Statistical Tests

The goal is to make a statement about how well the observed data is in agreement with a given prediction

This prediction is a hypothesis, and we would like to test the hypothesis in consideration, the will hypothesis, under the light of data

↳ H_0

H_0 would refer, e.g. to the concrete specification of a probability distribution

$f(\vec{x}|\theta)$ ← some value

e.g. the hypothesis that some data at the LHC comes from a BSM resonance, with mass M:

$$N(\vec{x}|\mu(M,\dots),\sigma^2)$$

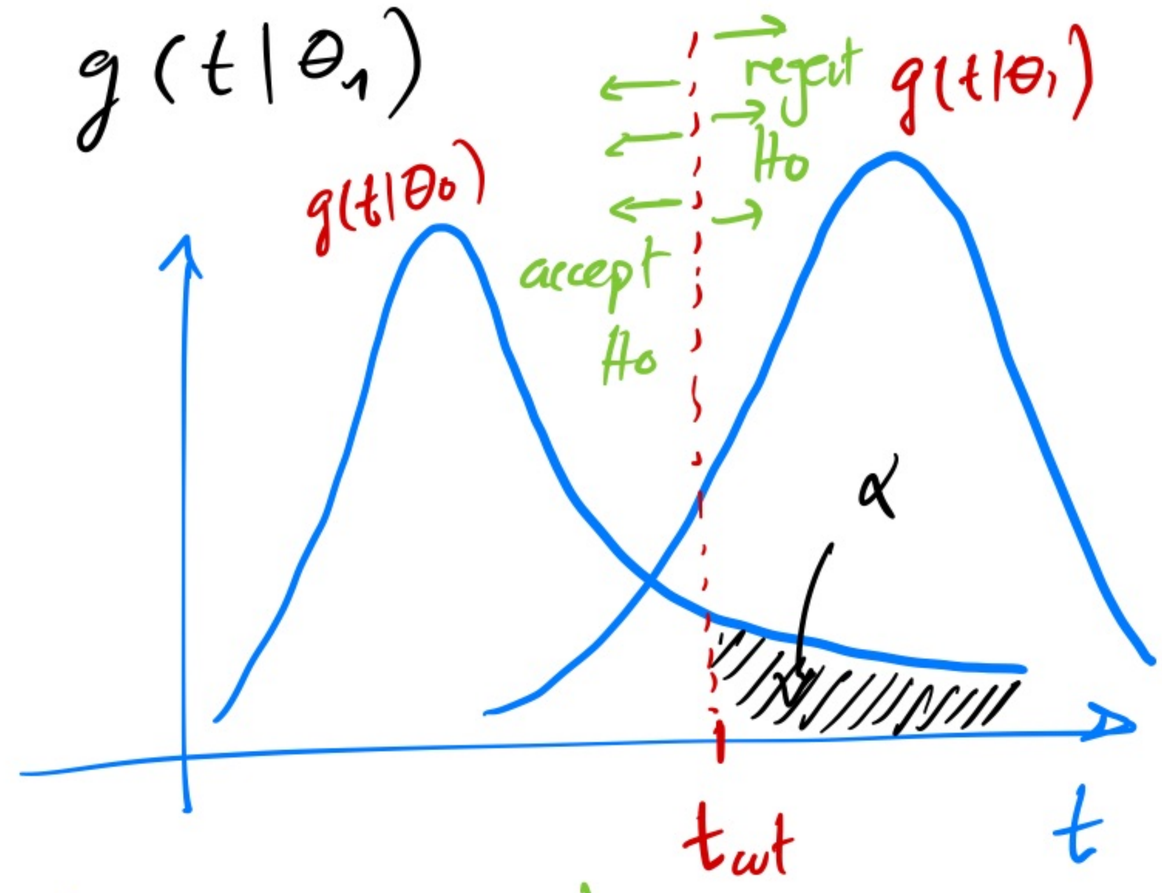
Commonly what we are doing in physics, as well as in statistical inference in general, is to estimate the mean of the distrib of some data

In order to construct a measure of agreement between data and our H_0, we build what it's called a test statistic $t(\vec{x})$, which is a (typically scalar) function of the data, and it attempts to summarise the data somehow

* In many situations we want to discriminate H_0 from an alternative hypothesis, called H_1
(it could be, e.g. another value of the parameter θ determining the distribution)

$t(\vec{x})$ is itself a random variable (since the data $\vec{x}$ is)
thus it will have its own PDF

$$g(t|\theta_0) \quad \text{or} \quad g(t|\theta_1)$$



The significance level of the test is defined as

$$\alpha = \int_{t_{cut}}^{\infty} g(t|\theta_0)\,dt$$

↳ prob. to reject H_0 while H_0 is true ("type I error")

Another complementary quantity is

$$\beta = \int_{-\infty}^{t_{cut}} g(t|\theta_1)\,dt$$

↳ prob. of accepting H_0 while H_0 is false ("type II error")

[Power of the test is: $1-\beta$]

- In the context of physics application. Suppose data is generated by some particle we are interested in

(this is our signal) while other part of data is not (background).

