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.

1. Bayesian interval estimation
2. Frequentist interval estimation
   a. The Normal Theory Approximation
   b. The Exact Method (Neyman)
   c. Likelihood-based Methods

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 very often the case. Elementary books present only this theory.
3. Exact Frequentist Theory was developed by Jerzy Neyman with the help of Karl 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.

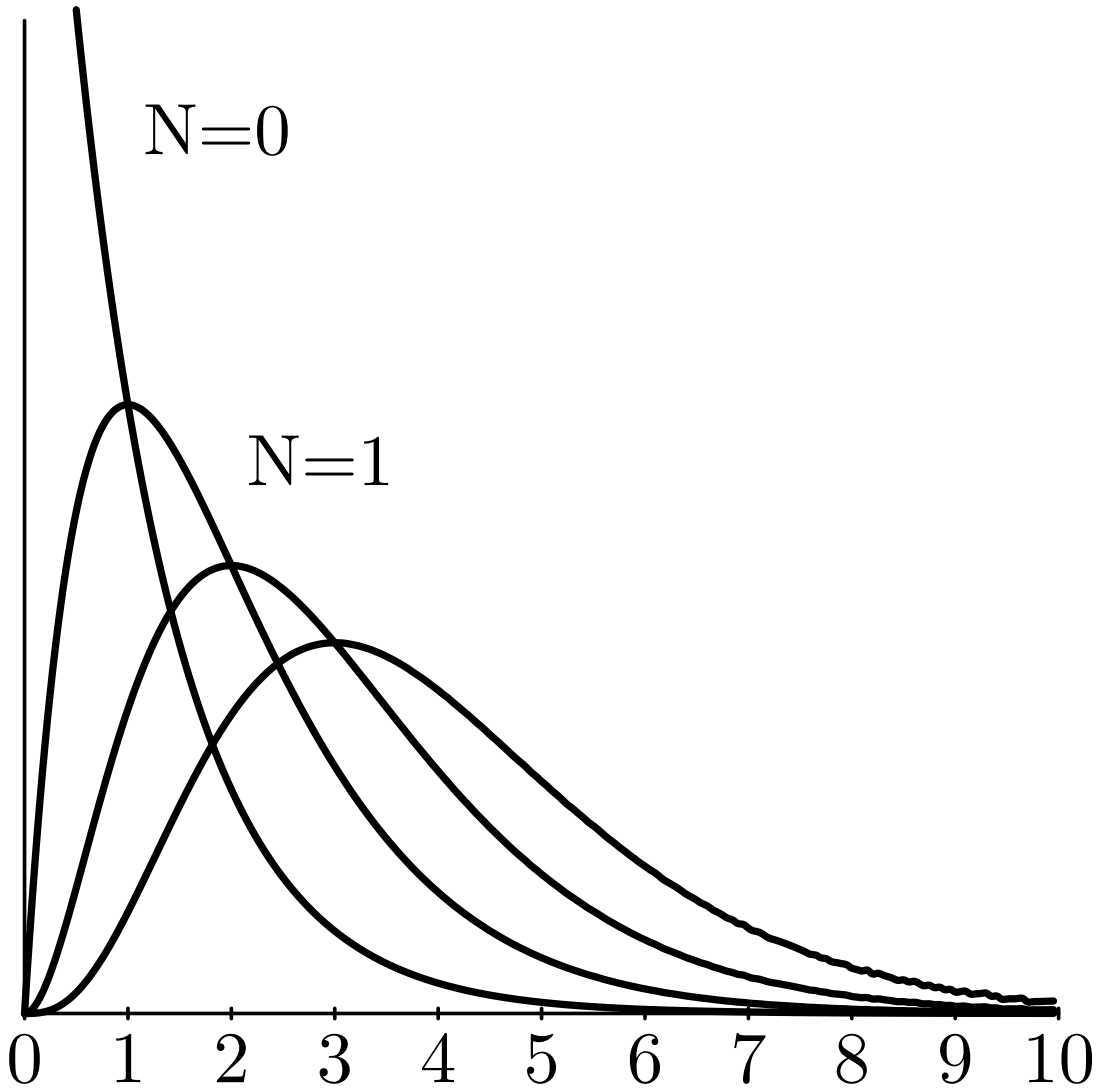
Interval Estimation - Bayesian

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 for Poisson, Uniform Prior

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



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.

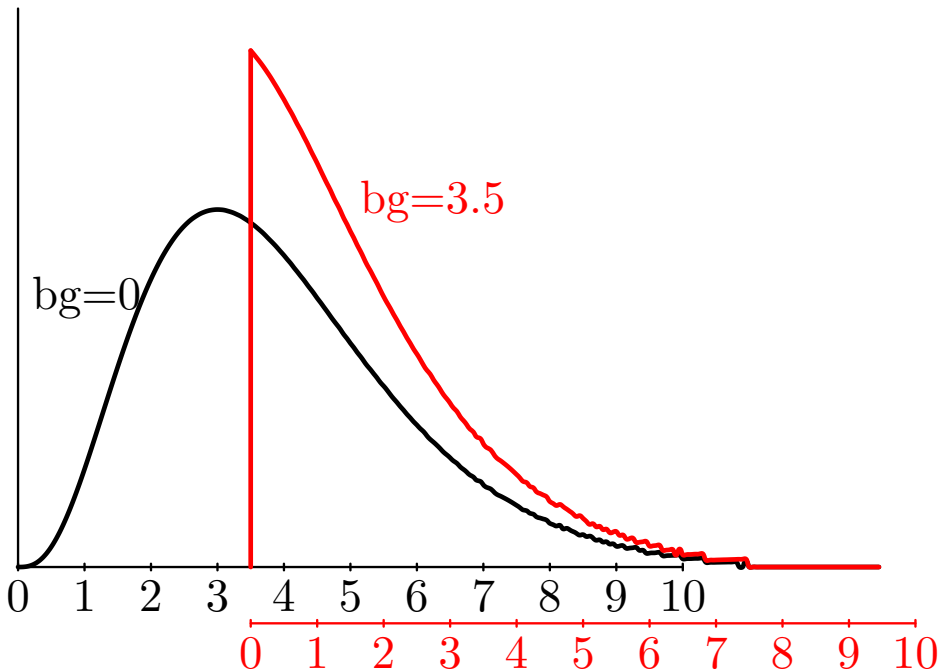
Background in Poisson Processes

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 Known Background

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



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:

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

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

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 doesn't solve the problem seen on the previous slide. The divergences (the zero upper limits) 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.

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. [The Ur-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 equation is said to possess the property of coverage.

Note that the random variables in this equation are (θ_a, θ_b), not θ_{true}.

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 $N(\mu, \sigma^2)$.
When μ and σ^2 are known, we can evaluate:

$$\beta = P(a \leq X \leq b) = \int_a^b N(\mu, \sigma^2) dX$$

When μ is unknown, one can no longer calculate the probability content of the interval $[a, b]$. Instead, one calculates the probability β that X lies in some interval relative to its unknown mean, say $[\mu + c, \mu + d]$. Letting $Y = (X' - \mu)/\sigma$, we have:

$$\begin{aligned} \beta &= P(\mu + c \leq X \leq \mu + d) = \int_{\mu+c}^{\mu+d} N(\mu, \sigma^2) dX' \\ &= \int_{c/\sigma}^{d/\sigma} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} Y^2\right] dY \end{aligned}$$

Now re-arrange the inequalities inside the probability to obtain:

$$\beta = P(X - d \leq \mu \leq X - c)\,.$$

Normal Theory Interval Estimation

The above magic works because:

- The data were Normally distributed, and the Normal pdf is symmetric in X and μ, being a function only of $(X-\mu)^2$.
- It was tacitly assumed that one could always integrate in both variables as far as we want in both directions. That means we encounter no physical boundaries.

According to the theory of Point Estimation, both of the above should be true asymptotically for the usual estimators:

maximum likelihood and **least squares**.

Therefore we already have

an asymptotic theory of interval estimation.

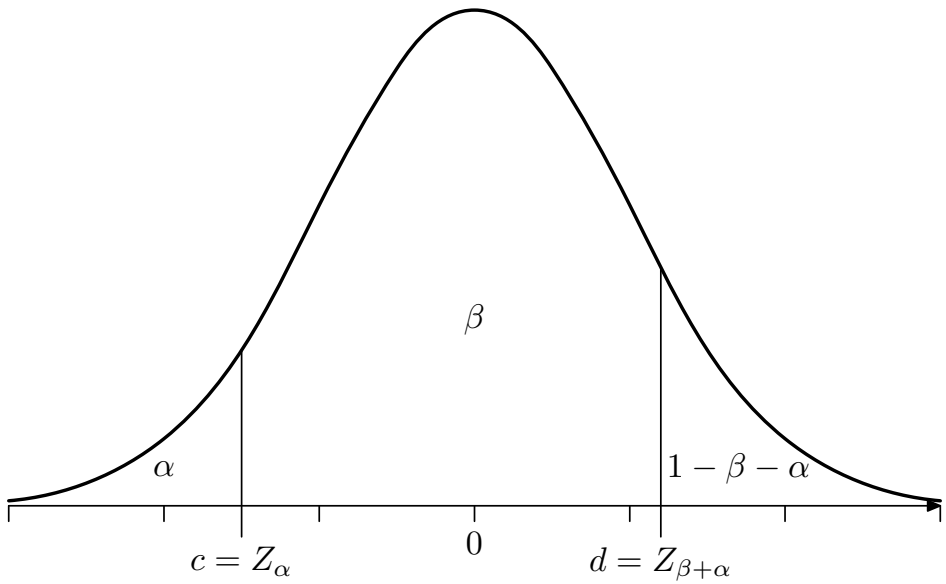
We shall see that the extension to many variables is straightforward.

