Statistics for Particle Physicists
Lecture 3: Hypothesis Tests, Confidence Intervals

Summer Student Lectures
CERN
4 – 7 July 2023

https://indico.cern.ch/event/1254879/timetable/

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

Outline

Lecture 1: Introduction, probability,

Lecture 2: Parameter estimation

→ Lecture 3: Hypothesis tests and confidence intervals
(some exercises here).

Lecture 4: Introduction to Machine Learning

Frequentist hypothesis tests

Suppose a measurement produces data $\boldsymbol{x}$; consider a hypothesis H_0 we want to test and alternative H_1

H_0, H_1 specify probability for $\boldsymbol{x}$: $P(\boldsymbol{x}|H_0)$, $P(\boldsymbol{x}|H_1)$

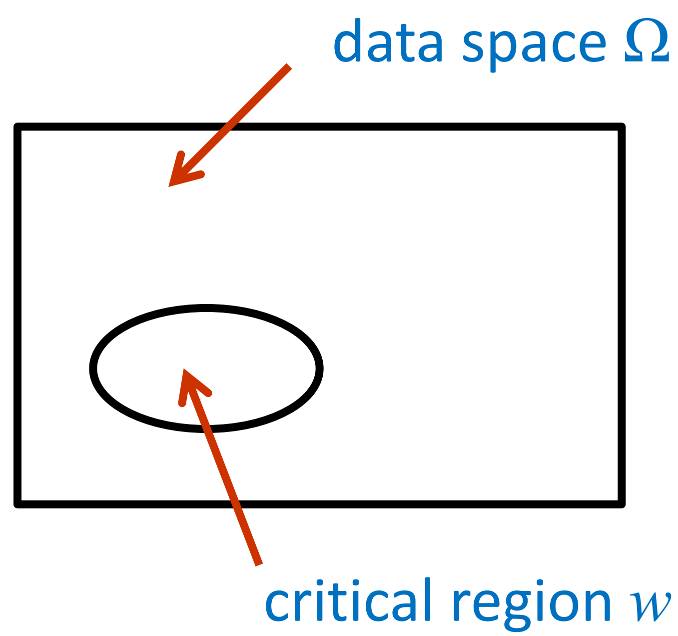
A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(\boldsymbol{x} \in w \mid H_0) \leq \alpha$$

Need inequality if data are discrete.

α is called the size or significance level of the test.

If $\boldsymbol{x}$ is observed in the critical region, reject H_0.

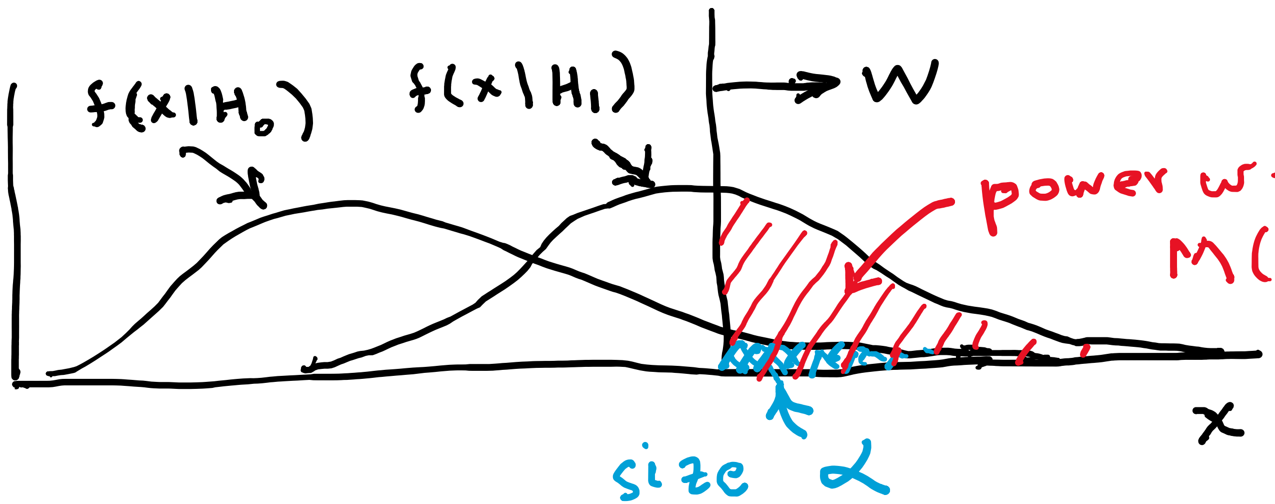


Definition of a test (2)

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

Use the alternative hypothesis H_1 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:



Classification viewed as a statistical test

Suppose events come in two possible types:

s (signal) and b (background)

For each event, test hypothesis that it is background, i.e., H_0 = b.

