Numerical Methods
Lesson 8

Dr. Jose Feliciano Benitez

Universidad de Sonora

Dr. Benitez Homepage: www.jfbenitez.science

Course page: http://jfbenitez.ddns.net:8080/Courses/MetodosNumericos

Goals for this Lesson

Statistical tests

In this chapter some basic concepts of statistical test theory are presented. As this is a broad topic, after a general introduction we will limit the discussion to several aspects that are most relevant to particle physics. Here one could be interested, for example, in the particles resulting from an interaction (an event), or one might consider an individual particle within an event. An immediate application of statistical tests in this context is the selection of candidate particles or events which are then used for further analysis. Here one is concerned with distinguishing events of interest (signal) from other types (background). These questions are addressed in Sections 4.2–4.4. Another important aspect of statistical tests concerns goodness-of-fit; this is discussed in Sections 4.5–4.7.

4.1 Hypotheses, test statistics, significance level, power

A statement about the validity of H_0 often involves a comparison with some **alternative hypotheses**, $H_1, H_2, \ldots$. Suppose one has data consisting of n measured values $\mathbf{x} = (x_1, \ldots, x_n)$, and a set of hypotheses, $H_0, H_1, \ldots$, each of which specifies a given joint p.d.f., $f(\mathbf{x}|H_0)$, $f(\mathbf{x}|H_1), \ldots$.[1] The values could, for example, represent n repeated observations of the same random variable, or a single observation of an n-dimensional variable. In order to investigate the measure of agreement between the observed data and a given hypothesis, one constructs a function of the measured variables called a **test statistic** $t(\mathbf{x})$. Each of the hypotheses will imply a given p.d.f. for the statistic t, i.e. $g(t|H_0)$, $g(t|H_1)$, etc.

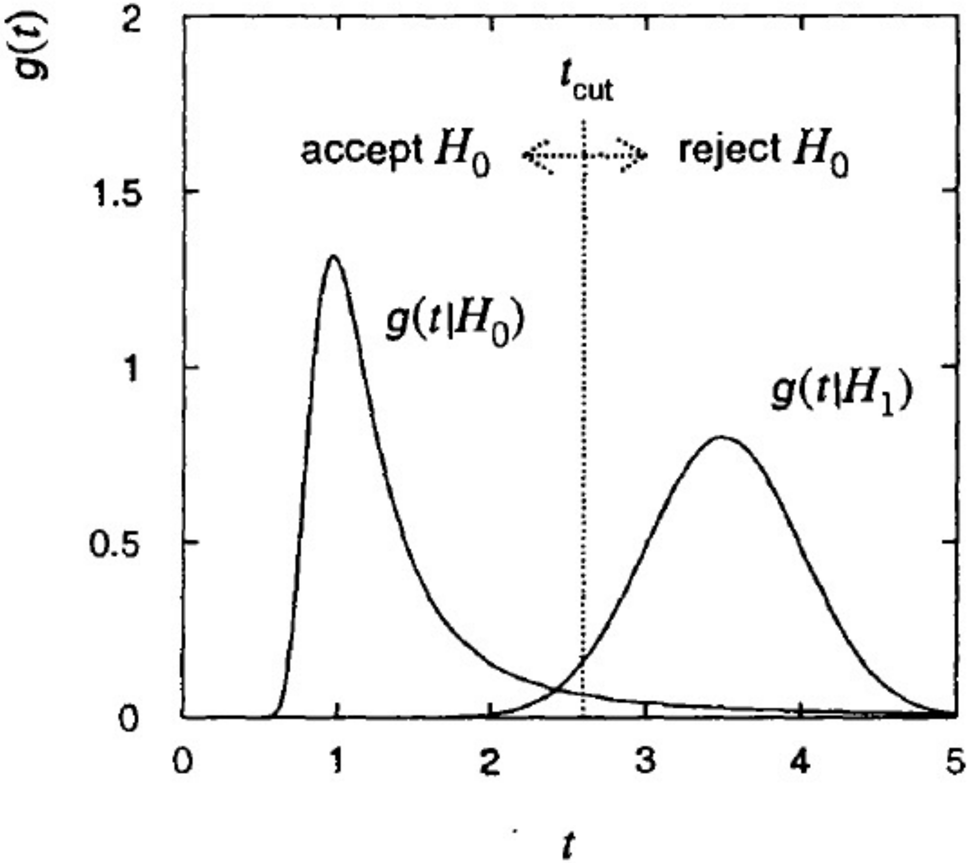


Fig. 4.1 Probability densities for the test statistic t under assumption of the hypotheses H_0 and H_1. H_0 is rejected if t is observed in the critical region, here shown as $t > t_{cut}$.