We want to test H_0 to be the presence of signal.

The signal efficiency

$$\varepsilon_{signal} = \int_{-\infty}^{t_{cut}} g(t|\theta_0)dt = 1-\alpha$$

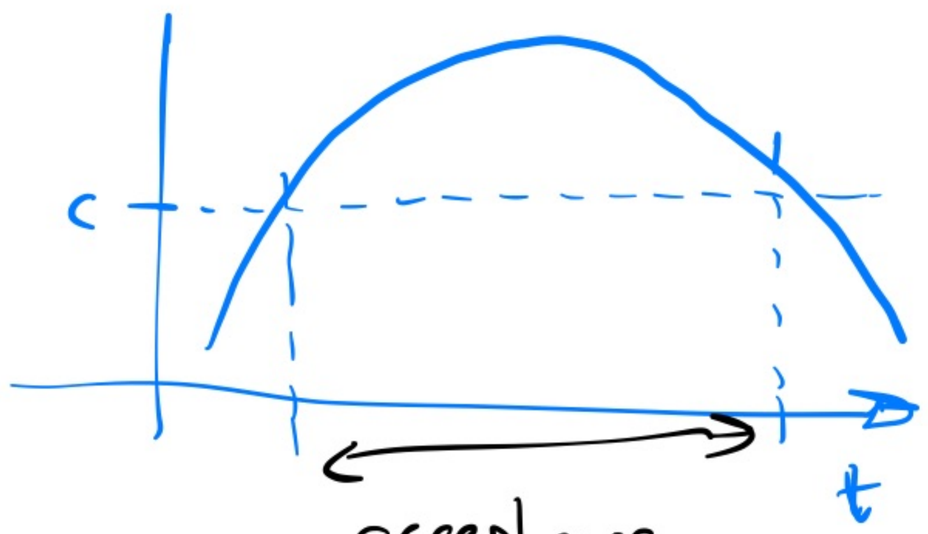
$$\varepsilon_{backg} = \int_{-\infty}^{t_{cut}} g(t|\theta_1)\,dt = \beta$$

Clearly the price to pay for increasing signal efficiency is to increase background efficiency as well

What is desirable, is that for a given α, the power of the test is as large as possible:

⇒ The Neyman–Pearson lemma says that the region (in t-space) giving the highest power is the region where

$$\frac{g(t|H_0)}{g(t|H_1)} > c$$



here c is chosen s.t.

$$\int g(t|\theta_0)\,\Theta\left(\frac{g(t|\theta_0)}{g(t|\theta_1)} - c\right)dt = \alpha$$

↳ step function

Let's see a concrete example of this:

Suppose a single observation x, where x is assumed to be Gaussian, with $\sigma^2 = 1$

$$N(x|\mu, 1)$$

we want to test $H_0 : \mu = 0$ against

$$H_1 : \mu = 2$$

In practice the NP lemma is implemented with the likelihood ratio

$$r = \frac{N(x|0,1)}{N(x|2,1)} > c = x_{cut}$$

Let's choose c s.t. $\alpha = 0.05$

$$\alpha = 0.05 = \int_{x_{cut}}^{\infty} N(x|0,1)\,dx$$

Remember

$$\Phi = \int_{-\infty}^{x_\alpha} N(x|\mu,\sigma^2)\,dx = \alpha$$

$\Phi(x_{cut}) = 0.95$ for a Gaussian

$x_{cut} \sim 1.64$

$$r = e^{-\frac{1}{2}x^2 + \frac{1}{2}(x-2)^2} = e^{-2x+2} \geq 1.64$$

$$-2x+2 \geq \ln 1.64$$

$$x \leq 1 - \frac{1}{2}\ln 1.64 = 0.75$$

$\therefore$ $x \leq 0.75$ → acceptance region

In the general case of many observations

$$\{\vec{x}_i\} \quad i = 1, \dots, N$$

The likelihood of the data under the H_0 for the above example is

$$\mathcal{L}_{H_0} = \prod_{i=1}^{N} N(\vec{x}_i \mid 0, 1)$$

and it is convenient to work with the log likelihood ratio

$$TS = \ln \frac{\mathcal{L}_{H_0}}{\mathcal{L}_{H_1}}$$

one of the most used quantities in statistics !

