Frequent problem: Decision making based on statistical information

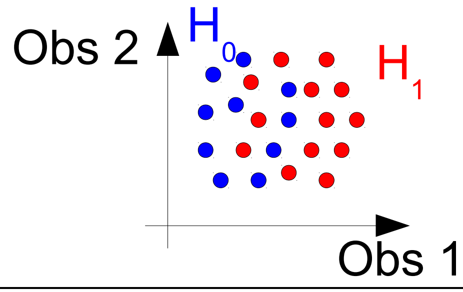
Examples:

a) Distinguishing different particle species in detector, e. g. e – π separation

Different signals in detector material :

Observables
- specific energy loss in tracking detector
- E_{ECAL} / E_{HCAL} ratio
- shower shapes in calorimeter
- …

P.d.f.s of observables are often different for different hypotheses but they overlap. In this case, the aim is: large efficiency, small background



b) Search for new particles

statistical separation of similar events: background and new particles (signal)

Concepts

Problem: How well do observed data agree with predicted probabilities (hypotheses)?

The hypothesis under consideration is traditionally called null hypothesis H_0.
Often comparison with alternative hypotheses H_1, H_2,…

n measured values $\vec{x} = (x_1, \ldots, x_n)$

Each hypothesis characterised through p.d.f. $f(\vec{x}|H_0), f(\vec{x}|H_1), f(\vec{x}|H_2), \ldots$

$(x_1, \ldots, x_n)$ could be :

- n measurements of the same random variable (n "events")
- n different observables of an event (E, p, …)
- Combination of both
- Same measurement from n different experiments

To test agreement between data and given hypothesis, one constructs a function

$$\vec{t}(\vec{x}) \qquad \vec{t} = (t_1, \ldots, t_m)$$

of $(x_1, \ldots, x_n)$ called test statistic (usually $t(\vec{x})$ is scalar function)

For each of the hypotheses, there is a p.d.f. for the statistic t: $g(t|H_0)$, $g(t|H_1)$, etc

Decision to accept or reject hypothesis H_0 by defining:

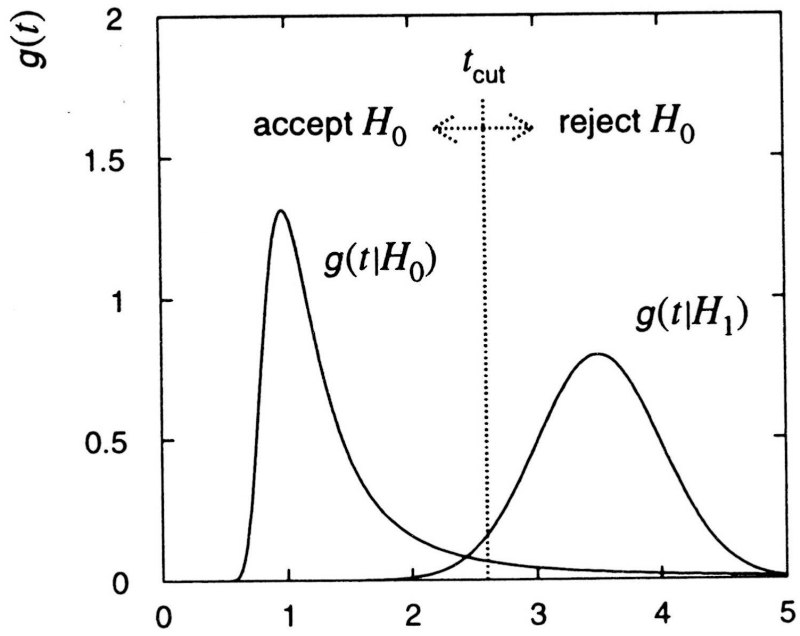
critical region (reject H_0)

acceptance region (accept H_0)

e. g. by defining a cut value t_{cut} :

$t < t_{cut}$ "accepted"

$t > t_{cut}$ "rejected"



Selecting t_{cut} (decision boundary), one defines significance level of the test:

$$\alpha = \int_{t_{cut}}^{\infty} g(t|H_0)\,dt$$

meaning that there is the probability α to reject the hypothesis H_0, if H_0 is true – called an "error of the first kind"

Probability to accept H_0 if an alternative hypothesis H_1 is true is given by :

$$\beta = \int_{-\infty}^{t_{cut}} g(t|H_1)\,dt$$

= "error of the second kind"

$1 - \beta$ is called the power of the test to discriminate against the alternative hypothesis H_1