Carry out test on many events, each is either of type s or b, i.e., here the hypothesis is the "true class label", which varies randomly from event to event, so we can assign to it a frequentist probability.

Select events for which where H_0 is rejected as "candidate events of type s". Equivalent Particle Physics terminology:

background efficiency $$\varepsilon_{\rm b} = \int_W f(\mathbf{x}|H_0)\,d\mathbf{x} = \alpha$$

signal efficiency $$\varepsilon_{\rm s} = \int_W f(\mathbf{x}|H_1)\,d\mathbf{x} = 1 - \beta = \text{power}$$

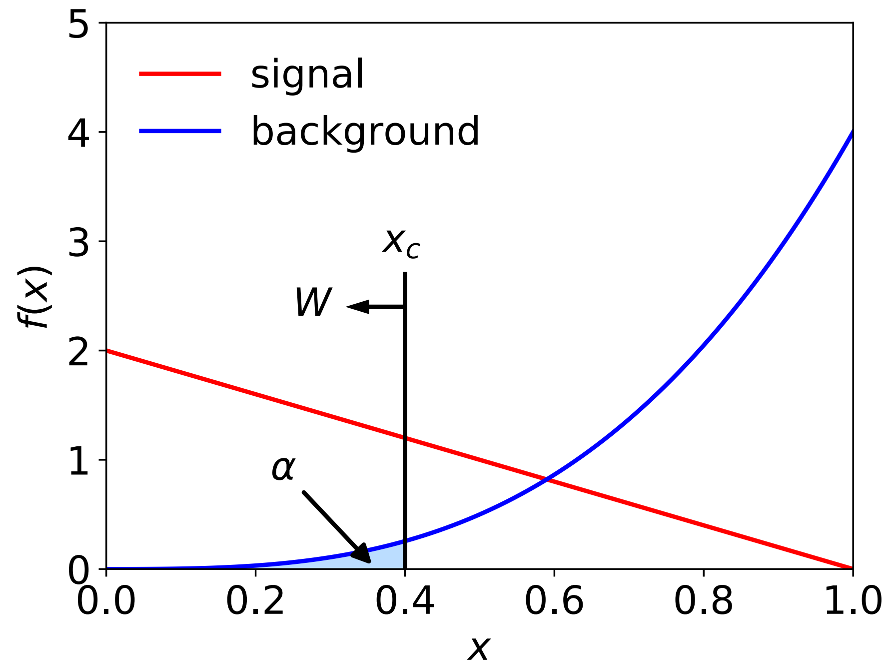
Example of a test for classification

Suppose we can measure for each event a quantity x, where

$$f(x|s) = 2(1-x)$$

$$f(x|b) = 4x^3$$

with $0 \leq x \leq 1$.



For each event in a mixture of signal (s) and background (b) test

H_0 : event is of type b

using a critical region W of the form: $W = \{x : x \leq x_{\rm c}\}$, where $x_{\rm c}$ is a constant that we choose to give a test with the desired size α.

Classification example (2)

Suppose we want $\alpha = 10^{-4}$. Require:

$$\alpha = P(x \leq x_{\mathrm{c}} | \mathrm{b}) = \int_0^{x_{\mathrm{c}}} f(x|b)\, dx = \left. \frac{4x^4}{4} \right|_0^{x_{\mathrm{c}}} = x_{\mathrm{c}}^4$$

and therefore $x_{\mathrm{c}} = \alpha^{1/4} = 0.1$

For this test (i.e. this critical region W), the power with respect to the signal hypothesis (s) is

$$M = P(x \leq x_{\mathrm{c}} | \mathrm{s}) = \int_0^{x_{\mathrm{c}}} f(x|\mathrm{s})\, dx = 2x_{\mathrm{c}} - x_{\mathrm{c}}^2 = 0.19$$

Note: the optimal size and power is a separate question that will depend on goals of the subsequent analysis.

Classification example (3)

Suppose that the prior probabilities for an event to be of type s or b are:

$$\pi_\mathrm{s} = 0.001$$
$$\pi_\mathrm{b} = 0.999$$

The "purity" of the selected signal sample (events where b hypothesis rejected) is found using Bayes' theorem:

$$P(\mathrm{s}|x \le x_\mathrm{c}) = \frac{P(x \le x_\mathrm{c}|\mathrm{s})\pi_\mathrm{s}}{P(x \le x_\mathrm{c}|\mathrm{s})\pi_\mathrm{s} + P(x \le x_\mathrm{c}|\mathrm{b})\pi_\mathrm{b}}$$

