Contents

1 Hypothesis and Significance Testing: an overview of Fisher's and Neyman - Pearson's approaches

The main objective of statistical inference consists in employing mathematical methods in order to obtain, starting from the information arising from a given sample, information regarding the distribution X from which such observations come from or a given parameter θ which characterizes it.

This is the only part of testing which is the same under both Fisher's and Neyman - Pearson's approaches. In the following part of the chapter, in order to generally present how does significance and hypothesis testing work, we will focus more in describing the latter approach, a choice made just because it is the mainstream approach both in theoretical teaching as well as in empirical applications in all fields, from economics to biology to psychology. Though, given that a lot of the misinterpretation of the p - value based results is also due to the silence in which this difference lays, before going deeper with the Neyman - Pearson's approach, we will also briefly present the Fisherian approach.

Fisher's significance testing and inductive inference

Let suppose we have observed a sample of size $n \geq 1$ arising from a random variable X having density $X \sim f(x \mid \theta)$, $\theta \in \Theta$. According to the Fisherian theory, we wish to test *one* only hypothesis - called H_0, the *Null Hypothesis* - about the distribution of the data. We can formalize such hypothesis as:

$$H_0\colon\ f \sim f_0 \quad (1)$$

or, if it is about only one parameter of the distribution:

$$H_0 : f_\theta \sim f_{\theta_0};\ H_0 : \theta \in \Theta_0 \quad (2)$$

An example of (1) might be H_0 : X $\sim$ $ga(\alpha, \beta)$, while an example of (2), in the same context, might be $H_0 : \alpha = \alpha_0$, where α_0 consists in a fixed value of α of interest. One of the main differences between Fisher's and Neyman - Pearson's approaches relies in the absence, in the former, of the so called *Alternative Hypothesis*, which instead covers a central role in the latter. Indeed, according to the Fisherian approach, the only goal of the researcher is to check whether the empirical evidence supports or not the Null Hypothesis, without specifying how he/she should behave in the case in which it doesn't. It is in this context that the $P - Value$ (or "observed significance level") is defined and employed.

Definition 1: P - Value Let consider a sample of $n \geq 1$ observations and $T(\boldsymbol{X})$ to be a test statistic, extreme values of which consist in strong evidence against the Null Hypothesis. Given an observed realization $\boldsymbol{x} \in \mathcal{X}$ and, consequently, an observed value of the statistic $T(\boldsymbol{x}) = t$, we define the $p - value$ as the probability of the chosen summary statistic $T(\boldsymbol{X})$ to assume values equal or more extreme of its observed realization t under the Null Hypothesis $H_0 : \theta \in \Theta_0$. Formally:

$$p - value\ (\boldsymbol{x}) = \sup_{\theta \in \Theta_0} P(\ |\ T(\boldsymbol{X})\ |\ \geq |\ t\ |\ )$$

It can be proved that if $T(\cdot)$ is continuous, then $p - value\ (\boldsymbol{x}) \sim U(0, 1)$.

We now go through a couple of relevant examples of its calculation.

Example 1 Let consider a sample $\boldsymbol{X} = (X_1,....,X_n)$ of $n \geq 1$ i.i.d. observations distributed according to a Normal distribution of unknown parameter θ and known σ^2, $X_i \sim N(\theta, \sigma^2)$. We will have that $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\theta, \frac{\sigma^2}{n})$. Hence, defining the pivotal quantity:

$$T(\overline{X}) = \frac{\sqrt{n}|\overline{X} - \theta|}{\sigma}$$

it follows that the p - value, under the Null Hypothesis $H_0 : \theta = \theta_0$ becomes:

$$
\begin{aligned}
P_{\theta_0}(\mid T(X) \mid \geq \mid t \mid) &= P(\ \tfrac{\sqrt{n}|\overline{X}-\theta_0|}{\sigma} \geq \mid t \mid) \\
&= 1 - P(\ \tfrac{\sqrt{n}|\overline{X}-\theta_0|}{\sigma} < \mid t \mid) \\
&= 1 - P(-\mid t \mid < \tfrac{\sqrt{n}(\overline{X}-\theta_0)}{\sigma} < \mid t \mid) \\
&= 2(\ 1 - \Phi(\mid t \mid)\ )
\end{aligned}
$$

where Φ is the standard normal cumulative distribution function.

Example 2 We will now consider the computation of the p - value in the context of the linear regression models (which will also be briefly presented). Such an example will turn out to be very useful latter to properly understand the practical consequences of the misinterpretation and misuse of the p - value.

A *Linear Regression Model* consists in a statistical model in which a *dependent variable* Y is expressed as the linear combination of $k \geq 1$ *independent variables* - each of them weighted through a coefficient β - and a random variable E. Formally, the model is:

$$\mathrm{Y} = \boldsymbol{X'\beta} + \mathrm{E}$$

Where $\boldsymbol{X} = (\mathrm{X}_1,....,X_k)'$ is a $k \geq 1$ dimensional random vector and $\boldsymbol{\beta} = (\beta_1,....,\beta_k)'$ is a $k \geq 1$ dimensional vector of parameters and E is a univariate random variable. In order to provide an unbiased estimation of these k parameters, we need to make some further assumptions about the model and the sampling mechanism. In particular, given a sample of $n \geq 1$ observations, we need to assume that the $k + 1$ dimensional random vectors in the sequence $\{(y_i, x_{i,1}, ..., x_{i,k})\}_{i=1}^{n}$ are independently distributed. At this point, defining X as an n x k matrix whose $x_{i,j}$ element represents the value of the j-th independent explanatory variable for the i-th individual of the sample, the model can be written as:

$$\boldsymbol{Y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{E}$$

Where $\boldsymbol{Y} = (Y_1,....,Y_n)$ is the vector of the observed dependent variable Y_i for each individual of the sample, X is the matrix previously described, $\boldsymbol{\beta} = (\beta_1,...,\beta_k)$ is the parameters vector

and $\boldsymbol{E} = (E_1,....,E_n)$ is the random vector containing the n realizations of the error variables E_i for each individual.

If we consider to hold the following three assumptions:

1) Linearity in the parameters

2) X to be *full column rank*, that is $rank(X) = k$

3) *Strict exogeneity* of the k variables in the matrix X, that is: $E(\boldsymbol{E} \mid X) = 0$

the so called *Ordinary Least Squares* estimator $\boldsymbol{b_{ols}}$, given by:

$$\boldsymbol{b_{ols}} = (X'X)^{-1}X'Y$$

is unbiased for $\boldsymbol{\beta}$. Indeed:

$$E(\boldsymbol{b_{ols}}|X) = E[(X'X)^{-1}X'Y|X] = E[(X'X)^{-1}X'(X\boldsymbol{\beta} + \boldsymbol{E})|X] = \boldsymbol{\beta} + (X'X)^{-1}X'E(\boldsymbol{E}|X)$$
$$= \boldsymbol{\beta}$$

and applying the Law of Iterated Expectations:

$$E(\boldsymbol{b_{ols}}) = E_X[E(\boldsymbol{b_{ols}}|X)] = E(\boldsymbol{\beta}) = \boldsymbol{\beta}.$$

If we also add two more assumptions:

4) *Conditional Homoskedasticity*: $Var(\boldsymbol{E}|X)=\sigma^2 I_n$ (where I_n is th n dimensional identity matrix)

5) *Conditional Normality*: $\boldsymbol{E}|X \sim N(0, \sigma^2 I_n)$

it can be proved that:

$$\boldsymbol{b_{ols}}|X \sim N(\boldsymbol{\beta}, \sigma^2(X'X)^{-1}) \Leftrightarrow \frac{\boldsymbol{b_{ols}}-\boldsymbol{\beta}}{\sqrt{\sigma^2(X'X)^{-1}}}|X \sim N(0,1)$$

and defining $(X'X)^{-1}_{ii}$ as the i - th element of the main diagonal of the k x k matrix $(X'X)^{-1}$, b_i as the i - th element of the k dimensional vector $\boldsymbol{b_{ols}}$ and similarly β_i for $\boldsymbol{\beta}$, it follows that:

$$b_i|X \sim N(\beta_i, \sigma^2(X'X)^{-1}_{ii}) \Leftrightarrow \frac{b_i-\beta_i}{\sqrt{\sigma^2(X'X)^{-1}_{ii}}}|X \sim N(0,1).$$

Hence, if σ^2 is known, a Null Hypothesis $H_0 : \beta_i = \beta_i^*$ can tested considering the previous quantity. However, if the variance σ^2 is not known, defining the *Ols Residual* as:

$$\boldsymbol{e} = \boldsymbol{y} \text{ - } X\boldsymbol{b_{ols}}$$

that is, an n dimensional vector containing the differences, for each individual, between the observed value of the variable Y and that estimated by the Ols coefficient ($\mathrm{X}\boldsymbol{b_{ols}}$), we can replace σ^2 with its unbiased estimator $s^2 = \frac{\boldsymbol{e'e}}{n-k}$, obtaining the quantity:

$$t_i = \frac{b_i - \beta_i^*}{\sqrt{s^2 (X'X)^{-1}_{ii}}}$$

which we wish to prove to be distributed according to a $T - \textit{Student}$ distribution of n - k degrees of freedom under the Null Hypothesis $\beta_i = \beta_i^*$ (note that since the distribution of the T does not depend on the sample information, we can directly consider t_i, since it and $t_i|X$ follow the same distribution).

What usually is investigated in research papers of all fields and is presented in econometric outputs is the p - value associated with the Null Hypothesis $H_0 : \beta_i^* = 0$, that is the hypothesis of absence of correlation between the dependent variable and a specific one of all the $k \geq 1$ regressors. If such a correlation exists, keeping fixed the values of all the other coefficients, we have that its density is measured by the magnitude of the Ols estimated coefficient presented in the same output. Formally, the computed (and presented) p - value is calculated as:

$$\begin{aligned} P_{\beta_i=0}(\ |\ T_{n-k}\ |\ \geq |\ t_i\ |\ ) &= P_{\beta_i=0}[(T_{n-k} \geq |\ t_i\ |) \cup (T_{n-k} \leq \text{ - } |\ t_i\ |)] \\ &= 1 \text{ - } P_{\beta_i=0}(\ \text{ - } |\ t_i\ | \leq T_{n-k} \leq |\ t_i\ |\ ) \\ &= 2P(\ T_{n-k} \geq |\ t_i\ |\ ) \end{aligned}$$

where the values of the latter probability are tabulated. For instance, if the degrees of freedom are n - $k = 10$ and $t_i = 1.812$, the p - value will be $2P(\ T_{10} \geq 1.812\ ) = 0.1$.