Often one formulates the statement about the compatibility between the data and the various hypotheses in terms of a decision to accept or reject a given null hypothesis H_0. This is done by defining **a critical region** for t. Equivalently, one can use its complement, called the **acceptance region**. If the value of t actually observed is in the critical region, one rejects the hypothesis H_0; otherwise, H_0 is accepted. The critical region is chosen such that the probability for t to be observed there, under assumption of the hypothesis H_0, is some value α, called the **significance level** of the test. For example, the critical region could consist of values of t greater than a certain value t_{cut}, called the **cut or decision boundary**, as shown in Fig. 4.1. The significance level is then

$$\alpha = \int_{t_{\mathrm{cut}}}^{\infty} g(t|H_0)\,dt. \tag{4.1}$$

One would then accept (or, strictly speaking, not reject) the hypothesis H_0 if the value of t observed is less than t_{cut}. There is thus a probability of α to reject H_0 if H_0 is true. This is called an **error of the first kind**. An **error of the second kind** takes place if the hypothesis H_0 is accepted (i.e. t is observed less than t_{cut}) but the true hypothesis was not H_0 but rather some alternative hypothesis H_1.

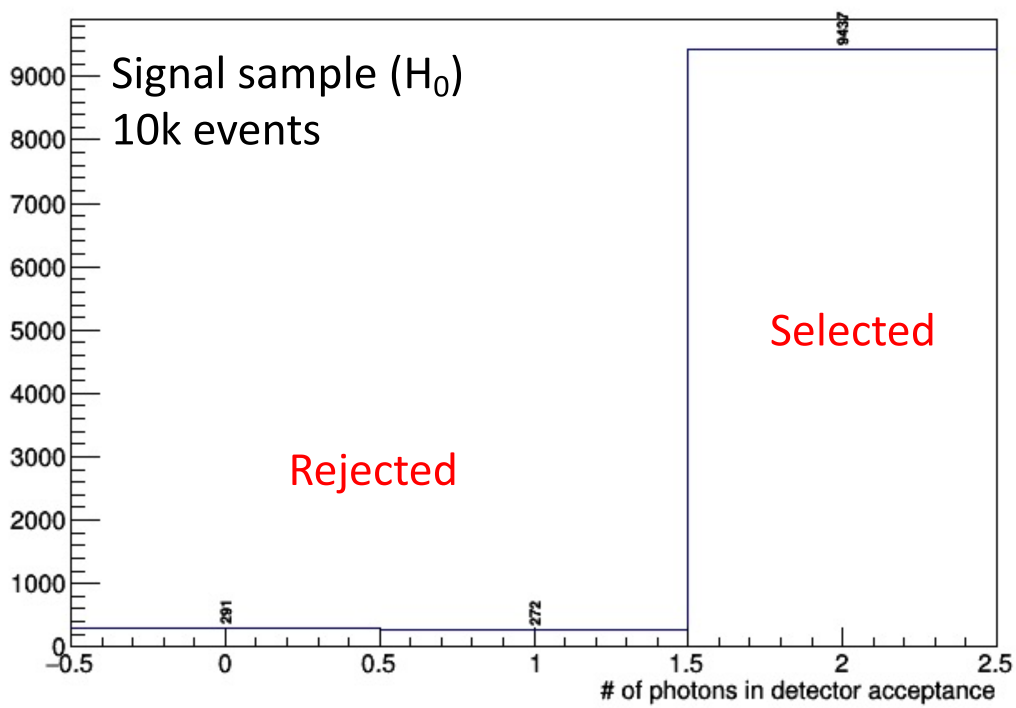
The probability for this is

$$\beta = \int_{-\infty}^{t_{\text{cut}}} g(t|H_1)dt. \tag{4.2}$$

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

Example

- Higgs to diphoton search experiment (Lesson 7)
- Number of photons in the detector acceptance (histogram below) is a *test statistic* which is used to categorize (select) the events.
- The *acceptance* region is the value 2 and the *critical* region are values 0 and 1.

