FORMAL REMARKS ON THE TEST OF HYPOTHESES

(DA_2022)

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

Lecture n. 15, Rome April 27, 2021

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

Outline L 15

- Parametric vs. non parametric tests
- Formal structure of a test: parameter space and the space of samples
- Role of H_0
- Critical regions
- Mapping beween parameter and sample space: type I and Type II errors
- Amplitude and power of a test
- Optimal critical region

Study materials:
Rosner's chap. 7 par. 7.1- and lecture notes DA_2020_L18_notes

GENERAL SCHEME FOR TESTING HYPOTHESES

STATISTICS THE SCIENCE OF TAKING DECISIONS UNDER UNCERTAINTY (ALTERNATIVE HYP)

- TRANSLATE A PROBLEM INTO A STATISTICAL MODEL (i.e. Postulate a probability distribution)
- GET INFORMATION FROM SAMPLING + MODEL
- TAKE A DECISION (ABOUT HYP)

BASIC DISTINCTION

PARAMETRIC TESTS

(A specific functional form of the pdf/MODEL is assumed e.g. $f(x,\underline{\theta}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[(x-\mu)^2/2\sigma^2\right]$

NON PARAMETRIC TESTS

(a pdf is assumed, some general features are assumed but not a specific family of functions. Eg discrete/continuous, unimodal/multimodal, simmetric/asymmetric, limited support,)...)

The distinction may be not so rigid

[it is not ideological, it depends on the context, on how much you know about the experimental situation and the pdf/model]

DEF A statistical test of hypothesis is a decision rule on a sample space about rejecting/not rejecting the null hypothesis H_0

REAL WORLD ① → observables, variables ② $\{X_1, X_2, \ldots, X_n\}$ →

→ $f(x, \underline{\theta})$ ③ MODEL

② + ③ ≈ statistical model (sample space)

$\Omega(\underline{\theta})$ (parameter space) the set of possible values for $\underline{\theta}$ in $f(x, \underline{\theta})$

$\omega_0 \subset \Omega(\underline{\theta})$ is the range of values corresponding to H_0 $\quad H_0$ true $\rightarrow \underline{\theta} \in \omega_0$

alternatively: $\quad H_1$ true $\rightarrow \underline{\theta} \notin \omega_0$

THE DECISION RULE HAS TO DO WITH CHOOSING

$H_0 = \underline{\theta} \in \omega_0$ vs $H_1 = \underline{\theta} \notin \omega_0$

ON THE BASIS OF THE OBSERVED SAMPLE $\{x_1, x_2, \ldots, x_n\}$

LEXICON: simple hypotheses ω_0 is a specific value θ_0

composite " ω_0 is an interval

unidirectional hypotheses $\theta \geq \theta_0$; $\theta \leq \theta_0$

bidirectional " $\theta \neq \theta_0$ $\theta = \theta_0$

EXAMPLE Let us suppose that, with reference to a sample $\{X_1, X_2, \dots, X_n\}$ we assume that the sample comes from a $f(x, \theta)$ with $\theta = K$ (a specific value)

then: a test might consist in checking if $H_0 = \theta = K$ is true against $H_1 : \theta > K$
In this case H_0 is simple and H_1 is composite and unidirectional
or it might consist in checking whether $H_0 : \theta = K$ or $H_1 : \theta \neq K$
In this case H_1 is composite and bidirectional

REM 1 In general, as said in the previous lect. H_0 is the hyp that represents the situation that is widely assumed as default, as true until proven otherwise (fino a prova contraria, we say in Italian)

REM 2 THE TWO HYP IN A TEST ARE NOT EQUIVALENT

THE TEST IS ALWAYS INCONCLUSIVE ABOUT H_1 IT ONLY CONCERNS THE DECISION ABOUT REJECT / NOT REJECT H_0

FORMAL REMARK

A STATISTICAL TEST OF HYPOTHESIS HAS TO DO (this is a pragmatic expression I evidently like) WITH A CRITICAL REGION of the sample space (spazio campionario) i.e. a critical subspace of $\mathbb{R}^n$ such that