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

Normal Theory Intervals in Many Variables

In more than one dimension, the Normal pdf becomes:

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

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

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

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

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

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

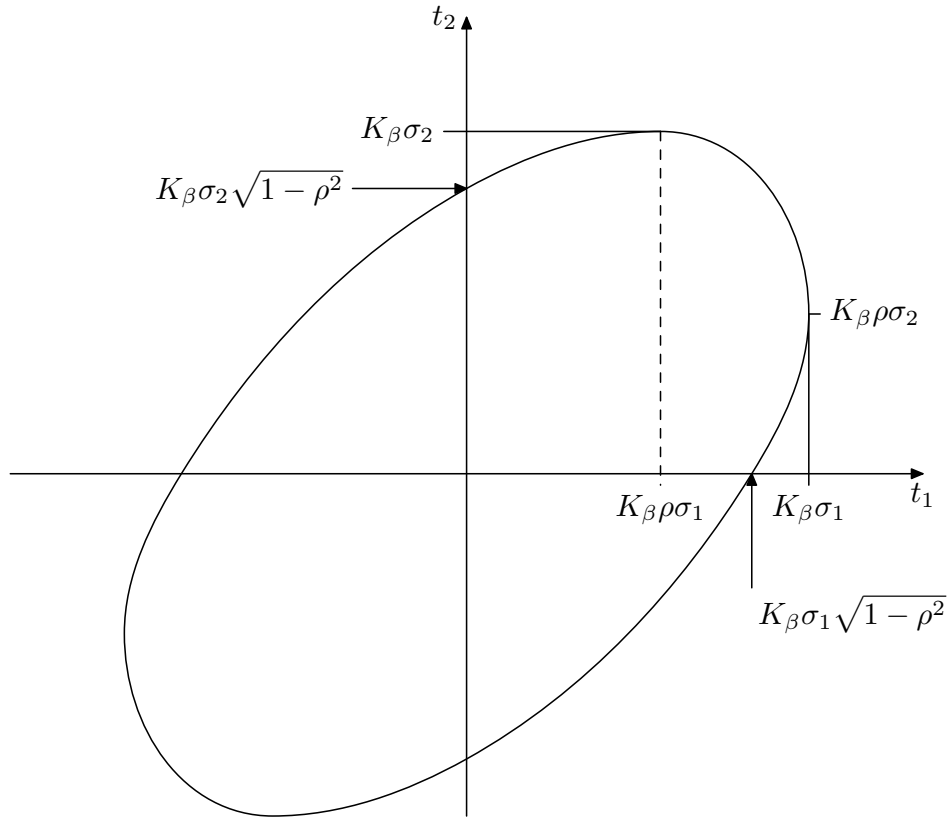
Then the confidence interval becomes a confidence region in $\mathbf{t}$-space with probability content β, defined by $Q(\mathbf{t},\boldsymbol{\theta}) \leq K_\beta^2$.
Q is a hyperellipsoid of constant probability density for the Normal pdf.

Normal Theory Intervals in Two Variables

For two Normally-distributed variables with covariance matrix

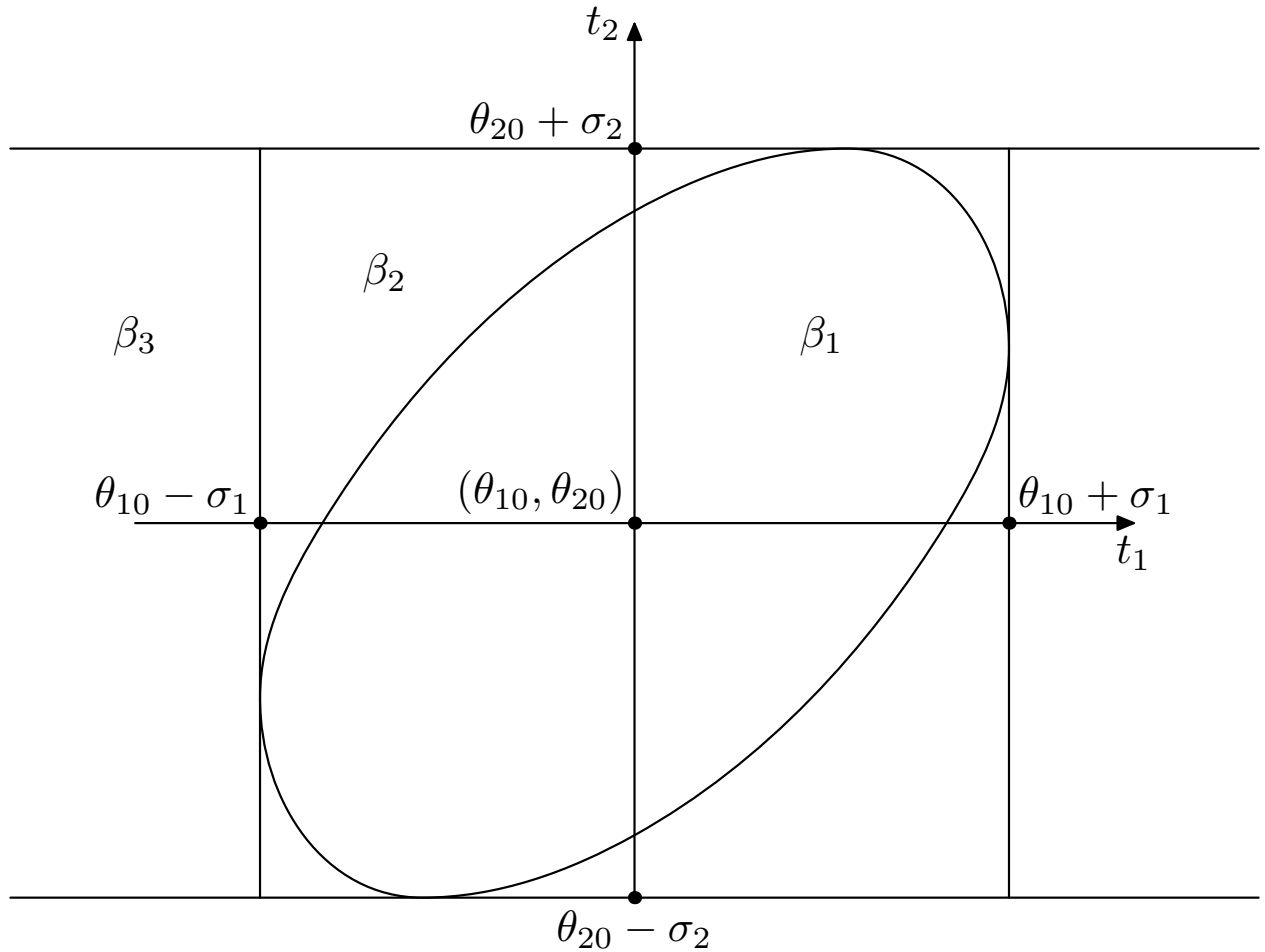
$$\underset{\sim}{V} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

the elliptical confidence region of probability content β will look like this:



Shown here is the case $\rho = 0.5$.
If ρ is negative, the major axis of the ellipse has a slope of -1.
If $\rho = 0$, the axes of the ellipse coincide with the coordinate axes.
If $\rho = 1$, the ellipse degenerates to a diagonal line.

Normal Theory Intervals in Two Variables



Confidence regions for the Normal estimators t_1, t_2, with $K_\beta = 1$, $\rho = 0.5$. The probability content is β_1 for the elliptic regions, β_2 for the circumscribed rectangle, and β_3 for the infinite horizontal band.

Normal Theory Intervals in Two Variables

We give below the probability contents of the three regions for different values of K_β and the correlation ρ, for two Gaussian-distributed variables. For the inner ellipse (region 1) and the infinite band (region 3), the probability content β does not depend on ρ.

	$K_\beta = 1$	$K_\beta = 2$	$K_\beta = 3$
inner ellipse β_1	0.393	0.865	0.989
square β_2 for $\rho = 0.00$	0.466	0.911	0.995
for $\rho = 0.50$	0.498	0.917	0.995
for $\rho = 0.80$	0.561	0.929	0.996
for $\rho = 0.90$	0.596	0.936	0.996
for $\rho = 0.95$	0.622	0.941	0.996
for $\rho = 1.00$	0.683	0.954	0.997
infinite band β_3	0.683	0.954	0.997

Exact Frequentist Intervals for the general case

Fisher got as far as the Normal Theory intervals, but his attempt to solve the general case produced the so-called **fiducial intervals** which were essentially Bayesian (just what he was trying to avoid).

Meanwhile Jerzy Neyman, a young Polish mathematician born in imperial Russia, had finished his studies in the Ukraine, and decided to go to London to work on statistics with Karl Pearson, discoverer of the Chi-square Test.

He was disappointed to find that Pearson did not know modern mathematics, but his son Egon Pearson did.
Neyman and Egon Pearson collaborated very fruitfully, producing two major contributions to statistics:

- The Neyman construction of confidence intervals presented here, and
- The Neyman-Pearson Test, to be treated later under Tests of Hypotheses.

Exact Frequentist Intervals

The first important step in finding an exact theory was to work in the right space: P(data|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 hypotheses and data live in different spaces.

Bayesian Parenthesis:

page 81 of O'Hagan, "Bayesian Inference" shows how Bayesians plot different kinds of functions on the same axes