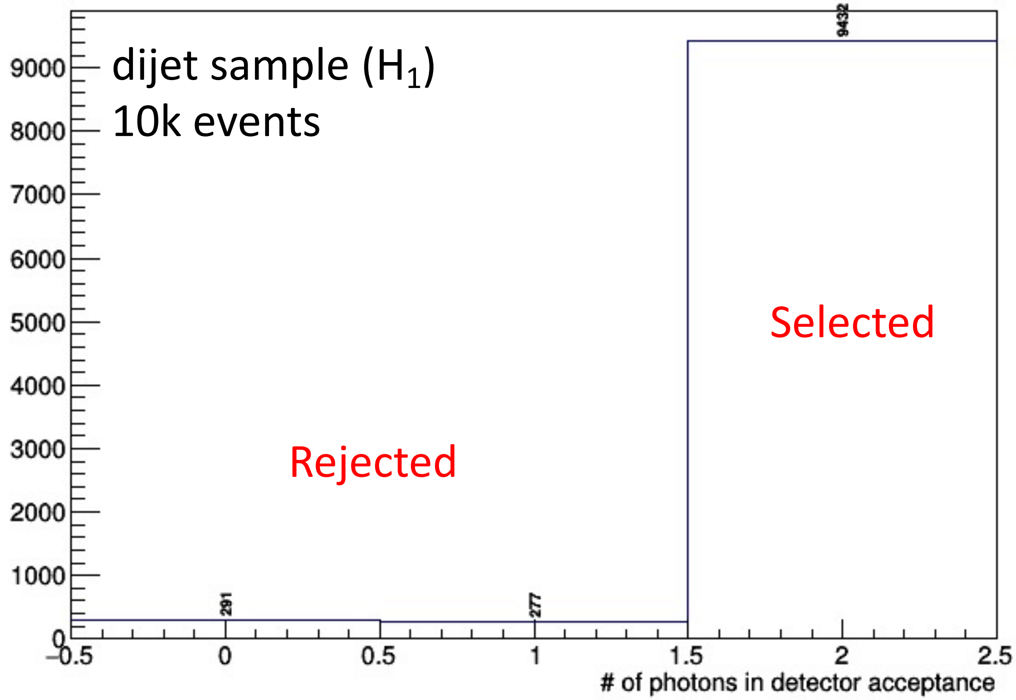


From the numbers in this histogram:
Significance level (alpha) = (291+272)/10000 = 5.63%

Error of first kind = alpha

Example continued

- For the same experiment in the previous slide, use the dijet (background) sample to study *error of the second kind*: number selected events in background.



From the numbers in this histogram:
Error of second kind
= 9432 /10000 = 94.32%

Power = 1 – 0.9432 = 5.68% is low

Note: in this example this variable is not a good one to discriminate events against background because the error of the second kind is large.

4.2 An example with particle selection

As an example, the test statistic t could represent the measured ionization created by a charged particle of a known momentum traversing a detector. The amount of ionization is subject to fluctuations from particle to particle, and depends (for a fixed momentum) on the particle's mass. Thus the p.d.f. $g(t|H_0)$ in Fig. 4.1 could correspond to the hypothesis that the particle is an electron, and the $g(t|H_1)$ could be what one would obtain if the particle was a pion, i.e. $H_0 = \mathrm{e}$, $H_1 = \pi$.

4.3 Choice of the critical region using the Neyman–Pearson lemma

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

The ***likelihood ratio*** to determine the acceptance region

4.4.1 Linear test statistics, the Fisher discriminant function

The simplest form for the statistic $t(\mathbf{x})$ is a linear function,

$$t(\mathbf{x}) = \sum_{i=1}^{n} a_i x_i = \mathbf{a}^T \mathbf{x},$$

Neural networks

4.4.2 Nonlinear test statistics, neural networks

If the joint p.d.f.s $f(\mathbf{x}|H_0)$ and $f(\mathbf{x}|H_1)$ are not Gaussian or if they do not have a common covariance matrix, then the Fisher discriminant no longer has the optimal properties seen above. One can then try a more general parametrization for