$$= 0.655$$

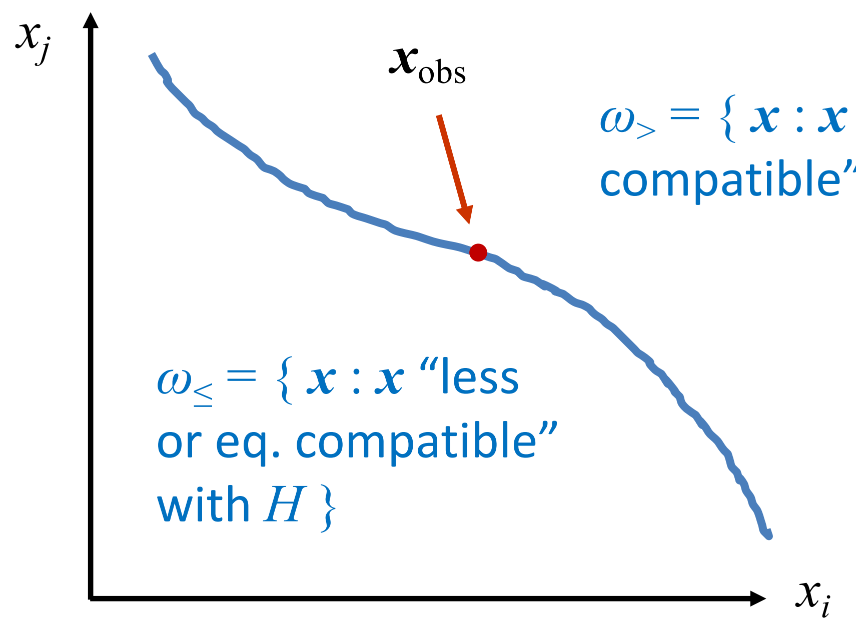
Testing significance / goodness-of-fit

Suppose hypothesis H predicts pdf $f(\boldsymbol{x}|H)$ for a set of observations $\boldsymbol{x} = (x_1,\ldots x_n)$.

We observe a single point in this space: $\boldsymbol{x}_{\text{obs}}$.

How can we quantify the level of compatibility between the data and the predictions of H?

Decide what part of the data space represents equal or less compatibility with H than does the point $\boldsymbol{x}_{\text{obs}}$. (Not unique!)



p-values

Express level of compatibility between data and hypothesis (sometimes 'goodness-of-fit') by giving the p-value for H:

$$p = P(\mathbf{x} \in \omega_{\leq}(\mathbf{x}_{\mathrm{obs}})|H)$$

= probability, under assumption of H, to observe data with equal or lesser compatibility with H relative to the data we got.

= probability, under assumption of H, to observe data as discrepant with H as the data we got or more so.

Basic idea: if there is only a very small probability to find data with even worse (or equal) compatibility, then H is "disfavoured by the data".

If the p-value is below a user-defined threshold α (e.g. 0.05) then H is rejected (equivalent to hypothesis test of size α as seen earlier).

p-value of H is not $P(H)$

The p-value of H is not the probability that H is true!

In frequentist statistics we don't talk about $P(H)$ (unless H represents a repeatable observation).

If we do define $P(H)$, e.g., in Bayesian statistics as a degree of belief, then we need to use Bayes' theorem to obtain

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

where $\pi(H)$ is the prior probability for H.

For now stick with the frequentist approach;
result is p-value, regrettably easy to misinterpret as $P(H)$.

The Poisson counting experiment

Suppose we do a counting experiment and observe n events.

Events could be from *signal* process or from *background* – we only count the total number.

Poisson model:

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

s = mean (i.e., expected) # of signal events

b = mean # of background events

Goal is to make inference about s, e.g.,

test $s = 0$ (rejecting $H_0 \approx$ "discovery of signal process")

test all non-zero s (values not rejected = confidence interval)

In both cases need to ask what is relevant alternative hypothesis.