Note, it remains the problem of knowing the likelihoods, in order to build this TS.

However in many situations the likelihood may be very complicated to extract, e.g. one would need to construct n-dimensional histograms (here n is the dimension of $\vec{x}$) and it may be computationally prohibitive

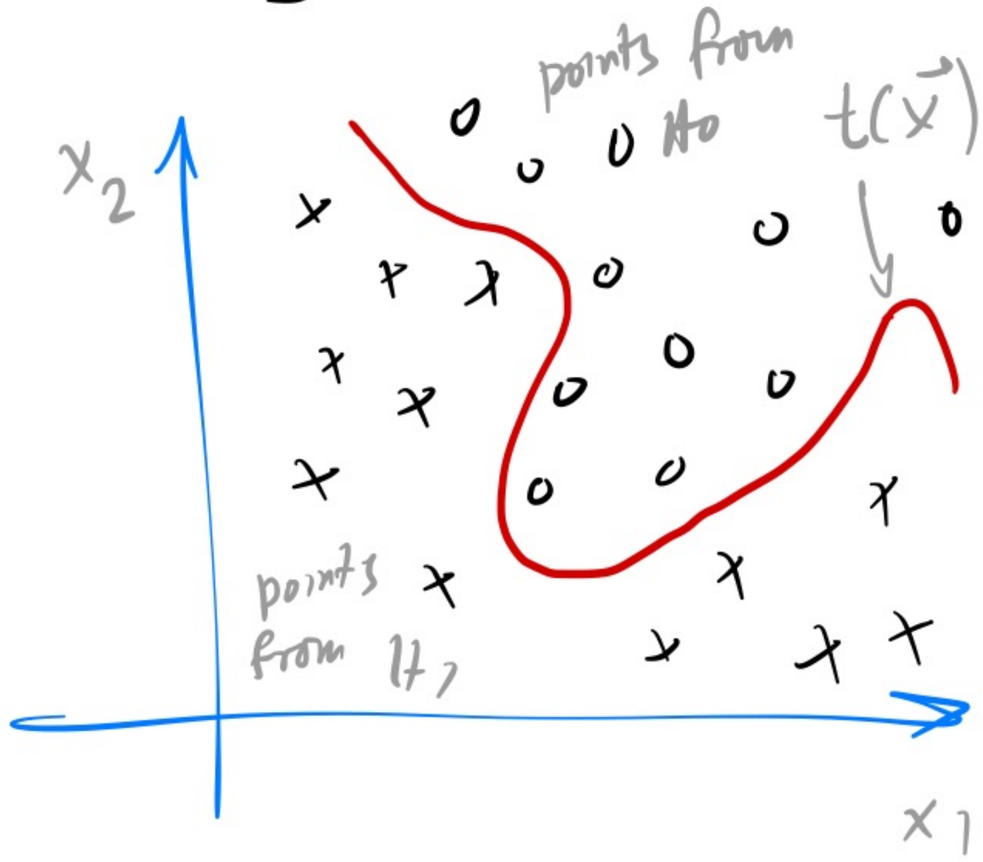
⇒ How to proceed then ?

↳ Approximate $t(\vec{x}) = \ln \frac{\mathcal{L}(\vec{x} \mid H_0)}{\mathcal{L}(\vec{x} \mid H_1)}$

by a function of $\vec{x}$, and fit its parameters such as to maximise the separation between the $g(t \mid H_0)$ and $g(t \mid H_1)$

The above can be rephrased as a Supervised Learning problem in ML, where the idea is to classify events (points) coming from either H_0 or H_1

e.g. points above this line would lead us to accept the H_0



* Goodness-of-fit tests

In some situations one may be just interested in how well a given hypothesis is compatible with data, without comparing to an alternative H_1

In these cases one can use the "p-value"

p-value : the prob., under H_0, of obtaining a result as compatible or less than the one actually observed

E.g. Suppose we have an experiment and observe some # of counts n_{obs}.

