \mathrm{E}(\hat{\theta}) - \theta$

An estimator is unbiased if $\mathrm{E}(\hat{\theta}) = \theta$

Classical example: Two consistent estimators for σ^2

$$S^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 \quad \text{biased estimator with} \quad b = -\frac{\sigma^2}{n}$$

$$s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 \quad \text{unbiased}$$

Efficiency

There can be numerous consistent unbiased estimators of θ in $f(x \,|\, \theta)$:

$\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$, with different variances.

There is a minimum attainable variance given by Cramer-Rao bound:

$$\forall\, \hat{\theta}(\boldsymbol{x}) \quad \text{with} \quad \mathrm{E}(\hat{\theta}) = \theta :$$

$$\mathrm{Var}(\hat{\theta}) \geq \sigma^2_{\min} = \frac{1}{\mathrm{E}\left[\left(\frac{\partial}{\partial\theta} \sum_i \log f(x_i \,|\, \theta)\right)^2\right]}$$

$$\text{Efficiency } \hat{\theta} \equiv \frac{\sigma^2_{\min}}{\mathrm{Var}(\hat{\theta})}$$

Example: $x_i \sim \mathrm{N}(\mu, \sigma_i)$

n measurements of same physical quantity, different errors.

Three unbiased estimators of μ:

$$\hat{\mu}_2(\boldsymbol{x}) = \frac{\sum(x_i/\sigma_i^2)}{\sum(1/\sigma_i^2)} \qquad \hat{\mu}_1(\boldsymbol{x}) = \frac{\sum(x_i/\sigma_i)}{\sum(1/\sigma_i)} \qquad \hat{\mu}_0(\boldsymbol{x}) = \frac{\sum x_i}{n}$$

$$\sigma(\hat{\mu}_2) \quad < \quad \sigma(\hat{\mu}_1) \quad < \quad \sigma(\hat{\mu}_0)$$

$\hat{\mu}_2$ is 100% Efficient only for x_i gaussian,

Sufficiency: we don't loose information when replacing the n measurements $\boldsymbol{x}$, by the sole number $\hat{\theta}(\boldsymbol{x})$.

Robustness: not unduly affected by small departures from model assumptions (e.g., insensitivity to what goes on at the tails of the distribution)

The likelihood function

Random variable that depends on θ: $f(x \,|\, \theta)$

The probability to obtain the n independent measurements $\{x_i\}$ is

$$f(\boldsymbol{x} \,|\, \theta) = \prod_{i=1}^{n} f(x_i \,|\, \theta)$$

The likelihood function is exactly this same expression,

but thought as a function of θ, given the measurements $\{x_i\}$

$$\mathscr{L}(\theta \,|\, \boldsymbol{x}) \quad \text{or} \quad \mathscr{L}(\boldsymbol{x} \,|\, \theta) \quad \equiv \quad \prod_{i=1}^{n} f(x_i \,|\, \theta)$$

The notation $\mathscr{L}$ stresses that we mean fixed data $\{x_i\}$.

$\mathscr{L}(\theta \,|\, \boldsymbol{x})$ in *not* a probability density for θ: $\int \mathscr{L}(\theta \,|\, \boldsymbol{x}) \, \mathrm{d}\theta \neq 1$

Maximum likelihood estimator

Obtain the estimator $\hat{\theta}$ by maximizing $\mathscr{L}$:

$$\left.\frac{\partial \mathscr{L}(\theta \,|\, x_i)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

Solution of this equation (analytical or numerical) yields $\hat{\theta} = \hat{\theta}(\boldsymbol{x})$.

Properties:

- ML estimators are consistent.
- ML will produce a sufficient, 100% efficient estimator, if it exists.
- ML estimators are asymptotically 100% efficient, sufficient and unbiased.

Method of least squares

When the probability $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ is gaussian, the maximum likelihood principle yields the method of least squares, also known as "minimizing" the χ^2 (square of a gaussian)

$$\mathscr{L}(\boldsymbol{x} \mid \theta) = \mathrm{C} \prod_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} \implies \log \mathscr{L} = -\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 + \mathrm{C'}$$

Maximizing $\mathscr{L}$ equals minimizing the sum of gaussians squared.

If $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ is not gaussian, one can still apply least squares.

Gauss-Markov Theorem: Among all unbiased estimators that are linear in the data (gaussian or not gaussian), the Least Squares method produces the estimator with smallest variance.

The second step in you job, is to find the error

on the parameter you have estimated

Confidence Interval

Confidence Interval: Simple gaussian case

Random variable x with gaussian distribution $N(x \mid \mu, \sigma)$

Assume that the precision of the instrument, σ is known.

Perform a measurement and obtain x. Probability then states

$$P(\ \mu - \sigma \ \leq \ x \ \leq \ \mu + \sigma) = 0.6827 \approx 0.68$$

But

$$\mu - \sigma \leq x \ \Rightarrow \ \mu \leq x + \sigma \quad \text{and} \quad x \leq \mu + \sigma \ \Rightarrow \ x - \sigma \leq \mu$$

Then

$$P(\ x - \sigma \ \leq \ \mu \ \leq \ x + \sigma) \ = \ 0.68$$

Last equation again:

$$P(\ x-\sigma \ \leq \ \mu \ \leq \ x+\sigma) \ = \ 0.68$$

This doesn't mean that μ has a 68% probability of being in $x \pm \sigma$.

μ is NO random variable, it is a FIXED parameter.

Here $[x-\sigma, x+\sigma]$ is a random interval,

that will contain the fixed parameter μ, 68% of the time .

This is the frequentist interpretation of “error”

We write $x \pm \sigma$ and $x \pm 2\sigma$ meaning 68% and 95% CL intervals.

Neyman's construction

It is not always possible to isolate analytically the parameter of interest.

For instance, we have a n measurements $x_i \sim \mathrm{N}(\mu, \sigma)$.

Want to estimate σ^2 with its error (confidence region at 68% CL)

Use the well known unbiased estimator

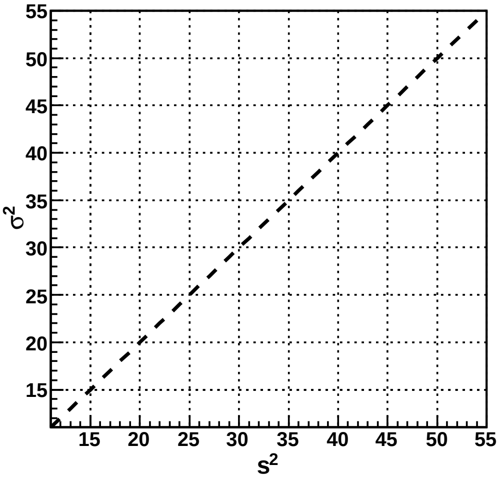
$$s^2 = \frac{1}{n-1} \sum_i^n \left(x_i - \frac{\sum x_i}{n} \right)^2$$

To get the error need the distribution of the random variable s^2.