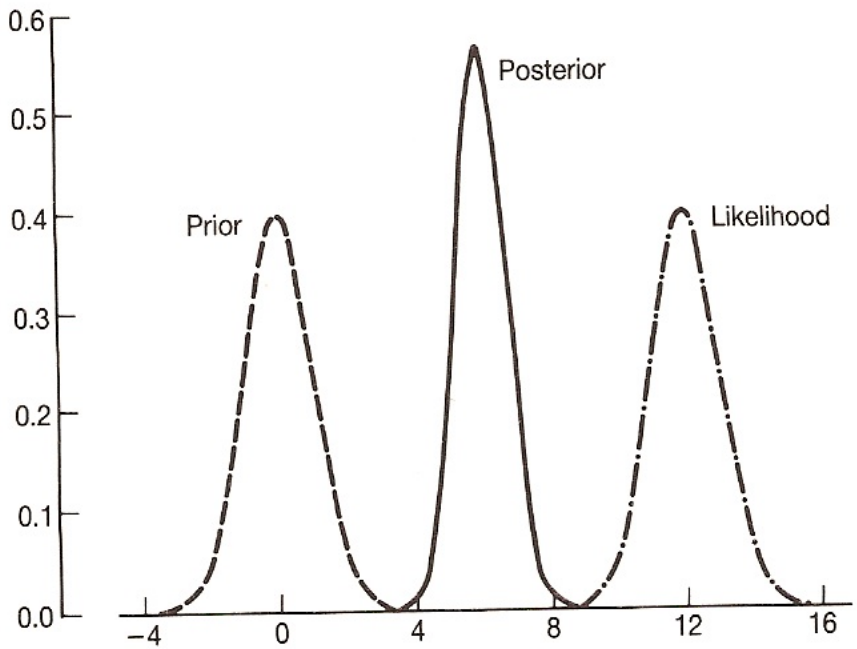
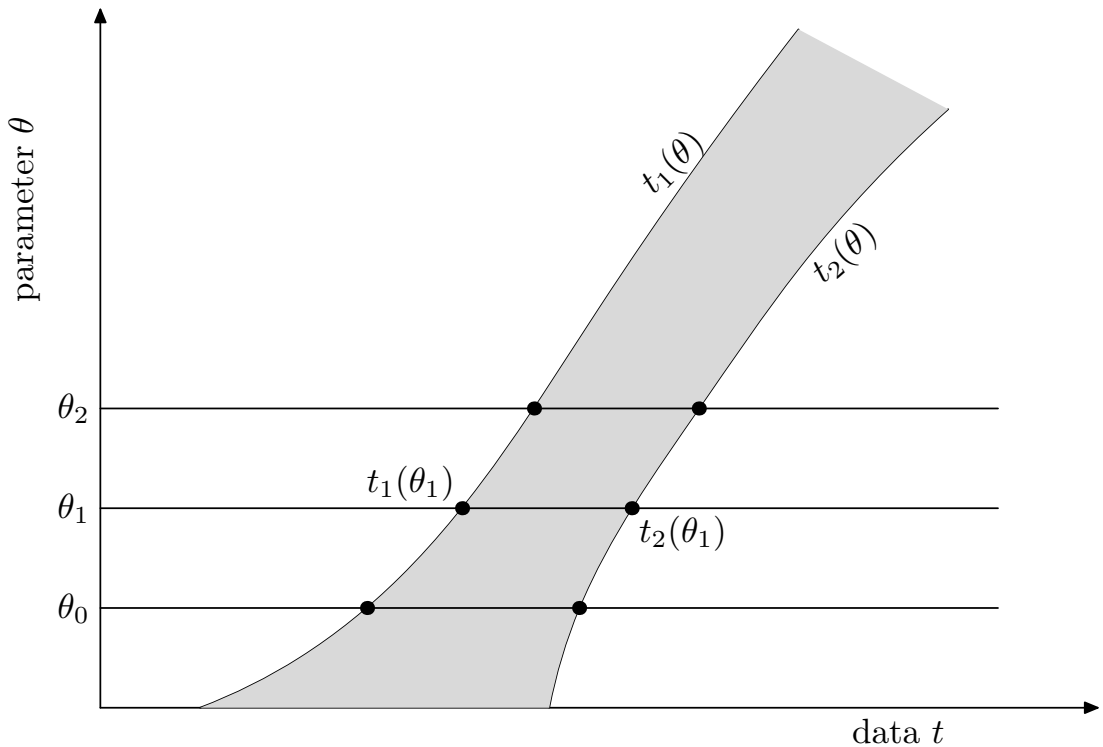


Fig. 3.1 Prior, likelihood and posterior in conflict

The Neyman Construction

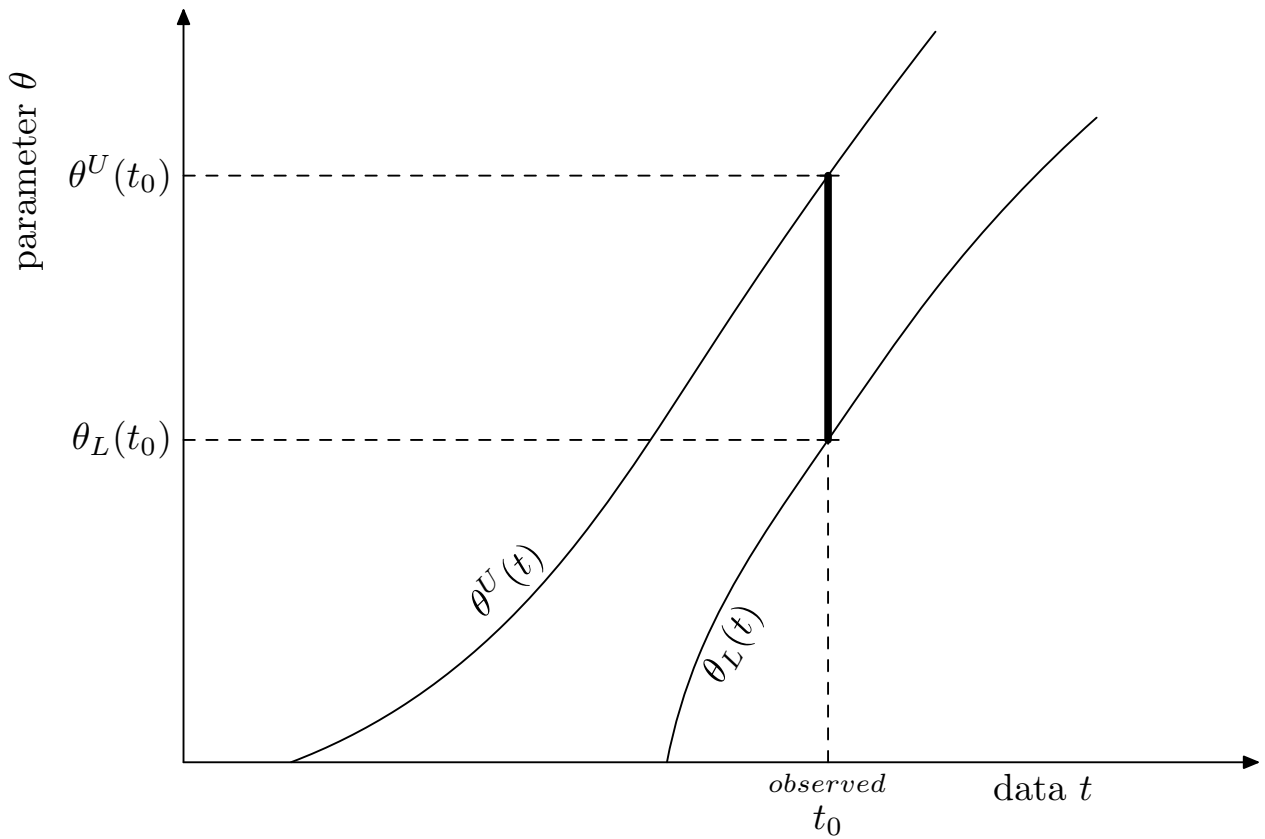
The confidence belt is constructed horizontally in the space of $P(t|\theta)$.



$t_1(\theta)$ 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

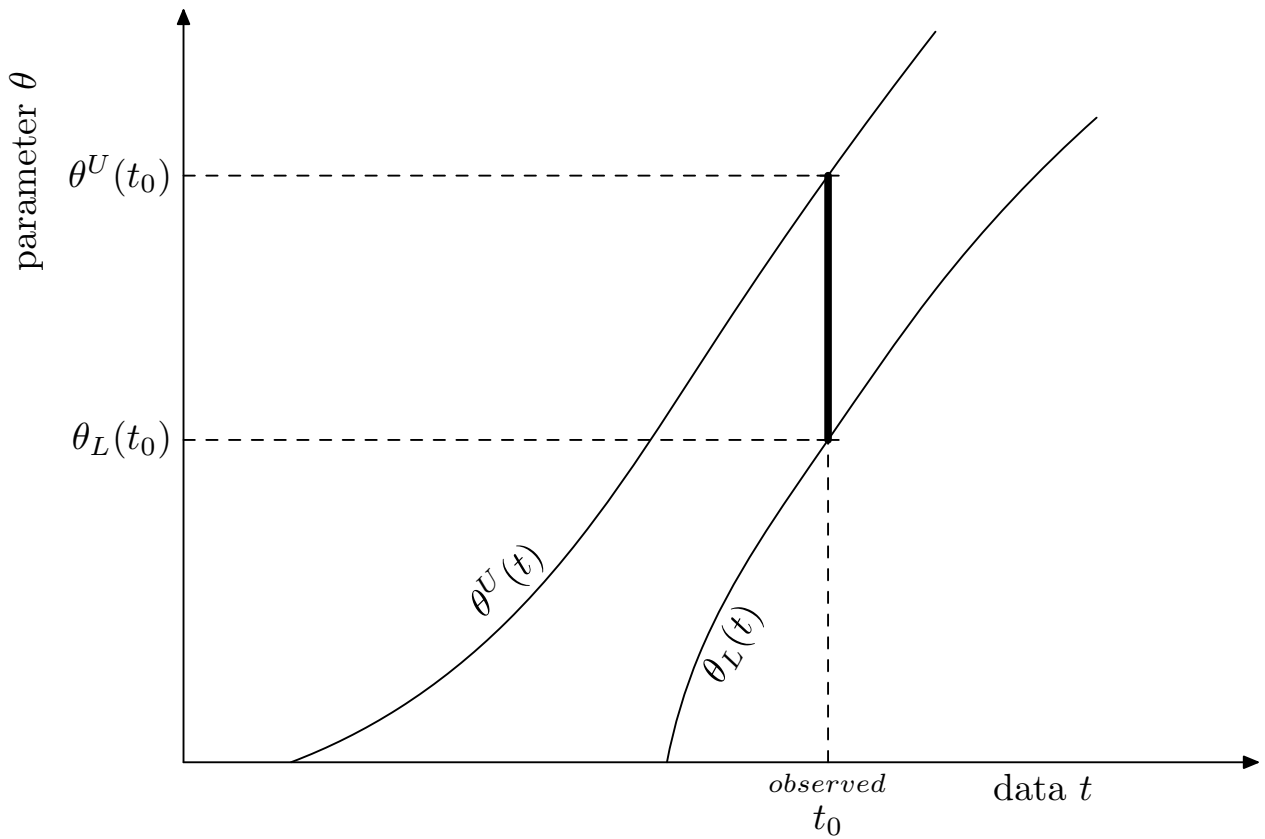
The two curves of $t(\theta)$ are re-labelled as $\theta(t)$, and the confidence limit is read vertically.



For observed data t_0, the confidence interval is $(\theta_L(t_0), \theta^U(t_0))$.