Example 1: Particle identification

test statistic = measured specific energy loss (for fixed momentum)

e. g. H_0 = e; H_1 = π

Assumption: Only electrons and pions are in sample

Task: Select a sample of electrons ("signal") by requiring $t < t_{cut}$

Selected pions are background. Probabilities to select electrons and pions ("efficiencies") are :

$$\varepsilon_e = \int_{-\infty}^{t_{cut}} g(t|e)\,dt = 1 - \alpha \qquad \varepsilon_\pi = \int_{-\infty}^{t_{cut}} g(t|\pi)\,dt = \beta$$

How to choose t_{cut}?

large t_{cut}: large signal efficiency, much background

small t_{cut}: small signal efficiency, little background (= large purity)

Number of accepted particles:

$$N_{acc} = \varepsilon_e N_e + \varepsilon_\pi N_\pi = \varepsilon_e N_e + \varepsilon_\pi (N_{tot} - N_e)$$

$$\Rightarrow N_e = \frac{N_{acc} - \varepsilon_\pi N_{tot}}{\varepsilon_e - \varepsilon_\pi}$$ (can only be used if $\varepsilon_e \neq \varepsilon_\pi$)

Probability that a particle with observed value of test statistic t is an electron:

$$h(e|t) = \frac{a_e g(t|e)}{a_e g(t|e) + a_\pi g(t|\pi)}$$

a_e, $a_\pi = 1 - a_e$ are prior probabilities for electrons and pions, respectively

→ must be known (for example from MC simulation)

Purity :

$$p_e = \frac{N_{\text{electrons}} \text{ with } t < t_{\text{cut}}}{N_{\text{all part.}} \text{ with } t < t_{\text{cut}}} = \frac{\int_{-\infty}^{t_{cut}} a_e g(t|e)\,dt}{\int_{-\infty}^{t_{cut}} a_e g(t|e)\,dt + \int_{-\infty}^{t_{cut}} (1-a_e) g(t|\pi)\,dt} = \frac{a_e \varepsilon_e N_{tot}}{N_{acc}}$$

Example 2: Counting experiment

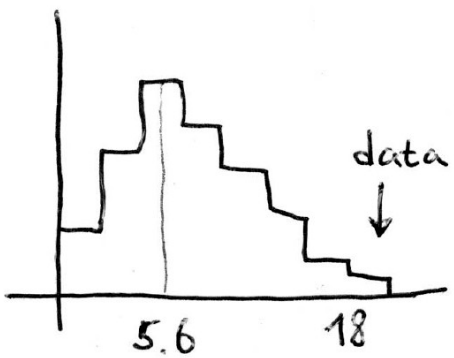
An experiment counts events of a certain type. Events comprise (on average) ν_B background and ν_S signal events

The observed number of events is n

Test statistic: Poisson probability $$f(n;\nu_S,\nu_B)=\frac{(\nu_S+\nu_B)^n}{n!}e^{-(\nu_S+\nu_B)}$$

null hypothesis : $f(n|H_0)=f(n;0,\nu_B)$

Example: $\nu_B = 5.6$, $n = 18$



Question: Probability to observe 18 or more events if one expects on average 5.6 and H_0 is true

$$p-value(CL_B)=\sum_{n=n_{data}}^{\infty} f(n;0,\nu_B)=1-\sum_{n=0}^{n_{data}-1} f(n;0,\nu_B)=2.4\cdot 10^{-5}$$

p-value is often expressed in terms of equivalent (Gaussian) std. deviations

$$P(1\sigma)=1-\int_{-1}^{1}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=0.3173$$

$$P(2\sigma)=1-\int_{-2}^{2}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=0.0455$$

$$P(3\sigma)=\ldots=2.7\cdot 10^{-3}$$

$$P(4\sigma)=\ldots=6.3\cdot 10^{-5}$$

$$P(5\sigma)=\ldots=5.7\cdot 10^{-7}$$

Convention: if p-value (CL_B) < "5σ" ($5.7\cdot 10^{-7}$), then the background hypothesis is rejected ("discovery")

Note:

- That does not mean that one detects a signal with 1-p=0.99…
  (no statement is made about H_1)
- Requires exact understanding of expected background

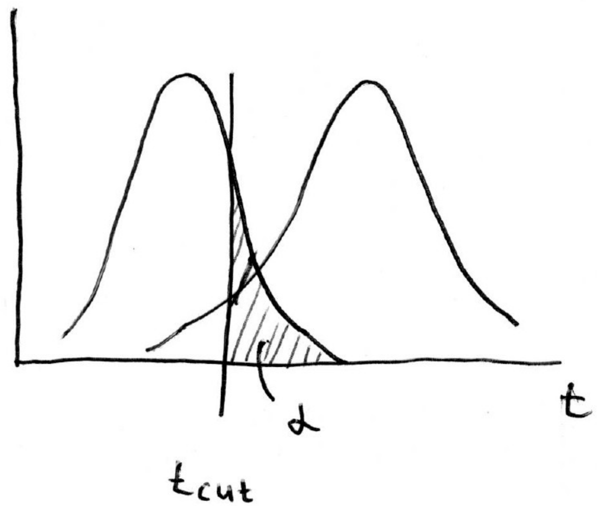
Choice of the critical region

How does one choose a critical region of a statistical test in an optimal way?