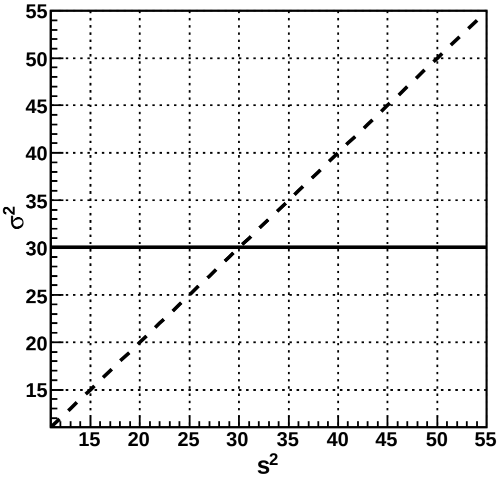
$$x_i \sim \mathrm{N}(\mu, \sigma) \quad \Longrightarrow \quad \frac{(n-1)\, s^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_i^n (x_i - \bar{x})^2 \;\sim\; \chi^2_{n-1}$$

Note that the distribution of s^2 depends on the unknown parameter σ^2

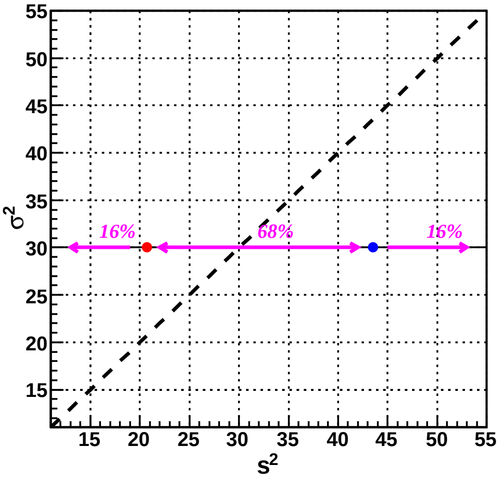
For each σ^2, get s_{d}^2 and s_{u}^2: $\int_0^{s_{\mathrm{d}}^2} \chi_{n-1}^2 \, \mathrm{d}u = 0.16 \quad \int_{s_{\mathrm{u}}^2}^{\infty} \chi_{n-1}^2 \, \mathrm{d}u = 0.16$



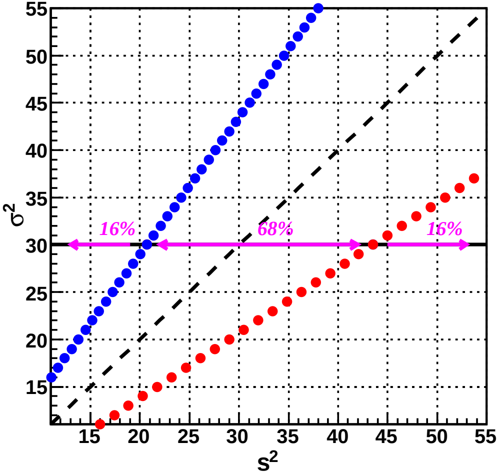
For each σ^2, get s_d^2 and s_u^2: $\int_0^{s_\mathrm{d}^2} \chi_{n-1}^2 \,\mathrm{d}u = 0.16 \quad \int_{s_\mathrm{u}^2}^{\infty} \chi_{n-1}^2 \,\mathrm{d}u = 0.16$



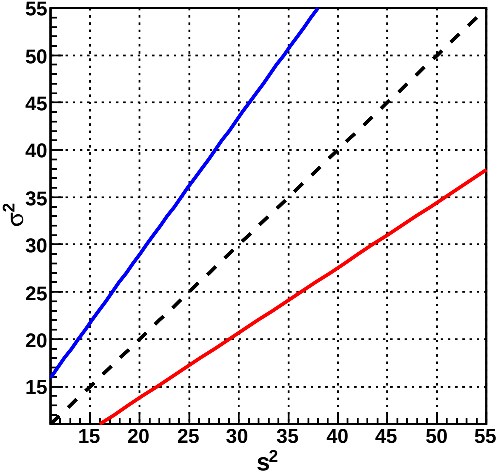
For each σ^2, get s_{d}^2 and s_{u}^2: $\int_0^{s_{\mathrm{d}}^2} \chi_{n-1}^2 \, \mathrm{d}u = 0.16 \quad \int_{s_{\mathrm{u}}^2}^{\infty} \chi_{n-1}^2 \, \mathrm{d}u = 0.16$



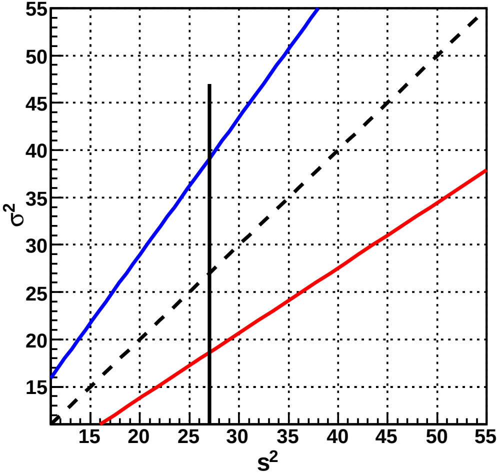
For each σ^2, get s_d^2 and s_u^2: $\int_0^{s_\mathrm{d}^2} \chi^2_{n-1}\, \mathrm{d}u = 0.16 \quad \int_{s_\mathrm{u}^2}^{\infty} \chi^2_{n-1}\, \mathrm{d}u = 0.16$



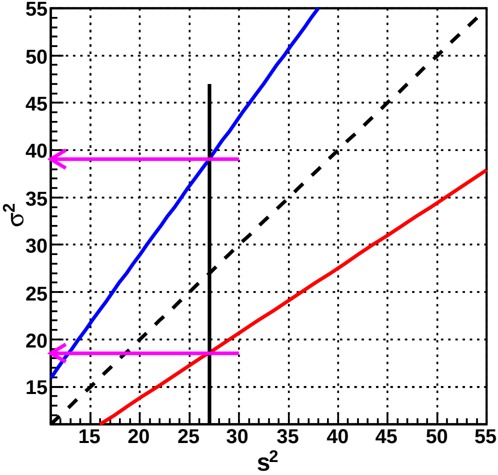
For each σ^2, get s_d^2 and s_u^2: $\int_0^{s_\mathrm{d}^2} \chi_{n-1}^2 \, \mathrm{d}u = 0.16 \quad \int_{s_\mathrm{u}^2}^{\infty} \chi_{n-1}^2 \, \mathrm{d}u = 0.16$



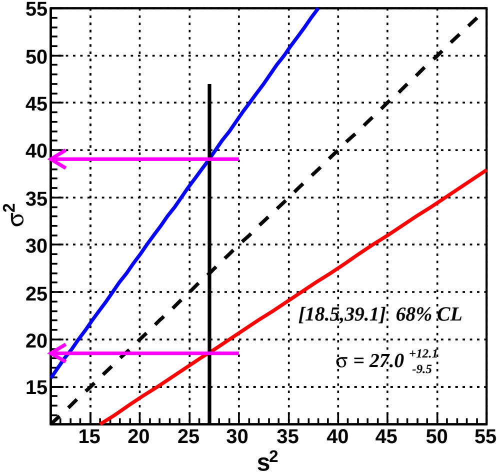
For each σ^2, get s_d^2 and s_u^2: $\int_0^{s_\mathrm{d}^2} \chi^2_{n-1}\,\mathrm{d}u = 0.16 \quad \int_{s_\mathrm{u}^2}^{\infty} \chi^2_{n-1}\,\mathrm{d}u = 0.16$



For each σ^2, get s^2_{d} and s^2_{u}: $\int_0^{s^2_{\mathrm{d}}} \chi^2_{n-1}\,\mathrm{d}u = 0.16 \quad \int_{s^2_{\mathrm{u}}}^{\infty} \chi^2_{n-1}\,\mathrm{d}u = 0.16$



For each σ^2, get s_{d}^2 and s_{u}^2: $\int_0^{s_{\mathrm{d}}^2} \chi^2_{n-1}\,\mathrm{d}u = 0.16 \quad \int_{s_{\mathrm{u}}^2}^{\infty} \chi^2_{n-1}\,\mathrm{d}u = 0.16$



Coverage

By construction, for all values of the unknown σ^2:

$$P\big(\sigma^2 \in [\sigma^2_{\mathrm{d}}, \sigma^2_{\mathrm{u}}]\big) = 0.68 \quad \forall\, \sigma^2$$

This expresses that the “confidence belt” we built has *coverage*:

A method is said to yield a $100\,\alpha\,\%$ Confidence Interval if,
were the experiment to be repeated many times,
the resulting intervals would include (or cover) the true parameter
at least $100\,\alpha\,\%$ of the time,
no matter what the value of the true parameter is.