The Neyman Construction 2

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

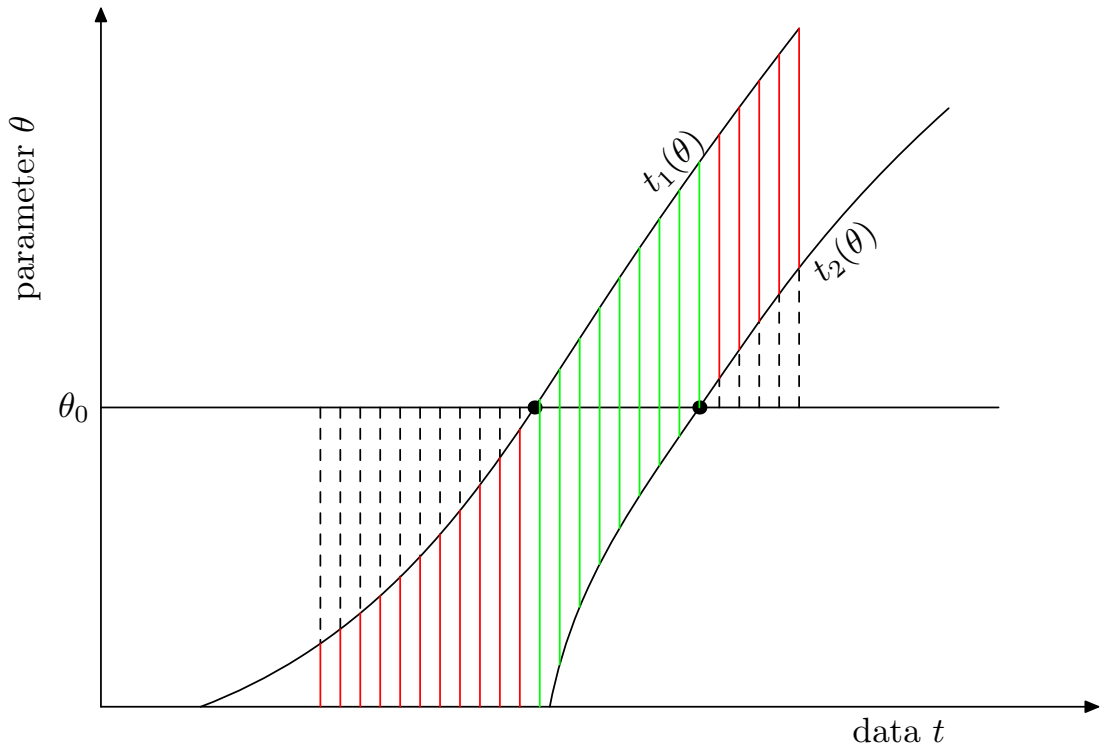


For observed data t_0, the confidence interval is $(\theta_L(t_0), \theta^U(t_0))$.

I now claim that $P(\theta_L < \theta_{\text{true}} < \theta^U) = \beta$

The Neyman Construction 3

Suppose the true value is θ_0. Then, depending on the observed data, we could get the intervals indicated as red and green vertical lines below:



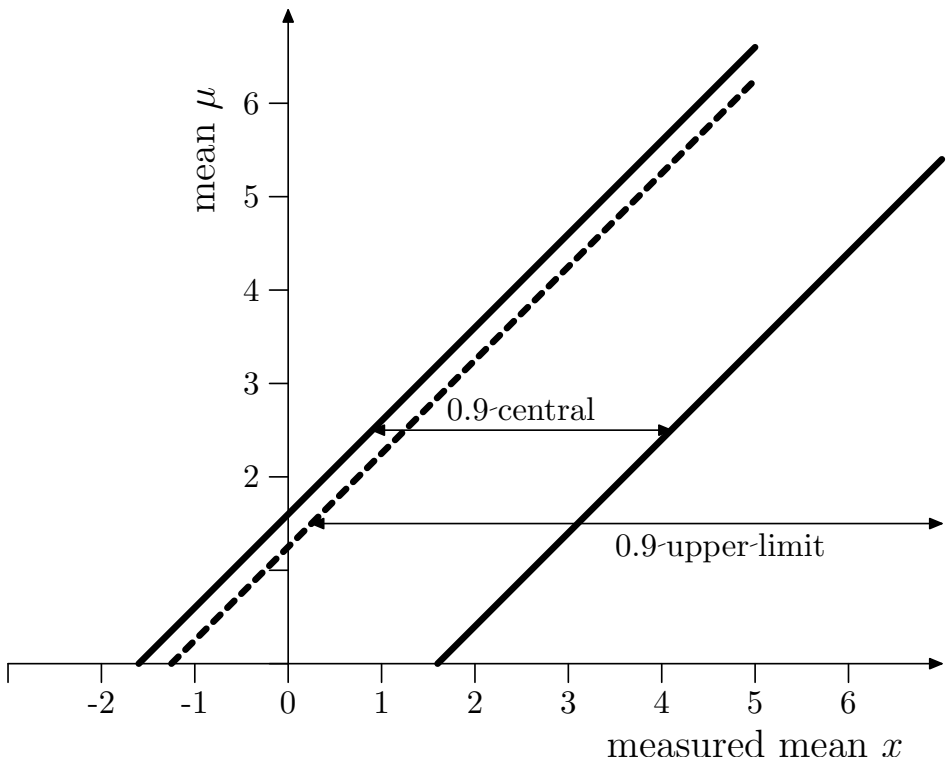
Only the green confidence intervals cover the true value.
The probability of getting a green confidence interval is β.
By construction, for any value θ_0, $P(\theta_L < \theta_0 < \theta^U) = \beta$.

Upper limits and Central Intervals

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

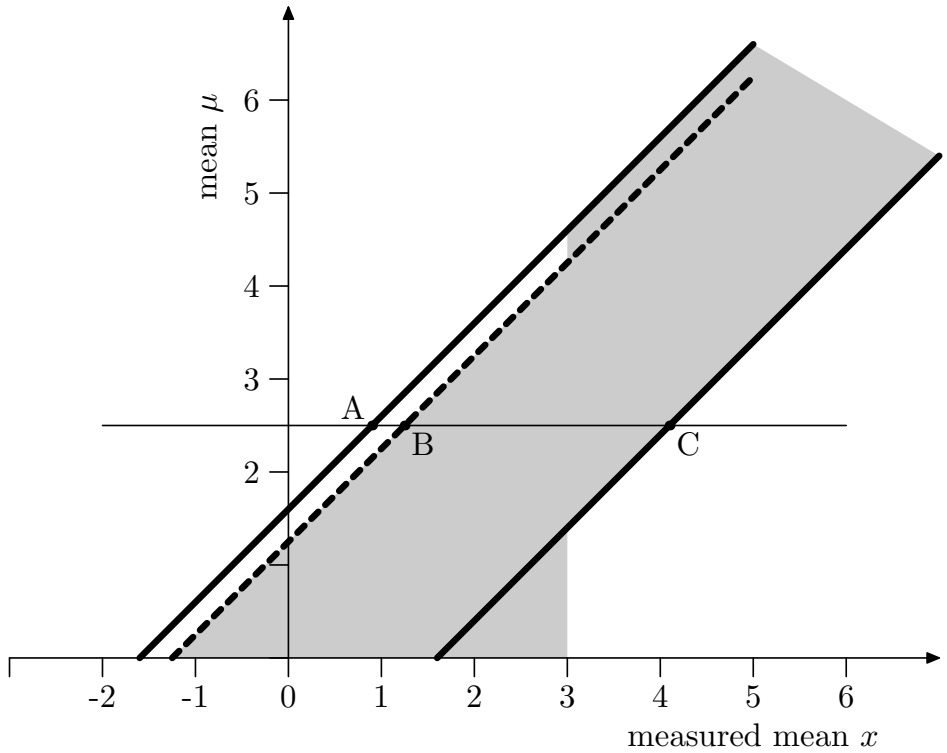
Confidence belts for a Gaussian measurement (μ unknown, $\sigma = 1$).

The solid lines delimit the central 90% confidence belt, the dashed line the 90% upper limit.



Flip-flopping and Empty Intervals

empty intervals
down here $\longrightarrow$



Flip-flopping for a Gaussian measurement. The shaded area represents the effective confidence belt resulting from choosing to report an upper limit only when the measurement is less than 3σ above zero. 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.

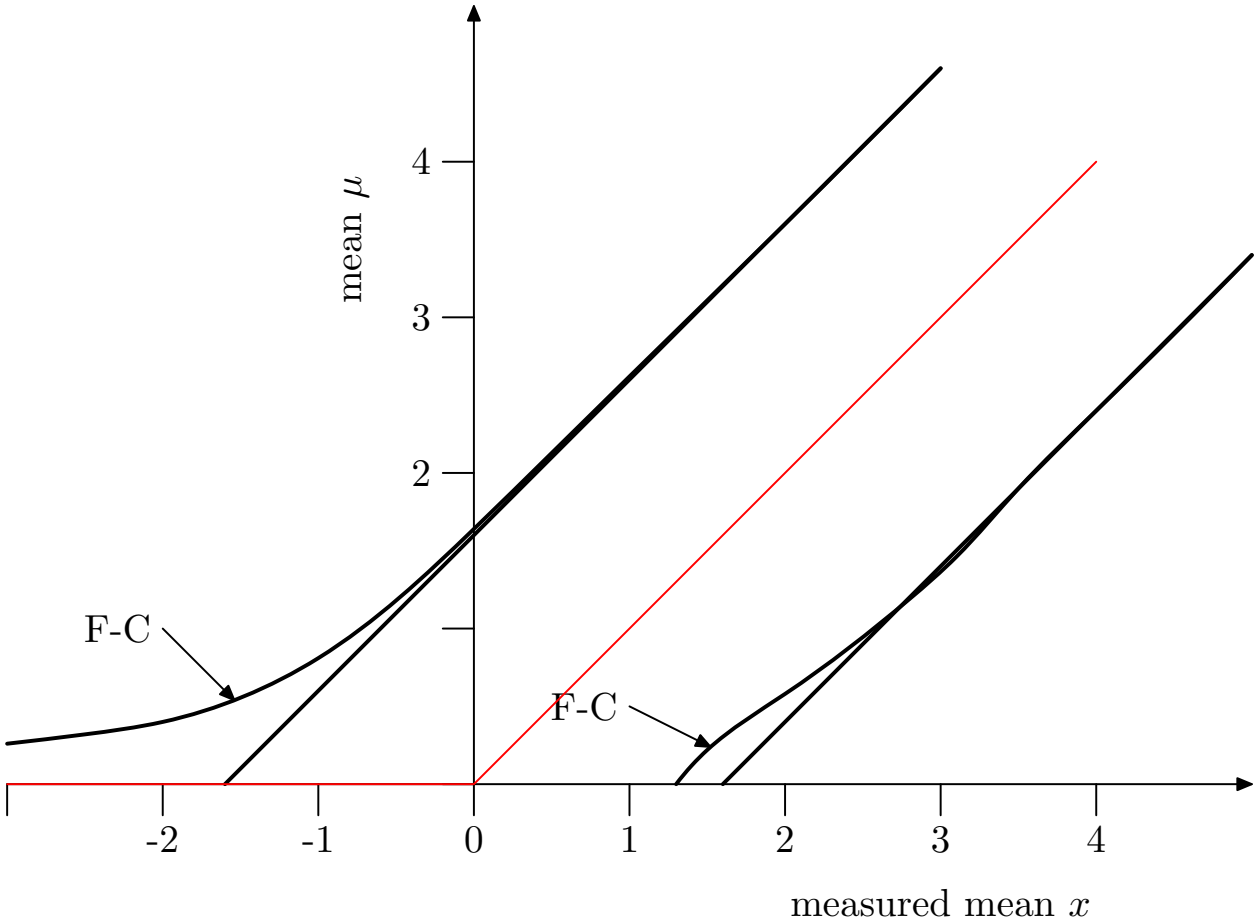
Inspired by an important result in hypothesis testing
which we will see in the next chapter,
Feldman and Cousins proposed the
likelihood ratio ordering principle:
see Feldman and Cousins, *Unified Approach ...*
Phys. Rev. D 57 (1998) 3873

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

Confidence Intervals for Multidimensional Data

In the 1937 paper in which Neyman described his construction, he used two dimensions for the data (and the third dimension for the parameter θ).

