

Classical interval estimation, limits systematics and beyond - Part I



IN2P3 School of Statistics
Zoom / 19 January 2021



Glen Cowan
Physics Department
Royal Holloway, University of London
`g.cowan@rhul.ac.uk`
`www.pp.rhul.ac.uk/~cowan`

Outline

- Interval estimation
- Confidence region using Wilks' theorem
- Limits for Poisson parameter

Recap of hypothesis tests

Consider test of a parameter μ , e.g., proportional to cross section.

Result of measurement is data \mathbf{x} , whose pdf depends on μ .

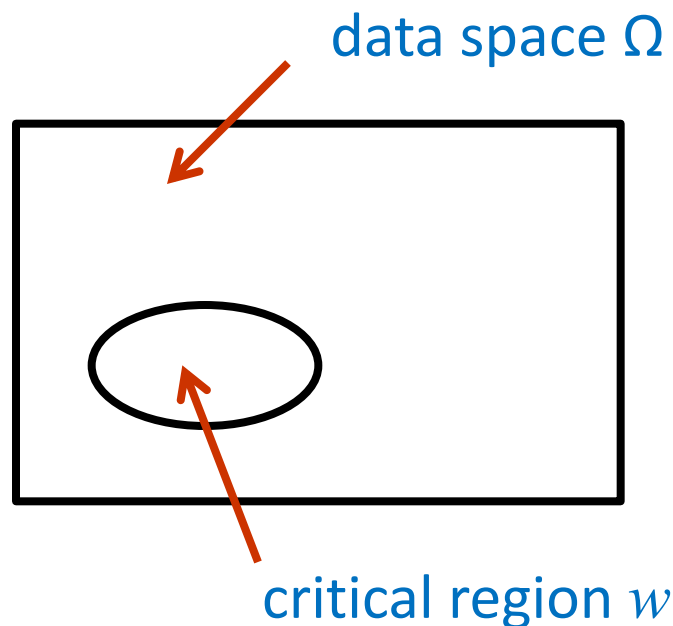
To define test of μ , specify *critical region* w_μ , such that probability to find $\mathbf{x} \in w_\mu$ is not greater than α (the *size* or *significance level*):

$$P(\mathbf{x} \in w_\mu | \mu) \leq \alpha$$

Often use, e.g., $\alpha = 0.05$.

If observe $\mathbf{x} \in w_\mu$, reject μ .

Equivalently define a *p-value* p_μ such that the critical region corresponds to $p_\mu \leq \alpha$.

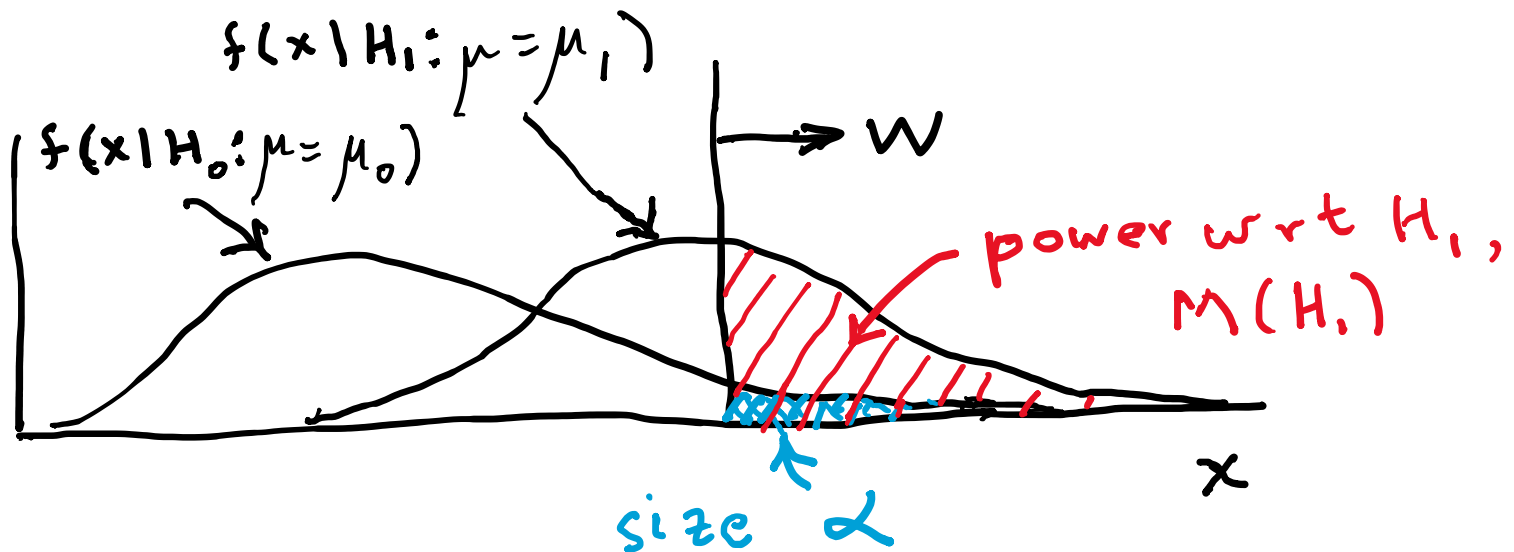


Power of test

In general there are an infinite number of possible critical regions that give the same size α .

To define the test of H_0 , consider a relevant alternative H_1 and use it to motivate where to place the critical region.

Roughly speaking, place the critical region where there is a low probability (α) to be found if H_0 is true, but high if H_1 is true:



Test statistic for p -value

Often define the test with a statistic $q_\mu(\mathbf{x})$ such that the boundary of the critical region is $q_\mu(\mathbf{x}) = c_\alpha$ for some constant c_α .

For examples of statistics based on the profile likelihood ratio, see, e.g., CCGV, EPJC 71 (2011) 1554; arXiv:1007.1727.

Usually define q_μ such that higher values represent increasing incompatibility between the data and the hypothesized μ , so that the p -value of μ is

$$p_\mu = \int_{q_{\mu,\text{obs}}}^{\infty} f(q_\mu|\mu) dq_\mu$$

observed value of q_μ

pdf of q_μ assuming μ

Equivalent formulation of test: reject μ if $p_\mu \leq \alpha$.

Confidence intervals by inverting a test

In addition to a 'point estimate' of a parameter we should report an interval reflecting its statistical uncertainty.

Confidence intervals for a parameter θ can be found by defining a test of the hypothesized value θ (do this for all θ):

Specify values of the data that are 'disfavoured' by θ (critical region) such that $P(\text{data in critical region} | \theta) \leq \alpha$ for a prespecified α , e.g., 0.05 or 0.1.

If data observed in the critical region, reject the value θ .

Now invert the test to define a confidence interval as:

set of θ values that are not rejected in a test of size α (confidence level CL is $1 - \alpha$).

Relation between confidence interval and p -value

Equivalently we can consider a significance test for each hypothesized value of θ , resulting in a p -value, p_θ .

If $p_\theta \leq \alpha$, then we reject θ .

The confidence interval at $CL = 1 - \alpha$ consists of those values of θ that are not rejected.