Coverage is, in the frequentist approach, the main property which confidence intervals have to fulfill.

The construction of the confidence belt is far from unique.

In the example we have built the "central" C.I.

one could choose the "shortest", or upper, or lower, limits.

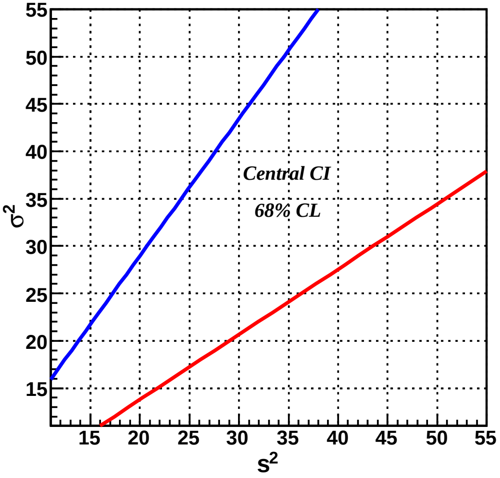
The confidence belt depends also on which estimator you choose for your measurement.

Some choices for classical confidence intervals

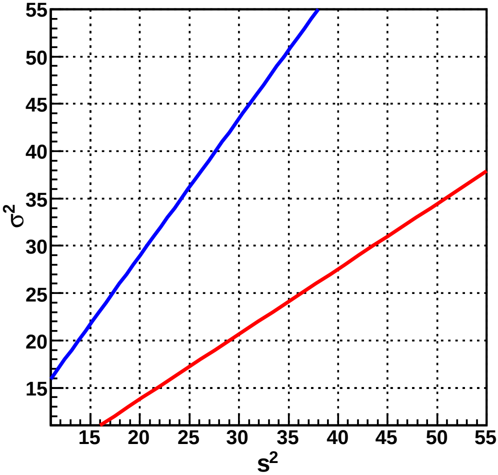
central interval	$P(x \leq x_\mathrm{d} \mid \theta) = P(x \geq x_\mathrm{u} \mid \theta) = (1-\alpha)/2$
equal probability densities	$f(x_\mathrm{d} \mid \theta) = f(x_\mathrm{u} \mid \theta)$
minimum size	$\theta_\mathrm{high} - \theta_\mathrm{low}$ is minimum
symmetric	$\theta_\mathrm{high} - \hat{\theta} = \hat{\theta} - \theta_\mathrm{low}$
upper limit	$\theta_\mathrm{low} = -\infty$
lower limit	$\theta_\mathrm{high} = +\infty$
likelihood ratio ordering	$f(x_\mathrm{d} \mid \theta)/f(x_\mathrm{d} \mid \theta_\mathrm{best}) = f(x_\mathrm{u} \mid \theta)/f(x_\mathrm{u} \mid \theta_\mathrm{best})$

A few more confidence belts for free ...

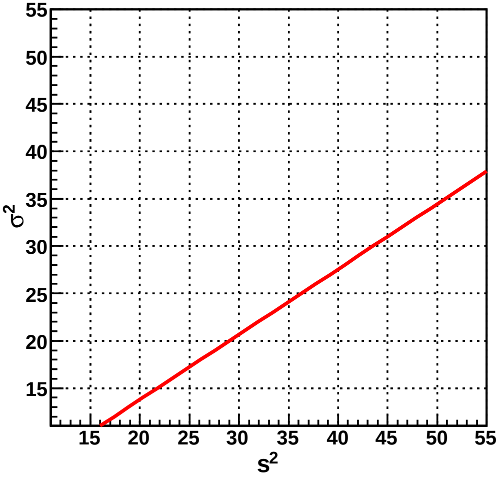
A few more confidence belts for free ...



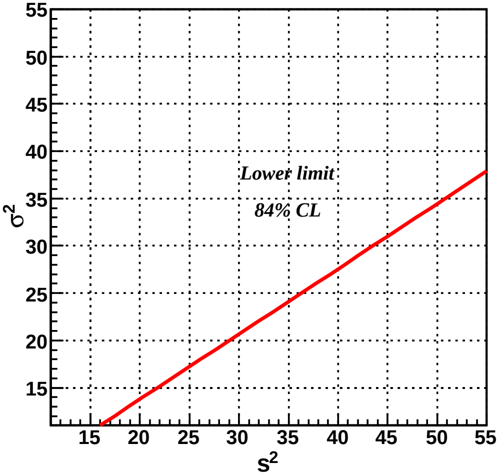
A few more confidence belts for free ...



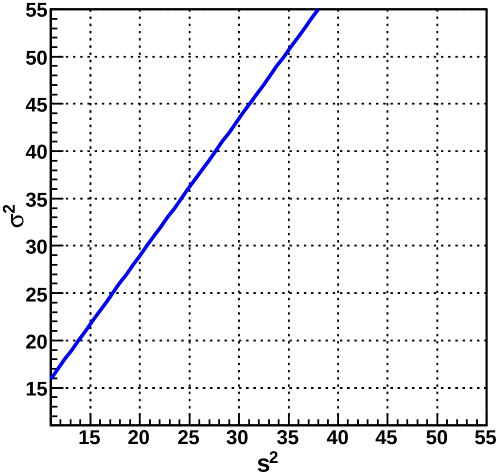
A few more confidence belts for free ...



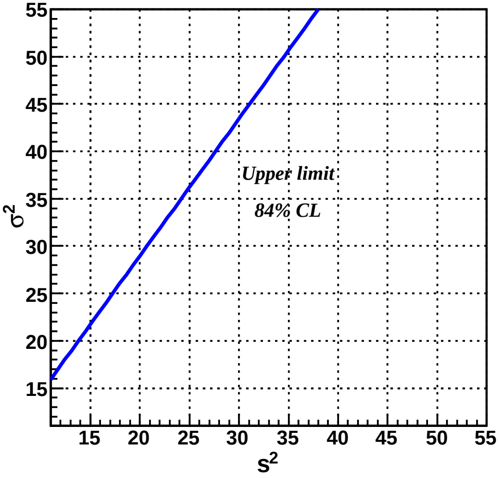
A few more confidence belts for free ...



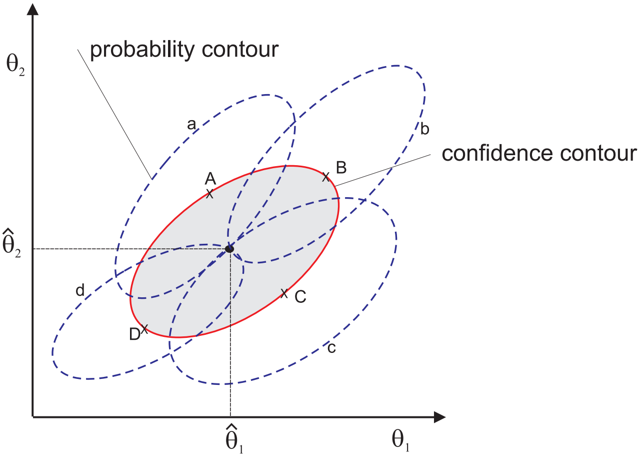
A few more confidence belts for free ...