Now the interval in t becomes a 2-d region in data space containing probability β. And the determination of the confidence interval requires finding all the values of θ for which the observed data lies inside this region.

Philosophical Transactions of the Royal Society of London. Series A, Mathematical an Sciences, Vol. 236, No. 767. (Aug. 30, 1937), pp. 333-380.

X—Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability

By J. Neyman

Reader in Statistics, University College, London

(Communicated by H. Jeffreys, F.R.S.—Received 20 November, 1936—Read 17 June, 1937)

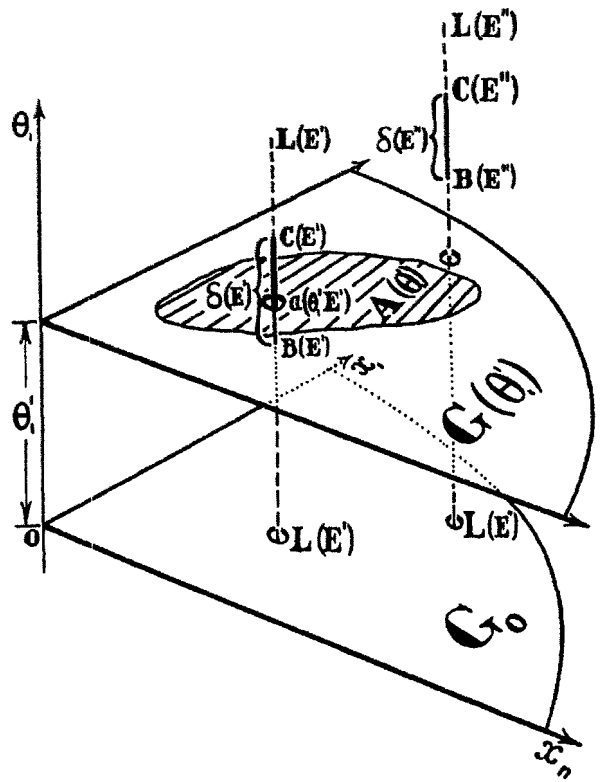


Fig. 1—The general space G.

Confidence Intervals for Discrete Data

In the Neyman construction, we have so far assumed that it would always be possible to accumulate exactly probability β within the confidence band, but this happens only when the data t are continuous. When the data are discrete, which will be the case for Poisson- and binomial-distributed data for example, we must replace the continuous t with discrete t_i, the integral becomes a summation, and unfortunately the equals sign must also go:

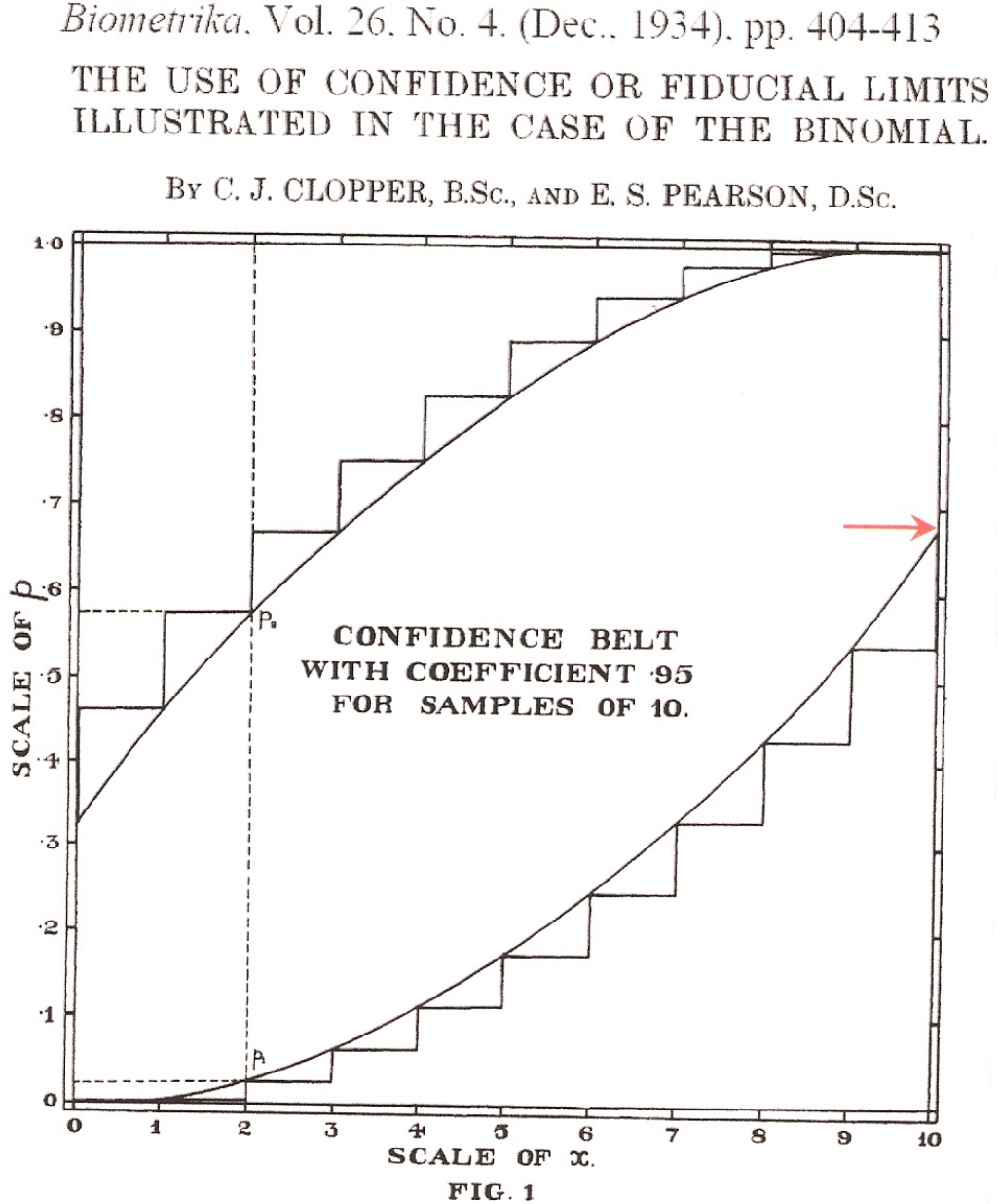
continuous data		discrete data
$\int_{t_1}^{t_2} f(t\|\theta)dt = \beta$	$\rightarrow$	$\sum_{i=L}^{U} P(t_i\|\theta) \geq \beta\,.$

Clopper and Pearson [Biometrika **34** (1934) 404]
showed how to do this for the binomial distribution.

Clopper-Pearson Intervals for Binomial Data

This diagram from the 1934 paper of Clopper and (Egon) Pearson shows the construction of 95% confidence limits for the binomial parameter with $N = 10$.

The data can take on only 11 discrete values, so the confidence belt takes the form of steps. It is important to join the inner edges of the steps in a smooth curve (as they do here) because these are the points that determine the confidence interval.


Biometrika. Vol. 26. No. 4. (Dec., 1934). pp. 404-413

THE USE OF CONFIDENCE OR FIDUCIAL LIMITS ILLUSTRATED IN THE CASE OF THE BINOMIAL.

By C. J. CLOPPER, B.Sc., and E. S. PEARSON, D.Sc.

FIG. 1

Coverage of Neyman Upper Limits for Poisson Data

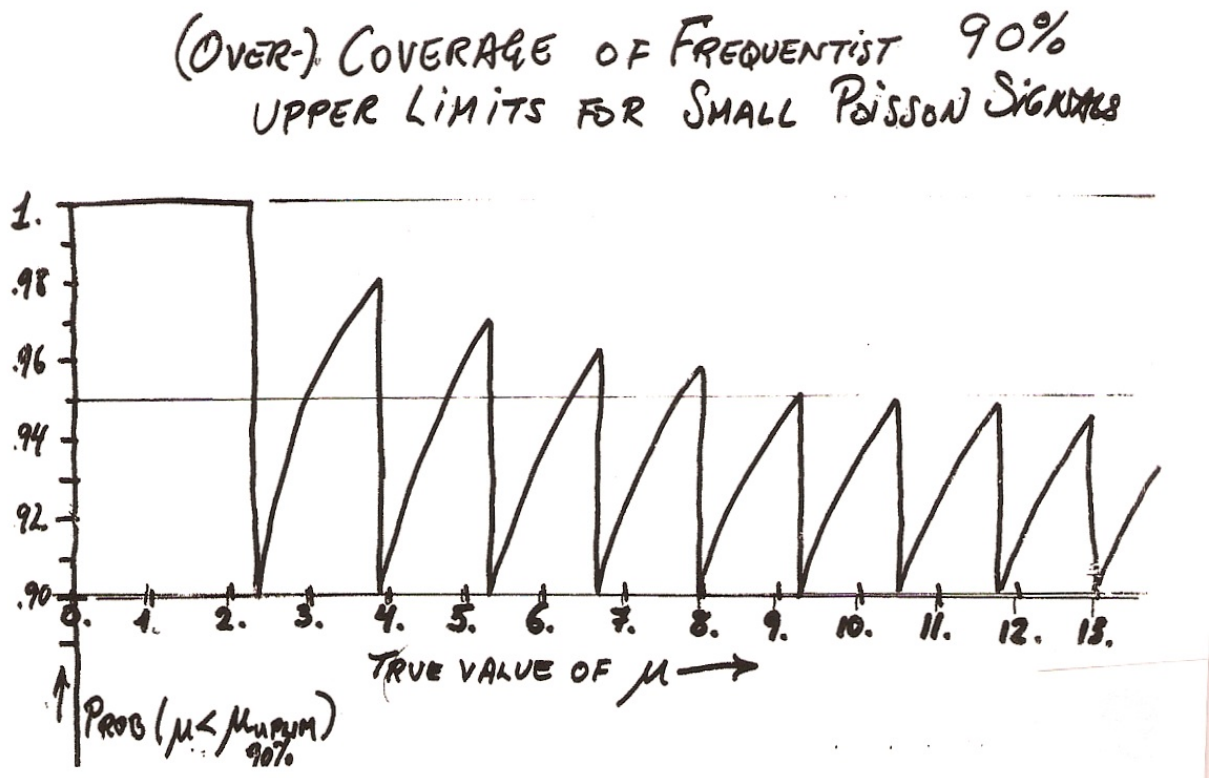
The confidence band for Poisson upper limits will be in the form of steps, as with the Clopper-Pearson confidence band shown above.