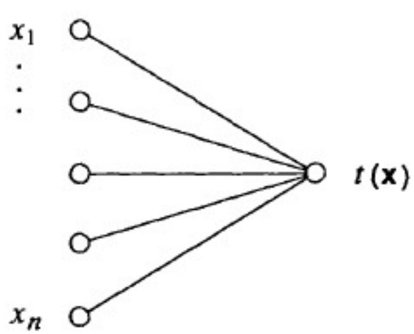
double layer

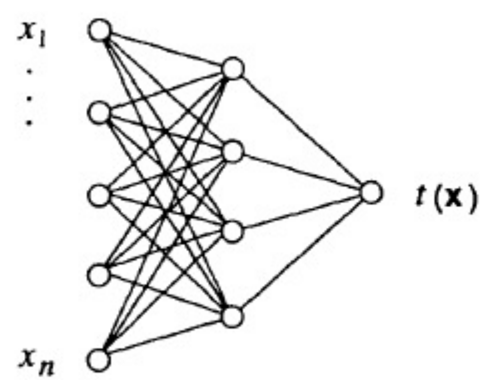
$$h_i(\mathbf{x}) = s\left(w_{i0} + \sum_{i=1}^{n} w_{ij} x_j\right).$$

$$t(\mathbf{x}) = s\left(a_0 + \sum_{i=1}^{m} a_i h_i(\mathbf{x})\right).$$

Single layer

$$t(\mathbf{x}) = s\left(a_0 + \sum_{i=1}^{n} a_i x_i\right).$$





4.5 Goodness-of-fit tests

Frequently one wants to give a measure of how well a given null hypothesis H_0 is compatible with the observed data without specific reference to any alternative hypothesis. This is called a test of the **goodness-of-fit**, and can be done by constructing a test statistic whose value itself reflects the level of agreement between the observed measurements and the predictions of H_0. Procedures for constructing appropriate test statistics will be discussed in Sections 4.7, 6.11 and 7.5. Here we will give a short example to illustrate the main idea.

The result of the goodness-of-fit test is thus given by stating the so-called ***P*-value**, i.e. the probability P, under assumption of the hypothesis in question H_0, of obtaining a result as compatible or less with H_0 than the one actually observed. The P-value is sometimes also called the **observed significance level** or **confidence level**[3] of the test. That is, if we had specified a critical region for the test statistic with a significance level α equal to the P-value obtained, then the value of the statistic would be at the boundary of this region. In a goodness-of-fit test, however, the P-value is a random variable. This is in contrast to the situation in Section 4.1, where the significance level α was a constant specified before carrying out the test.

Example

Suppose one tosses a coin N times and obtains n_h heads and $n_t = N - n_h$ tails. To what extent are n_h and n_t consistent with the hypothesis that the coin is 'fair', i.e. that the probabilities for heads and tails are equal? As a test statistic one can simply use the number of heads n_h, which for a fair coin is assumed to follow a binomial distribution (equation (2.2)) with the parameter $p = 0.5$. That is, the probability to observe heads n_h times is

$$f(n_h; N) = \frac{N!}{n_h!(N-n_h)!} \left(\frac{1}{2}\right)^{n_h} \left(\frac{1}{2}\right)^{N-n_h} . \tag{4.36}$$

Suppose that $N = 20$ tosses are made and $n_h = 17$ heads are observed. Since the expectation value of n_h (equation (2.3)) is $E[n_h] = Np = 10$, there is evidently a sizable discrepancy between the expected and actually observed outcomes. In order to quantify the significance of the difference one can give the probability of obtaining a result with the same level of discrepancy with the hypothesis or higher. In this case, this is the sum of the probabilities for $n_h = 0, 1, 2, 3, 17, 18, 19, 20$. Using equation (4.36) one obtains the probability $P = 0.0026$.