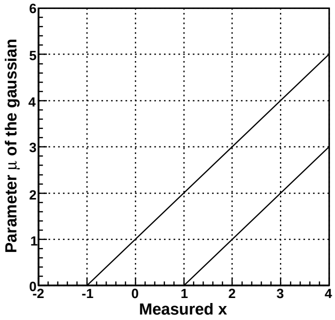
A few more confidence belts for free ...



Confidence Interval: Two-dimensional case

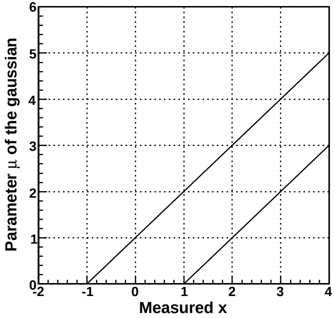


Confidence Interval near a bound

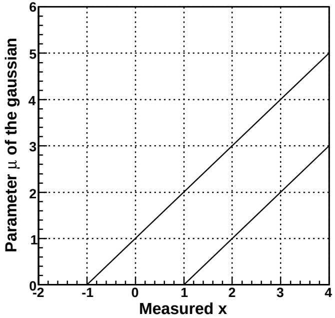


Central 68% confidence belt for a gaussian $N(\mu, 1)$ when for physics reasons we know $\mu \geq 0$ (like a mass or a production ratio)

$\forall\ \mu \geq 0$, obtain $[x_1(\mu), x_2(\mu)]$ as $P(x < x_1 \,|\, \mu) = P(x > x_2 \,|\, \mu) = 0.16$.



If measure: $x = +3.0$	$\Longrightarrow$	$2 < \mu < 4$	at 68% CL
If measure: $x = +0.8$	$\Longrightarrow$	$0 < \mu < 1.8$	at 68% CL
If measure: $x = -0.8$	$\Longrightarrow$	$0 < \mu < 0.2$	at 68% CL
If measure: $x = -1.5$	$\Longrightarrow$	Empty C.I	at 68% CL



If measure: $x = +3.0$	$\Longrightarrow$	$2 < \mu < 4$	at 68% CL
If measure: $x = +0.8$	$\Longrightarrow$	$0 < \mu < 1.8$	at 68% CL
If measure: $x = -0.8$	$\Longrightarrow$	$0 < \mu < 0.2$	at 68% CL
If measure: $x = -1.5$	$\Longrightarrow$	Empty C.I	at 68% CL

If you dislike these results, means you're a potential Bayesian!

IS THIS WRONG?

IS THIS WRONG? Nope.

IS THIS WRONG? Nope.

Frequentists say that in 68% of the cases your interval contains the true value of μ (remember *coverage*?)

This means 32% of the cases IT WILL NOT.

If you got an empty interval: TOO BAD, you fell in the unlucky 32%!

Trouble is you KNOW you were unlucky and you don't like it

And what about $0 < \mu < 0.2$ with 68% C.L.?

How come we got so precise in an experiment when $\sigma = 1$?

Answer: It's not supposed to mean that you have 68% *belief* that the true μ is in your interval.

It doesn't say anything about your particular interval.
It says something about the set of CI of experiments you didn't do.

In fact, in cases where μ is physically within a bounded domain, you could get a 68% CI that covers the whole domain!

Imagine publishing:

The branching ratio is between 0 and 1 with 68% CL !

The Bayesian way

Bayesians on the contrary do MEAN that

if you say $0 < \mu < 0.2$ (68% C.L.)

then it's because you are ready to bet

with odds 68/32 ($\sim$2/1) that μ IS in the interval.

And if your CI covers the whole domain,

for bayesians that is a 100% CL.

Of course in Bayesian statistics you can never get an empty interval.

Then ...

Then ...

Why isn't every physicist a Bayesian?

Robert D. Cousins
Department of Physics, University of California, Los Angeles, California 90024-1547

(Received 1 June 1994; accepted 3 November 1994)

Then ...

Why isn't every physicist a Bayesian?

Robert D. Cousins
Department of Physics, University of California, Los Angeles, California 90024-1547

(Received 1 June 1994; accepted 3 November 1994)

The price to pay is that you have to think of the charge of the electron as a random variable. But that's not the only price.

Then ...

Why isn't every physicist a Bayesian?

Robert D. Cousins
Department of Physics, University of California, Los Angeles, California 90024-1547

(Received 1 June 1994; accepted 3 November 1994)

The price to pay is that you have to think of the charge of the electron as a random variable. But that's not the only price.

"Frequentists use impeccable logic to deal with an issue of no interest to anyone"

Then ...

Why isn't every physicist a Bayesian?

Robert D. Cousins
Department of Physics, University of California, Los Angeles, California 90024-1547

(Received 1 June 1994; accepted 3 November 1994)

The price to pay is that you have to think of the charge of the electron as a random variable. But that's not the only price.

"Frequentists use impeccable logic to deal with an issue of no interest to anyone"

"Bayesians address the question everyone is interested in, by using assumptions no-one believes"

Discrete case: Poisson process with background

Observe n events, from unknown signal μ and background $b = 3$

$$P(n \,|\, \mu) = Poiss\,(n \,|\, \mu + b) = \frac{e^{-(\mu+b)}\,(\mu+b)^n}{n!}$$

Confidence belt at $100\,\alpha\,\%$ CL:

for each μ find $[n_1, n_2]$ such that $P\,(n \in [n_1, n_2] \,|\, \mu) = \alpha$

Central 90%: $P(n < n_1 \,|\, \mu) = 0.05$ and $P(n > n_2 \,|\, \mu) = 0.05$

Upper 90%: $P(n < n_1 \,|\, \mu) = 0.10$

Let's look at n_1 for the upper limit

$$0.10 = P\,(n < n_1 \,|\, \mu) = \sum_{n=0}^{n_1-1} \frac{e^{-(\mu+3)}\,(\mu+3)^n}{n!}$$

Since n_1 is discrete, only have exact solutions for certain μ.

$$
\begin{aligned}
& 0.10 = P\,(n < n_1 \mid \mu) & \\
n_1 = 1: \quad & 0.10 = e^{-(\mu+3)} \times\ 1 & \Longrightarrow \text{ no solution} \\
n_1 = 2: \quad & 0.10 = e^{-(\mu+3)} \times \left[1 + (\mu+3)\right] & \Longrightarrow \mu = 0.88972 \\
n_1 = 3: \quad & 0.10 = e^{-(\mu+3)} \times \left[1 + (\mu+3) + \tfrac{1}{2}(\mu+3)^2\right] & \Longrightarrow \mu = 2.32232
\end{aligned}
$$

Exact coverage is not possible: either "overcover" or "undercover".

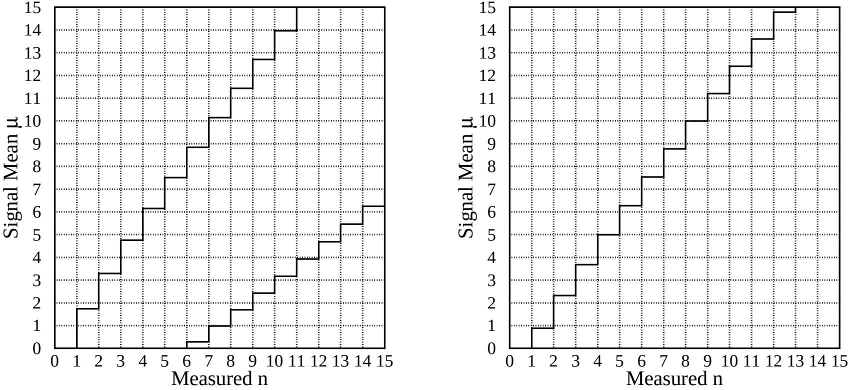
Avoid undercoverage by replacing

$$
P\,(n \in [n_1, \infty) \mid \mu) = 0.90 \quad \longrightarrow \quad P\,(n \in [n_1, \infty) \mid \mu) \geq 0.90
$$

Thus the choice is

$$
\begin{aligned}
0.0 \ \leq \mu < 0.88972 \quad & \Rightarrow \quad n_1 = 1 \\
0.88972 \ \leq \mu < 2.32232 \quad & \Rightarrow \quad n_1 = 2
\end{aligned}
$$

Minimum overcoverage 90% C.L. confidence belts for central confidence intervals and upper limit, for unknown Poisson signal mean and Poisson background $b = 3$.



