# Hypothesis testing

## 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, May 16, 2022, 4PM-5.30PM
  Statistical inference (overview), likelihood ratio test.
- **Lecture 2:** Tuesday, May 17, 2022, 11AM-12.30PM
  Goodness-of-fit.
- **Exercises:** Tuesday, May 17, 2022, 2PM-3.30PM
  To be held jointly with the exercises for Glen and Harrison's lectures.

## 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, 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 but, in both cases, the classical formulation of a statistical test entails 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) statistic.
   - Goodness-of-fit: e.g., Pearson $\chi^2$ test statistic.

   (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 or parametric bootstrap simulations. *Do you know what is the difference between the two?*

   - Monte Carlo: the data are simulated from a model which is fully specified
   - Parametric bootstrap: the data are simulated from a model for which some or all the parameters are unknown. And in the simulation we replace them with an estimate for them.

     Efron 1979

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

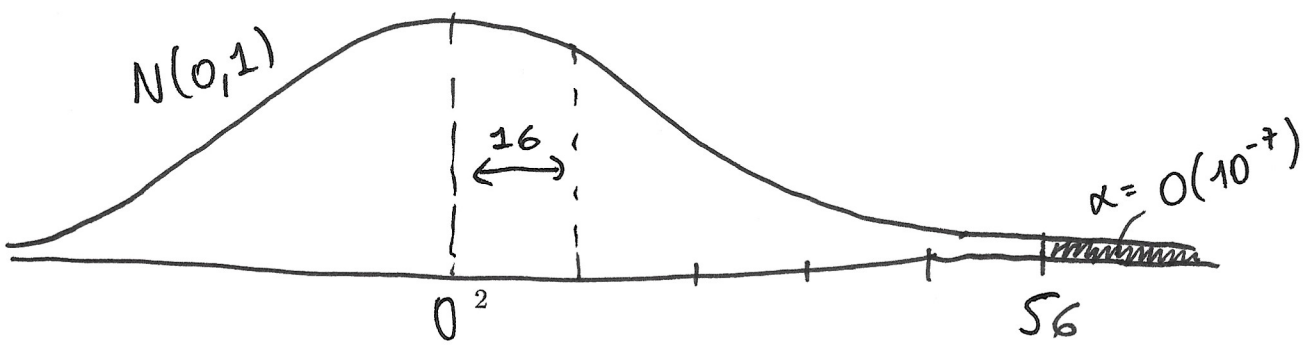
## 2 How do we assess if a test is good or not?

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 $\alpha$ 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.



# Simulating the null distribution of a GOF statistic using the parametric bootstrap

Inputs:
- $x_1, \ldots, x_N$ = data observed
- $N$ = sample size
- $B$ = # of bootstrap replicates
- $f(x, \theta)$ = the density under $H_0$
- $T(x, \theta)$ = test statistic

Step 0) Estimate $\theta$ on $x_1 \ldots x_N$ and we obtain $\hat{\theta}$

For $b = 1, \ldots, B$

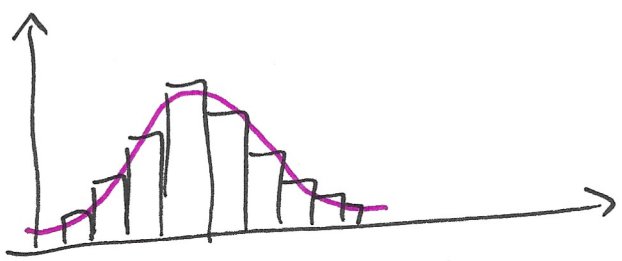
Step 1) Draw a sample of $N$ observations from $f(x, \hat{\theta})$ and call it

$$\underline{x}_N^{(b)} = (x_1^{(b)}, \ldots x_N^{(b)})$$

Step 2) Estimate $\theta$ using $\underline{x}_N^{(b)}$ and obtain $\hat{\theta}^{(b)}$