- if $(x_1, x_2, \ldots, x_n) \in CR$ H_0 is rejected
- if $(x_1, x_2, \ldots, x_n) \notin CR$ H_0 cannot be rejected

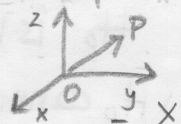
---

DOING A TEST IS, IN PRACTICE, FINDING IST CRITICAL REGION WHICH IS THE REJECTION REGION, ITS COMPLEMENT $\overline{RC}$ is the NO-REJECTION REGION

IN GENERAL WE EXPLORE THE SAMPLE SPACE THROUGH STATISTICS THAT REDUCE THE DIMENSION OF THE SAMPLE SPACE (Compression)

LET ME SPEND A LITTLE BIT OF TIME MAKING CLEAR METHODOLOGICAL POINTS ONCE IN A LIFETIME FOR THOSE OF YOU WHO WILDLY HATE MATHEMATICAL ABSTRACTION, BUT REMEMBER THERE IS NO FREE LUNCH = NO MATH $\longrightarrow$ NO SOUND DATA ANALYSIS)

- $X \sim f(x|\theta)$ stochastic variable $x \in \mathbb{R}$
- $x \in \mathbb{R}$ is an empirical determination of X

(much in the spirit of saying that to an abstract formal geometric object $\underline{V}$ vector can be represented by an empirical collection of numbers that require the specification of a reference frame 

) $= \underline{X}$ (random vector)

- $(X_1, X_2, \ldots, X_n)$ is a stochastic sample of dim n a collection of n independent, identically distributed stochastic variables whose 'realization' is an empirical (concrete) sample $(x_1, x_2, \ldots, x_n) = \underline{x}$ (replica of $f(x|\theta)$)
- EXTRACTION — SAMPLE — REALIZATION

  Before — stochastic $\underline{X}(X_1, X_2, \ldots, X_n)$

  after — empirical observed $\underline{x}(x_1, x_2, \ldots, x_n)$
- $T_n = T(X_1, X_2, \ldots, X_n)$ abstract statistic

  $t_n = T(x_1, x_2, \ldots, x_n)$ empirical observed statistic

Using the statistics is a way of compressing the information (dimensional reduction)

- ESTIMATORS ARE SUFFICIENT, UNBIASED, EFFICIENT, WITH MINIMUM MEAN SQUARE ERROR STATISTICS

- In most cases we are interested in 1-d statistics (e.g. sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$) whose realization spans the real line
- then critical regions for an hypothesis are in general intervals of the real line $\mathbb{R}$ [or discrete sets in the case of discrete variables]