With the choice $P\,(n \in [n_1, n_2] \,|\, \mu) \geq \alpha$

the intervals overcover and are conservative.

This is unavoidable for discrete distributions, but NO good.

A 90% C.I.interval *should* fail 10% of the time.

If want intervals that cover more than 90%, don't add conservatism, but rather go to higher confidence levels.

Flip-Flopping

Ideal Physicist	*Real Physicist*
Choose Strategy	Examine data
Examine data	Choose Strategy
Quote result	Quote Result

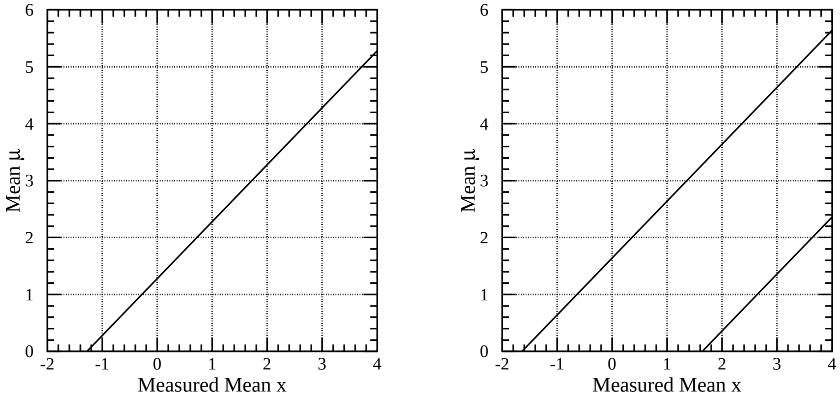
Example:

You have a background of 3.2

Observe 5 events? No discovery: Quote one-sided upper limit

Observe 25 events? Discovery: Quote two-sided confidende interval.

An experiment designed to measure a positive quantity;



Which one to use?

One may choose the following startegy:

If the result x is less than $3\,\sigma$ above zero, state an upper limit

If greater than $3\,\sigma$, state a central confidence interval

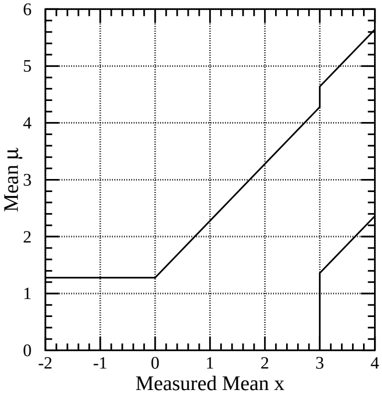
If measured value is negative, be conservative and pretend measured zero when calculating interval.

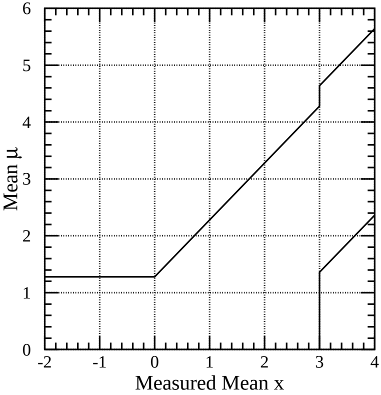
One may choose the following startegy:

If the result x is less than $3\,\sigma$ above zero, state an upper limit

If greater than $3\,\sigma$, state a central confidence interval

If measured value is negative, be conservative and pretend measured zero when calculating interval.





For $\mu = 2.0$, acceptance interval is $x_1 =$ 2-1.28 and $x_2 =$ 2+1.64,

$P(x_1 \leq x \leq x_2 \,|\, \mu = 2.0) = 85\% \,<\, 90\% \;\Rightarrow$ intervals *undercover*

They are *not* confidence intervals and certainly not "conservative" CI.

Problems:

- If you use the data to decide which plot to use, the hybrid method can undercover
- Your CI can be the empty set, or unreasonably “precise”.
- “Worse” experiment with larger expected background can get “better” CI.

Let’s discuss briefly this 3$^{\mathrm{rd}}$ point.

CASE I: Experiment expects no background, and observes no signal.

Frequentist 90% upper limit? Reject all values of μ for which

$$P\,(0\,|\,\mu) = Poiss\,(0\,|\,\mu) = \exp\,(-\mu) \quad \text{is less than } 10\%$$

$$\begin{aligned} P\,(0\,|\,\mu_{\text{reject}}) &< 0.10 \\ \exp\,(-\mu_{\text{reject}}) &< 0.10 \\ -\mu_{\text{reject}} &< \log 0.10 = -\log 10 \\ \mu_{\text{reject}} &> 2.30 \end{aligned}$$

CASE II: Experiment expects mean background b, observes no signal.

$$P\,(0\,|\,\mu) = Poiss\,(0\,|\,\mu + b) = \exp\left[-(\mu + b)\right]$$

$$\begin{aligned} P\,(0\,|\,\mu_{\text{reject}}) &< 0.10 \\ \exp\left[-(\mu_{\text{reject}} + b)\right] &< 0.10 \\ -(\mu_{\text{reject}} + b) &< \log 0.10 \\ \mu_{\text{reject}} &> 2.30 - b \end{aligned}$$

90% CL frequentist and Bayesian upper limits
for $n = 0$ observed events and background expectation b

	$b = 0$	$b = 1$	$b = 2$	$b = 3$	$b = 4$
Standard Classical	2.30	1.30	0.30	$\varnothing$	$\varnothing$
Unified Classical	2.44	1.61	1.26	1.08	1.01
Uniform Bayesian	2.30	2.30	2.30	2.30	2.30

The same problem that in the gaussian case.

If the experiment measures $n = 0$ it yields an empty set.

Should the experiment report "No result at 90% CL"?

The “unified” approach: Feldman-Cousins

Back to the confidence belt for a Poisson experiment with $b = 3$

Consider the horizontal acceptance interval at signal mean $\mu = 0.5$

The probability of obtaining $n = 0$ events is $\exp\left[-(0.5 + 3)\right] = 0.03$

Pretty low. But,compared to what?

If we got $n = 0$, our best bet for μ is $\mu_{\text{best}} = 0$

And for our best bet, the probability is $P(0 \,|\, \mu_{\text{best}}) = 0.05$

Now, 0.03 is *not* much smaller than 0.05, so $\mu = 0$ is not *that* bad.

Take the ratio 0.03/0.05=0.607 as figure of merit for $\mu = 0$ hypothesis.

For each n let μ_{best} be that value of μ which maximizes $P(n \mid \mu)$ within the physically allowed region (non-negative μ).

Thus, $\mu_{\text{best}} = \max(0, n - b)$.