Poisson counting experiment: discovery p-value

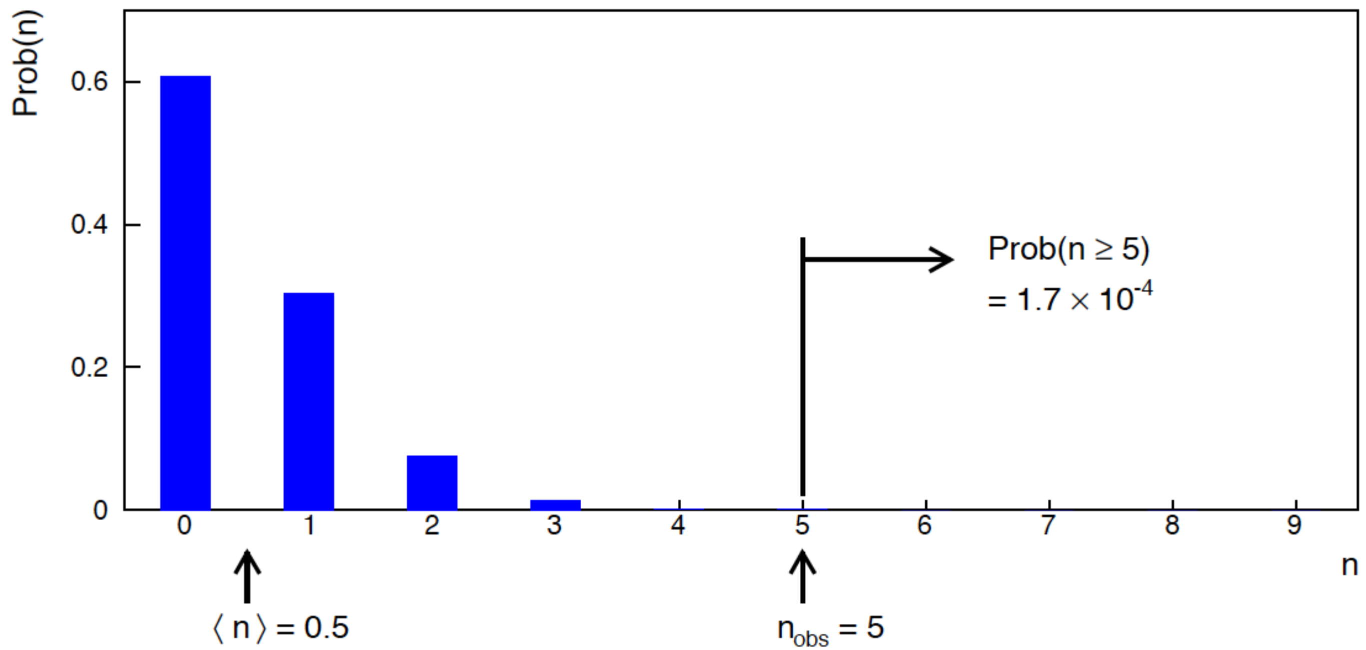
Suppose $b = 0.5$ (known), and we observe $n_{\text{obs}} = 5$.

Should we claim evidence for a new discovery?

Give p-value for hypothesis $s = 0$:

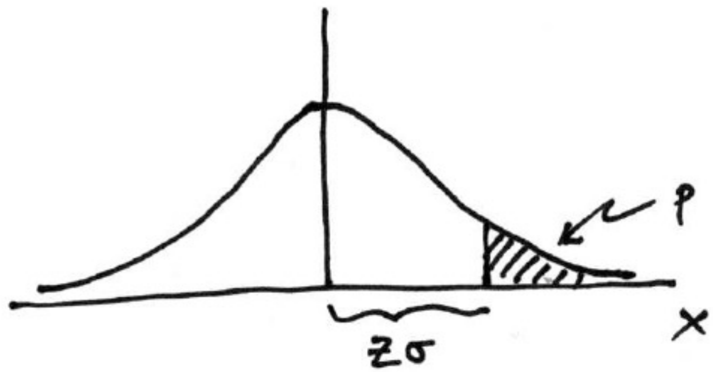
$$p\text{-value} = P(n \geq 5; b = 0.5, s = 0)$$

$$= 1.7 \times 10^{-4} \neq P(s = 0)!$$



Significance from p-value

Often define significance Z as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same p-value.



$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$

$$Z = \Phi^{-1}(1-p)$$

```
in ROOT:
p = 1 - TMath::Freq(Z)
Z = TMath::NormQuantile(1-p)
```