Let us go back to the tests

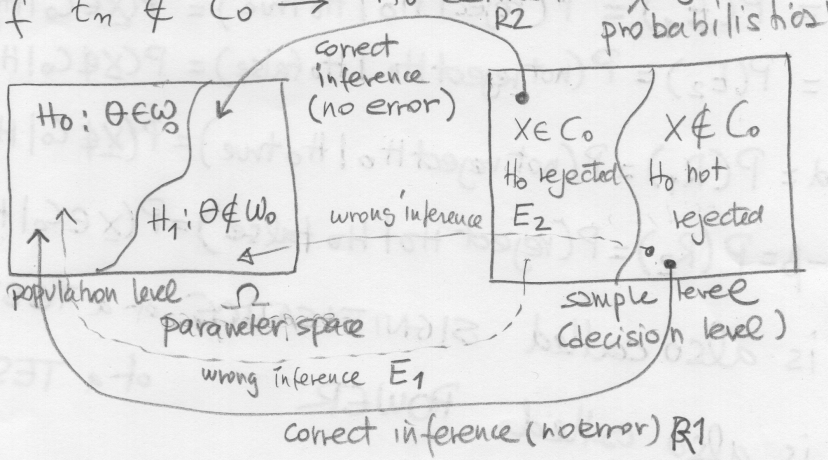
T_n abstract statistic

t_n empirical, evaluated statistic

C_0 critical interval (after effective dimens. reduction) probabilistically

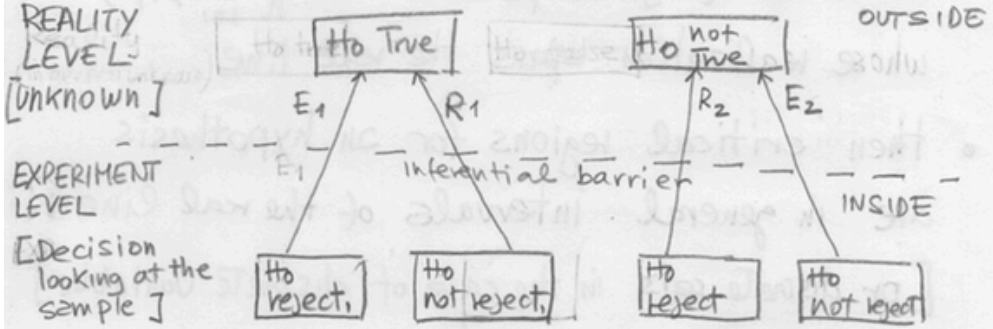
if $t_n \in C_0 \rightarrow H_0$ is rejected

if $t_n \notin C_0 \rightarrow H_0$ cannot be rejected probabilistically



IMPORTANT REMARKS

- H_0 is either true or false (tertium non datum)



We are in a probabilistic setting; we do not know if we are taking the right or the wrong choice. E_1 is a type I ERROR (H_0 rejected when it is true); E_2 is a type II ERROR (H_0 not rejected when it is not true). Then let us define:

$$\alpha = P(E_1) = P(\text{reject } H_0 \mid H_0 \text{ true}) = P(\underline{X} \in C_0 \mid H_0)$$

$$\beta = P(E_2) = P(\text{not reject } H_0 \mid H_0 \text{ false}) = P(\underline{X} \notin C_0 \mid H_1)$$

$$1-\alpha = P(R_1) = P(\text{not reject } H_0 \mid H_0 \text{ true}) = P(\underline{X} \notin C_0 \mid H_0)$$

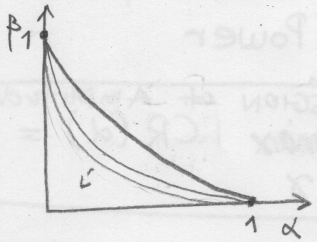
$$\gamma = 1-\beta = P(R_2) = P(\text{reject } H_0 \mid H_0 \text{ false}) = P(\underline{X} \in C_0 \mid H_1)$$

α is also called SIGNIFICANCE of a TEST

γ is also called POWER of a TEST

REMARK ON OPTIMALITY OF A TEST

- C_0 is the critical region of the sample space associated to the rejection of H_0 ($\underline{X} \in C_0 \rightarrow H_0$ rejected)
- α, β, γ are probabilities that depend on C_0 = $\alpha(C_0), \beta(C_0), \gamma(C_0)$.
- In a 'scatter' plot $\beta(C_0)$ vs $\alpha(C_0)$ we see that:



- it is impossible to minimize simultaneously both types of error; what we are looking for is to find a C_0 that is a good compromise taking into account that in general we are consider Type II errors more relevant than type II. Indeed, if I reject H_0 (which is the default, consensus accepted reality) and H_0 is true the whole community would say you were crazy in putting into doubt what all of us know it is true!

whereas failing to recognize a hard to be detected new aspect of reality is less shameful.

Then, as a general criterion, one decides the level α of risk of Type I error that one is prepared to accept, then, given α, one searches for the critical rejection region that has the minimum β, that is the maximum γ, Power

---

OPTIMAL CRITICAL REGION of Amplitude α

$$\max_{\gamma} RCR(\alpha) = OCR$$

USUALLY (this is a matter of taste and reputation there are no theorems ...)