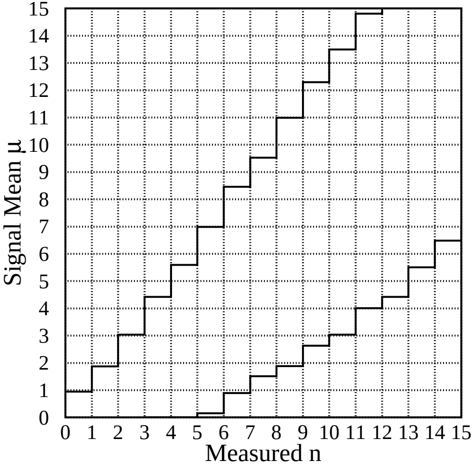
Choose what values of n to include in the confidence belt following a merit ordering based on the ratio of likelihoods

$$R = \frac{\mathscr{L}(n \mid \mu)}{\mathscr{L}(n \mid \mu_{\text{best}})}$$

Construction of confidence belt for signal mean $\mu = 0.5$ in the presence of known mean background $b = 3.0$.

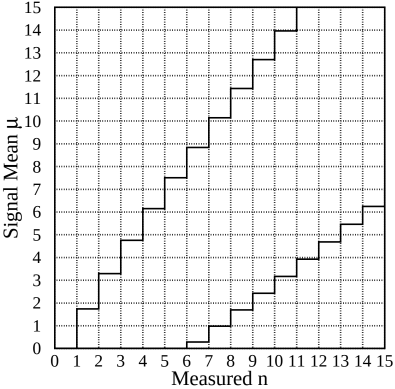
n	$P(n \mid \mu)$	$\mu_{\rm best}$	$P(n \mid \mu_{\rm best})$	R	rank	U.L.	central
0	0.030	0.	0.050	0.607	6		
1	0.106	0.	0.149	0.708	5	$\surd$	$\surd$
2	0.185	0.	0.224	0.826	3	$\surd$	$\surd$
3	0.216	0.	0.224	0.963	2	$\surd$	$\surd$
4	0.189	1.	0.195	0.966	1	$\surd$	$\surd$
5	0.132	2.	0.175	0.753	4	$\surd$	$\surd$
6	0.077	3.	0.161	0.480	7	$\surd$	$\surd$
7	0.039	4.	0.149	0.259		$\surd$	$\surd$
8	0.017	5.	0.140	0.121		$\surd$	
9	0.007	6.	0.132	0.050		$\surd$	
10	0.002	7.	0.125	0.018		$\surd$	
11	0.001	8.	0.119	0.006		$\surd$	

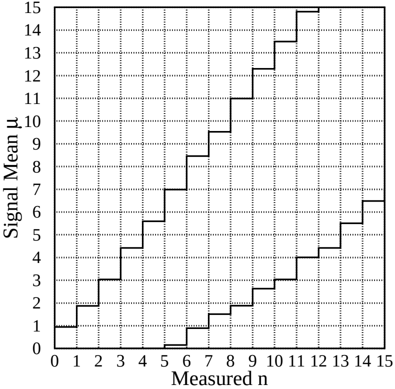
This process is repeated for each μ and yields



Because of the discreteness of *n*, the acceptance region contains a summed probability greater than 90%.

Comparison of standard and unified confidence belts





FC: Gaussian case near physical boundary

For a particular x, $\mu_{\rm best}$ is the physically allowed value of μ for which $P(x\,|\,\mu)$ is maximum. This is $\mu_{\rm best} = \max(0, x)$

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

And the likelihood ratio R :

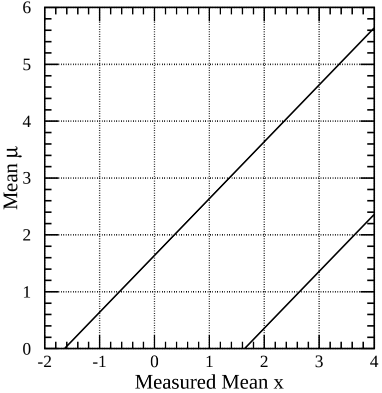
$$R(x) = \frac{P(x\,|\,\mu)}{P(x\,|\,\mu_{\rm best})} = \begin{cases} \exp\left(-(x-\mu)^2/2\right), & x \geq 0 \\ \exp\left(x\mu - \mu^2/2\right), & x < 0. \end{cases}$$

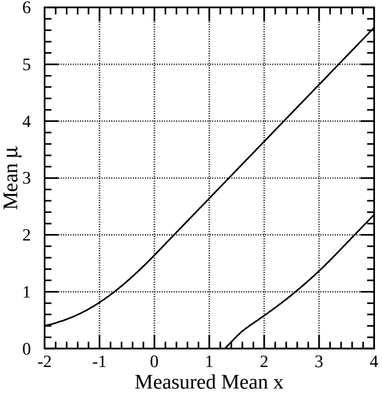
For a given μ, the acceptance interval $[x_1, x_2]$ satisfies

$$R(x_1) = R(x_2) \quad \text{and} \quad \int_{x_1}^{x_2} P(x\,|\,\mu)\,\mathrm{d}x = \alpha$$

Here the coverage is exactly 90% by construction.

Comparison of standard and unified confidence belts





FC does not solve the problem of shrinking CI for increasing background

90% CL frequentist and Bayesian upper limits
for $n = 0$ observed events and background expectation b

	$b = 0$	$b = 1$	$b = 2$	$b = 3$	$b = 4$
Standard Classical	2.30	1.30	0.30	$\varnothing$	$\varnothing$
Unified Classical	2.44	1.61	1.26	1.08	1.01
Uniform Bayesian	2.30	2.30	2.30	2.30	2.30

FC advocate to inform also the sensitivity of the experiment:

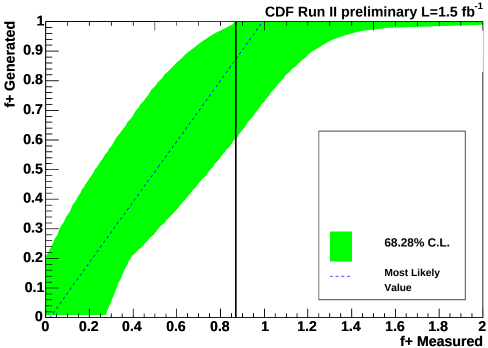
the average upper limit one would get from an ensemble of experiments with your expected background and no true signal.

Preliminary result from CDF on the top quark charge

f+ is fraction of pairs with top charge assigned to +2/3 via a jet charge algorithm using the charge of the tracks associated to the jet weighted by their momentum projection on the jet axis.

The measured value 0.87 yields a lower bound 0.6 @68% CL

Notice that a measurement above 1.2 would give extremely narrow confidence intervals.



Feldman and Cousins Summary

- Avoids forbidden regions and empty results in a Frequentist way
- Solves flip-flopping, it “unifies” central and upper limit belts
- Makes us more honest (a bit)
- Can lead to 2-tailed limits where you dont want claim discovery
- Not easy to calculate and extend to systematic errors
- Unphysically small CI still present
- Shrinking CI for increasing background
- Upper limits may tighten when including systematic errors

An example

Ratio of top quark branching fractions

$$R = \frac{\mathcal{B}(t \to Wb)}{\mathcal{B}(t \to Wq)}$$

with $q = b, s, d$.

$R > 0.61$ at 95% CL

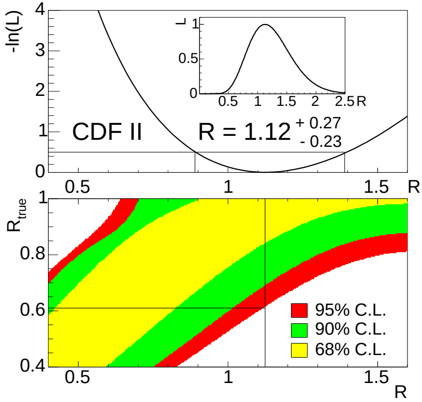


FIG. 2: The upper plot shows the likelihood as a function of R (inset) and its negative logarithm. The intersections of the horizontal line $\ln(L) = -0.5$ with the likelihood define the statistical 1σ errors on R. The lower plot shows 95% (outer), 90% (central), and 68% (inner) CL bands for R_{true} as a function of R. Our measurement of $R = 1.12$ (vertical line) implies $R > 0.61$ at the 95% CL (horizontal line).