Naive Frequentist 90% Upper Limits for Poisson with Background

events observed =	0	1	2	3
upper limit =	2.30	3.89	5.32	6.68

Dinosaur Plot

If the true value of μ is less than 2.30, the coverage will be 100 % because the upper limit cannot be less than 2.30.



Feldman-Cousins Intervals for Poisson Data with bgd

The Feldman-Cousins unified approach has had its greatest success when applied to Poisson data, where all previously used methods had some undesirable properties.

For all details of this method, the original publication
Feldman and Cousins, *Unified Approach ...* Phys. Rev. D 57 (1998) 3873
is recommended since it explains clearly how to use it to solve all the most common problems.

There is also an excellent set of slides available from Gary Feldman's website called Journeys of an Accidental Statistician.
The following slide is taken from Journeys.
It explains how the unified approach is applied to the problem of Poisson with background to determine the edges of the confidence belt for $\mu = 0.5$ and $b = 3.0$.

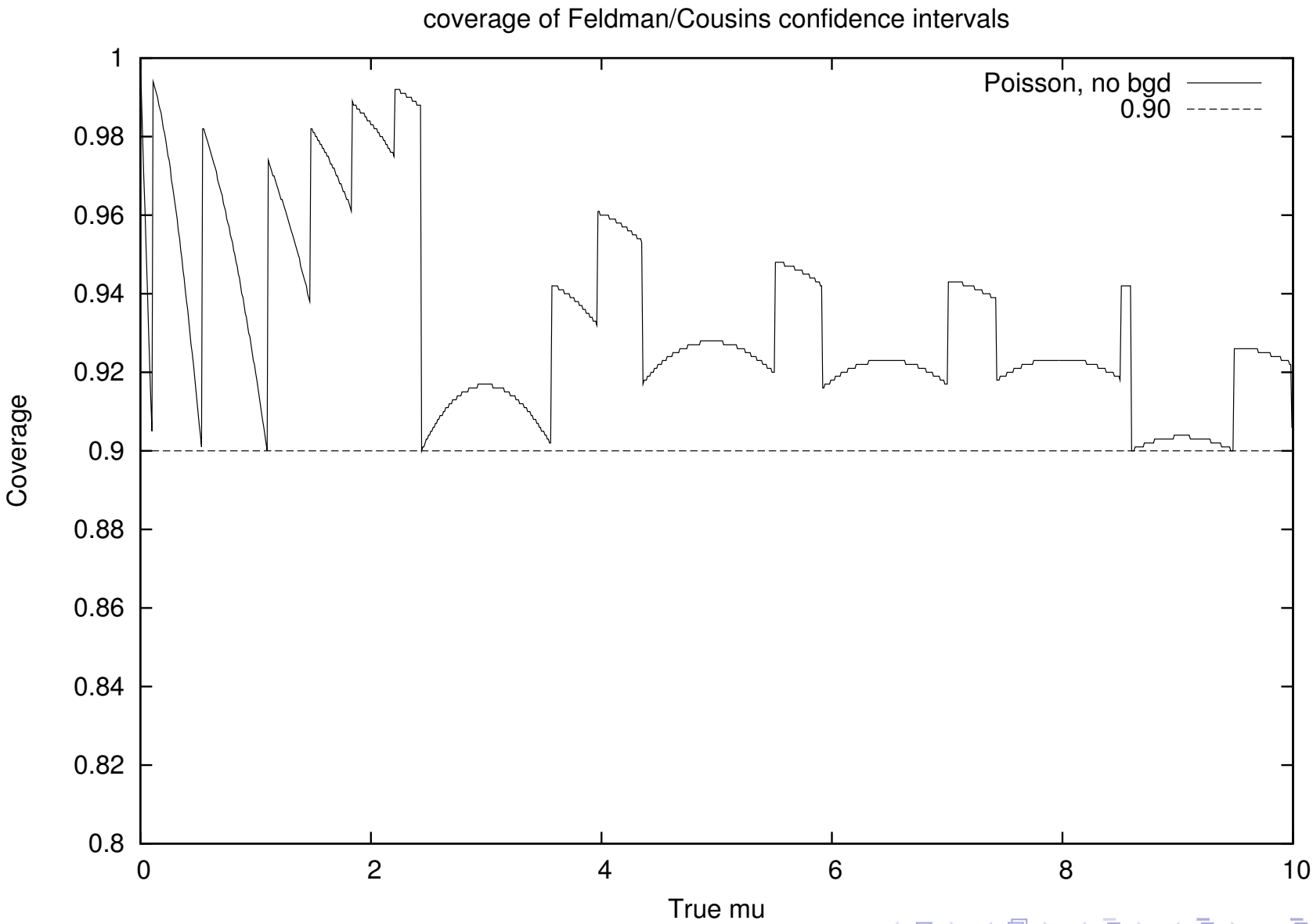
Therefore, we propose a new ordering principle based on the ratio of a given μ to the most likely μ, $\hat{\mu}$:

$$R = \frac{P(x \mid \mu)}{P(x \mid \hat{\mu})}$$

Example for μ = 0.5 and b = 3:

x	$P(x\|\mu)$	$\hat{\mu}$	$P(x\|\hat{\mu})$	R	rank	U.L.	C.L.
0	0.030	0.0	0.050	0.607	6 •		
1	0.106	0.0	0.149	0.708	5 •	•	•
2	0.185	0.0	0.224	0.826	3 •	•	•
3	0.216	0.0	0.224	0.963	2 •	•	•
4	0.189	1.0	0.195	0.966	1 •	•	•
5	0.132	2.0	0.175	0.753	4 •	•	•
6	0.077	3.0	0.161	0.480	7 •	•	•
7	0.039	4.0	0.149	0.259		•	•
8	0.017	5.0	0.140	0.121		•	

Feldman-Cousins coverage for Poisson, no background



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

The Unified Approach of Feldman and Cousins

The Unified Approach solves the major problems that we knew we had (empty intervals) and we didn't know we had (flip-flopping) in the framework of the classical Neyman construction by using an equally classical result from hypothesis testing (the Neyman-Pearson lemma, coming soon in this course).

Some other aspects also covered in the paper:

- Application to more complicated problems with two parameters.
- How to measure the sensitivity of the experiment.

And an important aspect not covered in the paper, but since resolved by Feldman:

- The inclusion of nuisance parameters, such as unknown background.

Unified Approach with 2 parameters

Feldman and Cousins show in their paper how to apply their method to the problem of neutrino oscillations with 2 parameters: $\sin^2(2\theta)$ and Δm^2. They also compare their method with other commonly used methods, and show the Unified Approach gives both tighter limits and correct coverage.

However, it is clear that as the complications increase, this method becomes very hard to use. I haven't seen it applied to three parameters.

Compare with the Bayesian method: With two or more parameters, there is no guarantee that Bayesian estimates are even consistent. Multidimensional priors pose a great problem, and tend to dominate the data, contrary to the usual assumption that the prior can be made "uninformative".

The Confidence Interval as a Measure of Sensitivity

We often use the size of the reported "errors" as a measure how good (sensitive) the experiment is. An experiment that reports smaller errors is supposed to be a better experiment.

However, in situations near a physical limit, it is possible to obtain a smaller interval estimate simply by "good luck": for example, observing fewer events than were expected from backgreound alone.
This can happen also in the Unified Approach.
Therefore, Feldman and Cousins propose an additional sensitivity measure:

The Feldman-Cousins sensitivity of the measurement of a small signal is the average upper limit that would be obtained by an ensemble of experiments with the expected background and no true signal.

It can be calculated beforehand since it is independent of the data observed.

Nuisance Parameters in the Unified Approach

It is known that in the Neyman construction, adding more parameters, even nuisance parameters in which we are not interested, is clumsy and always leads to overcoverage, since we require coverage for every possible value of the nuisance parameters.

At the time of their 1998 paper, Feldman had already found an elegant approximate solution to the problem of nuisance parameters, but it was not published because of one important detail that was not understood: The results did not tend to the expected limit as the uncertainty in the nuisance parameter approached zero.

This turns out to be a special feature of the Poisson problem which is now understood, so the method seems to be correct.
It is not yet published, but is available on the Web in a presentation Journeys of an Accidental Statistician [Google:feldman journeys]

Trick for Including Errors on Backgrounds

If one provides coverage for the "worst case" of the nuisance parameter, then perhaps one will have provided coverage for all possible values of the nuisance parameter.

Let μ = unknown true value of the signal parameter

β = unknown true value of the background parameter

x = measurement of $\mu + \beta$

b = measurement of β in the ancillary experiment

The rank R becomes

$$R(\mu, x, b) = \frac{P(x \mid \mu + \hat{\hat{\beta}})\, P(b \mid \hat{\hat{\beta}})}{P(x \mid \hat{\mu} + \hat{\beta})\, P(b \mid \hat{\beta})},$$

where

$\hat{\mu}$ and $\hat{\beta}$ maximize the denominator (as usual), and

$\hat{\hat{\beta}}$ maximizes the numerator.

Since R is only a function of μ and the data, one proceeds to construct the confidence belt as before.

Examples and Test of the Trick

Let x be a Poisson measurement of $\mu + \beta$ and

b be a Poisson measurement of β/r in an ancillary experiment (i.e., r = signal region/control region),

r / n, rb	0, 3	3, 3	6, 3	9, 3
0.0	0.00- 1.08	0.00- 4.42	0.15- 8.47	1.88-12.30
0.5	0.00- 1.11	0.00- 4.42	0.00- 8.47	1.75-12.30
1.0	0.00- 1.49	0.00- 4.73	0.00- 8.70	1.32-12.55
3.0	0.00- 1.57	0.00- 4.85	0.00- 9.36	0.00-13.03

A test of coverage for $r = 1.0$, $0 \le \mu \le 20$, $0 \le \beta \le 15$ at 10,000 randomly picked values of μ and β:

Median coverage:	90.25%
Fraction that undercovered:	4.10%
Median coverage for those that undercovered:	89.96%
Worst undercoverage:	89.46%

Comment from Harvard statisticians: "We have never seen a statistical approximation work this well."