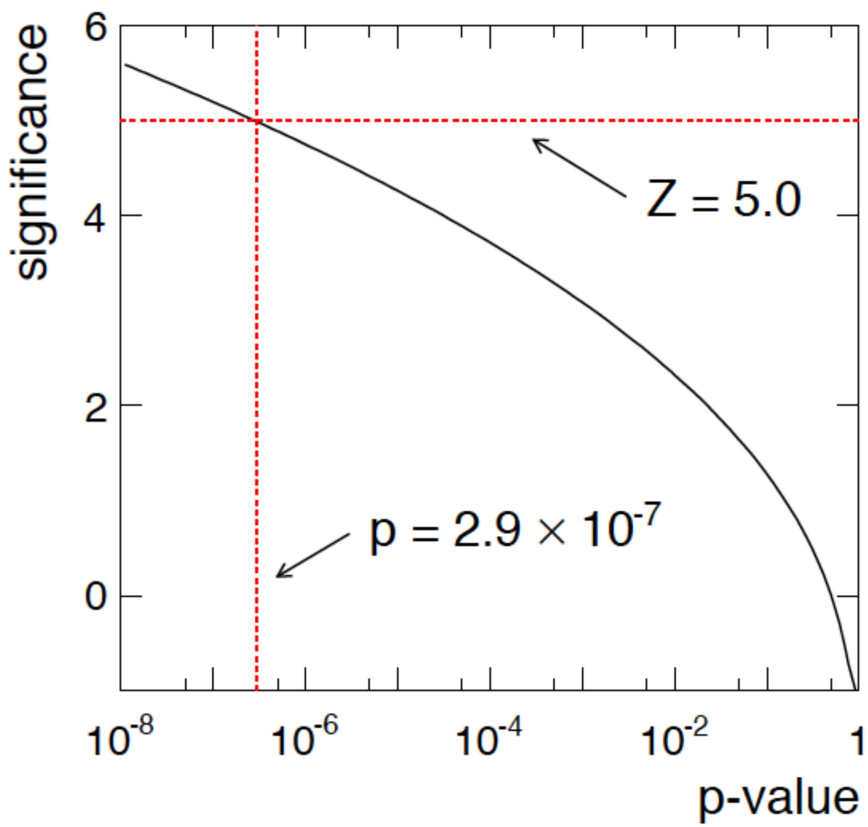
```
in python (scipy.stats):
p = 1 - norm.cdf(Z) = norm.sf(Z)
Z = norm.ppf(1-p)
```

Result Z is a "number of sigmas". Note this does not mean that the original data was Gaussian distributed.

Poisson counting experiment: discovery significance

Equivalent significance for $p = 1.7 \times 10^{-4}$: $Z = \Phi^{-1}(1-p) = 3.6$

Often claim discovery if $Z > 5$ ($p < 2.9 \times 10^{-7}$, i.e., a "5-sigma effect")



In fact this tradition should be revisited: p-value intended to quantify probability of a signal-like fluctuation assuming background only; not intended to cover, e.g., hidden systematics, plausibility signal model, compatibility of data with signal, "look-elsewhere effect" (~multiple testing), etc.

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 $p_\theta \le \alpha$, 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 $\boldsymbol{\theta} = (\theta_1, \ldots \theta_M)$ *use* same idea – result is a confidence "region" with boundary determined by $p_{\boldsymbol{\theta}} = \alpha$.

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 $P(\boldsymbol{x}|H,\theta)$.

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

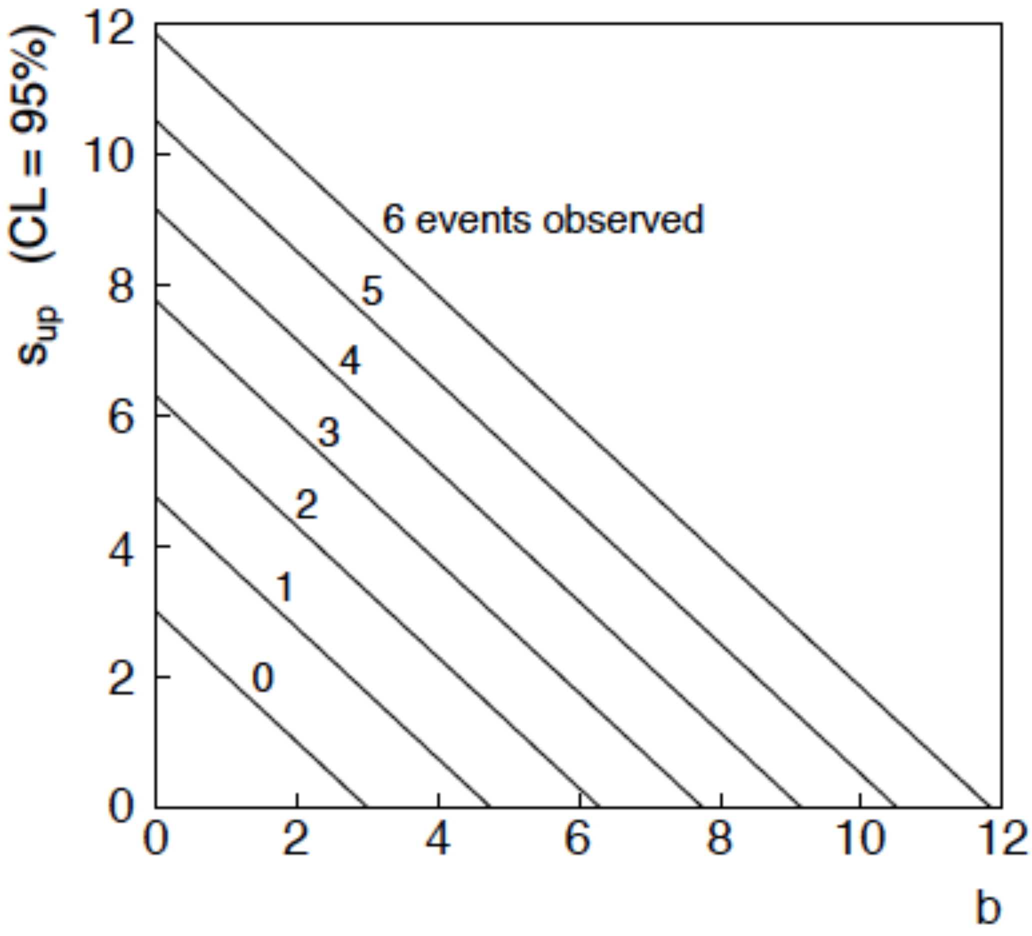
$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}}+b)}$$