α is set to three levels of risk propensión:

0.05	0.01	0.001
L	M	S

Theoretically, one can find the OCR using the Neyman-Pearson lemma.

Hypothesis testing: an example

To show you the basic concepts and terminology of hypothesis testing, we'll take you through all the steps by using an example. Our goal is to illuminate the basic process without distraction from the details of the probability calculations. We'll get to plenty of the details in later chapters.

Four basic steps are involved in hypothesis testing:

1. State the hypotheses.
2. Compute the test statistic.
3. Determine the *P*-value.
4. Draw the appropriate conclusions.

We'll define the new terms we just used in this section.

Example 6.2 tests a hypothesis about a proportion, but hypothesis testing can address a wide variety of quantities, such as means, variances, differences in means, correlations, and so on. We'll try to emphasize the general over the specific here. Further details of how to test hypotheses about proportions are discussed in Chapter 7.

Type I and Type II errors

There are two kinds of errors in hypothesis testing, prosaically named Type I and Type II.

Rejecting a true null hypothesis is a **Type I error**. Failing to reject a false null hypothesis is a **Type II error**. Both types of error are summarized in Table 6.3-1.

Type I error is rejecting a true null hypothesis. The significance level α sets the probability of committing a Type I error.

Type II error is failing to reject a false null hypothesis.

TABLE 6.3-1 Types of error in hypothesis testing.

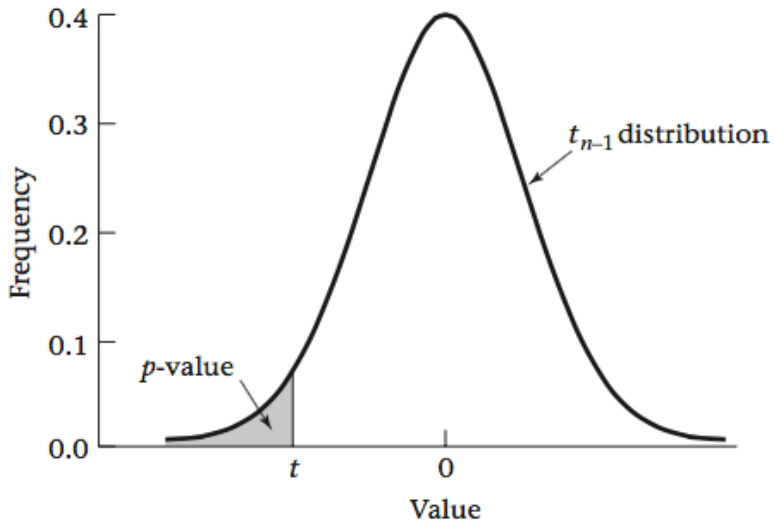
		Reality
Conclusion	H_0 true	H_0 false
Reject H_0	Type I error	Correct
Do not reject H_0	Correct	Type II error

The significance level, α, gives us the probability of committing a Type I error. If we go along with convention and use a significance level of $\alpha = 0.05$, then we reject H_0 whenever P is less than or equal to 0.05. This means that, if the null hypothesis were true, we would reject it mistakenly one time in 20. Biologists typically regard this as an acceptable error rate.

DEFINITION 7.13 The ***p*-value** for any hypothesis test is the α level at which we would be indifferent between accepting or rejecting H_0 given the sample data at hand. That is, the *p*-value is the α level at which the given value of the test statistic (such as t) is on the borderline between the acceptance and rejection regions.

DEFINITION 7.14 The ***p*-value** can also be thought of as the probability of obtaining a test statistic as extreme as or more extreme than the actual test statistic obtained, given that the null hypothesis is true.

Graphic display of a *p*-value



We know that under the null hypothesis, the t statistic follows a t_{n-1} distribution. Hence, the probability of obtaining a t statistic that is no larger than t under the null hypothesis is $Pr(t_{n-1} \leq t) = p$-value, as shown in Figure 7.1.

