Statistics for Rare Events

Lectures organization

In order to ease the learning process, some portions of this handout will be completed during the lectures. The complete version of this handout will be posted online after each lecture.

The intended plan for our time together is the following:

- **Lecture 1:** Monday July 12, 2021 2PM-3PM
  Statistical inference (overview), likelihood ratio test.
- **Discussion/Exercises:** Monday July 12, 2021 4.20PM-5.30PM.
- **Lecture 2:** Tuesday July 13, 2021 2PM-3PM
  Confidence intervals/upper limits, goodness-of-fit.

1 A brief overview on statistical inference

Statistical inference is the area of statistics which aims to develop and study reliable tools to make conclusion on the phenomenon/population under study based on what has been observed on a data sample. Among the main inferential tools we have *tests of hypotheses* or *goodness-of-fit tests*. Confidence intervals and upper limits are also widely used but more often than not, they can be obtained by simply inverting tests of hypotheses. Hence, for the moment, we focus on

- **Tests of hypotheses:** Given a postulated model for the data (e.g., background only), we test it against an alternative model (e.g., background+ signal we are looking for).
- **Goodness-of-fit tests:** Given a postulated model for the data we test it against all possible alternatives (e.g., do we have only background events or is there something else?).

Which type of test to use typically depends on the problem under study and the classical formulation of a statistical test consist in the following steps:

1. **Specification of the null and the alternative hypotheses, namely H_0 and H_1**

   - Tests of hypotheses: e.g., we expect that $X \sim N(\mu, 1)$, we test

   $$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu \neq 0.$$

   - Goodness-of-fit: e.g., we expect that $X \sim N(\mu, 1)$, we test

   $$H_0 : X \sim N(\mu, 1) \quad (\text{versus} \quad H_1 : X \not\sim N(\mu, 1)).$$

   we call the model specified under H_0 the "null model" and the model specified under H_1 the "alternative model".

2. **Specification of a test statistic**

   - Tests of hypotheses: e.g., the Likelihood Ratio Test (LRT) statistics.
   - Goodness-of-fit: e.g., Pearson χ^2 test statistics.

   (We will see both of them later in this handout.)

3. **Derivation of the distribution of the test statistic under H_0**
   This is typically done by relying on asymptotic results or via Monte Carlo simulations.

4. **Computation of p-values or quantiles of the null distribution.**

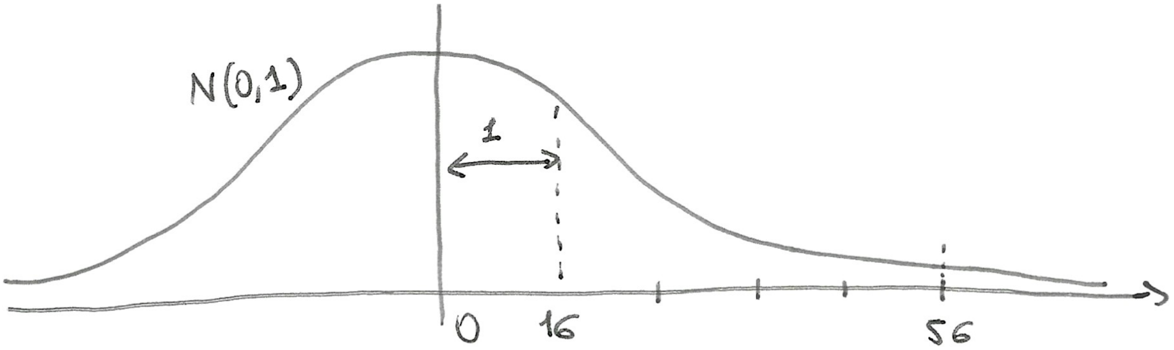
2 How do we a assess if a test is good or not?

In principle we could construct an infinite number of different test statistics for any model under study and perform a test of hypothesis by simulating their null distribution. However, not all tests are good tests and indeed, we can quantify how "good" a test is by looking at two main quantities the *probability of type I error* and the *power*.