E.g. an upper limit on θ is the greatest value for which $p_\theta > \alpha$.

In practice find by setting $p_\theta = \alpha$ and solve for θ .

For a multidimensional parameter space $\theta = (\theta_1, \dots, \theta_M)$ use same idea – result is a confidence “region” with boundary determined by $p_\theta = \alpha$.

Coverage probability of confidence interval

If the true value of θ is rejected, then it's not in the confidence interval. The probability for this is by construction (equality for continuous data):

$$P(\text{reject } \theta | \theta) \leq \alpha = \text{type-I error rate}$$

Therefore, the probability for the interval to contain or “cover” θ is

$$P(\text{conf. interval “covers” } \theta | \theta) \geq 1 - \alpha$$

This assumes that the set of θ values considered includes the true value, i.e., it assumes the composite hypothesis $P(x|H, \theta)$.

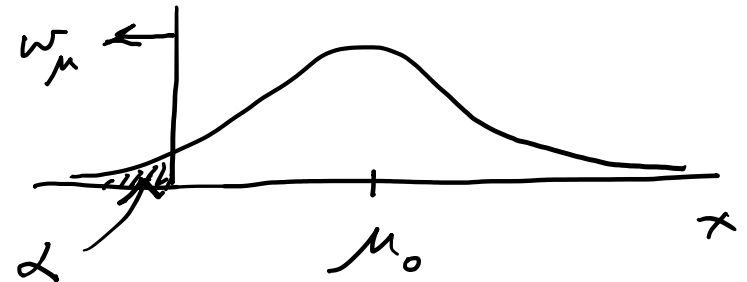
Example: upper limit on mean of Gaussian

When we test the parameter, we should take the critical region to maximize the power with respect to the relevant alternative(s).

Example: $x \sim \text{Gauss}(\mu, \sigma)$ (take σ known)

Test $H_0 : \mu = \mu_0$ versus the alternative $H_1 : \mu < \mu_0$

→ Put w_μ at region of x -space characteristic of low μ (i.e. at low x)

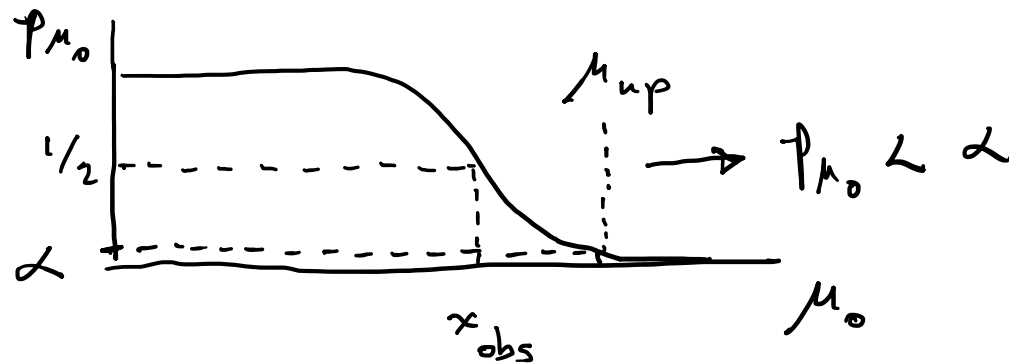


Equivalently, take the p -value to be

$$p_{\mu_0} = P(x \leq x_{\text{obs}} | \mu_0) = \int_{-\infty}^{x_{\text{obs}}} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_0)^2/2\sigma^2} dx = \Phi\left(\frac{x_{\text{obs}} - \mu_0}{\sigma}\right)$$

Upper limit on Gaussian mean (2)

To find confidence interval, repeat for all μ_0 , i.e., set $p_{\mu_0} = \alpha$ and solve for μ_0 to find the interval's boundary



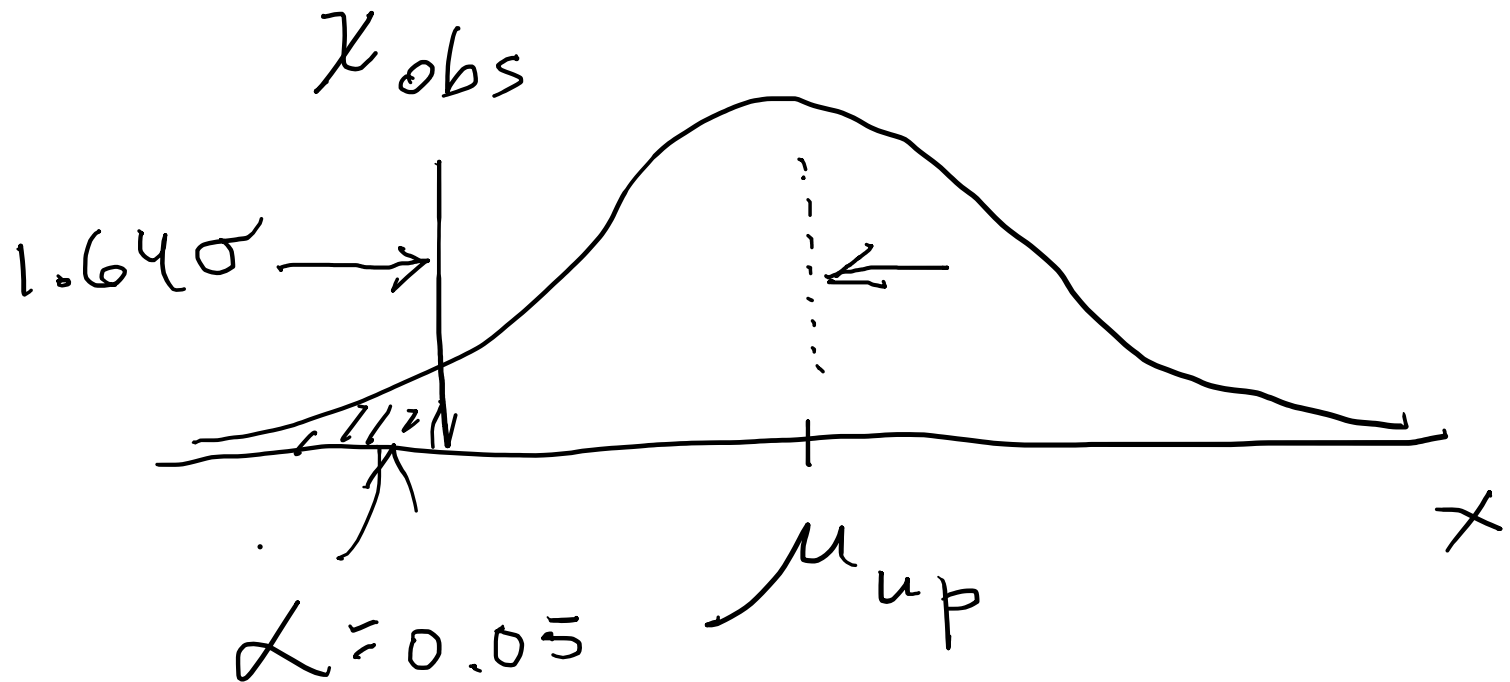
$$\mu_0 \rightarrow \mu_{up} = x_{obs} - \sigma \Phi^{-1}(\alpha) = x_{obs} + \sigma \Phi^{-1}(1 - \alpha)$$

This is an upper limit on μ , i.e., higher μ have even lower p -value and are in even worse agreement with the data.