EQUATION 7.4

Guidelines for Judging the Significance of a *p*-Value

If $.01 \le p < .05$, then the results are *significant.*
If $.001 \le p < .01$, then the results are *highly significant.*
If $p < .001$, then the results are *very highly significant.*
If $p > .05$, then the results are considered *not statistically significant* (sometimes denoted by NS).
However, if $.05 < p < .10$, then a trend toward statistical significance is sometimes noted.

EQUATION 7.5

Determination of Statistical Significance for Results from Hypothesis Tests

Either of the following methods can be used to establish whether results from hypothesis tests are statistically significant:

(1) The test statistic t can be computed and compared with the critical value $t_{n-1,\,\alpha}$ at an α level of .05. Specifically, if $H_0: \mu = \mu_0$ vs. $H_1: \mu < \mu_0$ is being tested and $t < t_{n-1,.05}$, then H_0 is rejected and the results are declared *statistically significant* ($p < .05$). Otherwise, H_0 is accepted and the results are declared *not statistically significant* ($p \ge .05$). We have called this approach the **critical-value method** (see Definition 7.12).

(2) The exact p-value can be computed and, if $p < .05$, then H_0 is rejected and the results are declared *statistically significant.* Otherwise, if $p \ge .05$, then H_0 is accepted and the results are declared *not statistically significant.* We will refer to this approach as the ***p*-value method.**

EXAMPLE 6.39 **Hypertension** An *Arteriosonde machine* "prints" blood-pressure readings on a tape so that the measurement can be read rather than heard. A major argument for using such a machine is that the variability of measurements obtained by different observers on the same person will be lower than with a standard blood-pressure cuff.

Suppose we have the data in Table 6.6, consisting of systolic blood pressure (SBP) measurements obtained on 10 people and read by two observers. We use the difference d_i between the first and second observers to assess interobserver variability. In particular, if we assume the underlying distribution of these differences is normal with mean μ and variance σ^2, then it is of primary interest to estimate σ^2. The higher σ^2 is, the higher the interobserver variability.

TABLE 6.6 **SBP measurements (mm Hg) from an Arteriosonde machine obtained from 10 people and read by two observers**

	Observer		
Person (i)	1	2	Difference (d)
1	194	200	−6
2	126	123	+3
3	130	128	+2
4	98	101	−3
5	136	135	+1
6	145	145	0
7	110	111	−1
8	108	107	+1
9	102	99	+3
10	126	128	−2

We have seen previously that an unbiased estimator of the variance σ^2 is given by the sample variance S^2. In this case,

$$\text{Mean difference} = (-6+3+\cdots-2)/10 = -0.2 = \bar{d}$$

$$\text{Sample variance} = s^2 = \sum_{i=1}^{n}(d_i-\bar{d})^2/9$$

$$= \left[(-6+0.2)^2+\cdots+(-2+0.2)^2\right]/9 = 8.178$$

How can an interval estimate for σ^2 be obtained?

7.8 ONE-SAMPLE χ^2 TEST FOR THE VARIANCE OF A NORMAL DISTRIBUTION

7.46 **Hypertension** Consider Example 6.39, concerning the variability of blood-pressure measurements taken on an Arteriosonde machine. We were concerned with the difference between measurements taken by two observers on the same person = $d_i = x_{1i} - x_{2i}$, where x_{1i} = the measurement on the ith person by the first observer and x_{2i} = the measurement on the ith person by the second observer. Let's assume this difference is a good measure of interobserver variability, and we want to compare this variability with the variability using a standard blood-pressure cuff. We have reason to believe that the variability of the Arteriosonde machine may differ from that of a standard cuff. Intuitively, we think the variability of the new

method should be lower. However, because the new method is not as widely used, the observers are probably less experienced in using it; therefore, the variability of the new method could possibly be higher than that of the old method. Thus a two-sided test is used to study this question. Suppose we know from previously published work that $\sigma^2 = 35$ for d_i obtained from the standard cuff. We want to test the hypothesis $H_0: \sigma^2 = \sigma_0^2 = 35$ vs. $H_1: \sigma^2 \neq \sigma_0^2$. How should we perform this test?