- **The probability of type I error** is the probability of incorrectly rejecting H_0 when it is true. It is typically denoted by α and is also called "significance level", i.e.,

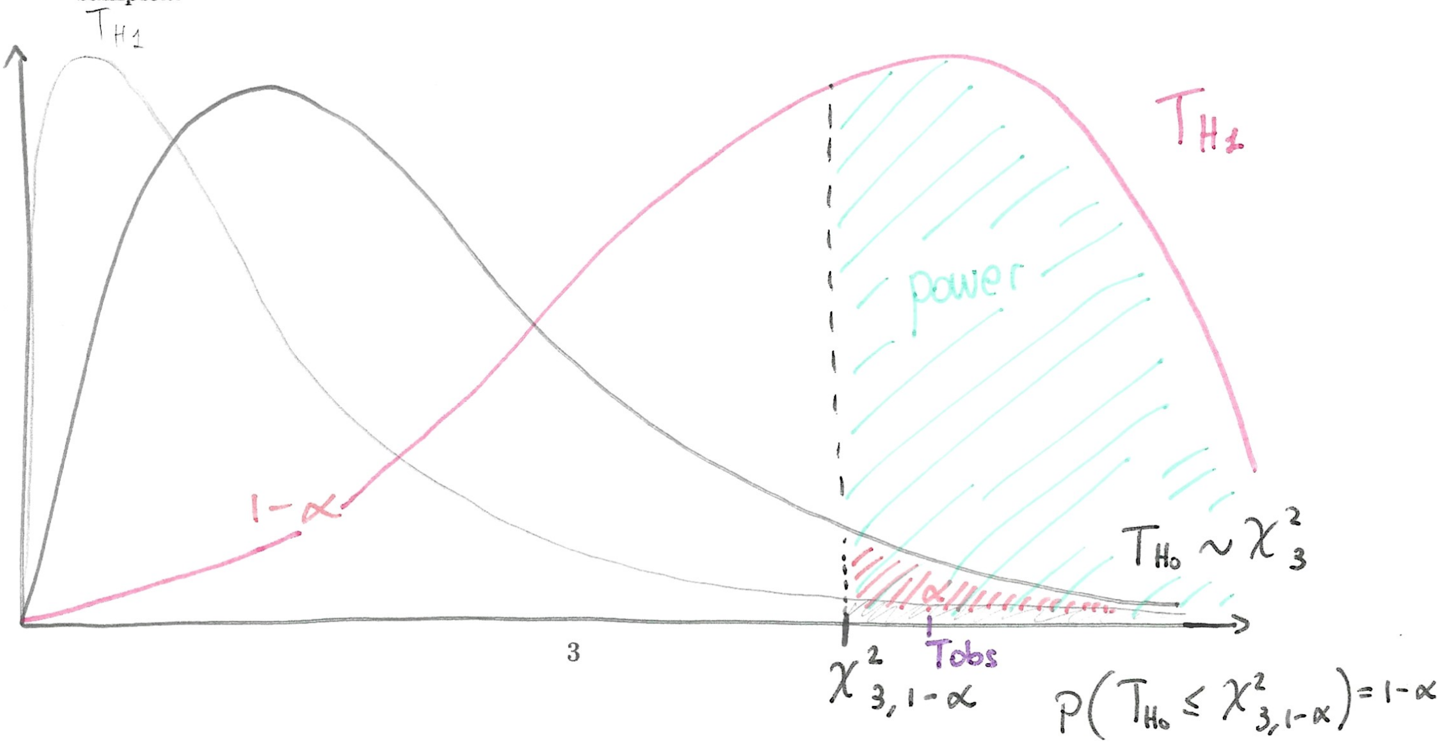
$$P(\text{Probability of Type I error}) = \alpha = 1 - \text{coverage probability}.$$

In physics, it is typically expressed in "number of sigmas", i.e., $\#\sigma = \Phi(1 - \alpha/2)$, with $\Phi(\cdot)$ being the cumulative density function of a standard normal.



- **The power** is the probability of correctly rejecting H_0 when H_1 is true.

In order to provide a provide an explicit visualization of what these two are, suppose we are testing two hypothesis H_0 and H_1 using a test statistics T which we know is asymptotically distributed as a χ^2_3. Suppose we have collected a sample of size $N = 5{,}000$ and so we are pretty sure we can trust the asymptotics. Call $\boxed{T_{obs}}$ the value of the test statistics T calculated on such sample...



Clearly, as α increases, the power increases. Moreover, we say that a test is *admissible* or *unbiased*, if its power is higher than its probability of type I error.