Visit to Harvard Statistics Department

Towards the end of this work, I decided to try it out on some professional statisticians whom I know at Harvard.

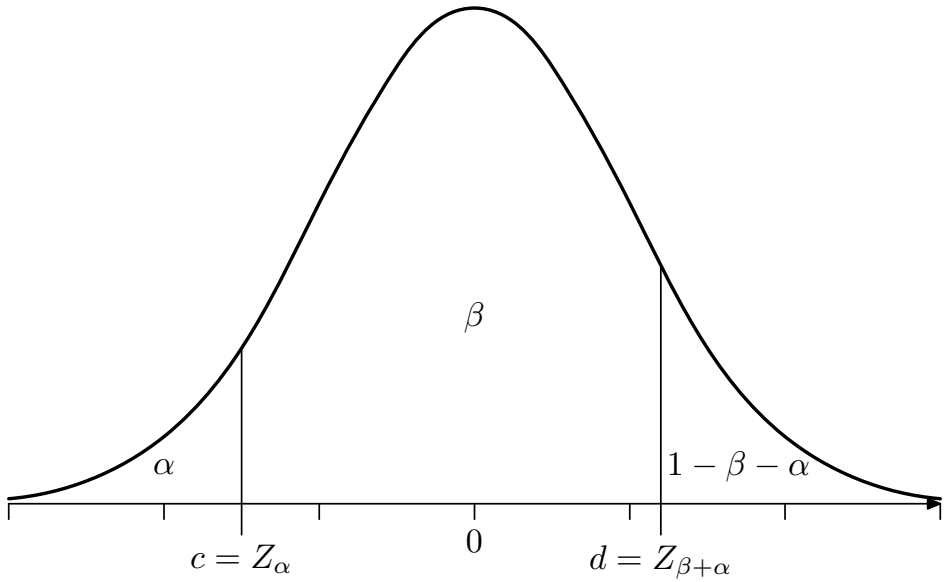
They told me that this was the standard method of constructing a confidence interval!

I asked them if they could point to a single reference of anyone using this method before, and they could not.

They explained that in statistical theory there is a one-to-one correspondence between a hypothesis test and a confidence interval. (The confidence interval is a hypothesis test for each value in the interval.) The Neyman-Pearson Theorem states that the likelihood ratio gives the most powerful hypothesis test. Therefore, it must be the standard method of constructing a confidence interval.

I decided to start reading about hypothesis testing…

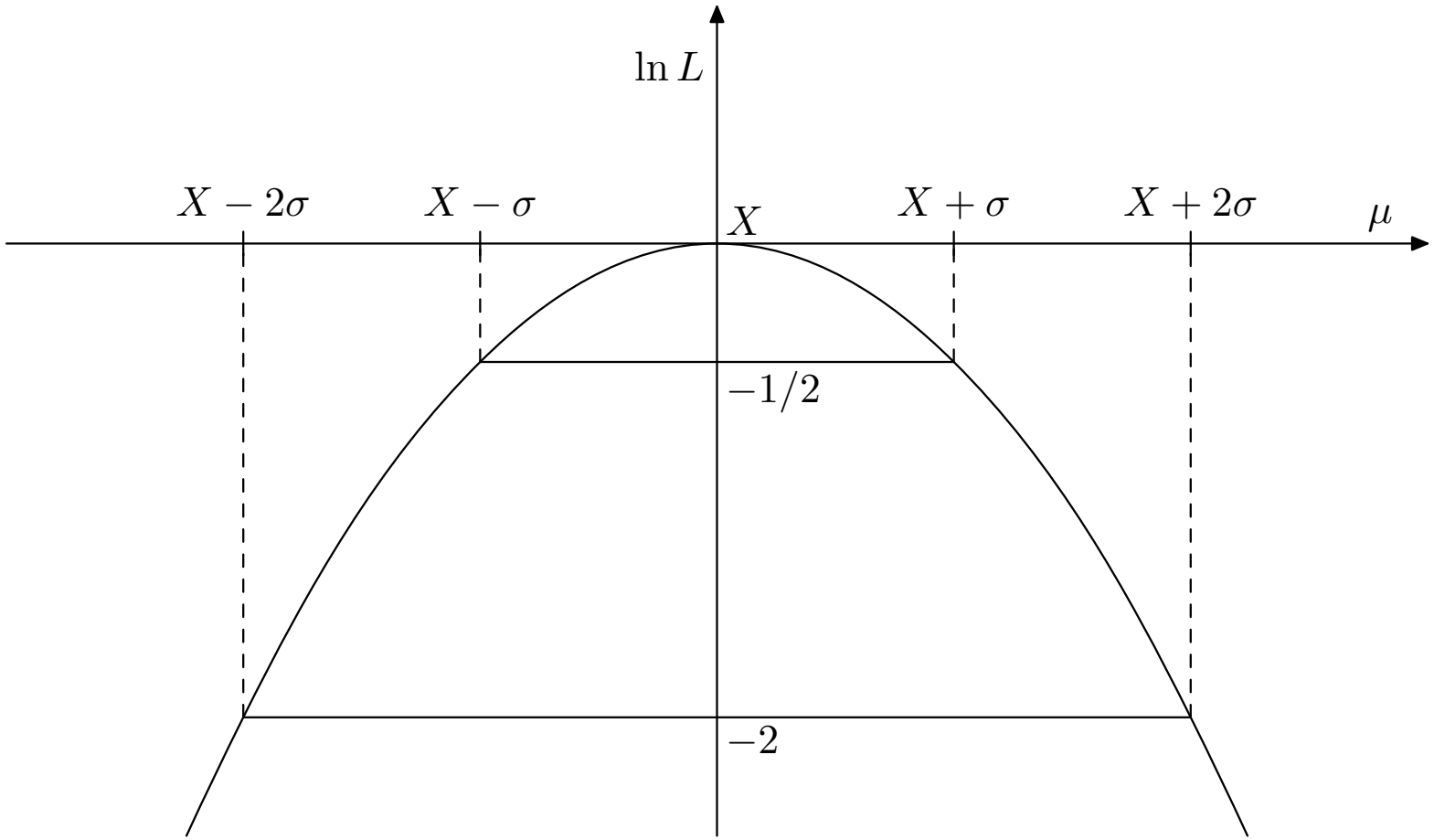
Recall the Normal Theory of Interval Estimation



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

In the Normal Theory, we showed how to convert the above pdf into a likelihood function by exchanging X and θ. Now if we take the logarithm of the above likelihood function, the Gaussian becomes a parabola as on the next slide.

Likelihood-based Confidence Intervals



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

Likelihood-based Confidence Intervals

When the data are Gaussian-distributed, the Normal Theory applies and confidence intervals can be calculated easily without the Neyman construction. In this case, the log-likelihood has a parabolic shape.

Let us now assume the inverse: that a parabolic log-likelihood function implies that the Normal Theory is applicable.

But if the log-likelihood is not parabolic, it can always (almost always) be transformed to a parabola by a (non-linear) transformation of the parameter μ.

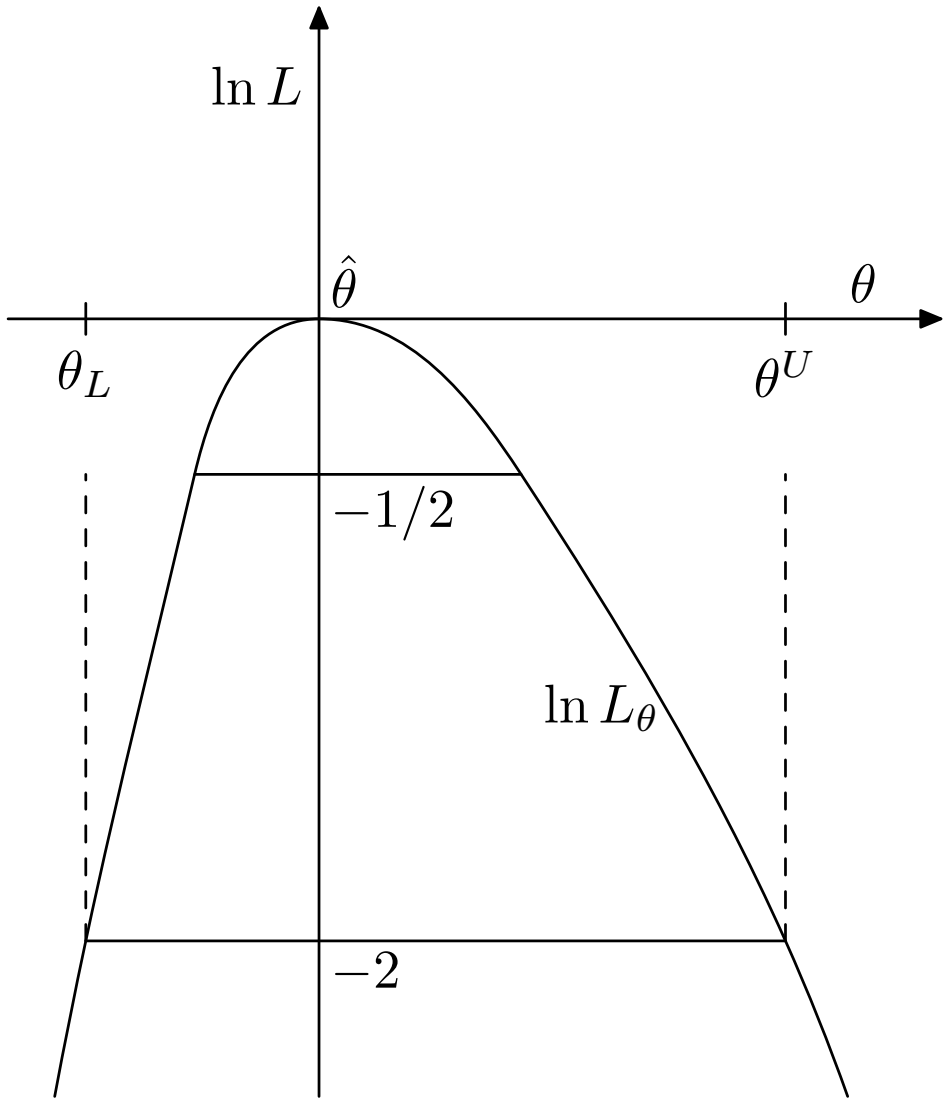
But, since the values of the likelihood function are invariant, it is not necessary to find the transformation that would make it Gaussian.

One only has to read off the parameter values for which
$\ln L = \ln L_{\max} - 1/2$ (for the one-sigma confidence interval).