Bayes' theorem

Conditional probability: given two events X and Y

$$P(X \mid Y) \equiv \frac{P(X \cap Y)}{P(X)}$$

Example, rolling dice:

$$P(n < 3 \mid n\,\text{even}) = \frac{P(n < 3 \;\cap\; n\,\text{even})}{P(n\,\text{even})} = \frac{1/6}{3/6} = \frac{1}{3}$$

Consider the sample space divided in exclusive events Y_i:

$$Y_i \cap Y_j = \varnothing,\; i \neq j \qquad \text{and} \qquad \sum_i P(Y_i) = 1$$

For any event X, Bayes theorem states:

$$P(Y_k \mid X) = \frac{P(X \mid Y_k)\,P(Y_k)}{\sum_i (X \mid Y_i)\,P(Y_i)}$$

Example: Particles entering a threshold Cerenkov can be e, π or K,

$$P(e) = 1\% \qquad P(\pi) = 70\% \qquad P(K) = 29\%$$

The probabilities that the detector fires (*efficiencies*) are

$$P(C \mid e) = 99\% \qquad P(C \mid \pi) = 2\% \qquad P(C \mid K) = 1\%$$

If a particle fired the detector, what's the probability that it's an e?

$$P(e \mid C) = \frac{P(C \mid e)P(e)}{P(C|e)P(e) + P(C|\pi)P(\pi) + P(C|K)P(K)}$$

$$= \frac{0.99 \times 0.01}{0.99 \times 0.01 \; + \; 0.02 \times 0.70 \; + \; 0.01 \times 0.29} = 37\%$$

Notice that is is a rather selective detector,

yet 63% of signals will be background (π and K).

- To invert probabilities, $P(A\,|\,B) \to P(B\,|\,A)$, need $P(B)$

  $P(C\,|\,e) \to P(e\,|\,C)$, need $P(e)$

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

  $P(C\,|\,e) \neq P(e\,|\,C)$

Or, with a real life example:

A = female or male	$P(\text{pregnant}\,|\,\text{female}) \approx 0.5\%$
B = pregnant or non-pregnant	$P(\text{female}\,|\,\text{pregnant}) \ggg 1\%$

Bayes' Theorem: Continuous version

Instead of discrete probabilities $P(Y)$, we have density functions $f(y)$

Conditional probability:

$$P(X \mid Y) \equiv \frac{P(X \cap Y)}{P(X)} \xrightarrow[\text{case}]{\text{Continuous}} f(x \mid y) \equiv \frac{f(x, y)}{f(x)}$$

Bayes Theorem:

$$P(Y_k \mid X) = \frac{P(X \mid Y_k)\, P(Y_k)}{\sum_i P(X \mid Y_i)\, P(Y_i)} \xrightarrow[\text{case}]{\text{Continuous}} f\,(y \mid x) = \frac{f(x \mid y)\, f(y)}{\int f(x \mid y)\, f(y)\, \mathrm{d}y}$$

Example: The 200 GeV CERN muon beam had an approximately gaussian energy distribution with $\mu_b = 200\,\text{GeV}$ and $\sigma_b = 5\,\text{GeV}$.

$$f(E_b) = \frac{1}{\sqrt{2\pi}\sigma_b}\,\exp\left[-\frac{1}{2}\left(\frac{E_b-\mu_b}{\sigma_b}\right)^2\right]$$

The EMC spectrometer measured the energy of *each* incoming muon with a gaussian uncertainty of 0.5% ($\sigma_b = 1$ GeV),

$$f(E_m \,|\, E_b) = \frac{1}{\sqrt{2\pi}}\,\exp\left[-\frac{(E_m-E_b)^2}{2}\right]$$

Question: For a given event the measured energy was $E_m = 208\,\text{GeV}$. What can we say of the true energy E_b *after* the measurement?

$$f(E_b \,|\, E_m) = \frac{f(E_m \,|\, E_b)\, f(E_b)}{\int f(E_m \,|\, E_b)\, f(E_b)\; \mathrm{d}E_b}$$

Answer: $f(E_b \,|\, E_m) \,\sim\, \mathrm{N}\,(207.5, 0.9)$.

Bayesian use of Bayes' Theorem

Parameter μ of an $f(x \mid \mu)$ is regarded as a random variable itself.

Apply Bayes:

$$f(\mu \mid x) = \frac{f(x \mid \mu)\, f(\mu)}{\int f(x \mid \mu)\, f(\mu)\, \mathrm{d}\mu}$$

to calculate how our knowledge of μ improves after measuring x

$$f(\mu) \xrightarrow{\text{Measurement}} f(\mu \mid x)$$

$f(\mu)$: "degree of belief" on the physical magnitud *before* experiment

Write it $\pi(\mu)$, and call it *prior*, to emphasize this interpretation

$f(\mu \mid x)$: *posterior*, describes knowledge after the experiment is done

Sometimes written as $p(\mu)$ to emphasize interpretation

$$p(\mu) = f(\mu \mid x) = \frac{f(x \mid \mu)\, \pi(\mu)}{\int f(x \mid \mu)\, \pi(\mu)\, \mathrm{d}\mu} \propto f(x \mid \mu)\, \pi(\mu) = \mathscr{L}(x \mid \mu)\, \pi(\mu)$$

Choice of prior

Informative (subjective):

Previous measurement is $\mu = a \pm b$: take $\pi(\mu) \sim N(\mu, b)$

Uninformative (objective):

$\pi(\mu) = \text{const}$

However there is arbitrariness in how ignorance is parametrized

- Should we choose $\pi(\mu)$ flat in μ, in $1/\mu$, or in μ^2?
- Use decay constant λ or the $\tau = 1/\lambda$?
- Use m_ν or m_ν^2, the actual observable?

Statisticians investigate theoretically motivated uninformative priors

(e.g., scale independence in Poisson if choose $1/\mu$)

The preferred choice in HEP is $\pi(\mu) = \text{const}$

$$f(\mu \mid x) \propto \mathscr{L}(x \mid \mu)\,\pi(\mu)$$

The posterior becomes the likelihood, suitable normalized.

$\mathscr{L}(x \mid \mu)$ is promoted to a probability density on x and on μ.

Note:

- how the symmetric notation $\mathscr{L}(x \mid \mu) \leftrightarrow \mathscr{L}(\mu \mid x)$ comes handy,
- the parallelism with conditional probability

What's the attitude of physicists?

Physicists want the data to "speak for themselves",
and choosing one's favorite prior is not precisely in this direction.

But even in frequentist procedures there is arbitrariness.
What estimator to choose? How to construct your confidence belt?
There are different frequentist results for the same data...

A growing attitude towards Bayesian approaches is:
Why not?, if one can show that it provides adequate coverage...

This is the "pragmatic" approach. After all, Bayesian methods:

1. easily account for boundaries: set $\pi(\mu) = 0$ for μ unphysical
2. are handy for treating uncertainty in nuisance parameters.