Finally, power and probability of type I error allow us to tell a bit more about the differences between goodness-of-fit tests and tests of hypotheses. Specifically, the former will have some power against all possible alternative models. Whereas, the latter will have high power only against the alternative model we have specified under H_1. This is the main distinction among these two classes of tests.

3 Testing hypotheses using the Likelihood Ratio Test

Given two plausible models for the data, an hypothesis testing procedure aims to select the one which is more consistent with the data collected. But what does "more consistent" mean exactly? In statistics, we often refer to the *likelihood function* to quantify how plausible a model is for the data being collected.

3.1 The likelihood function and maximum likelihood estimation

≥ 0

$\int f(y;\theta)dy = 1$

Let $\overset{\text{cdF}}{F(y;\boldsymbol{\theta})}$ be the postulated distribution for the sample $\boldsymbol{y} = (y_1, \dots, y_N)$, and denote with $f(y;\boldsymbol{\theta})$ the respective probability density function. (pdf) The vector $\boldsymbol{\theta}$ collects the p parameters which characterize the distribution F. For instance, if our sample was generated from a Normal distribution with mean μ and variance σ^2 we would have $\boldsymbol{\theta} = (\mu, \sigma^2)$, $p = 2$, and

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

whereas, $F(y;\mu,\sigma^2) = \Phi(y;\mu,\sigma^2) = \int_{-\infty}^{y} f(\overset{t}{y};\mu,\sigma^2)d\overset{t}{y}$. The (continuous/unbinned) likelihood function specifies as:

$$L(\boldsymbol{\theta};\boldsymbol{y}) = \prod_{i=1}^{N} f(y_i;\mu,\sigma^2).$$

We will see in Section 7 how the likelihood specifies if the data are binned. The likelihood function is particularly useful for the purpose of estimating the parameters in $\boldsymbol{\theta}$. Specifically, if the parameters in $\boldsymbol{\theta}$ are unknown, we can estimate them by maximizing the likelihood function, and obtain the so-called *Maximum Likelihood Estimate* (MLE), i.e.:

$$\widehat{\theta} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta};\boldsymbol{y})$$

where $l(\boldsymbol{\theta};\boldsymbol{y}) = \log\{L(\boldsymbol{\theta};\boldsymbol{y})\}$ is called *log-likelihood function*, and we rely on it just because it is typically much easier to maximize $l(\boldsymbol{\theta};y)$ rather than $L(\boldsymbol{\theta};\boldsymbol{y})$.

Finally, under suitable regularity conditions we have that

$$\widehat{\boldsymbol{\theta}} \approx N(\boldsymbol{\theta}, I^{-1}(\boldsymbol{\theta})), \quad \text{as } n \to \infty, \tag{1}$$

where $I^{-1}(\boldsymbol{\theta})$ is the inverse of the so-called *Fisher's Information Matrix*, i.e., $I(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{\partial\theta^2} l(\boldsymbol{\theta};\boldsymbol{y})\right]$.

3.2 The Likelihood Ratio Test

The likelihood ratio test consists in comparing the null and the alternative models specified under H_0 and H_1 in terms of their log-likelihood.

- Specification of the null and alternative hypotheses.
  Typically, we consider tests such as:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{(simple hypothesis)} \quad \text{versus} \quad H_1 : \boldsymbol{\theta} > \boldsymbol{\theta}_0 \quad \text{(one sided hypothesis)} \qquad (2)$$

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{(simple hypothesis)} \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \quad \text{(two sided hypothesis)} \qquad (3)$$

- Specification of the test statistic:

$$\Lambda(\boldsymbol{\theta}_0) = -2 \log \frac{L(\boldsymbol{\theta}_0; y)}{L(\widehat{\boldsymbol{\theta}}; y)} = -2 \sum_{i=1}^{N} [l(y_i; \boldsymbol{\theta}_0) - l(y_i; \widehat{\boldsymbol{\theta}})]$$

Neyman
pearson

- Derivation of the distribution of the test statistic under H_0.
  Under suitable regularity conditions, as $N \to \infty$, under H_0,