We have already said that the misinterpretation and misuse of the p - value become particularly relevant in the regression contexts similar to that presented above. Indeed, in this context, the wrong interpretation of the p - value yields to an overestimation, by the researchers, of the evidence against the Null Hypothesis, that is the absence of correlations between the dependent variable Y and the explanatory variable X_j. As a proof of this, we will see a re - analysis of a work made by a team of researchers - the Open Science Collaboration - whose goal was to test the reproducibility of the econometric results based on the p - value in psychology. What they found out was that of 100 experiments in which statistically significant results were reported in the 97% of the original studies, once replicated, only the 36% reported again statistically significant results, and that the magnitude of the replicated estimations was half of that in the original study. Their conclusion was that a p - value of 0.05, commonly considered as a good threshold to reject the Null Hypothesis, consists essentially in a *substantial support* in favor of the Null Hypothesis.

However, it has to be clarified that the poor performances of the p - value based tests are not due to the p - value itself as a measure of significance, but to the fact that it has no long - run frequentist implications. Indeed, consider that the p - value assumes a central role in the Fisherian theory, where Fisher himself rejected inverse probability methods assessing the probability of a an hypothesis to be true given an observed sample. As a consequence, he was eager to develop a more objective approach, based on assessing the probability of observing a given sample assuming an hypothesis H to be true. Behind this approach, there was his personal view that it is possible to "argue from consequences to causes, from observations to hypotheses" (Fisher 1966). Hence, $significance\ testing$ essentially consists in a procedure to compute the probability of observing a given sample under a Null Hypothesis set by the researcher which *totally* specifies the distribution from which such sample is supposed to arise. If, in particular, the test is focused on testing the existence of an effect or a relationship between two variables, the significance testing procedure will assess the probability of

the observed sample realization to arise under the hypothesis of no relationship.

The logic behind the computation of the p - value is that the Null Hypothesis will be "disproved" if the sample estimate deviates from the mean of the sampling distribution under H_0 by a more than a specified criterion, which Fisher (1966) suggested to be a 0.05 probability. Hence, according to such a suggestion, the Rejection Region of the Null Hypothesis consists in the set of all those observed samples $\boldsymbol{x}$ which produce a $p - value$ smaller or equal than 0.05. However, despite Fisher himself suggested such a threshold, he always clarified that in some particular contexts, whose evaluation is discretionary and left to the researchers, a threshold of 0.02 or 0.01 might be preferable. It can now be understood why trying to interpret the p - value as the Posterior Probability of the Null Hypothesis (which consists in another typical misinterpretation of it) consists in a theoretical mistake, which (we will see) in particular yields to an overestimation of the evidence against H_0. The p - value represents the keystone of the Fisherian approach to testing, which born as an *opposite* alternative to the inverse probability methods. At the same time, given that Fisher never talked about an Alternative Hypothesis and, therefore, did not consider any kind of error due to the rejection of a true Null or to the not rejection of a false Null (two concepts developed in the Neyman - Pearson theory and that will be properly described later), trying to interpret the p - value also as the probability of committing one of these mistakes (in particular rejecting a true Null Hypothesis) consists in another theoretical mistake, which also contributes to the reproducibility issues which have been previously mentioned. Indeed, as we will see, in the Neyman - Pearson theory α consists in the $ex\ ante\ fixed$ maximum probability of rejecting a true Null Hypothesis and therefore it can'te be associated with the $data\ based\ p - value(\boldsymbol{x})$. At this point, in order to further discuss the misinterpretation of the p - value and its consequences, it becomes necessary to discuss the Neyman - Pearson test theory and its underlying inductive behavior.

Neyman - Pearson hypothesis testing and inductive behavior

The Neyman - Pearson statistical methodology originally born as an attempt to improve Fisher's theory, relying though on a totally different approach. The main difference between Neyman - Pearson's and Fisher's theories consists in the clarification, in the former, of an *Alternative Hypothesis* H_1, which instead is totally absent in the Fisherian context and works. Indeed, according to Neyman, such a clarification consisted in a necessary extension of the Fisherian Null Hypothesis since, according to him, despite Fisher has never made any explicit reference to the Alternative Hypothesis, it was in a sense *implied* by the theory and *subconsciously* taken in consideration by Fisher himself. As a point of fact, when describing the $p - value$ and assessing that values of it which fall below the 0.05 threshold "disprove" the Null Hypothesis, it can't be denied that indeed a "somehow implicit" alternative is considered. However, it remains something that Fisher never explicited, nor was interested in testing: his only goal was to assess whether the data supported or not the Null Hypothesis, without expliciting what instead they support if the they disproved it. Intuitively, in the Fisherian approach, an implied - however not specified nor tested for - alternative hypothesis might consist in the *complementary* model to the one tested under the Null Hypothesis. However, it is impossible to formalize the idea of a complementary model. If, for instance, the computed p - value disproves and provides empirical support against the Null Hypothesis of the data to be distributed according to a standard normal, what should be considered as the complementary model? A model with a difference mean and/or variance? A gamma distribution? Clearly, it is impossible to say it. As a consequence, it appears that if Fisher had in mind two possible models for the data, he would have thought two $different$ tests in which the two competing models represented the Null Hypothesis, and then assessing in which of the two cases the p - value was lower. In the previous example, if the p - value provided empirical support against a standard normal model and his personal idea was that

this was because the mean of the population is in reality different from 0, he would have elaborated a new test with a Null Hypothesis H_0 : X $\sim N(\mu,1)$, $\mu \neq 0$. Instead, Neyman and Pearson would have immediately considered a *single* test in which the two competing models represented the two different hypotheses, with the model believed to be the most likely, the one which is true until otherwise proven, under the Null Hypothesis. Formally, according to the example, the test Neyman and Pearson would have performed is:

$$H_0 : \text{X} \sim N(0,1) \text{ vs. } H_1 : \text{X} \sim N(\mu,1),\ \mu \neq 0.$$

Hence, reconnecting to what as been said above, in the Neyman - Pearson context the *Null Hypothesis* becomes the one which is true until otherwise proven, representing somehow the *status quo*, while the competing *Alternative Hypothesis* consists in the one which requires strong empirical evidence against H_0 to be accepted.

At this point, in order to be able to dig deeper with the analysis of the Neyman - Pearson testing approach and its differences with the Fisherian one, some definition about the nature of the different possible kinds of hypotheses have to be done.

1.1 Definitions and properties of the tested hypotheses

We now provide some definitions necessary to fully understand test theory. First, let consider X to be a random variable having density X $\sim f(\text{x} \mid \theta)$, $\theta \in \Theta \subseteq \mathcal{R}^k$, $k \in \mathcal{N}$. A *Parametric Hypothesis* consists in a statement about the value of one or more parameters, while a *Non Parametric Hypothesis* consists in a statement about the distribution itself.

An example of the former, in a Null vs. Alternative Hypothesis framework, can be described as:

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1\ \theta \in \Theta_1, \text{ with } \Theta_0 \cup \Theta_1 = \Theta,\ \Theta_0 \cap \Theta_1 = \emptyset$$

while, a general example for the latter might consist in:

$$H_0 : \mathrm{X} \sim f_0 \text{ vs. } H_1 : \mathrm{X} \sim f_1,\ f_0 \neq f_1.$$

Finally, in the parametric context, we define as *Simple* the Hypothesis which totally specifies the distribution of the observations, while *Composite* the Hypothesis which doesn't, and that just tells if the parameter is greater or smaller of a given value (*Unilateral Composite Hypothesis*) or just different from it (*Bilateral Composite Hypothesis*).
Formally, assuming the Null Hypothesis to be of the kind $\theta \in \Theta_0 \subseteq (-\infty, \theta_0]$ (resp. $\Theta_0 \subseteq [\theta_0, \infty)$), if the Alternative is a Unilateral Composite Hypothesis, then the tested hypotheses will become:

$$H_0\text{: } \theta \leq \theta_0 \text{ vs. } H_1 : \theta > \theta_0 \ (resp.\ H_0 : \theta \geq \theta_0 \text{ vs. } H_1 : \theta < \theta_0)$$

while, instead, assuming the Null Hypothesis to be of the kind $\theta \in \Theta_0 = \{\theta_0\}$ (or, more easily, $\theta = \theta_0$), if the Alternative consists of a Bilateral Composite Hypothesis, the tested hypotheses will be:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0$$

which consist essentially in the type of tested hypotheses that have already been presented previously in the Normal model example. Assuming that we're considering a sample $\boldsymbol{x}$ of size $n \geq 1$, and defining the sample space as $\mathcal{X}$, we will have that the test will partition the sample space in two areas, one such that if our observed sample falls in it we reject H_0, called *Rejection Region* R, and a complementary one which is called *Acceptance Region* A: the two regions will be such that $R \cap A = \emptyset$. A test is customary defined by its Rejection Region.