Usually use $\Phi^{-1}(\alpha) = -\Phi^{-1}(1-\alpha)$ so as to express the upper limit as x_{obs} plus a positive quantity. E.g. for $\alpha = 0.05$, $\Phi^{-1}(1-0.05) = 1.64$.

Upper limit on Gaussian mean (3)

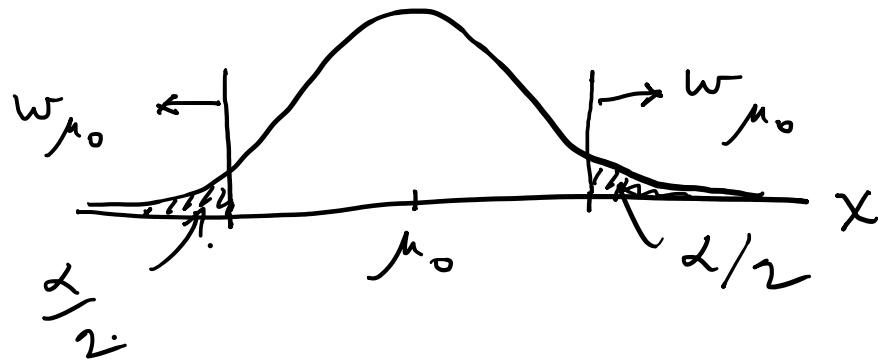
μ_{up} = the hypothetical value of μ such that there is only a probability α to find $x < x_{\text{obs}}$.



1- vs. 2-sided intervals

Now test: $H_0 : \mu = \mu_0$ versus the alternative $H_1 : \mu \neq \mu_0$

I.e. we consider the alternative to μ_0 to include higher and lower values, so take critical region on both sides:



Result is a “central” confidence interval $[\mu_{\text{lo}}, \mu_{\text{up}}]$:

$$\mu_{\text{lo}} = x_{\text{obs}} - \sigma \Phi^{-1} \left(1 - \frac{\alpha}{2} \right)$$

E.g. for $\alpha = 0.05$

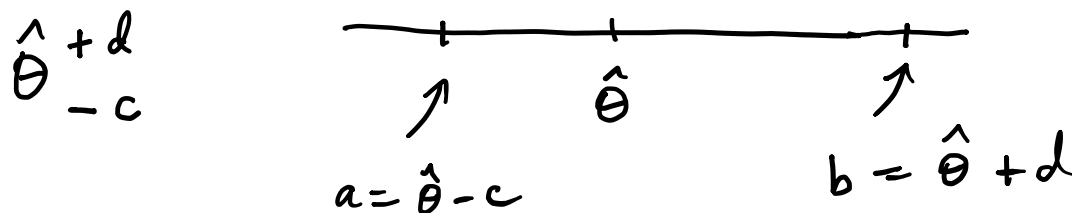
$$\mu_{\text{up}} = x_{\text{obs}} + \sigma \Phi^{-1} \left(1 - \frac{\alpha}{2} \right)$$

$$\Phi^{-1} \left(1 - \frac{\alpha}{2} \right) = 1.96 \approx 2$$

Note upper edge of two-sided interval is higher (i.e. not as tight of a limit) than obtained from the one-sided test.

On the meaning of a confidence interval

Often we report the confidence interval $[a, b]$ together with the point estimate as an “asymmetric error bar”, e.g.,


$$\hat{\theta} + d$$
$$-c$$
$$a = \hat{\theta} - c$$
$$b = \hat{\theta} + d$$

E.g. (at $CL = 1 - \alpha = 68.3\%$):

$$\hat{\theta} = 80.25^{+0.31}_{-0.25}$$

Does this mean $P(80.00 < \theta < 80.56) = 68.3\%$? No, not for a frequentist confidence interval. The parameter θ does not fluctuate upon repetition of the measurement; the endpoints of the interval do, i.e., the endpoints of the interval fluctuate (they are functions of data):

$$P(a(x) < \theta < b(x)) = 1 - \alpha$$

Approximate confidence intervals/regions from the likelihood function

Suppose we test parameter value(s) $\theta = (\theta_1, \dots, \theta_n)$ using the ratio

$$\lambda(\theta) = \frac{L(\theta)}{L(\hat{\theta})} \quad 0 \leq \lambda(\theta) \leq 1$$

Lower $\lambda(\theta)$ means worse agreement between data and hypothesized θ . Equivalently, usually define

$$t_\theta = -2 \ln \lambda(\theta)$$

so higher t_θ means worse agreement between θ and the data.

p -value of θ therefore

$$p_\theta = \int_{t_{\theta, \text{obs}}}^{\infty} f(t_\theta | \theta) dt_\theta$$

need pdf

Confidence region from Wilks' theorem

Wilks' theorem says (in large-sample limit and provided certain conditions hold...)

$$f(t_\theta|\theta) \sim \chi_n^2$$

chi-square dist. with # d.o.f. =
of components in $\theta = (\theta_1, \dots, \theta_n)$.

Assuming this holds, the p -value is

$$p_\theta = 1 - F_{\chi_n^2}(t_\theta) \quad \leftarrow \text{set equal to } \alpha$$

To find boundary of confidence region set $p_\theta = \alpha$ and solve for t_θ :

$$t_\theta = F_{\chi_n^2}^{-1}(1 - \alpha)$$

Recall also

$$t_\theta = -2 \ln \frac{L(\theta)}{L(\hat{\theta})}$$

Confidence region from Wilks' theorem (cont.)

i.e., boundary of confidence region in θ space is where

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2} F_{\chi_n^2}^{-1}(1 - \alpha)$$