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

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

$n \sim$ 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 $s_{\rm up}$

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

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 $s = 0$

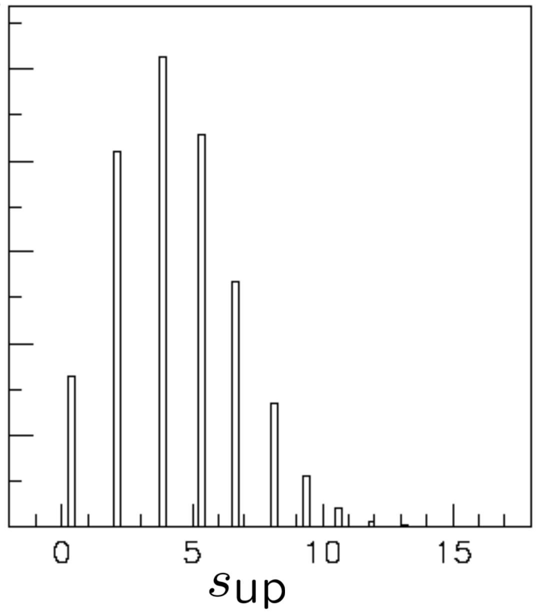
Physicist: I should have used CL = 0.95 — then $s_{\mathrm{up}} = 0.496$

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

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



Extra slides

Approximate confidence intervals/regions from the likelihood function

Suppose we test parameter value(s) $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ using the ratio

$$\lambda(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})} \qquad 0 \le \lambda(\boldsymbol{\theta}) \le 1$$

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

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

so higher t_θ means worse agreement between $\boldsymbol{\theta}$ and the data.

p-value of $\boldsymbol{\theta}$ therefore

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

need pdf

Confidence region from Wilks' theorem

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

$$f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta}) \sim \chi^2_n$$

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

Assuming this holds, the p-value is

$$p_{\boldsymbol{\theta}} = 1 - F_{\chi^2_n}(t_{\boldsymbol{\theta}})$$

← set equal to α

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

$$t_\theta = F^{-1}_{\chi^2_n}(1-\alpha)$$

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 $\boldsymbol{\theta}$ space is where

$$\ln L(\boldsymbol{\theta}) = \ln L(\hat{\boldsymbol{\theta}}) - \tfrac{1}{2} F^{-1}_{\chi^2_n}(1-\alpha)$$