Given the above provided definitions, it can now be argued why the main characteristic of the Neyman - Pearson testing theory is the *inductive behavior* on which it is based, which is translated in a *decisional* approach towards the hypotheses. Indeed, rejecting or not an

hypothesis becomes just a *decision* - made according to an *objective* mathematical decisional rule - which, when is made, does not reflect in any way the final personal idea that the researcher has developed toward the hypotheses. In other words, a Neyman - Pearson test consists in a rule based over the sample which tells whether to reject or not H_0, which does not imply that the researcher believes such an hypothesis to be false or not. Hence, we talk about *inductive behavior* since, given the results provided by the test, the researcher adjusts his/her personal opinion toward the hypotheses. Neyman (1950) himself stressed the fact that such an adjustment is partly conscious and partly subconscious. In particular, it is conscious in the part in which the mathematical rule prescribes to reject/not reject the hypothesis, while it is subconscious in the part in which the researcher inevitably adjusts his/her personal belief toward the hypotheses, despite such an adjustment is totally unrelated with the final decision to reject or not. This feature of the Neyman - Pearson testing approach consists in something totally disruptive with respect to the Fisherian approach in which, once that the p - value has been observed, the researcher is $free$ to reject or not the Null Hypothesis. Indeed, 0.05 is just a suggested threshold and not a rule, since Fisher himself clarified that it can be changed according to the researcher's beliefs. Another difference between the two approaches consists is more epistemological: we have already said that the Fisherian approach is an *inductive* one, since, according to it, we start from the particular, that is the observed sample, and then argue about the general, that is the hypothesis and, therefore, the model. In the Neyman - Pearson approach, instead, the above described inductive behavior implies a *deductive* approach, since the researcher starts arguing from the general, that is his/her personal belief about the hypotheses which make possible to define the Null and Alternative hypotheses, and make conclusions about the particular case, applying the decision rule.

As a consequence of its structure, the Neyman - Pearson approach produces the possibility of committing two possible errors which can be made when deciding whether to reject or not

the Null Hypothesis. We define as *I Type Error* the error consisting in rejecting the Null Hypothesis when it is true, while *II Type Error* the error consisting in not rejecting the Null Hypothesis when it is false. The I Type Error is usually regarded as the most severe, and it is indeed the one whose probability of being committed is ex ante decided and kept very low, customary no more than 0.10.

Decision	True H_0	False H_0
Reject H_0	I Type Error	correct
Do not Reject H_0	correct	II Type Error

Table 1: Decisions and possible errors

In the particular case in which the hypotheses are of the kind $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, we will define as α the probability of committing the I Type Error and β as the probability of committing the other. Formally:

$$\alpha = Pr\{\boldsymbol{X} \in R \mid \theta = \theta_0\} \Leftrightarrow 1 - \alpha = Pr\{\boldsymbol{X} \in A \mid \theta = \theta_0\}$$

and similarly:

$$\beta = Pr\{\boldsymbol{X} \in A \mid \theta = \theta_1\} \Leftrightarrow 1 - \beta = Pr\{\boldsymbol{X} \in R \mid \theta = \theta_1\}$$

Decision	True H_0	False H_0
Reject H_0	α	$1 - \beta$
Do not Reject H_0	$1 - \alpha$	β

Table 2: Decisions and possible errors in case of two simple hypotheses

It can be observed that the two type error probabilities are inversely related, in the sense that if our aim is to reduce one of the two, then necessarily we will have to increase the other. In order to observe this phenomenon more precisely, we first have to introduce the *Power Function* and then formalize the concept of *Test Function*.

Definition 2: Power Function The function $Q : \Theta \to [0,1]$ which defines the probability of rejecting H_0 depending on θ, that is:

$$Q(\theta) = \Pr_\theta\{\boldsymbol{X} \in R\}$$

is called *Power Function*. If the hypotheses are of the type $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, then the Power Function becomes:

$$Q : \{\theta_0, \theta_1\} \to \{\alpha, 1-\beta\} \Leftrightarrow Q(\theta_0) = \alpha,\ Q(\theta_1) = 1-\beta.$$

At this point, it can be observed why the two error probabilities, for a fixed sample size, are inversely related: the more a test tends to be conservative, that is the more it tends to not reject H_0 (hence lowering the probability of rejecting it when it is true), the more it is exposed to the risk of not rejecting it when it is false.

To conclude this introductory part about test theory, we define the *Test Function* ψ as a function which tells us the probability of rejecting the Null Hypothesis when $\boldsymbol{x}$ is observed. More precisely, we will define a *Non Randomized Test* a test which is characterized by a test function $\psi : \mathcal{X} \to \{0,1\}$ which assumes value 1 if $\boldsymbol{x}$ belongs to the Rejection Region R and 0 otherwise. That is:

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \boldsymbol{x} \in R \\ 0 & \boldsymbol{x} \in A \end{cases}$$

where $R \cup A = \mathcal{X}$. Conversely, a *Randomized Test* is defined as a test which is characterized by a test function $\psi : \mathcal{X} \to [0,1]$ which assumes value 1 if $\boldsymbol{x}$ belongs to the Rejection Region R, value 0 if $\boldsymbol{x}$ belongs to the Acceptance Region A and $c \in (0,1)$ otherwise. Clearly, in this case, $R \cup A \subset \mathcal{X}$, that is:

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \boldsymbol{x} \in R \\ 0 & \boldsymbol{x} \in A \\ c & otherwise \end{cases}$$

The idea is that if the $n \geq 1$ observations are extracted according to a discrete distribution it might be not possible to obtain the fixed desired level α. However, given that conceptually almost anything changes between the two types of tests, for the sake of simplicity we will mainly focus on non randomized tests.

1.2 Error probability and UMP tests

According to the Neyman - Pearson decisional approach it becomes possible, for the researcher, to ex ante fix the maximum probability level of committing the I Type Error he/she is disposed to suffer, and then to choose the test which, among all those whose probability of committing the I Type Error is at most the fixed α, guarantees the lowest probability of committing the II Type Error. First, let define $\mathcal{C}_\alpha$ as the class of all the tests whose probability of committing the I Type Error is at most α. Formally:

$$\mathcal{C}_\alpha = \{\psi : \mathcal{X} \to [0,1] : \sup_{\theta \in \Theta_0} Q_\psi(\theta) \leq \alpha\}.$$

The test ψ^* in this class which also minimize the probability of committing the II Type Error for each $\theta \in \Theta_1$ is called $Uniformly\ Most\ Powerful$ level α test, and can be formally defined as the test ψ^* such that:

$$\psi^* \in \mathcal{C}_\alpha \wedge Q_{\psi^*}(\theta) \geq Q_\psi(\theta), \forall\ \theta \in \Theta_1, \forall\ \psi \in \mathcal{C}_\alpha.$$

At this point, it can be understood why the Neyman - Pearson behavioral approach has become the most popular among researchers and scholars. Indeed, in this context, the researcher not only is provided with an objective decision rule which tells him/her whether to

reject or not the Null Hypothesis, but also can find and employ (in some particular cases) a test which, if some particular properties are satisfied, is also the most powerful for the ex ante fixed level α he/she has set. In regard with this last consideration, we now present the Neyman - Pearson Lemma which states which is the $Most\ Powerful$ test. Then, we will describe the Karlin - Rubin test, which always consists in the $Uniformly\ Most\ Powerful$ (UMP) test in case of two unilateral composite hypotheses and we will conclude presenting the Likelihood Ratio Test, which provides a reasonable test in those cases in which a UMP test can't be found.

Lemma 1: The Neyman - Pearson Lemma: Let ψ be a test for the hypotheses $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ characterized by Rejection and Acceptance Regions given by:

$$R_\alpha = \{\boldsymbol{x} \in \mathcal{X} : k_\alpha \mathcal{L}(\boldsymbol{x}; \theta_0) < \mathcal{L}(\boldsymbol{x}; \theta_1)\} \Leftrightarrow A_\alpha = \{\boldsymbol{x} \in \mathcal{X} : k_\alpha \mathcal{L}(\boldsymbol{x}; \theta_0) > \mathcal{L}(\boldsymbol{x}; \theta_1)\}$$

where $\boldsymbol{x}$ is a sample of size n ≥ 1 and k_α is such that $Q_\psi(\theta_0) = \alpha$. Then, it can be proved that ψ is the Most Powerful Test of level α.
We now go through an example, in which it is shown how to determine the *Rejection Region*, the necessary value of k for an ex ante fixed α and is also formally observed the above mentioned existing inverse relation between the two error probabilities.

Example 3 Let consider a sample $\boldsymbol{X} = (X_1,....,X_n)$ of $n \geq 1$ i.i.d. random variables with $X_i \sim ga(\zeta, \lambda)$ where ζ is known, and that we wish to test:

$$H_0 : \lambda = \lambda_0 \text{ vs. } H_1 : \lambda = \lambda_1,\ \lambda_1 > \lambda_0$$

observing that:

$$\frac{\mathcal{L}(\boldsymbol{x};\lambda_1)}{\mathcal{L}(\boldsymbol{x};\lambda_0)} > k \Leftrightarrow \frac{\lambda_1^{n\zeta}[\Gamma(\zeta)]^{-n}\prod_{i=1}^n x_i^{\zeta-1}e^{-\lambda_1\sum_{i=1}^n x_i}I(x_{(1)}>0)}{\lambda_0^{n\zeta}[\Gamma(\zeta)]^{-n}\prod_{i=1}^n x_i^{\zeta-1}e^{-\lambda_0\sum_{i=1}^n x_i}I(x_{(1)}>0)} > k$$

$$\Leftrightarrow (\tfrac{\lambda_1}{\lambda_0})^{n\zeta} \mathrm{e}^{(-\lambda_1+\lambda_0)\sum_{i=1}^{n} x_i} > k$$

$$\Leftrightarrow e^{(-\lambda_1+\lambda_0)\sum_{i=1}^{n} x_i} > k(\tfrac{\lambda_1}{\lambda_0})^{-n\zeta}$$

$$\Leftrightarrow (\text{-}\lambda_1 + \lambda_0) \sum_{i=1}^{n} x_i > \ln[k(\tfrac{\lambda_1}{\lambda_0})^{-n\zeta}]$$

$$\Leftrightarrow \sum_{i=1}^{n} x_i < (-\lambda_1 + \lambda_0)^{-1} \ln[k(\tfrac{\lambda_1}{\lambda_0})^{-n\zeta}] \equiv k$$

we can finally define the Rejection Region as:

$$R = \{\boldsymbol{x} \in \mathcal{X} : \textstyle\sum_{i=1}^{n} X_i < k\}.$$

Hence, we will have that k will be such that:

$$\alpha = Pr\{\textstyle\sum_{i=1}^{n} X_i < k \mid \lambda = \lambda_0\} \Leftrightarrow 1 \text{ - } \alpha = Pr\{\sum_{i=1}^{n} X_i \geq k \mid \lambda = \lambda_0\}$$

$$\beta = Pr\{\textstyle\sum_{i=1}^{n} X_i > k \mid \lambda = \lambda_1\} \Leftrightarrow 1 \text{ - } \beta = Pr\{\sum_{i=1}^{n} X_i \leq k \mid \lambda = \lambda_1\}.$$

Once that k has been fixed such that the maximum probability of committing the I Type Error is at most α, β can residually determined. Moreover, it can be observed that the two error probabilities are inversely related.