For example, for $1 - \alpha = 68.3\%$ and $n = 1$ parameter,

$$F_{\chi_1^2}^{-1}(0.683) = 1$$

and so the 68.3% confidence level interval is determined by

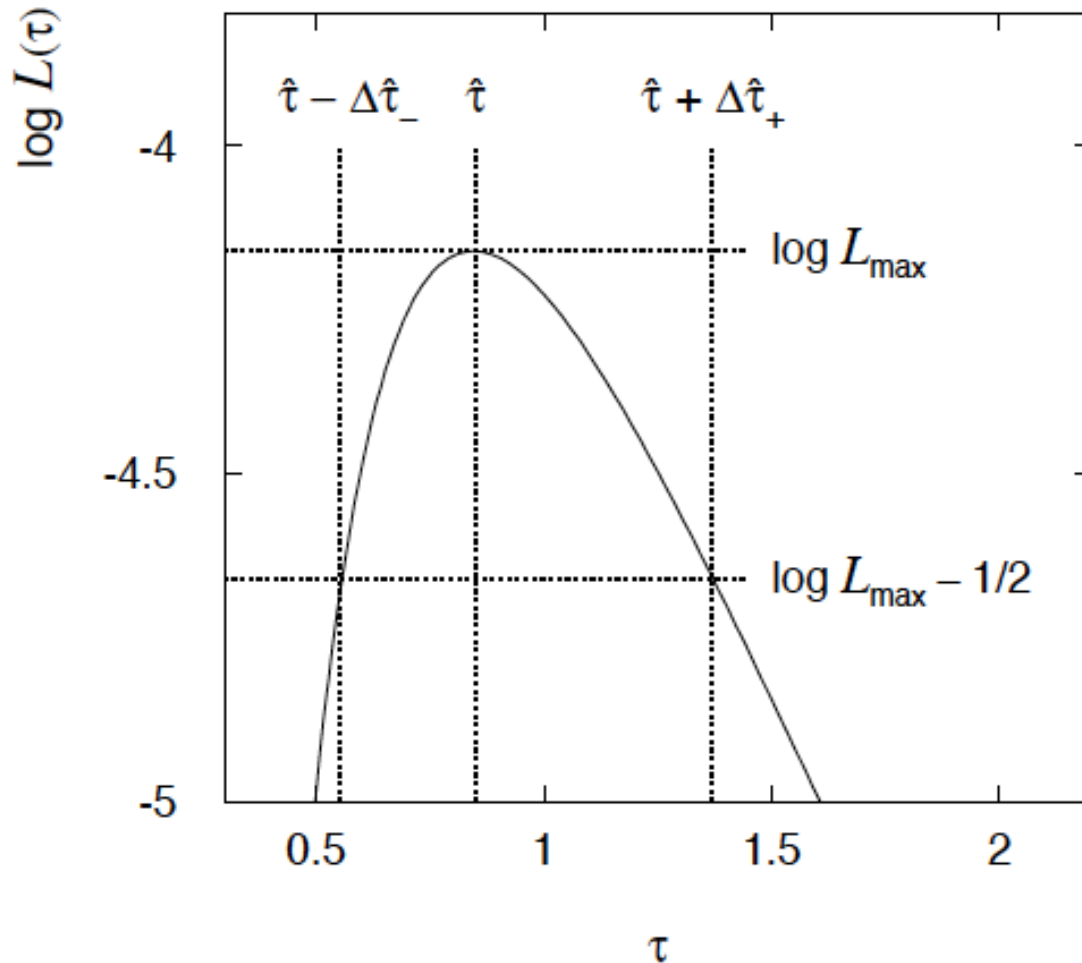
$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

Same as recipe for finding the estimator's standard deviation, i.e.,

$[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$ is a 68.3% CL confidence interval.

Example of interval from $\ln L(\theta)$

For $n=1$ parameter, $\text{CL} = 0.683$, $Q_\alpha = 1$.



Our exponential example, now with only $n = 5$ events.

Can report ML estimate with approx. confidence interval from $\ln L_{\max} - 1/2$ as “asymmetric error bar”:

$$\hat{\tau} = 0.85^{+0.52}_{-0.30}$$

Multiparameter case

For increasing number of parameters, $CL = 1 - \alpha$ decreases for confidence region determined by a given

$$Q_\alpha = F_{\chi_n^2}^{-1}(1 - \alpha)$$

Q_α	$1 - \alpha$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

Multiparameter case (cont.)

Equivalently, Q_α increases with n for a given $CL = 1 - \alpha$.

$1 - \alpha$	\hat{Q}_α				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

Frequentist upper limit on Poisson parameter

Consider again the case of observing $n \sim \text{Poisson}(s + b)$.

Suppose $b = 4.5$, $n_{\text{obs}} = 5$. Find upper limit on s at 95% CL.

Relevant alternative is $s = 0$ (critical region at low n)

p -value of hypothesized s is $P(n \leq n_{\text{obs}}; s, b)$

Upper limit s_{up} at $\text{CL} = 1 - \alpha$ found from

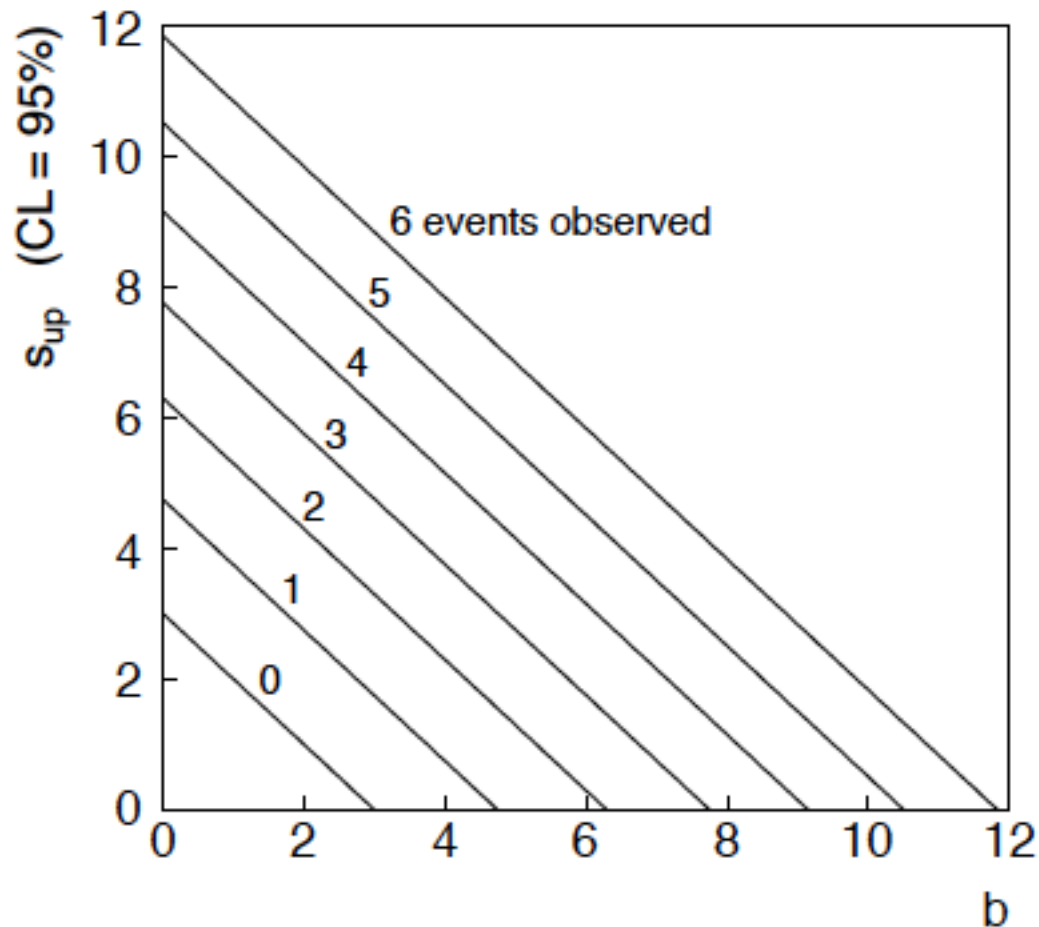
$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}} + b)}$$

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1}(1 - \alpha; 2(n_{\text{obs}} + 1)) - b$$

$$= \frac{1}{2} F_{\chi^2}^{-1}(0.95; 2(5 + 1)) - 4.5 = 6.0$$

$n \sim \text{Poisson}(s+b)$: frequentist upper limit on s

For low fluctuation of n , formula can give negative result for s_{up} ; i.e. confidence interval is empty; all values of $s \geq 0$ have $p_s \leq \alpha$.