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

$$F^{-1}_{\chi^2_1}(0.683) = 1$$

and so the 68.3% confidence level interval is determined by

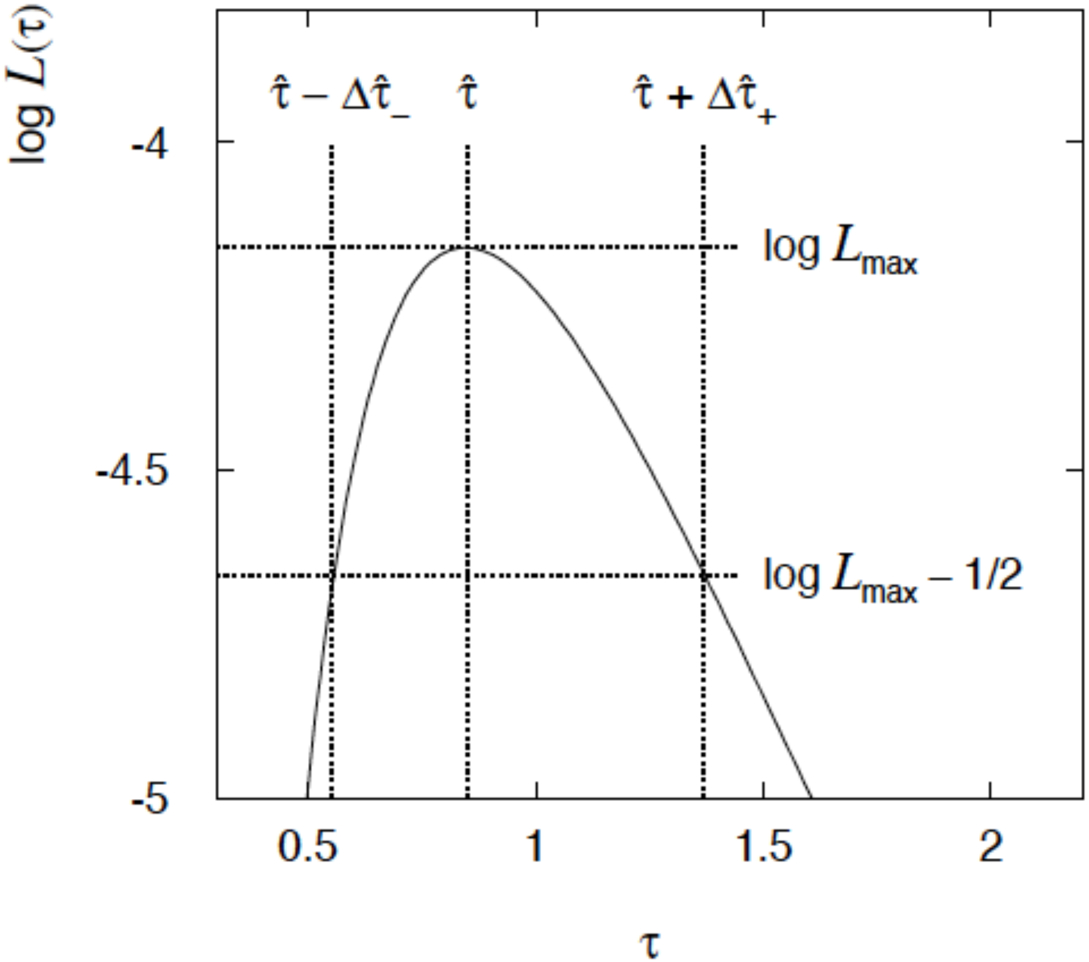
$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

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, 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

$$Q_\alpha = F^{-1}_{\chi^2_n}(1-\alpha)$$

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 $= 1 - \alpha$.

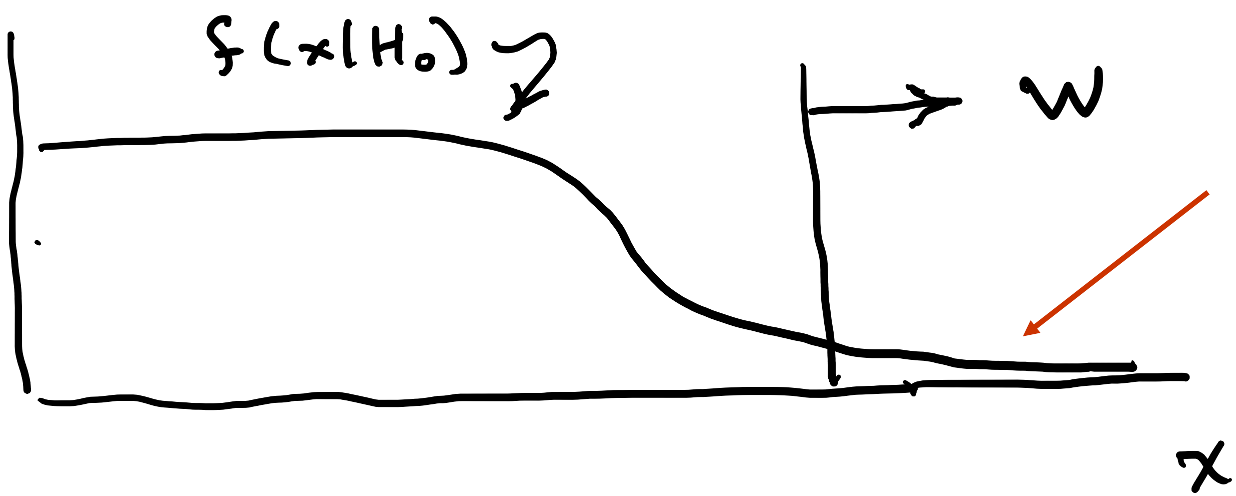
$1-\alpha$	Q_α				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

Obvious where to put W?

In the 1930s there were great debates as to the role of the alternative hypothesis.

Fisher held that one could test a hypothesis H_0 without reference to an alternative.