In this example, knowing that:

$$T = \textstyle\sum_{i=1}^{n} X_i \sim ga(n\zeta,\, \lambda)$$

we have that:

$$\alpha = \textstyle\int_0^k \frac{\lambda_0^{n\zeta}}{\Gamma(n\zeta)} t^{n\zeta-1} e^{-t\lambda_0} dt \Leftrightarrow 1 \text{ - } \alpha = \int_k^{+\infty} \frac{\lambda_0^{n\zeta}}{\Gamma(n\zeta)} t^{n\zeta-1} e^{-t\lambda_0} dt$$

$$\beta = \textstyle\int_k^{+\infty} \frac{\lambda_1^{n\zeta}}{\Gamma(n\zeta)} t^{n\zeta-1} e^{-t\lambda_1} dt \Leftrightarrow 1 \text{ - } \beta = \int_0^k \frac{\lambda_1^{n\zeta}}{\Gamma(n\zeta)} t^{n\zeta-1} e^{-t\lambda_1} dt.$$

In this case, the problem can be solved without explicit computations exploiting the existing relation between the gamma distribution and the χ^2 distribution (whose values are tabulated), that is:

$$\sum_{i=1}^{n} X_i \sim ga(\mathrm{n}\zeta,\, \lambda) \Leftrightarrow 2\lambda \sum_{i=1}^{n} X_i \sim ga(\tfrac{2n\zeta}{2},\, \tfrac{1}{2}) = \chi^2_{2n\zeta}.$$

Defining $\chi^2_{n,\alpha}$ as the quantity such that:

$$P(\chi^2_n > \chi^2_{n,\alpha}) = \alpha$$

we will have that in order to determine the value of β starting from a for a fixed value of α, first we explicit the value of k:

$$1 - \alpha = Pr\{\textstyle\sum_{i=1}^{n} X_i \geq k \mid \lambda = \lambda_0\}$$

$$= Pr_{\lambda_0}\{2\lambda_0 \textstyle\sum_{i=1}^{n} X_i \geq 2\lambda_0 k\}$$

$$= Pr_{\lambda_0}\big\{2\lambda_0 \textstyle\sum_{i=1}^{n} X_i \geq \chi^2_{2n\zeta,1-\alpha}\big\}$$

so we can conclude that:

$$\alpha = Pr\big\{2\lambda_0 \textstyle\sum_{i=1}^{n} X_i < \chi^2_{2n\zeta,1-\alpha}\big\} \Leftrightarrow 1 - \alpha = Pr\big\{2\lambda_0 \textstyle\sum_{i=1}^{n} X_i \geq \chi^2_{2n\zeta,1-\alpha}\big\}$$

$$\beta = Pr\big\{2\lambda_1 \textstyle\sum_{i=1}^{n} X_i > 2\lambda_1 \chi^2_{2n\zeta,1-\alpha}\big\} \Leftrightarrow 1 - \beta = Pr\big\{2\lambda_1 \textstyle\sum_{i=1}^{n} X_i \leq 2\lambda_1 \chi^2_{2n\zeta,1-\alpha}\big\}$$

where both $\mathrm{W} = 2\lambda_0 \sum_{i=1}^{n} X_i$ and $\mathrm{V} = 2\lambda_1 \sum_{i=1}^{n} X_i$ are distributed according to a chi - squared distribution with $2\mathrm{n}\zeta$ degrees of freedom. In order to assess the impact that the sample size has on the two error probabilities, let consider the case in which $n = 4$ and $\zeta = \frac{1}{2}$, and the values of the parameter λ under the two hypotheses to be:

$$H_0 : \lambda = 2 \text{ vs. } H_1 : \lambda = 4.$$

For instance, if $\alpha = 0.01$, then:

$$0.99 = P(\chi^2_4 > \chi^2_{4,0.99}) \Leftrightarrow \chi^2_{4,0.99} = 0.297$$

from which it follows that:

$$\beta = P(\chi^2_4 > 2.376) = 0.67$$

It will follow that the four error probabilities will be, for three different ex ante fixed values (0.01, 0.05, 0.10) will be:

α	1 - α	β	1 -β
0.01	0.99	0.67	0.33
0.05	0.95	0.22	0.78
0.10	0.90	0.07	0.93

Table 3: Error probabilities in the case in which n = 4, $\zeta = \frac{1}{2}, \lambda_1 = 4$

It is worth observing how the values of β, for the same fixed values of α, sharply decrease if the sample size is doubled:

α	1 - α	β	1 -β
0.01	0.99	0.10	0.90
0.05	0.95	0.005	0.995
0.10	0.90	0.0005	0.9995

Table 4: Error probabilities in the case in which n = 8, $\zeta = \frac{1}{2}, \lambda_1 = 4$

The test suggested by Neyman and Pearson, we have said, is Most Powerful for the hypotheses $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$. However, it remains such also in the case in which one of two hypotheses is composite (provided that the rejection area doesn't depend on the value of θ_1) or of two non parametric hypotheses of the kind $H_0 : \mathrm{X} \sim f_0$ vs. $H_1 : \mathrm{X} \sim f_1$. Indeed, the basic idea is that of comparing the likelihood under the two hypotheses, and then choosing the hypothesis yielding the greater likelihood, with, in particular, for the Alternative Hypothesis to be chosen, being k_α times greater than the likelihood under the Null Hypothesis, something which remarks the generally preferred approach toward more conservative tests. Moreover, note that Neyman - Pearson Lemma based test can be employed also in the case in which the observations are not identically and independently distributed, provided that we know the joint distribution of the sample.
Now, in order to describe the *Karlin - Rubin Test*, it is necessary to previously define when a family of distributions presents a *Monotone Likelihood Ratio*.

Definition 3: Monotone Likelihood Ratio: Let $\{f_\theta, \theta \in \Theta \subseteq \mathcal{R}\}$ be a family of distributions. Then, such family is said to have a *monotone likelihood ratio* in the unidimensional statistic $T(\boldsymbol{X})$ if:

$$\forall\ \boldsymbol{x} \in \mathcal{X},\ \forall\ (\theta^*,\ \theta) \in \Theta^2 \text{ s.t. } \theta^* > \theta,\ \frac{\mathcal{L}(\boldsymbol{x};\theta^*)}{\mathcal{L}(\boldsymbol{x};\theta)} \text{ is not increasing/not decreasing in } T(\boldsymbol{x})$$

Definition 4: Karlin - Rubin Test Let consider a family of distributions $\{f_\theta, \theta \in \Theta \subseteq \mathcal{R}\}$ having a not decreasing monotone likelihood ratio in the unidimensional statistic $T(\boldsymbol{X})$. Then, if we wish to test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$ (resp. $H_1 : \theta < \theta_0$), the test ψ characterized by the Rejection and Acceptance Regions defined as:

$$R_\alpha = \{\boldsymbol{x} \in \mathcal{X} : T(\boldsymbol{X}) > k_\alpha\} \Leftrightarrow A_\alpha = \{\boldsymbol{x} \in \mathcal{X} : T(\boldsymbol{X}) < k_\alpha\}$$

$$(\text{resp. } R_\alpha = \{\boldsymbol{x} \in \mathcal{X} : T(\boldsymbol{X}) < k_\alpha\} \Leftrightarrow A_\alpha = \{\boldsymbol{x} \in \mathcal{X} : T(\boldsymbol{X}) > k_\alpha\}\ )$$

where $\boldsymbol{x}$ is a sample of size n ≥ 1 and k_α is such that $Q_\psi(\theta_0) = \alpha$, is Uniformly Most Powerful of level α. Rejection and Acceptance Regions have to be exchanged if the the likelihood ratio is not increasing in the unidimensional statistic $T(\boldsymbol{X})$.

One of the most interesting features of the latter category of tests is that they are UMP also in the case of two composite hypotheses of the type $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta \geq \theta_0$. Indeed:

$$\alpha = \sup_{\theta \in \Theta_0} Q_\psi(\theta) = \sup_{\theta \in \Theta_0} P(T(\boldsymbol{X}) > k_\alpha) = P_{\theta_0}(T(\boldsymbol{X}) > k_\alpha) = Q_\psi(\theta_0).$$

There are, however, cases in which a Uniformly Most Powerful test can't be found. Intuitively, such cases will be those where a simple Null Hypothesis is compared against a two - sided Alternative Hypothesis, that is, in the case of a unidimensional parameter $\theta \in \Theta \subseteq \mathcal{R}$:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0.$$

Defining the test statistic Λ as:

$$\Lambda = \frac{\mathcal{L}(\boldsymbol{x}, \theta_0)}{\sup_{\theta \in \Theta} \mathcal{L}(\boldsymbol{x}, \theta)}$$

a test yielding good results, known as the *Likelihood Ratio Test* is the test characterized by the decision rule:

$$\text{Reject } H_0 \text{ if and only if } \Lambda < k \in (0,1).$$

It is immediate to observe that if k is such that the probability of committing the I Type Error is exactly α and that the hypotheses are both simple, the Likelihood Ratio Test coincides with the test described in the Neyman - Pearson Lemma, and as such is MP.