Limits near a boundary of the parameter space

Suppose e.g. $b = 2.5$ and we observe $n = 0$.

If we choose $CL = 0.9$, we find from the formula for s_{up}

$$s_{\text{up}} = -0.197 \quad (CL = 0.90)$$

Physicist:

We already knew $s \geq 0$ before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma when testing parameter values for which one has very little experimental sensitivity, e.g., very small s .

Expected limit for $s = 0$

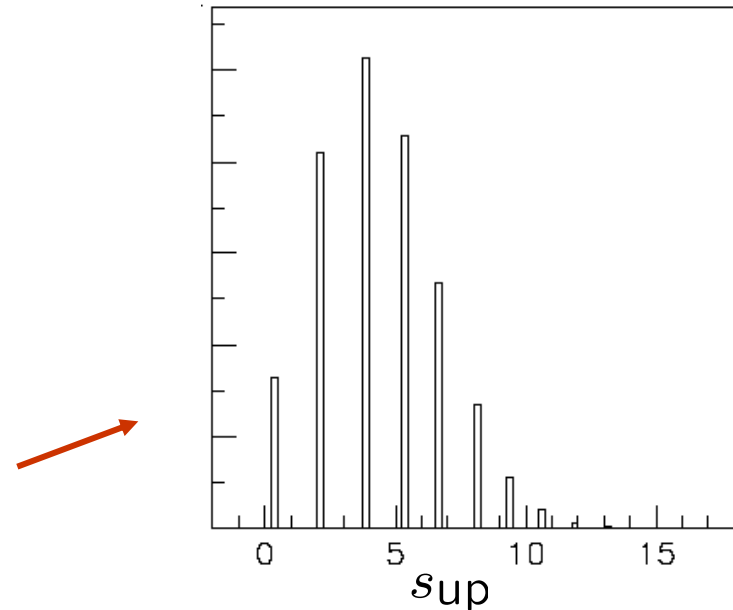
Physicist: I should have used $\text{CL} = 0.95$ — then $s_{\text{up}} = 0.496$

Even better: for $\text{CL} = 0.917923$ we get $s_{\text{up}} = 10^{-4}$!

Reality check: with $b = 2.5$, typical Poisson fluctuation in n is at least $\sqrt{2.5} = 1.6$. How can the limit be so low?

Look at the mean limit for the no-signal hypothesis ($s = 0$) (sensitivity).

Distribution of 95% CL limits
with $b = 2.5$, $s = 0$.
Mean upper limit = 4.44



The Bayesian approach to limits

In Bayesian statistics need to start with ‘prior pdf’ $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Bayes’ theorem tells how our beliefs should be updated in light of the data x :

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta') d\theta'} \propto L(x|\theta)\pi(\theta)$$

Integrate posterior pdf $p(\theta|x)$ to give interval with any desired probability content.

For e.g. $n \sim \text{Poisson}(s+b)$, 95% CL upper limit on s from

$$0.95 = \int_{-\infty}^{s_{\text{up}}} p(s|n) ds$$

Bayesian prior for Poisson parameter

Include knowledge that $s \geq 0$ by setting prior $\pi(s) = 0$ for $s < 0$.

Could try to reflect ‘prior ignorance’ with e.g.

$$\pi(s) = \begin{cases} 1 & s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Not normalized; can be OK provided $L(s)$ dies off quickly for large s .

Not invariant under change of parameter — if we had used instead a flat prior for a nonlinear function of s , then this would imply a non-flat prior for s .

Doesn’t really reflect a reasonable degree of belief, but often used as a point of reference; or viewed as a recipe for producing an interval whose frequentist properties can be studied (e.g., coverage probability, which will depend on true s).

Bayesian upper limit with flat prior for s

Put Poisson likelihood and flat prior into Bayes' theorem:

$$p(s|n) \propto \frac{(s+b)^n}{n!} e^{-(s+b)} \quad (s \geq 0)$$

Normalize to unit area:

$$p(s|n) = \frac{(s+b)^n e^{-(s+b)}}{\Gamma(b, n+1)}$$

← upper incomplete gamma function

Upper limit s_{up} determined by requiring

$$1 - \alpha = \int_0^{s_{\text{up}}} p(s|n) ds$$

Bayesian interval with flat prior for s

Solve to find limit s_{up} :

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1} [p, 2(n+1)] - b$$

where

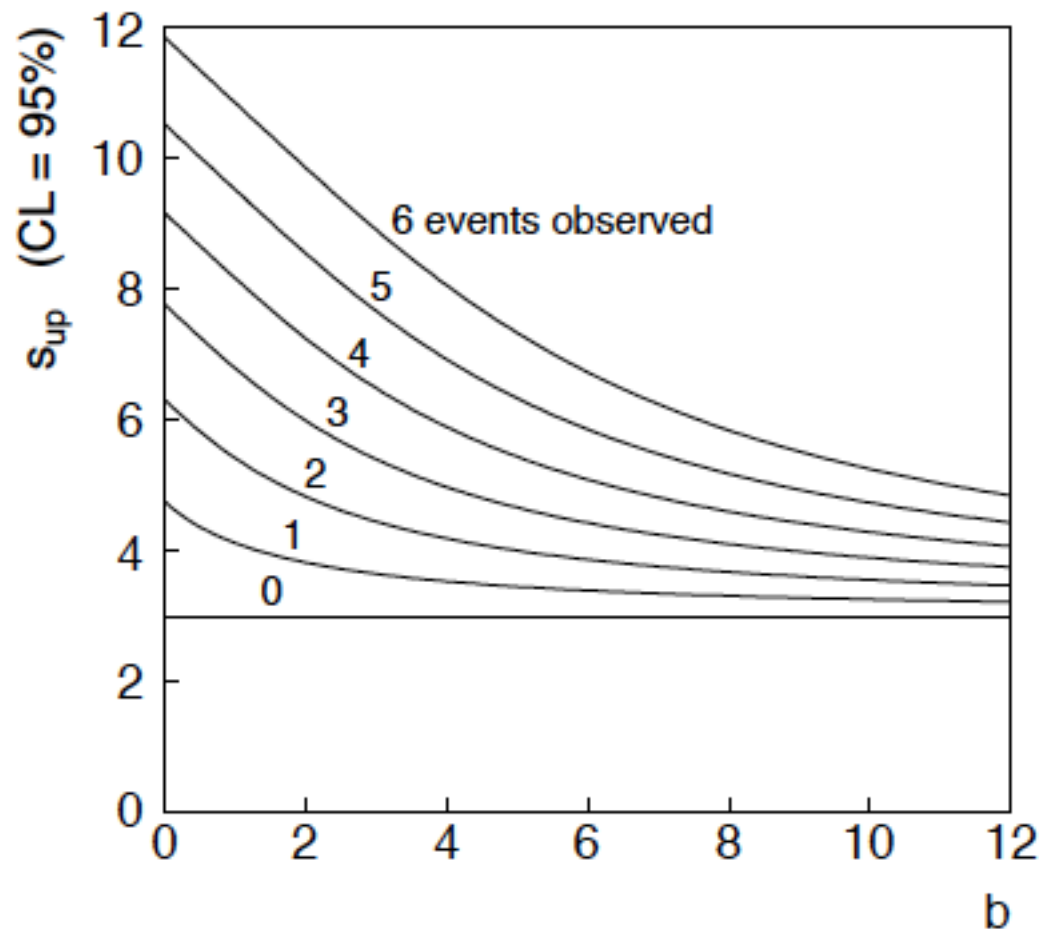
$$p = 1 - \alpha \left(1 - F_{\chi^2} [2b, 2(n+1)] \right)$$