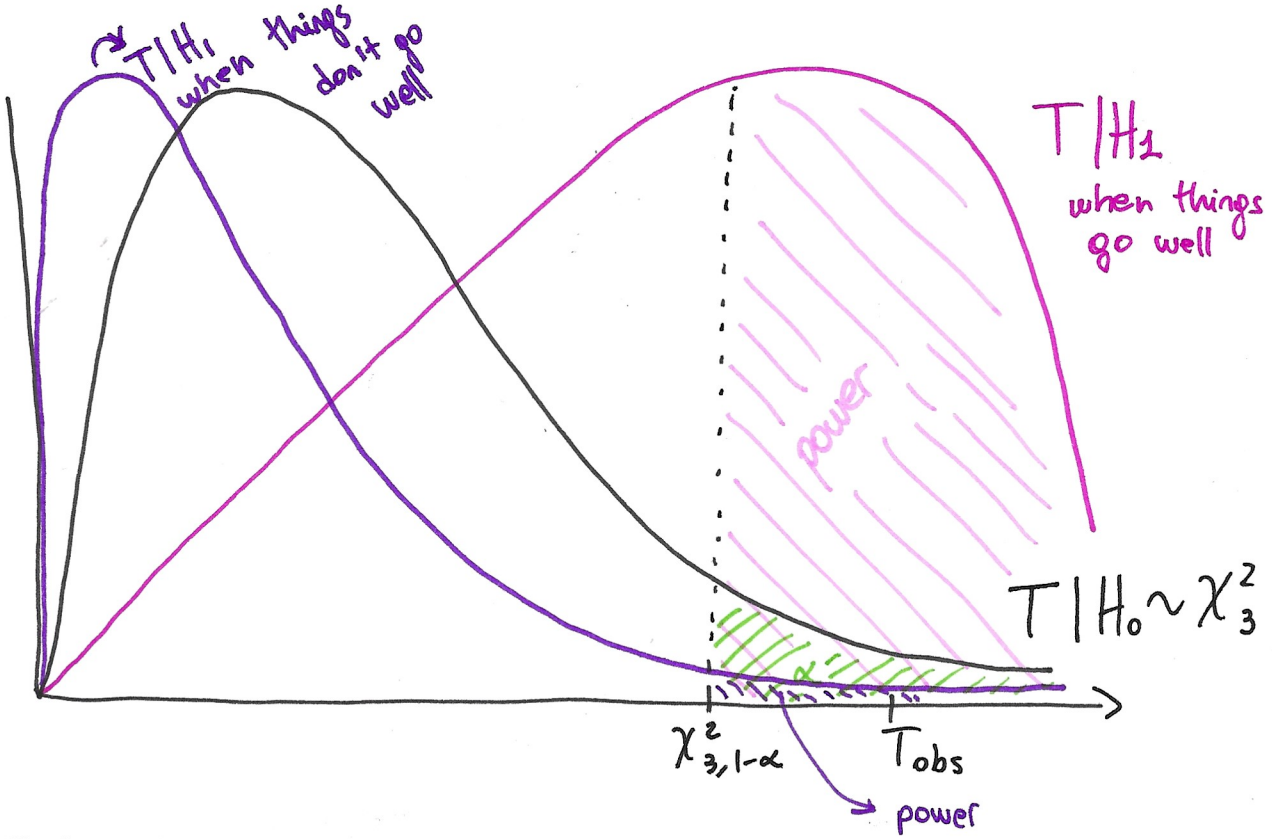
Step 3) Evaluate $T$ at $\underline{x}_N^{(b)}$ and $\hat{\theta}^{(b)}$ and obtain $T^{(b)}$

$$T^{(1)} \quad \ldots \quad T^{(B)}$$



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

In order to provide an explicit visualization of what these two are, suppose we are testing two hypothesis $H_0$ and $H_1$ using a test statistic $T$ which we know is asymptotically distributed as a $\chi^2_3$. Suppose we have collected a sample of size $N = 500,000$ and so we are pretty sure we can trust the asymptotics. Call $T_{obs}$ the value of the test statistic $T$ calculated on such sample...




Clearly, as $\alpha$ 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

Let $F(y;\boldsymbol{\theta})$ be the postulated cumulative distribution function (cdf) for the sample $\boldsymbol{y} = (y_1,\ldots,y_N)$, 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 $\mu$ and variance $\sigma^2$ 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, $F(y;\mu,\sigma^2) = \Phi(y;\mu,\sigma^2) = \int_{-\infty}^{y} f(t;\mu,\sigma^2)\mathrm{d}t$. 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 4 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.:

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

where $l(\boldsymbol{\theta};\boldsymbol{y}) = \log\{L(\boldsymbol{\theta};\boldsymbol{y})\}$ is the *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., $I(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{\partial^2\boldsymbol{\theta}} l(\boldsymbol{\theta};\boldsymbol{y})\right]$.

### 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:

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

- Specification of the test statistic:

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

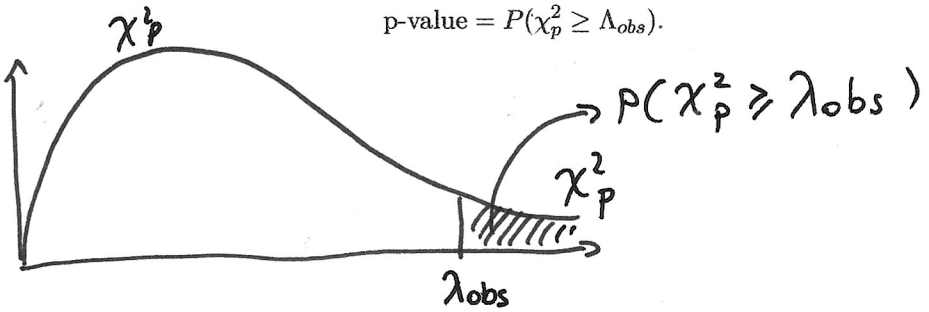
- Derivation of the distribution of the test statistic under $H_0$.
  Under suitable regularity conditions, as $N \to \infty$, under $H_0$,

$$\Lambda(\boldsymbol{\theta}_0) \approx \chi^2_p.$$

  Alternatively, one can simulate the distribution of $\Lambda$ under $H_0$ via Monte Carlo.