How to choose α in a decisional framework

Given this brief description of some of the most useful tests which find place within the Neyman - Pearson approach, we now go through the analysis of how the ex ante value of α is chosen and how therefore β is consequently determined in order to minimize the probability of committing the II Type Error. Despite such choice is - as said - arbitrary, some customary values for α are 0.01, 0.05 and, less frequently, 0.10. The choice of such value should also (if not primarily) be made according to which are the expectations and the priorities of the researcher toward the evaluation of the Null Hypothesis. In order to formalize this concept, if we define $L(H_i,H_j)$ as the function yielding the loss derived from accepting H_i being H_j true, (with, $i = 0,1$ and $j = 0,1$), and letting its four possible values to be:

$L(H_i, H_j)$	True H_0	False H_0
Accept H_0	0	$l_{0,1}$
Reject H_0	$l_{1,0}$	0

Table 5: Possible values of $L(H_i,H_j)$

then, it customary holds that:

$$L(H_0,H_0) = L(H_1,H_1) = 0 < L(H_0,H_1) = l_{0,1} < L(H_1,H_0) = l_{1,0}$$

that is, committing the I Type Error (Rejecting true H_0) is regarded as more severe than committing the II Type Error (Not Rejecting - Accepting a false H_0).

Before moving on (whereas we will get back on this loss - based framework when we will compare both Fisher's and the Neyman - Pearson's approaches against the Bayesian one), it is worth noting that if we consider the case of two one sided composite hypotheses:

$$H_0 : \theta \leq \theta_0 \text{ vs. } H_1 : \theta > \theta_0$$

it sounds reasonable to assume that Rejecting H_0 when $\theta = \theta_0 + \varepsilon$ is a lot less severe than Rejecting H_0 when, say, $\theta = \theta_0 + 50\varepsilon$, $\varepsilon > 0$. Hence, it would make sense to consider the two losses as functions of θ itself - that is $l_{0,1} = l_{0,1}(\theta)$ and $l_{1,0} = l_{1,0}(\theta)$ - with, in particular, the first to be increasing in (θ - θ_0) and the former in (θ_0 - θ). In other words, in the former case, if we accept the hypothesis that θ is smaller than (or equal to) θ_0 while the opposite is true, the greater is the real value of θ with respect to θ_0, the higher is the loss we suffer in accepting such wrong hypothesis. A similar arguments holds for the converse case.

In order to understand why the hypotheses are structured so that committing the I Type Error is regarded as more severe, we need once again to recall that the Null Hypothesis is considered the one holding and considered true unless a strong empirical evidence against it is provided by the data. For instance, in the previously formally described econometric framework, the Null Hypothesis states the absence of any form of (linear) relation among the variables. In this kind of context, fixing α very low means that accepting that exists some kind of relation (whose magnitude is assessed through the estimated value of the coefficient) while in reality there's not is much more severe than doing the opposite. The reason why the first kind of error is considered to be more severe than the second has nothing to do with the statistical theory, and has to be found just within the considered context, and is usually related with different valuations.
As an example, let consider the general case in which a new drug has been developed to cure a disease and the researcher is interested in assessing whether it actually works or not. In this case, the Null Hypothesis will be:

$$H_0 : \text{The new cure has no effect in curing the disease.}$$

Moreover, if we define as Y_i the potential outcome of individual i with unobservable characteristics E_i and D_i as a *dummy variable* assuming value 1 if the individual is treated (receives

the new drug) and 0 otherwise, considering the regression:

$$Y_i = \beta D_i + E_i$$

the previous Null Hypothesis of absence of effectiveness of the treatment can be summarized through:

$$H_0 : \beta = 0.$$

Similar examples can for instance be found in economics, where the treatment variable can be considered, for instance, as the implementation/not implementation of an expansionary monetary policy and the dependent variable the GDP, where the goal is trying to assess an eventual empirical relation between the two variables using, as sample, an historical dataset of monetary policies and GDPs. Consider now a different case: a researcher is testing whether a new drug - which has already been proved, by strong empirical evidence, to be effective, and it's ready to be produced and sold - it's related with some dangerous collateral effects which are considered as possible given the nature of the drug itself and of the disease it is expected to cure. In this case, the probability of committing the I Type Error - which would consist in accepting that the drug doesn't produce collateral damages while, in reality, it does - should be kept extremely low.

With regard with this last example,, Berger and Delampady (1987) argue that in some contexts it is actually very hard to assume, as Null Hypothesis, a *total* absence of effectiveness of the new drug or *total* absence of collateral damages. What makes a lot more sense to assume is that actually the new drug has anyway *some* effect, and that produces anyway *some* collateral damages. It follows that what it might be more reasonable to investigate is whether the new drug has more or less than a given level of effectiveness or if produces more or less a given level of collateral damages. Under a statistical point of view, what follows is that instead of a Point Null Hypothesis and a Bilateral Composite Alternative, the authors suggest to consider:

$$H_0 : \mid \theta \text{ - } \theta_0 \mid \leq \varepsilon \text{ vs. } H_1 : \mid \theta \text{ - } \theta_0 \mid > \varepsilon$$

with ε to be small. They show that, in a classical context, it is not difficult to find those values such that these two hypotheses can be approximated by:

$$H_0^* : \theta = \theta_0 \text{ vs. } H_1^* : \theta \neq \theta_0.$$

Indeed, considering the definition of p - value we have given, we will find out that the p - value for testing H_0 vs. H_1 will be:

$$p \text{ - } value(\boldsymbol{x}, \varepsilon) = \sup_{\theta : |\theta - \theta_0| \leq \varepsilon} P_\theta \ (\mid T(\boldsymbol{X}) \mid \geq \mid t \mid).$$

The task, at this point, becomes to find under which conditions $p \text{ - } value(\boldsymbol{x}, \varepsilon) \approx p \text{ - } value(\boldsymbol{x})$. If, for instance, we have $\overline{X} \sim N(\theta, \sigma^2)$ and we consider $T(\overline{X}) = \sqrt{n}\sigma^{-1}(\overline{X} \text{ - } \theta_0)$, it can be proved that:

$$p \text{ - } value(\boldsymbol{x}, \varepsilon) = 2 \text{ - } [\ \Phi(\mid t \mid \text{ - } \tfrac{\varepsilon\sqrt{n}}{\sigma} \ ) + \Phi(\mid t \mid + \tfrac{\varepsilon\sqrt{n}}{\sigma} \ ) \ ]$$

where Φ, as before, represents the cumulative distribution function of a standard normal distribution. If, for instance, our goal is to approximate the interval null by a point one with no more than a 10% error, we will have to determine when:

$$p \text{ - } value(\boldsymbol{x}) \geq 0.9 p \text{ - } value(\boldsymbol{x}, \varepsilon).$$

Assuming to have observed, testing the precise point null, $p \text{ - } value(\boldsymbol{x}) = 0.1$ (which is implied by $t = 1{,}645$) it will follow that:

$$p \text{ - } value(\boldsymbol{x}) = 0.1 \geq 0.9 \left\{ 2 - [\Phi(1.645 - \tfrac{\varepsilon\sqrt{n}}{\sigma}) + \Phi(1.645 + \tfrac{\varepsilon\sqrt{n}}{\sigma})] \right\}$$

$$\Leftrightarrow \tfrac{1}{9} \geq 2 \text{ - } \left[\Phi(1.645 - \tfrac{\varepsilon\sqrt{n}}{\sigma}) + \Phi(1.645 + \tfrac{\varepsilon\sqrt{n}}{\sigma}) \right]$$

$$\Leftrightarrow \tfrac{1}{9} \geq 2 - \left[\Phi(1.645 - \tfrac{\varepsilon\sqrt{n}}{\sigma}) + 1 - \Phi(-1.645 - \tfrac{\varepsilon\sqrt{n}}{\sigma})\right]$$

$$\Leftrightarrow \tfrac{1}{9} \geq 1 - \Phi(1.645 - \tfrac{\varepsilon\sqrt{n}}{\sigma}) + \Phi(-1.645 - \tfrac{\varepsilon\sqrt{n}}{\sigma})$$

$$\Leftrightarrow \Phi(1.645 - \tfrac{\varepsilon\sqrt{n}}{\sigma}) - \Phi(-1.645 - \tfrac{\varepsilon\sqrt{n}}{\sigma}) \geq \tfrac{8}{9}$$

$$\Leftrightarrow \mathrm{P}\left(-1.645 - \tfrac{\varepsilon\sqrt{n}}{\sigma} < Z < 1.645 - \tfrac{\varepsilon\sqrt{n}}{\sigma}\right) \geq \tfrac{8}{9}$$

$$\Leftrightarrow \int_{-1.645-\frac{\varepsilon\sqrt{n}}{\sigma}}^{1.645-\frac{\varepsilon\sqrt{n}}{\sigma}} \tfrac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \geq \tfrac{8}{9}$$

where the latter integral can only be approximated through the Midpoint Rule, the Trapezoid Rule or the Simpson Rule. The value the author report, in this case, is $\frac{\varepsilon\sqrt{n}}{\sigma} = 0.257$. The other values they report are:

t	1.645	1.96	2.576	2.807	3.29	3.89
p - value	0.10	0.05	0.01	0.005	0.001	0.0001
Bound on $\varepsilon\sqrt{n}\sigma^{-1}$	0.257	0.221	0.173	0.160	0.138	0.117

Table 6: Bounds on $\varepsilon\sqrt{n}\sigma^{-1}$ yielding a 10% error in the P - value, Berger and Delampady, (1987).

We will get back on the approximation involving Point Null Hypotheses when we will present the Bayes Factor and how the Posterior Probabilities of the Null Hypothesis perform with respect to the p - value.

1.3 P - value as the lower bound on the I Type Error

In order to go deeper in the analysis of the discrepancies produced by the wrong interpretation of the p - value as a frequentist measure, now it is provided the definition which is given in Lehmann and Romano, (2005), which reflects the mainstream use and interpretation of it.