For special case $b = 0$, Bayesian upper limit with flat prior numerically same as one-sided frequentist case ('coincidence').

Bayesian interval with flat prior for s

For $b > 0$ Bayesian limit is everywhere greater than the (one sided) frequentist upper limit.

Never goes negative. Doesn't depend on b if $n = 0$.



Priors from formal rules

Last time we took the prior for a Poisson mean to be constant to reflect a lack of prior knowledge; we noted this was not invariant under change of parameter.

Because of difficulties in encoding a vague degree of belief in a prior, one often attempts to derive the prior from formal rules, e.g., to satisfy certain invariance principles or to provide maximum information gain for a certain set of measurements.

Often called “objective priors”

Form basis of Objective Bayesian Statistics

The priors do not reflect a degree of belief (but might represent possible extreme cases).

In Objective Bayesian analysis, can use the intervals in a frequentist way, i.e., regard Bayes’ theorem as a recipe to produce an interval with a given coverage probability.

Priors from formal rules (cont.)

For a review of priors obtained by formal rules see, e.g.,

Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996).

Formal priors have not been widely used in Particle Physics, but there has been interest in this direction, especially the reference priors of Bernardo and Berger; see e.g.

L. Demortier, S. Jain and H. Prosper, *Reference priors for high energy physics*, Phys. Rev. D 82 (2010) 034002, arXiv:1002.1111.

D. Casadei, *Reference analysis of the signal + background model in counting experiments*, JINST 7 (2012) 01012; arXiv:1108.4270.

Jeffreys prior

According to *Jeffreys' rule*, take prior according to

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}$$

where

$$I_{ij}(\boldsymbol{\theta}) = -E \left[\frac{\partial^2 \ln L(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right] = - \int \frac{\partial^2 \ln L(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} L(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}$$

is the Fisher information matrix.

One can show that this leads to inference that is invariant under a transformation of parameters in the following sense:

Start with the Jeffreys prior for θ : $\pi_{\theta}(\theta) \sim \sqrt{\det I(\theta)}$

Use it in Bayes' theorem to find:

$$P(\theta|\mathbf{x}) \propto P(\mathbf{x}|\theta)\pi_{\theta}(\theta)$$

Jeffreys prior (2)

Now consider a function $\eta(\theta)$. The posterior for η is

$$P(\eta|\mathbf{x}) = P(\theta|\mathbf{x}) \left| \frac{d\theta}{d\eta} \right|$$

Alternatively, start with η and use its Jeffreys' prior:

$$\pi_{\eta}(\eta) \propto \sqrt{\det I(\eta)}$$

Use this in Bayes' theorem: $P(\eta|\mathbf{x}) \propto P(\mathbf{x}|\eta)\pi_{\eta}(\eta)$

One can show that Jeffreys' prior results in the same $P(\eta|\mathbf{x})$ in both cases. For details (single-parameter case) see:

<http://www.pp.rhul.ac.uk/~cowan/stat/notes/JeffreysInvariance.pdf>

Jeffreys prior for Poisson mean

Suppose $n \sim \text{Poisson}(\mu)$. To find the Jeffreys' prior for μ ,

$$L(n|\mu) = \frac{\mu^n}{n!} e^{-\mu} \qquad \frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\mu^2}$$

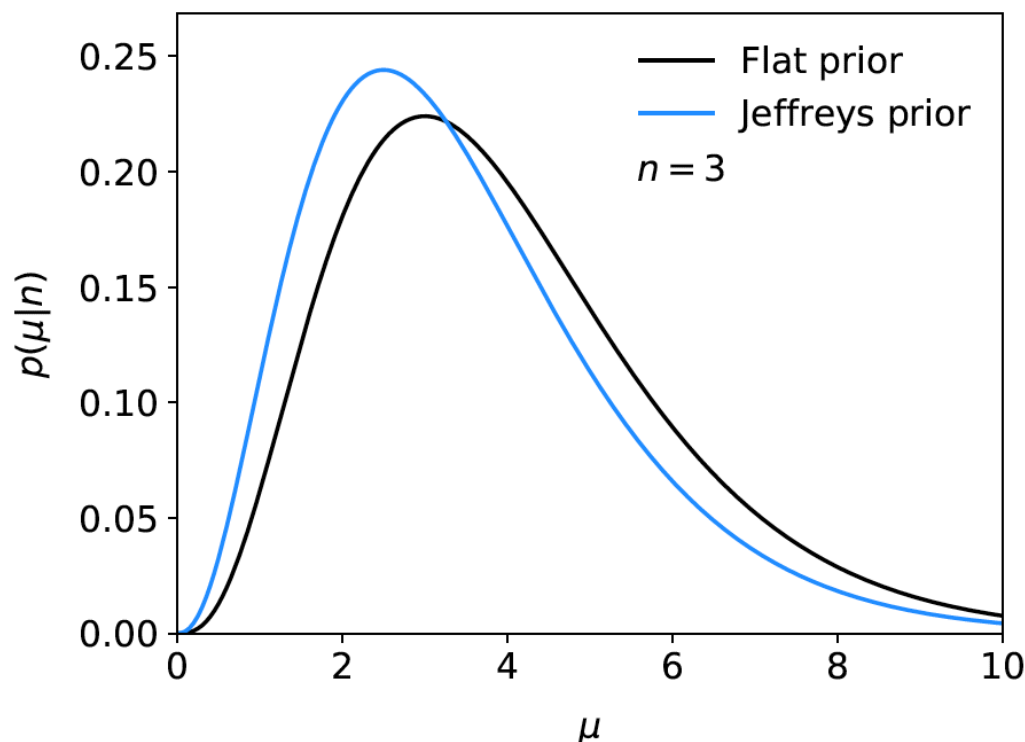
$$I = -E \left[\frac{\partial^2 \ln L}{\partial \mu^2} \right] = \frac{E[n]}{\mu^2} = \frac{1}{\mu}$$

$$\pi(\mu) \propto \sqrt{I(\mu)} = \frac{1}{\sqrt{\mu}}$$

So e.g. for $\mu = s + b$, this means the prior $\pi(s) \sim 1/\sqrt{s + b}$, which depends on b . But this is not designed as a degree of belief about s .