Poisson upper limit

We observe n events from a Poisson distribution with $\mu = s\,\varepsilon + b$

$$\mathscr{L}(n\,|\,s) = e^{-(s\varepsilon+b)}(s\varepsilon + b)^n/n!$$

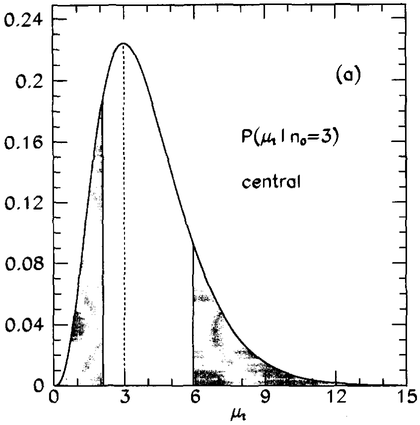
The posterior results $$p\,(s\,|\,\varepsilon, b, n) = \frac{1}{\mathcal{N}}\,e^{-\varepsilon s}(\varepsilon s + b)^n\,\pi(s)$$

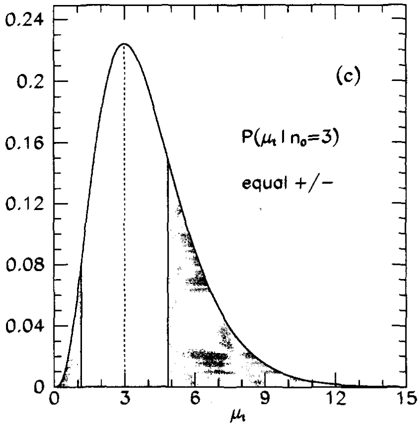
With normalization $$\mathcal{N} = \int_0^{\infty} e^{-\varepsilon s}(\varepsilon s + b)^n\,\pi(s)\ \mathrm{d}s$$

Note that for $n = 0$, the posterior becomes independent of ε and b, and for uniform prior ($\alpha = 1$) it is simply the exponential.

For uniform prior, $\varepsilon = 1$ and $b = 0$, Bayesian upper limits are identical to those obtained with Neyman's frequentist construction.

For Confidence Intervals there is the usual freedom to decide how to divide your $(1-\alpha)\%$ probability between the lower and upper tails





Binomial confidence interval

Estimate efficiency $\epsilon = n/N$, from N trials and n successes

$$p\,(\epsilon\,|\,n, N) = \frac{Binom\,(n\,|\,\epsilon, N)\;\pi(\epsilon)}{\int Binom\,(n\,|\,\epsilon, N)\;\pi(\epsilon)\;\mathrm{d}\epsilon}$$

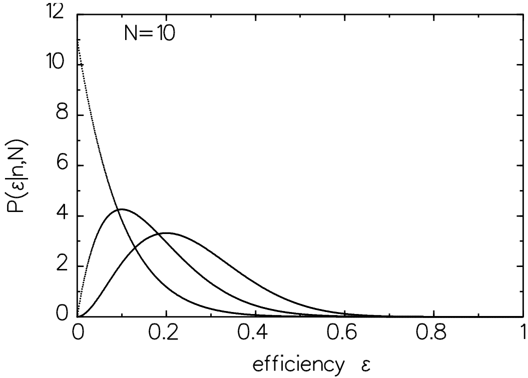
For uniform prior $\pi(\epsilon) = 1$ the integral in the denominator is

$$\int_0^1 \frac{N!}{(N-n)!\,n!}\,\epsilon^n\,(1-\epsilon)^{N-n}\;\mathrm{d}\epsilon = \frac{1}{N+1}$$

yielding the posterior

$$p\,(\epsilon\,|\,n, N) = (N+1)\,Binom\,(n\,|\,\epsilon, N)$$

The posterior distribution for $N = 10$, and $n = 0, 1, 2$



Obtain the CI $[\varepsilon_{\rm d}, \varepsilon_{\rm u}]$ at $100\,\alpha$ % CL via $\displaystyle\int_{\varepsilon_{\rm d}}^{\varepsilon_{\rm u}} f(\varepsilon \,|\, n, N)\, {\rm d}\varepsilon = \alpha$

Other methods

Bayesians or frequentists claim some self-consistent justification for their approach, Other methods are more ad hoc. Hence, they do not usually achieve either coverage or Bayesian credibility.

The two method most used are those implemented by MIGRAD/HESSE and by MINOS in the MINUIT package.

It is interesting to see how the statistics requirements of the HEP community evolved since the early 90s, as ahown in this excerpt from the MINUIT writeup:

```
MINOS is designed to calculate the correct errors in
all cases, especially when there are non-linearities
as described above...
```

Log-likelihood intervals

Have n data points x_i with p.d.f. $f(x_i \mid \theta_j)$ depending on k parameters θ_j.

The ML estimators satisfy $\mathscr{L}(\hat{\theta}_j \mid x_i) = \mathscr{L}_{\max}$.

The ratio of likelihoods is a random variable

$$\lambda(\theta_j) \equiv \frac{\mathscr{L}(\theta_j \mid x_i)}{\mathscr{L}(\hat{\theta}_j \mid x_i)}$$

The distribution of $-2 \ln \lambda(\theta)$ tends asymptotically to χ^2_k

$$-2 \ln \lambda(\boldsymbol{\theta}) = Q^2$$

$$\ln \mathscr{L}(\boldsymbol{\theta}) = \ln \mathscr{L}(\hat{\boldsymbol{\theta}}) - \frac{Q^2}{2} \quad \text{with} \quad Q^2 \sim \chi^2_k$$

Any departure of θ_j from $\hat{\theta}_j$ causes Q^2 to increase from 0.

We can calculate this probability

$$P\,(0 \leq Q^2 \leq a) = \int_0^a \chi_k^2(u)\,\mathrm{d}u = \alpha$$

Then, the α% CL interval is the region in $\boldsymbol{\theta}$ space that satisfies

$$\ln \mathscr{L}(\boldsymbol{\theta}) \;\geq\; \ln \mathscr{L}_{\max} - \frac{a}{2}$$

For one parameter the limits of the interval $[\theta_u, \theta_d]$ are the solution of

$$\ln \mathscr{L}(\theta) = \ln \mathscr{L}_{\max} - \frac{a}{2} \quad \text{where} \quad \int_0^a \chi_1^2\,\mathrm{d}u = \alpha$$

$a = 1, 4, 9$ for $\alpha = 68.27, 95.45, 99.73$, that is 1σ, 2σ, 3σ errors

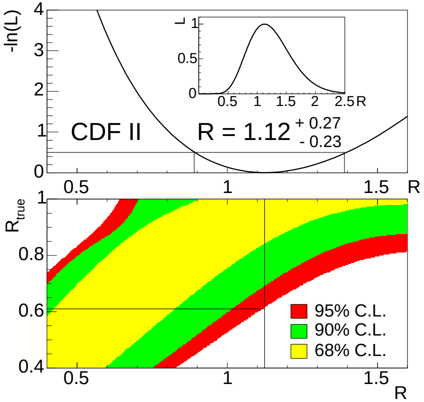


FIG. 2: The upper plot shows the likelihood as a function of R (inset) and its negative logarithm. The intersections of the horizontal line $\ln(L) = -0.5$ with the likelihood define the statistical 1σ errors on R. The lower plot shows 95% (outer), 90% (central), and 68% (inner) CL bands for R_{true} as a function of R. Our measurement of $R = 1.12$ (vertical line) implies $R > 0.61$ at the 95% CL (horizontal line).

One equation, two interpretations: $\boxed{\mu_{\rm d} \leq \mu \leq \mu_{\rm u}}$

$$\text{Frequentist} \begin{cases} \mu_{\rm d} \text{ and } \mu_{\rm u} \text{ known,} & \text{but random} \\ \mu \text{ unknown} & \text{but fixed} \end{cases}$$

$$\text{Bayesian} \begin{cases} \mu_{\rm d} \text{ and } \mu_{\rm u} \text{ known,} & \text{and fixed} \\ \mu \text{ unknown} & \text{and random} \end{cases}$$