Definition 5 P - value as a I Type Error probability: Let consider a non randomized test ψ for the single pair of hypotheses $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$, $\Theta = \Theta_0 \cup \Theta_1$, $\Theta_0 \cap \Theta_1 = \emptyset$. Moreover, let assume the test ψ to be based on a real valued statistics $T : \mathcal{X} \rightarrow \mathcal{R}$ (where the sample $\boldsymbol{x} \in \mathcal{X}$ can be of size $n \geq 1$) and to be characterized by a Rejection Region $R_\alpha \subset \mathcal{R}$ depending on α such that:

$$\forall\ \alpha \in (0,1),\ \forall\ \boldsymbol{x} \in \mathcal{X},\ \psi(\boldsymbol{x}) = 1 \Leftrightarrow T(\boldsymbol{x}) \in R_\alpha$$

defining $P^*(T(\boldsymbol{x}) \in R_\alpha)$ as:

$$P^*(T(\boldsymbol{x}) \in R_\alpha) = \sup_{\theta \in \Theta_0} P(T(\boldsymbol{x}) \in R_\alpha)$$

then, the p - $value$ of an observation $\boldsymbol{x} \in \mathcal{X}$ is defined as:

$$p\text{ - }value(\boldsymbol{x}) = \inf_{\{\alpha : T(\boldsymbol{x}) \in R_\alpha\}} P^*(T(\boldsymbol{x}) \in R_\alpha).$$

The main conceptual problem with a definition like this is that, when we have presented the Neyman - Pearson Lemma and the Karlin - Rubin theorem, we have already said that:

$$\alpha = \sup_{\theta \in \Theta_0} Q_\psi(\theta) = \sup_{\theta \in \Theta_0} P(T(\boldsymbol{x}) \in R_\alpha)$$

from which, defining the p - value as the *smallest* probability of committing the I Type Error consists in a theoretical contradiction, since, as we have already siad, in the Neyman - Pearson framework, such probability is ex ante fixed, and therefore can't be data dependent as the p - value. In particular, Fisher himself strongly denied the supposed equivalence of

the p - values and the Neyman - Pearson α level, since the latter consists in a *long run frequency measure* of rejecting H_0 when it is true, a feature the p - value doesn't have.

Another wrong definition of the p - value, similar to the one provided before, is for instance that reported by Gibbson and Pratt (1975), who state that "Reporting a p - value, whether exact or within an interval, in effect permits each individual to choose his/her level of significance as the maximum tolerable probability of the I Type Error". In this regard, Berger and Delampady (1987) strongly insisted that the p - value is *not* a repetitive error rate and that *only* the ex ante fixed level α has the actual frequentist interpretation that a long series of tests of α level will reject a true H_0 no more than $\alpha 100\%$ times, an interpretation absolutely not valid for the *data dependent* p - value. That is, once that the p - value is observed to be smaller than α, its specific value becomes irrelevant.

Note that, since for a fixed prespecified α the Neyman - Pearson decision rule is based on a totally specified Rejection Region determined by the sample - and, in turn, by a quantity which is function of it - it is indeed *correct* to say that we reject H_0 when the p - value is smaller than α. However, in this context, only the Neyman - Pearson interpretation remains valid, and, independently from how much the p - value is smaller than α, the only thing the researcher will be able to conclude is that in repeating testing he/she will have at most $\alpha 100\%$ cases of rejection of a true Null Hypothesis.

Nonetheless, such a procedure might induce the researcher to provide the latter, wrong, interpretation of the p - value, since the temptation to use its value as a way to reinforce the argument and the thesis that he/she desired to prove/disprove (by rejecting the Null Hypothesis) could be very strong. A possible argument - in favor of the good faith of the researchers who confuse the p - value and the Neyman- Pearson α - is that both measures are commonly referred as the *level of significance* of a test. Indeed, the significance provided by the p - value refers to the probability under the Null Hypothesis of observing values equal or more extreme than the computed one, and therefore it consists in a measure of inductive evidence

in a single experiment. Instead, in the Neyman - Pearson context, the aim of the researcher is to minimize the probability of committing the II Type Error given the maximum level of probability he/she is disposed to suffer of committing the I Type Error, whose value α is also referred as the level of significance.

To conclude this part among the differences existing - and too often ignored - between the Fisherian p - value and the Neyman - Pearson's α I Type Error probability, we discuss some of the main advantages and drawbacks of the two approaches.

The main advantage of the Neyman - Pearson approach relays in the objective interpretation that can be done of the results: if a test is of level α, it means that of all the tests performed at the same level no more than the $\alpha 100\%$ of them will prescribe to reject a true Null Hypothesis; this frequentist interpretation, moreover, does not require to repeat the experiment. However, its main drawback is in the fact that α is ex ante fixed, and so the same interpretation holds for all the samples, independently from how much the information provided by the data appear compatible/incompatible with the Null Hypothesis. This, instead, consists in the major advantage of the Fisherian p - value: despite being only a *relative* measure of evidence, it provides more detailed information of *how much* the data are compatible/incompatible with the Null Hypothesis. However, its main drawback consists in the very difficult evaluation in *absolute* terms of the results: such a difficult interpretability, jointly with its wrong frequentist evaluation in the Neyman - Pearson theory, generally ends up consisting in an wrong estimation of the evidence that it provides against the Null Hypothesis (more in particular in an overestimation of the evidence against it). In order to assess how much the misinterpretation and misuse of the p -value can bias the evaluation of the Null Hypothesis, we will proceed comparing it with the *Bayes Factor* and the *Bayesian Posterior Probability* of the Null Hypothesis.

2 Bayesian Analysis and Hypothesis Testing

We have said that the main consequence of the misinterpretation and misuse of the p - value consists in an overestimation of the evidence against the Null Hypothesis. In order to verify it and to find a way to adjust the p - value so that a frequentist interpretation becomes possible, it is necessary an overview of Bayesian Testing. We now describe the general framework in which many possible models are compared, and will then go through the analysis of the Posterior Probability of the Null Hypothesis.

2.1 Bayes Factor and Bayes Tests

Let start with the case in which different models are considered, all belonging to the set $\{\mathcal{M} : M_i, i \in I\}$, which also contains the true one. Defining $P(M_i)$ the *prior belief* about the model M_i, $p(\boldsymbol{x} \mid M_i)$ the integrated likelihood of the data under the model M_i and $p(\boldsymbol{x})$ the marginal density of X, it follows that:

$$\frac{P(M_i|\boldsymbol{x})}{P(M_j|\boldsymbol{x})} = \frac{p(\boldsymbol{x}|M_i)P(M_i)[p(\boldsymbol{x})]^{-1}}{p(\boldsymbol{x}|M_j)P(M_j)[p(\boldsymbol{x})]^{-1}} = \frac{p(\boldsymbol{x}|M_i)P(M_i)}{p(\boldsymbol{x}|M_j)P(M_j)}.$$

In other words, the ratio between the Posterior Probabilities (updated given the information provided by the observed sample) of the models M_i and M_j is equal to the ratio between the integrated likelihoods of the data under the models M_i and M_j times the ratio between the Prior Probabilities of the models M_i and M_j.

Definition 6: Bayes Factor Let consider two hypotheses H_0, H_1 corresponding - respectively - to two possible different models M_0, M_1. Given the observed data $\boldsymbol{x} \in \mathcal{X}$, the *Bayes Factor* in favor of H_0 against H_1 is given by the posterior to prior odds ratio. Formally:

$$B_{0,1}(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|M_0)}{p(\boldsymbol{x}|M_1)} = \frac{P(M_0|\boldsymbol{x})}{P(M_1|\boldsymbol{x})}\left[\frac{P(M_0)}{P(M_1)}\right]^{-1}.$$

Observe that:

$$\frac{P(M_0|\boldsymbol{x})}{P(M_1|\boldsymbol{x})} = B_{0,1}(\boldsymbol{x})\frac{P(M_0)}{P(M_1)}$$

and therefore, the Bayes Factor tells whether the observed data $\boldsymbol{x}$ have increased or decreased the a priori odds of H_0 vs. H_1. Hence, if $B_{i,j} > 1$, it means that the observed data have increased the likelihood of H_i with respect to H_j, and the contrary if $B_{i,j} < 1$. Indeed, it can be immediately observed that:

$$B_{0,1}(\boldsymbol{x}) = \frac{1}{B_{1,0}(\boldsymbol{x})}$$

where $B_{1,0}$ is the Bayes Factor in favor of H_1 against H_0. Let now recall the loss function we have already presented when we have discussed the different relative weight that is customary assigned to the two error probabilities. We have defined $L(H_i,H_j)$ as the function yielding the level of loss obtained through accepting H_i being H_j true, from which it is customary assumed that:

$$L(H_0,H_0) = L(H_1,H_1) = 0 < L(H_0,H_1) = l_{0,1} < L(H_1,H_0) = l_{1,0}.$$

Defining now $\overline{L}(M_i|\boldsymbol{x})$ as the *expected loss* deriving from choosing model M_i having observed data $\boldsymbol{x} \in \mathcal{X}$, it will follow that, for all the possible models in $\{\mathcal{M} : M_i, i \in I\}$, it will be formally defined as:

$$\overline{L}(M_i|\boldsymbol{x}) = \textstyle\sum_{j\in I} L(M_i, M_j)P(M_j \mid \boldsymbol{x})$$

where, in the particular case of two possible models - M_0 and M_1 - with their associated hypotheses, we will have that the expected loss deriving from choosing the model associated with the Null Hypothesis will be:

$$\overline{L}(H_0|\boldsymbol{x}) = 0P(H_0|\boldsymbol{x}) + l_{0,1}P(H_1|\boldsymbol{x}) = l_{0,1}P(H_1|\boldsymbol{x})$$

while the expected loss deriving from choosing the model associated with the Alternative Hypothesis will be:

$$\overline{L}(H_1|\boldsymbol{x}) = 0P(H_1|\boldsymbol{x}) + l_{1,0}P(H_0|\boldsymbol{x}) = l_{1,0}P(H_0|\boldsymbol{x})$$

from which it will follow that we will choose the Null Hypothesis if and only if it yields a smaller expected loss, that is:

$$\text{Accept } H_0 \text{ if and only if } \overline{L}(H_0|\boldsymbol{x}) < \overline{L}(H_1|\boldsymbol{x})$$

where:

$$\overline{L}(H_0|\boldsymbol{x}) < \overline{L}(H_1|\boldsymbol{x}) \Leftrightarrow l_{0,1}P(H_1|\boldsymbol{x}) < l_{1,0}P(H_0|\boldsymbol{x}) \Leftrightarrow \frac{l_{0,1}}{l_{1,0}} < \frac{P(H_0|\boldsymbol{x})}{P(H_1|\boldsymbol{x})}$$

and similarly, we will have:

$$\text{Reject } H_0 \text{ if and only if } \overline{L}(H_0|\boldsymbol{x}) > \overline{L}(H_1|\boldsymbol{x})$$

where, as before:

$$\overline{L}(H_0|\boldsymbol{x}) > \overline{L}(H_1|\boldsymbol{x}) \Leftrightarrow l_{0,1}P(H_1|\boldsymbol{x}) > l_{1,0}P(H_0|\boldsymbol{x}) \Leftrightarrow \frac{l_{0,1}}{l_{1,0}} > \frac{P(H_0|\boldsymbol{x})}{P(H_1|\boldsymbol{x})}$$

$$\Leftrightarrow \frac{l_{1,0}}{l_{0,1}} < \frac{P(H_1|\boldsymbol{x})}{P(H_0|\boldsymbol{x})}$$