Likelihood-based Confidence Intervals

Non-parabolic log-likelihood

θ^U, θ_L are invariant



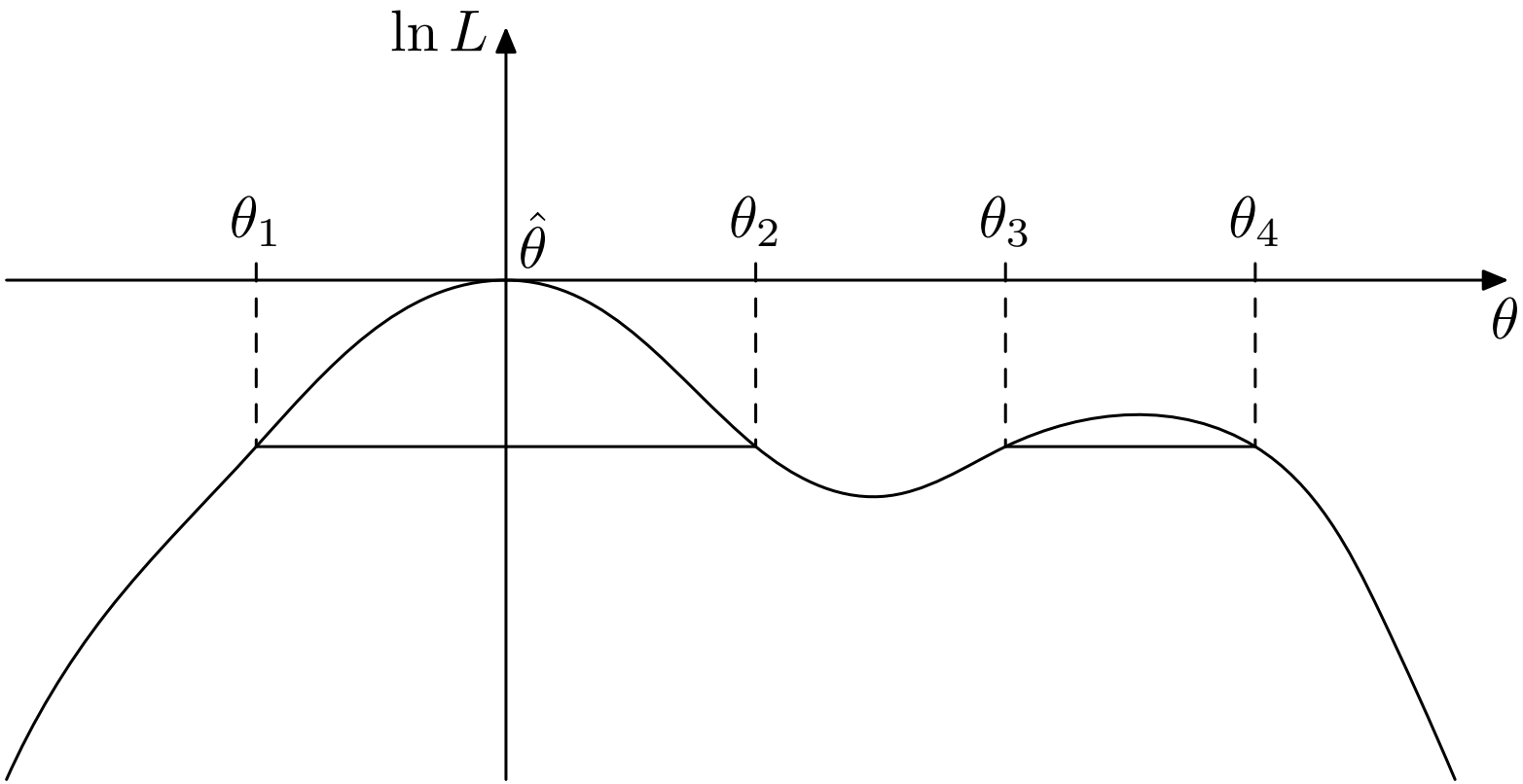
Likelihood-based Confidence Intervals

This method was suggested by a statistician working at CERN (D. Hudson) in 1964, and we included it already in the early versions of Minuit (1966). At the time, the properties of the method were known only for a few simple (one-parameter) problems.

Physicists know this as the method of MINOS, since that is the Minuit command that calculates this confidence interval. Much later, statisticians started to study it for the multiparameter case (which they called profile likelihood). It turns out to have surprisingly good coverage.

We knew it did not have good coverage for the simplest problem (Poisson with very few events and no background), but we thought it should be good for other problems.
It turns out to have excellent coverage for large numbers of parameters, so it is good for handling nuisance parameters.

Likelihood-based Confidence Intervals



"Pathological" log-likelihood function.

Likelihood-based Confidence Intervals

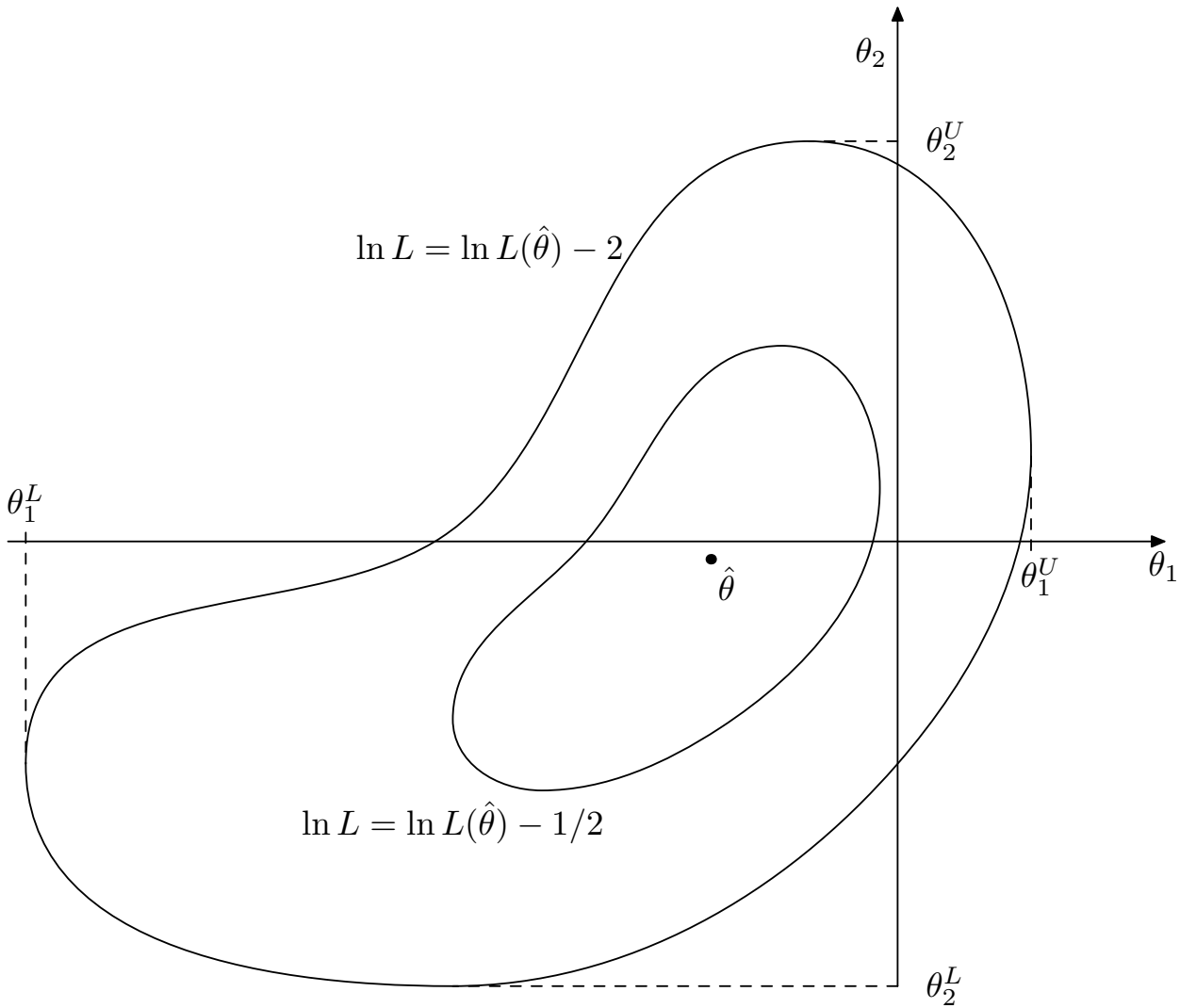
The Minuit command MINOS finds the intersection of the likelihood function with the value $\ln L_{\max} - 1/2$. This gives in general an asymmetric confidence interval, whereas the Normal Theory always gives an interval symmetric about the M.L. estimate.

When there are several parameters, the MINOS error on the k^{th} parameter is found by looking at the function g:

$$g(x_k) = \max_{x_i, i \neq k} \ln L(\mathbf{X})$$

and finding the two points where $g(x_k) = \ln L_{\max} - 1/2$.

Likelihood-based Confidence Intervals



Log-likelihood contours in two variables for non-linear problem.

Measurements outside the physically allowed region.

Suppose one measures the neutrino mass squared by estimating E^2 and p^2 and subtracting. Assume that the measurements E^2 and p^2 are Gaussian-distributed. Suppose also that $m^2 \approx 0$.

If such an experiment is repeated many times, one would expect half the measurements of m^2 to be negative. What should the unlucky experimenter report?

1. Using the methods of frequentist point estimation and neglecting the unphysicalness, one would report the unbiassed estimate, even if it is negative, and the standard deviation of the estimate. This result can be averaged with others, and also is a good measure of the sensitivity of the experiment.

Measurements outside the physically allowed region.

2. Using the methods of frequentist interval estimation one would report the Feldman-Cousins upper limit, which is always physical and has correct coverage, but cannot be used for averaging or measuring sensitivity.

Measurements outside the physically allowed region.

2. Using the methods of frequentist interval estimation one would report the Feldman-Cousins upper limit, which is always physical and has correct coverage, but cannot be used for averaging or measuring sensitivity.
3. Using Bayesian theory, one would report the Bayesian posterior or some interval obtained from it, which is always physical but depends on arbitrary choices including the prior. It cannot be used for averaging and does not in general have frequentist coverage.

Measurements outside the physically allowed region.

2. Using the methods of frequentist interval estimation one would report the Feldman-Cousins upper limit, which is always physical and has correct coverage, but cannot be used for averaging or measuring sensitivity.
3. Using Bayesian theory, one would report the Bayesian posterior or some interval obtained from it, which is always physical but depends on arbitrary choices including the prior. It cannot be used for averaging and does not in general have frequentist coverage.
4. Reporting the likelihood function is often proposed as the universal solution, but it is not really defined outside the physical region.