Counting experiment

4.6 The significance of an observed signal

A simple type of goodness-of-fit test is often carried out to judge whether a discrepancy between data and expectation is sufficiently significant to merit a claim for a new discovery. Here one may see evidence for a special type of signal event, the number n_s of which can be treated as a Poisson variable with mean ν_s. In addition to the signal events, however, one will find in general a certain number of background events n_b. Suppose this can also be treated as a Poisson variable with mean ν_b, which we will assume for the moment to be known without error. The total number of events found, $n = n_s + n_b$, is therefore a Poisson variable with mean $\nu = \nu_s + \nu_b$. The probability to observe n events is thus

$$f(n; \nu_s, \nu_b) = \frac{(\nu_s + \nu_b)^n}{n!} e^{-(\nu_s + \nu_b)}. \tag{4.37}$$

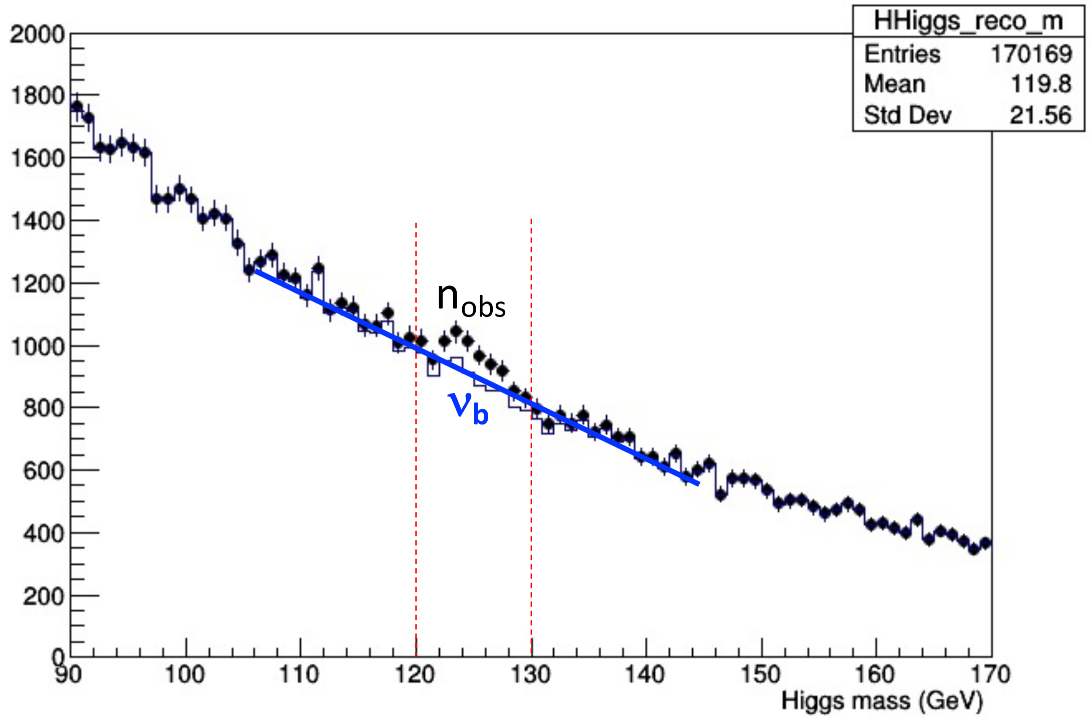
Suppose we have carried out the experiment and found n_{obs} events. In order to quantify our degree of confidence in the discovery of a new effect, i.e. $\nu_s \neq 0$, we can compute how likely it is to find n_{obs} events or more from background alone. This is given by

$$\begin{aligned} P(n \geq n_{obs}) &= \sum_{n=n_{obs}}^{\infty} f(n; \nu_s = 0, \nu_b) = 1 - \sum_{n=0}^{n_{obs}-1} f(n; \nu_s = 0, \nu_b) \\ &= 1 - \sum_{n=0}^{n_{obs}-1} \frac{\nu_b^n}{n!} e^{-\nu_b}. \end{aligned} \tag{4.38}$$

For example, if we expect $\nu_b = 0.5$ background events and we observe $n_{obs} = 5$, then the P-value from (4.38) is 1.7×10^{-4}. It should be emphasized that

Example

- In Higgs to diphoton search (Lesson 7)
- We could evaluate the significance of the small peak by estimating the background from the sidebands using a simple interpolation, and counting the observed events in a signal window.



Exercise for Lesson 8

- Evaluate the significance (p-value) of the signal in the previous slide. Use the distributions produced from the Lesson 7 code in the git repository.
- Use a signal window 120 to 125 GeV
- Do not apply the linear fit interpolation, just use the dijet distribution histogram in the code to get the ν_b