If one defines a significance level *α* for a null hypothesis, one wants the largest possible power $1 - \beta$.

In other words: For a fixed efficiency $\varepsilon = 1 - \alpha$, we want the largest possible purity

For a one dimensional test statistic:



trivial, if $\frac{\rho(t|H_0)}{\rho(t|H_1)}$ is monotonic function

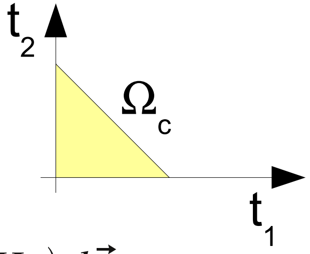
In that case the value of t_{cut} fixes $1 - \beta$

More difficult:

a) more complicated p.d.f. of test statistic

a simple cut might not be optimal

b) multidimensional test statistic



$$\frac{1-\beta}{\alpha}=\frac{\int_{\Omega_c} g(\vec{t}|H_1)\,d\vec{t}}{\int_{\Omega_c} g(\vec{t}|H_0)\,d\vec{t}}=\frac{\int_{\Omega_c}\frac{g(\vec{t}|H_1)\,d\vec{t}}{g(\vec{t}|H_0)\,d\vec{t}}\,g(\vec{t}|H_0)\,d\vec{t}}{\int_{\Omega_c} g(\vec{t}|H_0)\,d\vec{t}}=E\left(\frac{g(\vec{t}|H_1)}{g(\vec{t}|H_0)}\right)_{|\Omega_c}$$

→ $1 - \beta$ will be maximal, if one selects the region Ω_c such that the ratio $\frac{g(\vec{t}|H_1)}{g(\vec{t}|H_0)}$ is maximised.

Equivalently: Select acceptance region $1 - \Omega_c$ such that $\frac{g(\vec{t}|H_0)}{g(\vec{t}|H_1)}$ is minimal

Neyman-Pearson lemma:

The acceptance region giving the highest power (and signal purity) for a given significance level α (or selection efficiency $\varepsilon = 1 - \alpha$) is the region in t-space such that:

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

where t_c is determined by the desired efficiency

The quantity

$$r = \frac{g(\vec{t}|H_0)}{g(\vec{t}|H_1)}$$

is called likelihood ratio

→ reduction of n-dimensional test statistic $\vec{t}$ to one-dim. statistic *r*

<u>Important property of LR</u>:

n independent measurements with test statistics $t_1, \ldots, t_n$:

$$g(\vec{t}|H_0) = g_1(t_1|H_0) \ldots g_n(t_n|H_0)$$

and

$$g(\vec{t}|H_1) = g_1(t_1|H_1) \ldots g_n(t_n|H_1)$$

→ $$r = \frac{g(\vec{t}|H_0)}{g(\vec{t}|H_1)} = \frac{g_1(t_1|H_0)}{g_1(t_1|H_1)} \cdot \ldots \cdot \frac{g_n(t_n|H_0)}{g_n(t_n|H_1)} = r_1 \cdot \ldots \cdot r_n$$

<u>Example: Counting experiment</u>

$$r_i = \frac{g(N_i|s+b)}{g(N_i|b)} = \frac{\frac{(s_i+b_i)^{N_i}}{N_i!} \exp(-(s_i+b_i))}{\frac{b_i^{N_i}}{N_i!} \exp(-b_i)}$$

$$= \left(\frac{s_i+b_i}{b_i}\right)^{N_i} \exp(-s_i)$$

Combination of many counting experiments with different purities:

$$r_{tot} = \exp(-s_{tot}) \prod_{i=1}^{n} \left(\frac{s_i + b_i}{b_i} \right)^{N_i} \qquad (s_{tot} = \sum s_i)$$

or

$$\log r_{tot} = -s_{tot} + \sum_{i=1}^{n} N_i \log\left(1 + \frac{s_i}{b_i}\right)$$

= weighted sum of events