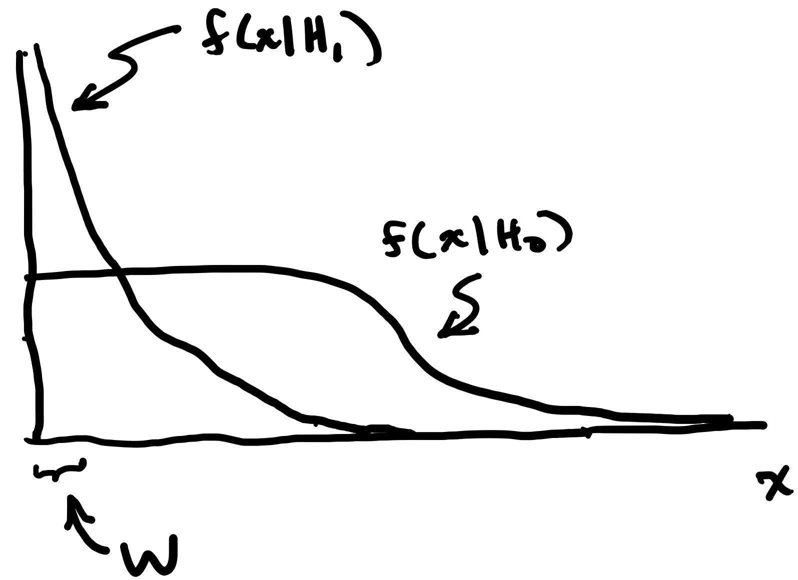
Suppose, e.g., H_0 predicts that x (suppose positive) usually comes out low. High values of x are less characteristic of H_0, so if a high value is observed, we should reject H_0, i.e., we put W at high x:



If we see x here, reject H_0.

Or not so obvious where to put W?

But what if the only relevant alternative to H_0 is H_1 as below:



Here high x is more characteristic of H_0 and not like what we expect from H_1. So better to put W at low x.

Neyman and Pearson argued that "less characteristic of H_0" is well defined only when taken to mean "more characteristic of some relevant alternative H_1".

Type-I, Type-II errors

Rejecting the hypothesis H_0 when it is true is a Type-I error.

The maximum probability for this is the size of the test:

$$P(x \in W \mid H_0) \leq \alpha$$

But we might also accept H_0 when it is false, and an alternative H_1 is true.

This is called a Type-II error, and occurs with probability

$$P(x \in S - W \mid H_1) = \beta$$

One minus this is called the power of the test with respect to the alternative H_1:

$$\text{Power} = 1 - \beta$$

Distribution of the p-value

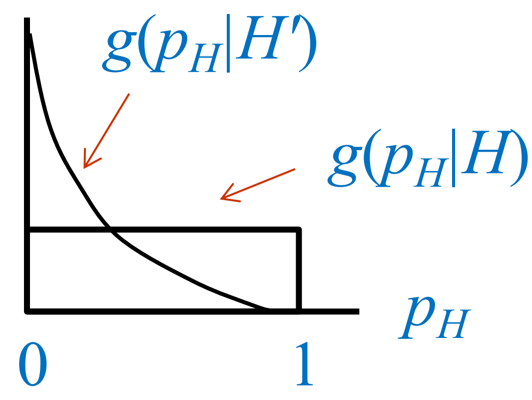
The p-value is a function of the data, and is thus itself a random variable with a given distribution. Suppose the p-value of H is found from a test statistic $t(\mathbf{x})$ as

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

The pdf of p_H under assumption of H is

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

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

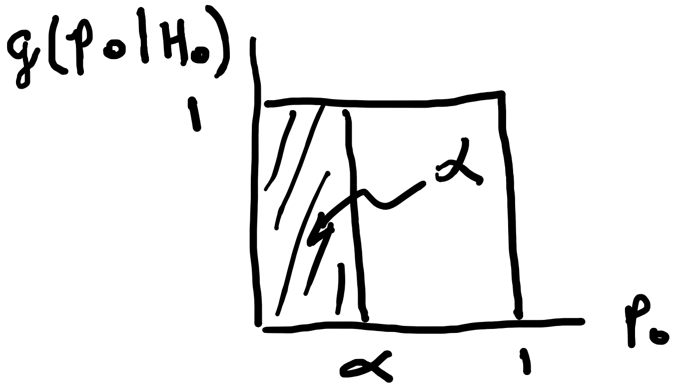


Using a p-value to define test of H_0

One can show that under assumption of a hypothesis H_0, its p-value, p_0, follows a uniform distribution in [0,1].

So the probability to find p_0 less than a given α is

$$P(p_0 \leq \alpha | H_0) = \alpha$$



So we can define the critical region of a test of H_0 with size α as the set of data space where $p_0 \leq \alpha$.

Formally the p-value relates only to H_0, but the resulting test will have a given power with respect to a given alternative H_1.