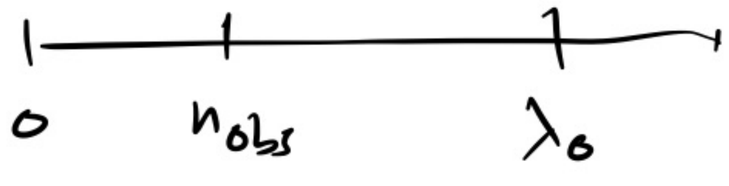
what is the goodness-of-fit of H_0 : $\lambda = \lambda_0$ where λ is the parameter of the Poisson?

↳ imagine λ depends on your BSM parameters

Then it depends:

a) 

$$\text{p-value} = \sum_{n=n_{obs}}^{\infty} \text{Pois}(n|\lambda_0)$$

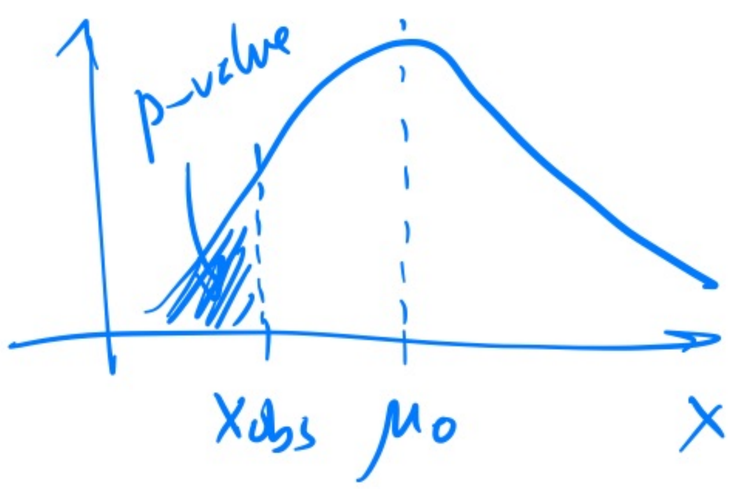
b) 

$$\text{p-value} = \sum_{n=0}^{n_{obs}} \text{Pois}(n|\lambda_0)$$

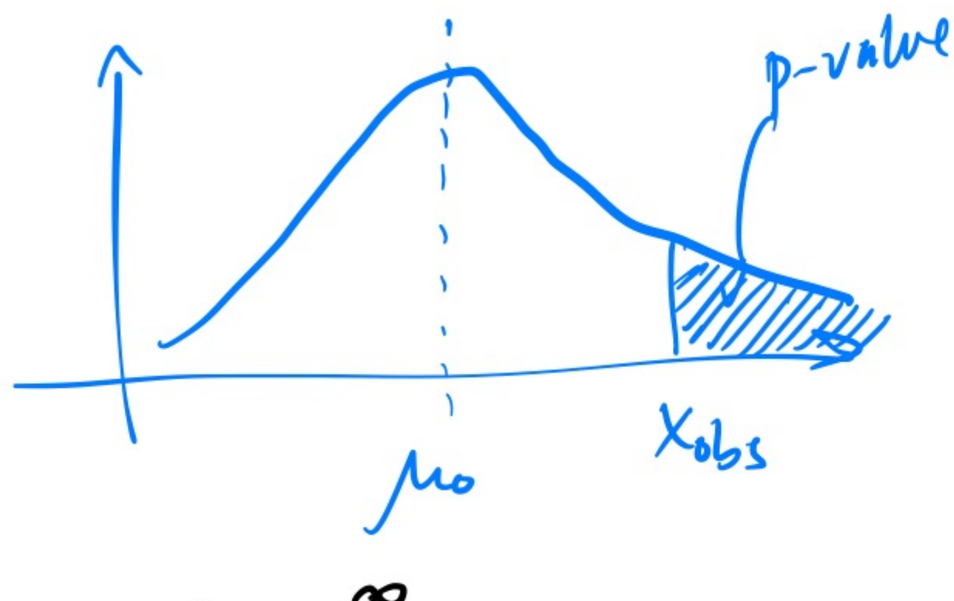
eg. $n_{obs} = 7$, $\lambda_0 = 3$

$$\text{p-value} = \sum_{n=7}^{\infty} \text{Pois}(n|3) \approx 0.033$$

- Analogously for a Gaussian variable X:



$$\text{p-val} = \int_{-\infty}^{x_{obs}} N(x|\mu_0, \sigma^2)\, dx$$



$$\text{p-val} = \int_{x_{obs}}^{\infty} N(x|\mu_0, \sigma^2)\, dx$$

- Arguably the most popular GoF test is the χ^2 test.