Posterior pdf for Poisson mean

From Bayes' theorem, $p(\mu|n) \propto \mu^n e^{-\mu} \pi(\mu)$



Flat, $\pi(\mu) = \text{const.}$

$$p(\mu|n) = \frac{\mu^n e^{-\mu}}{\Gamma(n+1)}$$

Jeffreys, $\pi(\mu) \sim 1/\sqrt{\mu}$

$$p(\mu|n) = \frac{\mu^{n-\frac{1}{2}} e^{-\mu}}{\Gamma(n+\frac{1}{2})}$$

In both cases, posterior is special case of gamma distribution.

Upper limit for Poisson mean

To find upper limit at $CL = 1 - \alpha$, solve

$$1 - \alpha = \int_0^{\mu_{\text{up}}} p(\mu|n) d\mu$$

Jeffreys prior: $\mu_{\text{up}} = P^{-1}(n + \frac{1}{2}, 1 - \alpha) = 7.03$

Flat prior: $\mu_{\text{up}} = P^{-1}(n + 1, 1 - \alpha) = 7.75$

$n=3,$
 $CL=0.95$

where P^{-1} is the inverse of the normalized lower incomplete gamma function (see `scipy.special`)

$$P(a, \mu_{\text{up}}) = \frac{1}{\Gamma(a)} \int_0^{\mu_{\text{up}}} \mu^{a-1} e^{-\mu} d\mu$$

Extra slides

Choosing a critical region

To construct a test of a hypothesis H_0 , we can ask what are the relevant alternatives for which one would like to have a high power.

Maximize power wrt H_1 = maximize probability to reject H_0 if H_1 is true.

Often such a test has a high power not only with respect to a specific point alternative but for a class of alternatives.

E.g., using a measurement $x \sim \text{Gauss}(\mu, \sigma)$ we may test

$H_0 : \mu = \mu_0$ versus the composite alternative $H_1 : \mu > \mu_0$

We get the highest power with respect to any $\mu > \mu_0$ by taking the critical region $x \geq x_c$ where the cut-off x_c is determined by the significance level such that

$$\alpha = P(x \geq x_c | \mu_0).$$

Test of $\mu = \mu_0$ vs. $\mu > \mu_0$ with $x \sim \text{Gauss}(\mu, \sigma)$

Standard Gaussian
cumulative distribution

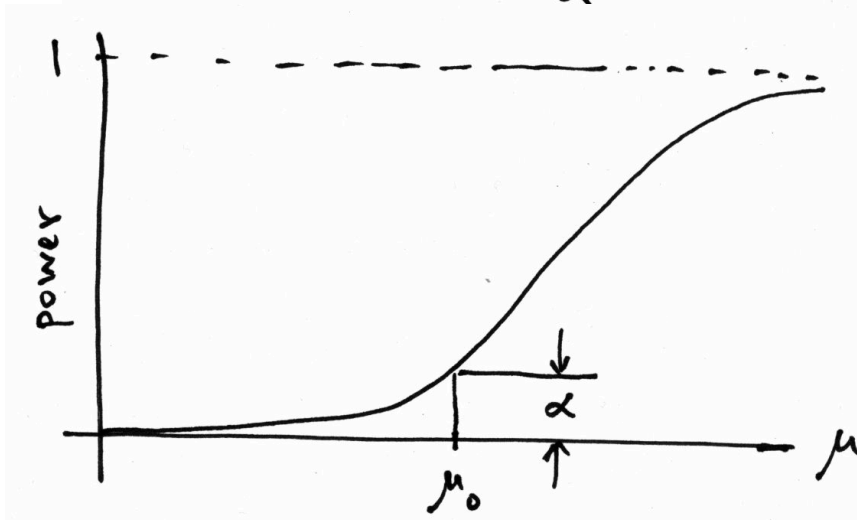
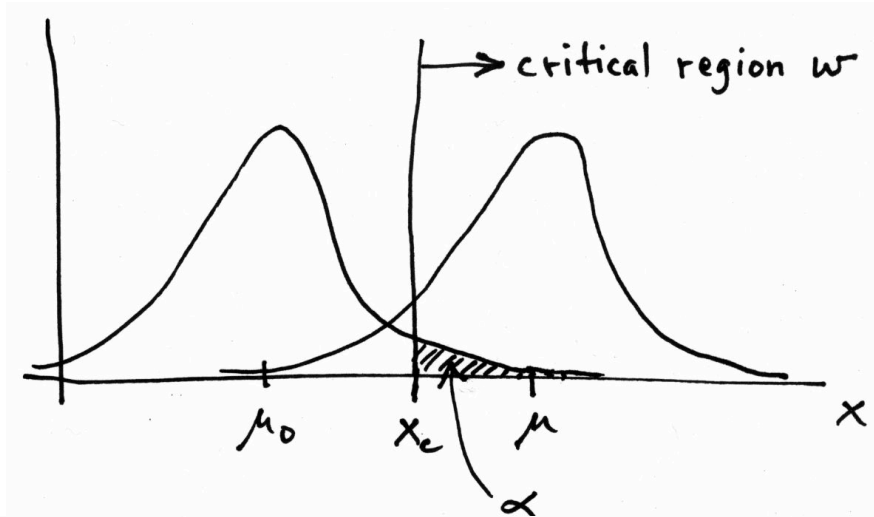
$$\alpha = 1 - \Phi\left(\frac{x_c - \mu_0}{\sigma}\right)$$