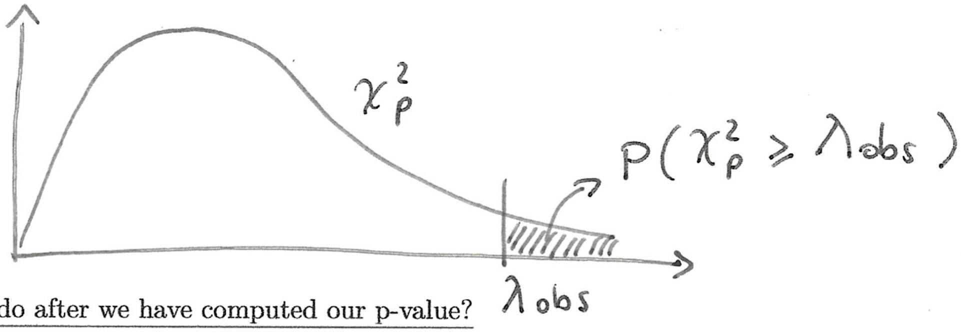
$$\overset{(\theta_0)}{\Lambda} \approx \chi^2_p.$$

Wilks 1938

  Alternatively, one can simulate the distribution of Λ under H_0 via Monte Carlo.

- Computation of the p-value.
  Let Λ_{obs} be the value of the test statistic Λ evaluated on the data,

$$\text{p-value} = P(\chi^2_p \geq \Lambda_{obs}).$$



- What do we do after we have computed our p-value?
  - If p-value$\geq \alpha$: we "fail to reject" $H_0 \Rightarrow$ our null model fits the data well.
  - If p-value$< \alpha$: we reject $H_0 \Rightarrow$ our null model does not fit the data well, the alternative model is better.

  This is a legitimate decision rule because, if you think about, the p-value is nothing but the probability of obtaining a value of our test statistics Λ which is more extreme than the value we have observed on the data, Λ_{obs}, when H_0 is true. So if this event is very unlikely (small p-value), that means that probably, the true distribution of Λ is different from the one we expect under H_0, and consequently, it is unlikely that H_0 is valid.

3.3 The Likelihood Ratio Test using the profile likelihood

We may encounter situations where, in addition to the parameter $\boldsymbol{\theta}$ which we are interested in, we may also have another set of parameters, namely τ, which are not interesting to us and play the role of nuisance parameters. The latter may correspond, for instance, to systematic uncertainties. The question then is: how do we deal with those? $H_0: \theta = \theta_0$ $H_1: \theta \neq \theta_0$

Interestingly, we can proceed exactly as we did before but this time, instead of the likelihood functions under H_0 and H_1 we consider the respective *profile likelihood* functions. The latter consist in replacing τ with its MLE obtained by fixing $\boldsymbol{\theta}$ to its value specified under the hypotheses,

$$\Lambda(\boldsymbol{\theta}_0) = -2\log\frac{L(\boldsymbol{\theta}_0;\widehat{\boldsymbol{\tau}}_{\boldsymbol{\theta}_0},y)}{L(\widehat{\boldsymbol{\theta}};\widehat{\boldsymbol{\tau}}_{\widehat{\boldsymbol{\theta}}},y)} = -2\sum_{i=1}^{N}[l(y_i;\widehat{\boldsymbol{\tau}}_{\boldsymbol{\theta}_0},y) - l(y;\widehat{\boldsymbol{\tau}}_{\widehat{\boldsymbol{\theta}}},y)]$$

and the asymptotic distribution of Λ under H_0 remains unchanged, i.e., we still have $\Lambda \approx \chi^2_p$ (provided that the regularity conditions needed hold). Alternatively, one can proceed by simulating the null distribution of $\Lambda(\boldsymbol{\theta}_0)$ under H_0, that is, when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\tau = \widehat{\tau}_{\boldsymbol{\theta}_0}$. Since the value of τ used is an estimate for it, in statistics, we refer to this process with *parametric bootstrap*.

(Efron , 1979)

3.4 Some warnings

In the previous sections, when deriving the asymptotic distribution of $\widehat{\theta}$ and $\Lambda(\boldsymbol{\theta}_0)$ we have said that, those results are valid "under some regularity conditions". Despite the latter are quite technical, we can identify a few conditions necessary for the validity of our χ^2 approximation. Specifically,

C1. The true value of the parameters must not be on the boundaries of their parameter space.

C2. The parameters are identifiable. That is, different values of the parameters specify distinct models.

C3. The model under H_0 is a special case of the model under H_1.

C4. The true model is one between those specified under H_0 or under H_1.

A discussion on how to identify situations of non-regularity in a more practical setting is postponed to Section 4.2.