- Computation of the p-value.
  Let $\Lambda_{obs}$ be the value of the test statistic $\Lambda$ 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 statistic $\Lambda$ which is more extreme than the value we have observed on the data, $\Lambda_{obs}$, when $H_0$ is true. So if this event is very unlikely (small p-value), that means that probably, the true distribution of $\Lambda$ 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 $\boldsymbol{\tau}$, which are of no interest 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?
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 of replacing $\boldsymbol{\tau}$ with its MLE obtained by fixing $\boldsymbol{\theta}$ to its value specified under the hypotheses,

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

The asymptotic distribution of $\Lambda$ 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 $\Lambda_{\boldsymbol{\tau}}(\boldsymbol{\theta}_0)$ under $H_0$, that is, when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\boldsymbol{\tau} = \widehat{\boldsymbol{\tau}}_{\boldsymbol{\theta}_0}$ via the *parametric bootstrap*.

### 3.4 Some warnings

In the previous sections, when deriving the asymptotic distribution of $\widehat{\theta}$ and $\Lambda(\theta_0)$ 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 $\chi^2$ approximation. Specifically,

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

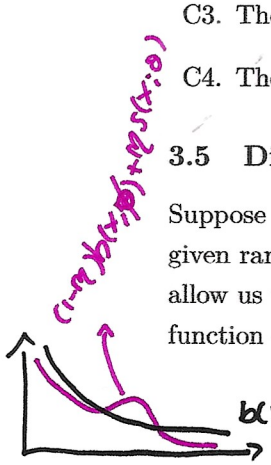
C2. The parameters are identifiable. That is, different values of the parameters specify distinct models.

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 $H_1$.

### 3.5 Discussion: Identifying situations of non-regularity

Suppose $X$ is the number of events detected by our instrument (e.g., photon emissions over a given range of energy) and let's assume that the resolution of the latter is sufficiently high to allow us to treat the data as a continuous stream of data (no binning). The probability density function of $X$ specifies as follows:




$$f(x;\eta,\phi,\theta) = (1-\eta)b(x,\phi) + \eta s(x,\theta), \quad \eta \in [0,1) \tag{2}$$

where $b(x,\phi)$ is the density of the background and $\phi$ is a parameter characterizing it. For instance, if $b$ is a power-law, $\phi$ is its slope. Whereas, $s(x,\theta)$ is the density of the signal and $\theta$ the parameter characterizing it. For instance, if we may consider $s$ to be the pdf of a Gaussian with variance 1 and mean $\theta$. Finally, $\eta$ is the relative intensity of the signal, i.e., the proportion of events that we expect to come from the signal and it is assumed to be non-negative. Finally, you are given a sample of $N = 10,000$ observations which you may use to assess the validity of the model in (2). A few questions for you...

1. **How would you specify your hypotheses in order to assess if a signal is present or not?**

$H_0: \eta = 0 \qquad H_1: \eta > 0$

2. **Suppose $\theta$ is known to be 125. Which among the necessary conditions needed for the $\chi^2$-approximation of the LRT fail (see Section 3.4)? Why?**

C1 Fails because $\eta$ is on the boundary of its parameter space

SOLUTION: Chernoff (1954)

$\Lambda(\theta_0) \not\sim \chi^2_1 \qquad \lambda(\theta_0) \approx \frac{1}{2}\chi^2_1 + \frac{1}{2}\delta(0)$

$$\text{p-value} = \frac{P(\chi^2_1 > \lambda_{obs})}{2}$$

3. **Suppose $\theta$ is unknown. Which among the necessary conditions needed for the $\chi^2$-approximation of the LRT fail (see Section 3.4)? Why?**

- Condition C1 fails (for the same reasons we have seen in Q2)
- Condition C2 fails because whenever $\eta=0$ for any value of $\theta$ we get the same model => we have NON-IDENTIFIABILITY

"LOOK-ELSEWHERE EFFECT"

Solution: Gross and Vitells (2010) EPJ-C

4. **Suppose you decided to test $H_0 : b(x, \phi)$ versus $H_1 : s(x, \theta)$. Which among the necessary conditions needed for the $\chi^2$-approximation of the LRT fail?**

C3 fails because we are testing non-nested models

Solution: Algeri et al. (2016) MNRAS Letters.

You can find a self-contained review on solutions for the problems above in

Algeri et al., 2020. *Searching for new phenomena with profile likelihood ratio tests.* Nature Reviews Physics. https://www.nature.com/articles/s42254-020-0169-5.

You may watch a seminar on this at https://www.youtube.com/watch?v=e0Y9QszyO1E.