$$x_c = \mu_0 + \sigma \Phi^{-1}(1 - \alpha)$$

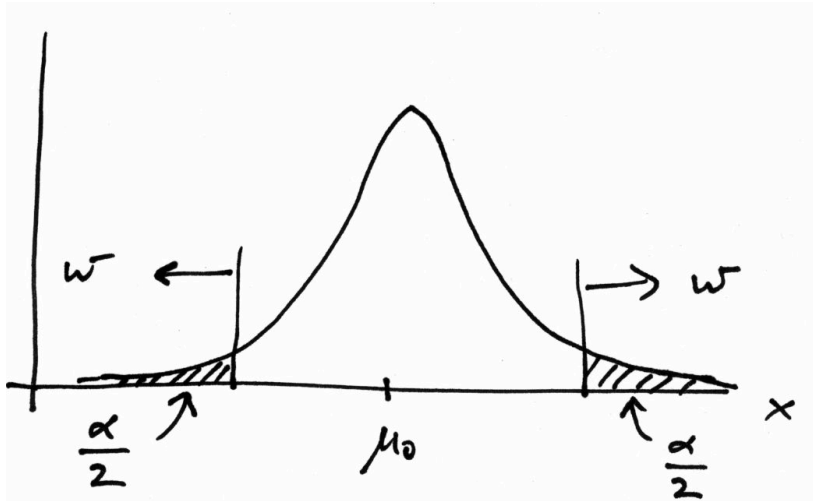
Standard Gaussian quantile

$$\text{power} = 1 - \beta = P(x > x_c | \mu) =$$

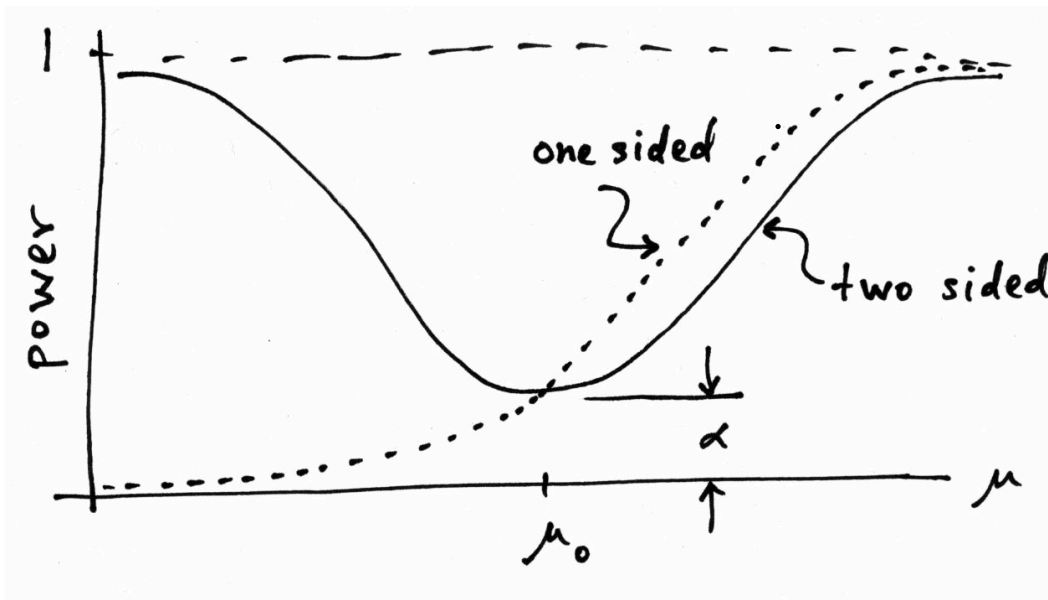
$$1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma} + \Phi^{-1}(1 - \alpha)\right)$$



Choice of critical region based on power (3)



But we might consider $\mu < \mu_0$ as well as $\mu > \mu_0$ to be viable alternatives, and choose the critical region to contain both high and low x (a two-sided test).



New critical region now gives reasonable power for $\mu < \mu_0$, but less power for $\mu > \mu_0$ than the original one-sided test.

No such thing as a model-independent test

In general we cannot find a single critical region that gives the maximum power for all possible alternatives (no “Uniformly Most Powerful” test).

In HEP we often try to construct a test of

H_0 : Standard Model (or “background only”, etc.)

such that we have a well specified “false discovery rate”,

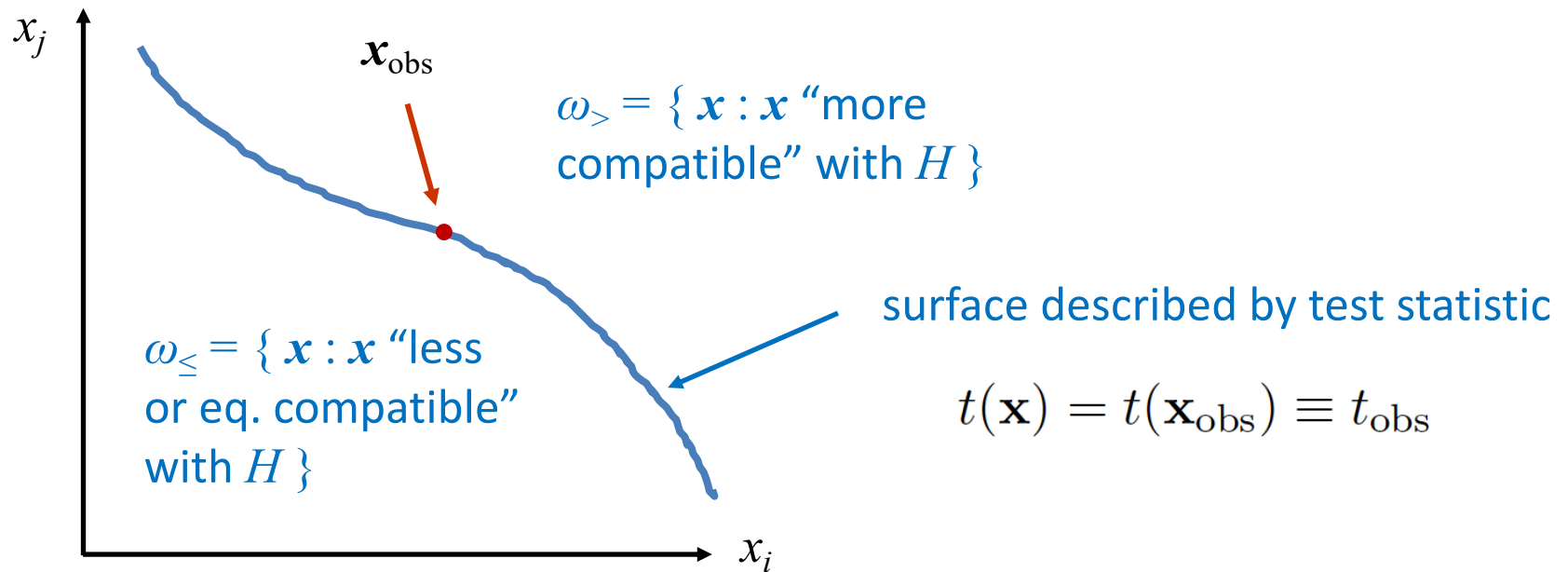
α = Probability to reject H_0 if it is true,

and high power with respect to some interesting alternative,

H_1 : SUSY, Z' , etc.

But there is no such thing as a “model independent” test. Any statistical test will inevitably have high power with respect to some alternatives and less power with respect to others.

p -value from test statistic



If e.g. we define the region of less or eq. compatibility to be $t(\mathbf{x}) \geq t_{\text{obs}}$ then the p -value of H is

$$p_H = \int_{t_{\text{obs}}}^{\infty} f(t|H) dt = \int_{\{\mathbf{x} : t(\mathbf{x}) \geq t_{\text{obs}}\}} f(\mathbf{x}|H) d\mathbf{x}$$

Distribution of the p -value

The p -value is a function of the data, and is thus itself a random variable with a given distribution. Suppose the p -value of H is found from a test statistic $t(\mathbf{x})$ as

$$p_H = \int_t^\infty f(t'|H) dt'$$

The pdf of p_H under assumption of H is

$$g(p_H|H) = \frac{f(t|H)}{|\partial p_H / \partial t|} = \frac{f(t|H)}{f(t|H)} = 1 \quad (0 \leq p_H \leq 1)$$

In general for continuous data, under assumption of H , $p_H \sim \text{Uniform}[0,1]$ and is concentrated toward zero for some (broad) class of alternatives.