Before getting back to our specific problem, a couple of observations can be made:

I) If $l_{0,1} = l_{1,0}$, then the decision rule simply becomes:

$$\overline{L}(H_0|\boldsymbol{x}) < \overline{L}(H_1|\boldsymbol{x}) \Leftrightarrow P(H_0|\boldsymbol{x}) > P(H_1|\boldsymbol{x}).$$

$$(>) \qquad\qquad (<)$$

Indeed, if committing the two types of error is considered as equally severe, the hypothesis the researcher will choose, in a Bayesian framework, will simply be the one with the highest Posterior Probability to be true given the observed data.

If this is not the case - as how we have assumed - the greater will be $l_{1,0}$ with respect to $l_{0,1}$, that is, the greater will be the loss associated with rejecting a true Null Hypothesis with respect to the loss associated with not rejecting a false Null Hypothesis, the smaller will have to be the positive difference between the Posterior Probability of the Null Hypothesis and the Posterior Probability of the Alternative, id est the less the Posterior Probability of the Null Hypothesis will have to be greater than the Posterior Probability of the Alternative. Intuitively, the greater is the loss associated with the wrong rejection of a true Null Hypothesis, the more the researcher will be cautious in rejecting it, hence he/she won't reject it unless the Posterior Probability of H_1 won't be fairly greater than the Posterior Probability of H_0. It might also happen that the researcher won't reject H_0 even though the Posterior Probability of the Alternative is greater. If, for instance, we fix, $l_{1,0} = 9 > l_{0,1} = 3$, which means that the loss suffered due to the rejection of a true Null Hypothesis is three times greater than the loss due to the not rejection of it when is false, we will have that:

$$\overline{L}(H_0|\boldsymbol{x}) > \overline{L}(H_1|\boldsymbol{x}) \Leftrightarrow \tfrac{3}{9} > \tfrac{P(H_0|\boldsymbol{x})}{P(H_1|\boldsymbol{x})}$$

which, in other words, means that we will reject H_0 (since it yields a greater expected loss) if and only if the Posterior Probability of H_1 will be more than three times than the Posterior Probability of H_0.

II) It can be observed that:

$$\overline{L}(H_0|\boldsymbol{x}) > \overline{L}(H_1|\boldsymbol{x}) \Leftrightarrow \frac{l_{0,1}}{l_{1,0}} > \frac{P(H_0|\boldsymbol{x})}{P(H_1|\boldsymbol{x})}$$

$$\Leftrightarrow 1 > \frac{l_{1,0}}{l_{0,1}} \frac{P(H_0|\boldsymbol{x})}{P(H_1|\boldsymbol{x})}$$

$$\Leftrightarrow 1 < \frac{l_{0,1}}{l_{1,0}} \frac{P(H_1|\boldsymbol{x})}{P(H_0|\boldsymbol{x})}$$

$$\Leftrightarrow 1 < \frac{l_{0,1}}{l_{1,0}} \frac{P(\boldsymbol{x}|H_1)P(H_1)}{P(\boldsymbol{x}|H_0)P(H_0)}$$

$$\Leftrightarrow \frac{P(\boldsymbol{x}|H_0)}{P(\boldsymbol{x}|H_1)} < \frac{l_{0,1}}{l_{1,0}} \frac{P(H_1)}{P(H_0)}$$

$$\Leftrightarrow B_{0,1}(\boldsymbol{x}) < \frac{l_{0,1}}{l_{1,0}} \frac{P(H_1)}{P(H_0)}.$$

Let assume $l_{0,1} < l_{1,0}$ and $P(H_0) = P(H_1) = \frac{1}{2}$. Even though in Bayesian testing do not exist the two types of errors, if we consider these two assumptions to hold, it becomes possible a rejoinder between the frequentist and the Bayesian approach. Indeed, if committing the I Type Error is regarded as more severe than committing the II Type Error and the a Prior Probability of the Null Hypothesis is equal to the Prior Probability of the Alternative, the quantity on the right hand side is smaller than one, so that it might happen that even though the Posterior Probability of the Alternative Hypothesis (having observed data $\boldsymbol{x}$) is greater than the Posterior Probability of the Null Hypothesis (which implies $B_{0,1}(\boldsymbol{x}) < 1$), we still do not reject the latter.

In order to conclude this introduction to Bayesian testing, observe that the values the Bayes Factor can assume different forms depending on the type of hypotheses that we are considering. In particular, defining g_i as the prior density for θ conditional on H_i being true, we will have:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta = \theta_1 \Rightarrow B_{0,1}(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta_0)}{p(\boldsymbol{x}|\theta_1)}, \Theta = \{\theta_0, \theta_1\}$$

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0 \Rightarrow B_{0,1}(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta_0)}{\int_\Theta p(\boldsymbol{x}|\theta) g_1(\theta) d\theta}, \Theta = \{\theta_0\} \cup \Theta_1$$

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1 \Rightarrow B_{0,1}(\boldsymbol{x}) = \frac{\int_{\Theta_0} p(\boldsymbol{x}|\theta) g_0(\theta) d\theta}{\int_{\Theta_1} p(\boldsymbol{x}|\theta) g_1(\theta) d\theta}, \Theta_0 \cap \Theta_1 = \emptyset, \Theta_0 \cup \Theta_1 = \Theta$$

where such definitions can be straightforwardly extended to the discrete case. Now, we can discuss how Bayes Factor and Posterior Probability of the Null perform against the p - value, so that it will become possible to quantify the consequences due to misinterpretation of the p - value.

3 P - Value and Bayesian Posterior: irreconcilability and a possible calibration

The aim of this chapter is to present the huge discrepancy existing between the Posterior Probability of the Null Hypothesis and the p - value, discussing possible arguments that can be carried on against both. The focus will be on the case in which a Simple Null Hypothesis is tested against a Bilateral Composite Alternative. We will start the analysis with the case studied by Berger and Delampady in (1987) and Berger and Sellke (1987). While the authors consider testing a simple Null Hypothesis versus a Bilateral Composite Alternative for the mean of a Normal distribution, we will also provide a partial extension considering the same types of hypotheses for the scale parameter of a Gamma distribution.

3.1 P - Value and Posterior Probability in the Normal model

Suppose $\boldsymbol{X} = (X_1,....,X_n)$ to be a sample of n i.i.d. random variables distributed according to Normal with mean θ and known variance σ^2. Knowing that $\overline{X} \sim N(\theta, \frac{\sigma^2}{n})$, if our aim is to test $H_0 : \theta = \theta_0$ vs. $H_0 : \theta \neq \theta_0$, we have already said that the p - value is:

$$p\text{ - }value(\boldsymbol{x}) = 2[\ 1\text{ - }\Phi(\ |\ t\ |\ )]$$

where Φ is the standard cumulative distribution function.

Now, supposing π_0 to be the Prior Probability of H_0 and g_1 to be the prior density for θ conditional on H_1 being true, we can define the *marginal density of* $\boldsymbol{X}$ as:

$$m(\boldsymbol{x}) = \pi_0 f(\boldsymbol{x}\ |\ \theta_0) + (1\text{ - }\pi_0)m_{g_1}(\boldsymbol{x})$$

where:

$$m_{g_1}(\boldsymbol{x}) = \int_\Theta f(\boldsymbol{x}\ |\ \theta)g_1(\theta)d\theta.$$

Hence, the Posterior Probability of H_0 will be given by:

$$P(H_0\ |\ \boldsymbol{x}) = \frac{\pi_0 f(\boldsymbol{x}|\theta_0)}{m(\boldsymbol{x})}$$

$$= \frac{\pi_0 f(\boldsymbol{x}|\theta_0)}{\pi_0 f(\boldsymbol{x}|\theta_0)\ +\ (1-\pi_0)m_{g_1}(\boldsymbol{x})}$$

$$= \left[\frac{\pi_0 f(\boldsymbol{x}|\theta_0)\ +\ (1-\pi_0)m_{g_1}(\boldsymbol{x})}{\pi_0 f(\boldsymbol{x}|\theta_0)}\right]^{-1}$$

$$= \left[1 + \frac{(1-\pi_0)m_{g_1}(\boldsymbol{x})}{\pi_0 f(\boldsymbol{x}|\theta_0)}\right]^{-1}$$

$$= \left\{1 + \frac{(1-\pi_0)}{\pi_0}\,[B_{0,1}(\boldsymbol{x})]^{-1}\right\}^{-1}.$$

In the Normal example, the prior g_1 considered by Berger and Delampady (1987) is the $N(\mu,\ \tau^2)$ with, in particular, $\mu = \theta_0$ and $\tau = \sigma$. Such a choice is justified by the authors since, given that a likelihoodist might interpret g_1 as a mere weight function necessary to compute an average likelihood for H_1, in this way it would be symmetric. They find that:

$$B_{0,1}(\boldsymbol{x}) = \sqrt{1+n}\ exp\left\{-\frac{t^2}{2(1+n^{-1})}\right\}.$$

Hence, for various values of t, n and π_0 it becomes possible to make a first comparison between the Posterior Probability of the Null Hypothesis and the p - value. Assuming $\pi_0 = 0.5$, the results are presented in the following Table.