If $x_1, \ldots, x_n$ are a random sample, then we can reasonably base the test on s^2 because it is an unbiased estimator of σ^2. We know from Equation 6.15 that if $x_1, \ldots, x_n$ are a random sample from an $N(\mu, \sigma^2)$ distribution, then under H_0,

$$X^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Therefore,

$$Pr\left(X^2 < \chi^2_{n-1,\alpha/2}\right) = \alpha/2 = Pr\left(X^2 > \chi^2_{n-1,1-\alpha/2}\right)$$

Hence, the test procedure is given as follows.

One-Sample χ^2 Test for the Variance of a Normal Distribution (Two-Sided Alternative)

We compute the test statistic $X^2 = (n-1)s^2/\sigma_0^2$.

If $X^2 < \chi^2_{n-1,\alpha/2}$ or $X^2 > \chi^2_{n-1,1-\alpha/2}$, then H_0 is rejected.

If $\chi^2_{n-1,\alpha/2} \leq X^2 \leq \chi^2_{n-1,1-\alpha/2}$, then H_0 is accepted.

The acceptance and rejection regions for this test are shown in Figure 7.10.

Alternatively, we may want to compute a p-value for our experiment. The computation of the p-value will depend on whether $s^2 \leq \sigma_0^2$ or $s^2 > \sigma_0^2$. The rule is given as follows.

Acceptance and rejection regions for the one-sample χ^2 test for the variance of a normal distribution (two-sided alternative)

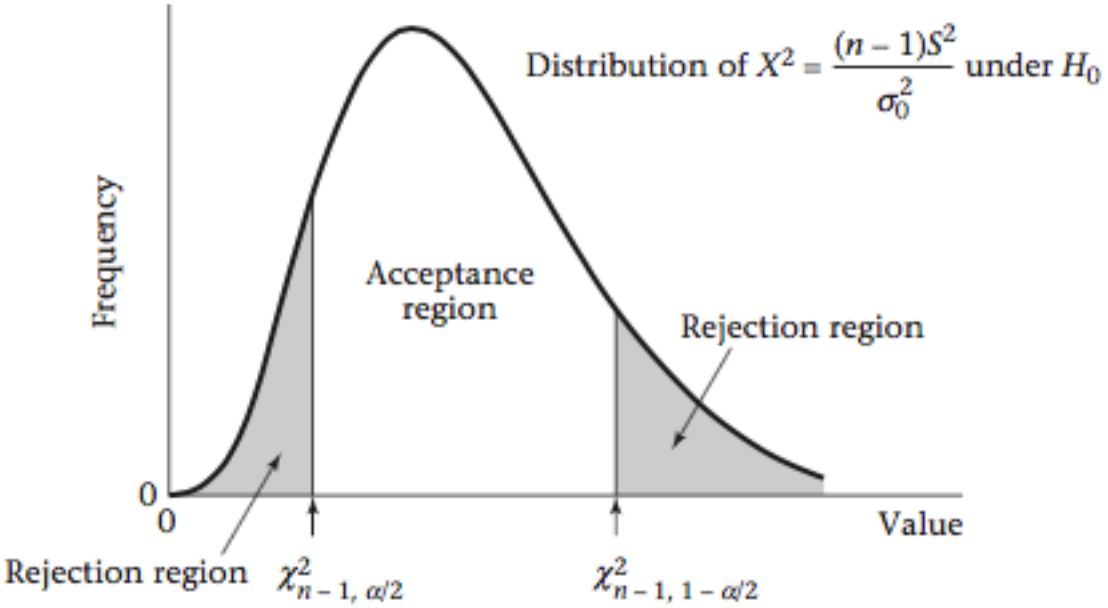


FIGURE 7.11 Illustration of the p-value for a one-sample χ^2 test for the variance of a normal distribution (two-sided alternative)