Suppose having a counting experiment (or register events) and for each event one measures a variable x (e.g. the angle in some detector)

Typically one then proceeds by constructing a histogram of X

in each bin one has n_i events

Assuming all measured values are independent, then

$$n_i \sim \text{Pois}(n_i \mid \lambda_i)$$

↳ expected # of entries

Then the following TS:

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \lambda_i)^2}{\lambda_i}$$

{ for a Poisson variable, this is the squared of the deviations, in units of the stds

follows, in the large sample limit ($n_i \geq 5 \ \forall i$) a χ^2-distrib:

$$p(z \mid k) = \frac{1}{2^{k/2}\,\Gamma(k/2)}\, z^{k/2-1}\, e^{-z/2}$$

k: D.O.F. $\quad k = 1, 2, \ldots$

$z \in \mathbb{R} \quad 0 \leq z < \infty$

↳ i.e. if one were to repeat the experiment many times, the value χ^2 would be distributed in this way

(regardless of the distrib. of the variable x)

Here DOF = # bins { shall the λ_i depend on some parameters we are interested in, this would be different

with this,

$$\text{p-val} = \int_{\chi^2}^{\infty} p(z \mid k)\, dz$$

{ This can be checked in tables

corresponding to $H_0: \{\lambda_i\}$

An issue with the χ^2-test is that the result will depend in general on the binning (particularly when the data samples are small)

The bin size should be large enough as for containing $n_i \geq 5$, while if too large, it will clearly throw away information about the distribution of x.

- An alternative GoF test, which doesn't require binning is the Kolmogorov–Smirnov test:

Assume having N observations $\{x_i\}_{i=1}^{N}$

If we sort them in increasing order and construct

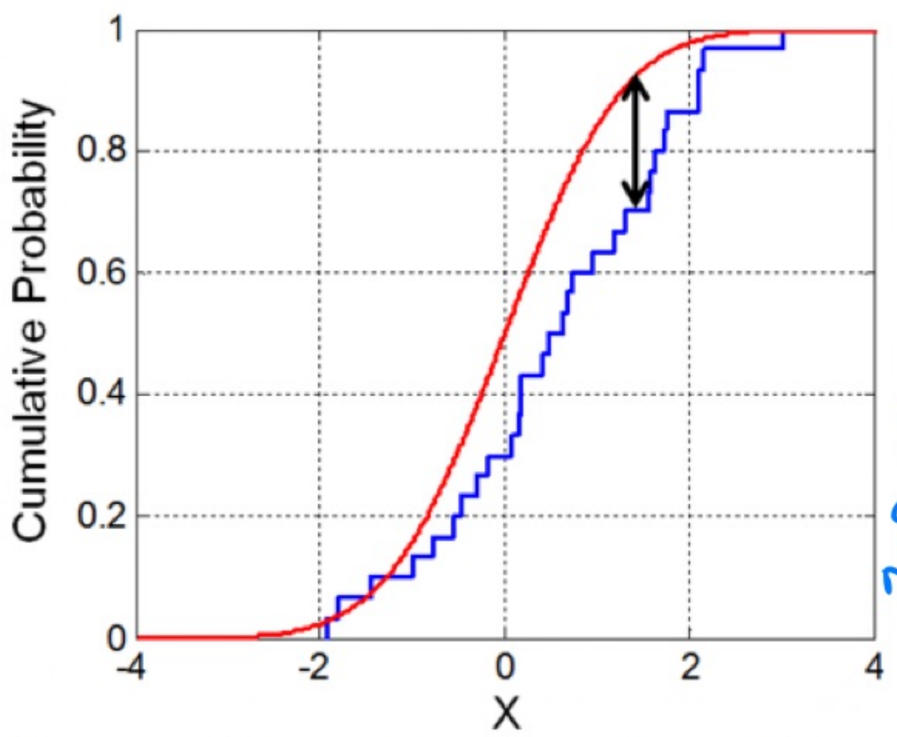
$$F_N(x) = \frac{1}{N} \sum_{i=1}^{N} \Theta(x - x_i)$$

↳ "empirical distrib. function" (cumulative)

gives the fraction of observations where $x_i \leq x$

Then the KS statistic for a reference cumulative $F(x)$ is:

$$D_N = \sup_x \left| F_N(x) - F(x) \right|$$



$D_N\sqrt{N}$ follows a Kolmogorov distrib in the large N limit

⇒ look up at tables: if value is too large H_0 can be rejected