Data evidences		n					
t	p - *value*	1	5	10	20	50	100
1.645	0.10	0.72 (0.42)	0.79 (0.44)	0.89 (0.47)	1.27 (0.56)	1.86 (0.65)	2.57 (0.72)
1.960	0.05	0.54 (0.35)	0.49 (0.33)	0.59 (0.37)	0.72 (0.42)	1.08 (0.52)	1.50 (0.60)
2.576	0.01	0.27 (0.21)	0.15 (0.13)	0.16 (0.14)	0.19 (0.16)	0.28 (0.22)	0.37 (0.27)
3.291	0.001	0.10 (0.09)	0.03 (0.03)	0.02 (0.02)	0.03 (0.03)	0.03 (0.03)	0.05 (0.05)

Table 7: Berger and Delampady (1987). Measures of evidence in the normal example: the Posterior Probability of the Null Hypotheses is in parentheses, while the other value is its associated Bayes Factor

As how can be seen, such results are quite strong, and represent a first valid argument against some commonly held opinions regarding hypothesis testing and in particular Classical and Bayesian testing, (where with "Classical" we refer to the testing procedure based on the p - value). To start, it can be observed that even for an equal prior probability on the two hypotheses, the Bayesian and the Classical answers strongly disagree. For instance, when the sample is composed by 100 units, the Posterior Probability of the Null Hypothesis associated with a p - value of 0.05 is 0.60, and when the sample size is 50, it is 0.52. In other words, if Fisher would have rejected the Null Hypothesis with a significance level of 0.05 in both cases, Jeffreys (English mathematician considered among the main contributors of the revival of Bayesian Statistics and, in particular, of the Bayesian interpretation of Probability) would have concluded that the evidences are actually *in favor* of the Null Hypothesis. It is worth observing that, even though Jeffreys himself suggested to consider as prior for θ conditional on H_1 being true a $Cauchy(\theta_0, \tau^2)$ distribution, using the Normal prior doesn't produce sharp differences for low values of t.

A possible argument which could be carried against the Bayesian analysis and its comparison with the p - value might rely in the lack of objectivity when it has to be specified the Prior

Probability of the Null Hypothesis, π_0, and the prior density for θ conditional on H_1 being true, g_1. Against such objections, it can be immediately argued that $\pi_0 = \frac{1}{2}$ is indeed the objective choice for the Prior Probability of Null Hypothesis, or that the problem can be avoided by considering the Bayes Factor (which, differently from $P(H_0|\boldsymbol{x})$, does not depend on π_0). Though, it is impossible to come down with an objective choice of g_1. For some, it may be more relevant to choose it to be symmetric around θ_0, which might be argued to be a good choice when the parameter space is the entire real line. For others, instead, it may be more important to choose a prior nonincreasing in $\mid \theta$ - $\theta_0 \mid$ in order to avoid treating θ_0 as a special value.

On one side, in some cases, the explicit choice of g_1 may not be very relevant: indeed, as in the Normal case, choosing g_1 as a $N(\theta_0, \tau^2)$ or a $Cauchy(\theta_0, \tau^2)$ it has already been said that, at least for small values of t, does not produce strongly different results. Though, on the other side, the choice of the parameters of g_1 can strongly influence the Bayesian answer. In order to understand why, suppose, in the previously described Normal framework, that g_1 is a $N(\theta_0, \tau^2)$, where this time $\tau \neq \sigma$. In this case, the Bayes Factor will be:

$$B_{0,1}(\boldsymbol{x}) = \sigma^{-1}\sqrt{\sigma^2 + n\tau^2}\ exp\left\{-\frac{t}{2(1+\frac{\sigma^2}{n\tau^2})}\right\}$$

and, since:

$$\pi_0 = \tfrac{1}{2} \Rightarrow P(H_0|\boldsymbol{x}) = \left\{1 + \frac{1}{B_{0,1}(\boldsymbol{x})}\right\}^{-1}$$

then, in the considered case:

$$P(H_0|\boldsymbol{x}) = \frac{1}{1 + \frac{1}{\sigma^{-1}\sqrt{\sigma^2\ +\ n\tau^2}\ exp\left\{-\frac{t}{2(1\ +\ \frac{\sigma^2}{n\tau^2})}\right\}}}$$

from which it can be observed that τ plays a fundamental role in determining $P(H_0|\boldsymbol{x})$. Moreover, $P(H_0|\boldsymbol{x}) \to 1$ as $\tau^2 \to \infty$, making g_1 noninformative, which makes no sense.
Robert (2014) argues that this phenomenon consists in a reinterpretation (mathematically equivalent) of the famous Jeffreys - Lindley's paradox, according to which the Posterior Probability of the Null Hypothesis converges to 1 as the sample size n goes to infinity. What argues Robert is that, in the very end, the phenomenon is not paradoxical, nor lacking sense. Indeed, the greater is the diffuseness of the prior g_1 (the prior under H_1), the more the only relevant information becomes that the prior itself is centered on θ_0. Quoting Lindley (1957) himself, "the value θ_0 is fundamentally different form any value $\theta \neq \theta_0$, however near θ_0 it might be".
A solution to this problem is proposed by Robert (1993), where it is considered to make π_0 dependent on σ itself (considering the case in which $\tau = \sigma$). Indeed, even though choosing $\pi_0 = 1$ - $\pi_0 = \frac{1}{2}$ appears to be the fairest choice, given that in the Bayesian framework the hypotheses we are actually testing are:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \sim N(\theta_0, \sigma^2)$$

it sounds reasonable that the prior probability of the two hypotheses may vary depending on σ^2. In particular, fixing $\pi_0 = \frac{1}{2}$ ignores the fact that the greater is σ^2, the larger is the set of values of θ that g_1 considers as likely (that is, the greater is σ^2, the less the distribution is peaked around $\theta = \theta_0$). For instance, defining in the 100(1 - α)% *Higher Posterior Density* region for θ as the subset $\mathcal{HPD} \subset \Theta$ such that:

$$\mathcal{HPD} = \{\theta \in \Theta : h(\theta \mid \boldsymbol{x}) \geq k(\alpha)\}$$

where $k(\alpha)$ is such that:

$$\int_{\mathcal{HPD}} h(\theta \mid \boldsymbol{x}) d\theta = 1 \text{ - } \alpha$$

the greater is σ^2 the larger is the 99% $\mathcal{HPD}$ region of g_1, which can be considered as its "effective support". For this reason, Robert argues that the Prior Probability of H_1 should *increase* in σ^2 (more in general, in the variance of the prior under H_1). In order to better describe Robert's intuition, let assume, without loss of generality, $\theta_0 = 0$ and $X_i \sim N(\theta,1)$. A first natural requirement is that the prior should give more weight to those values of θ which made of interest to test $H_0 : \theta = 0$, that is those θ's in a neighborhood of 0 from which arise those x_i which could also originate from a $N(0,1)$. Hence, fixing an arbitrary ε and for σ sufficiently small, since:

$$\Phi(\tfrac{\varepsilon}{\sigma}) \text{ - } \Phi(\text{-}\tfrac{\varepsilon}{\sigma}) = \tfrac{1}{\sqrt{2\pi}} \int_{-\frac{\varepsilon}{\sigma}}^{\frac{\varepsilon}{\sigma}} e^{-\frac{x_i^2}{2}} dx_i$$

and applying the Midpoint Rule, considering one only interval, that is $(\text{-}\frac{\varepsilon}{\sigma}, \frac{\varepsilon}{\sigma})$, we obtain:

$$\int_{-\frac{\varepsilon}{\sigma}}^{\frac{\varepsilon}{\sigma}} e^{-\frac{x_i^2}{2}} dx_i \simeq [\tfrac{\varepsilon}{\sigma} \text{ - } (\text{-}\tfrac{\varepsilon}{\sigma})] \; exp\Big\{-\tfrac{1}{2}(\tfrac{\frac{\varepsilon}{\sigma} - \frac{\varepsilon}{\sigma}}{2})\Big\} = \tfrac{2\varepsilon}{\sigma}$$

we find that:

$$g_1([\text{-}\varepsilon,0) \cup [0, \varepsilon)) = (1 \text{ - } \pi_0) \; [\Phi(\tfrac{\varepsilon}{\sigma}) \text{ - } \Phi(\text{-}\tfrac{\varepsilon}{\sigma})] \simeq (1 \text{ - } \pi_0) \; \tfrac{2\varepsilon}{\sigma} \; \phi(0) = (1 \text{ - } \pi_0) \; \tfrac{2\varepsilon}{\sigma\sqrt{2\pi}}$$

being $\phi(0) = (2\pi)^{-\frac{1}{2}}$. Given the previous similarity, it appears reasonable to restrict $\pi_0 \equiv \pi_0(\sigma)$:

$$\pi_0(\sigma) = 1 \text{ - } \sigma c \Leftrightarrow [1 \text{ - } \pi_0(\sigma)] = \sigma c$$

where c is a constant that has to be determined. Such a constraint, however, can be seen to be quite stringent: indeed, as σ goes to infinity, the prior probability under H_1 of any fixed interval centered on 0 would converge to 0. Hence, it becomes more reasonable to assume that the Prior Probability of the Null Hypothesis and the Prior Probability under the Alternative of a given interval are proportional, so that their ratio remains constant as σ goes to infinity. Formally, what we are requiring is that:

$$g_1([\text{-}\varepsilon,0) \cup [0, \varepsilon)) = (1 \text{ - } \pi_0) \ [\Phi(\tfrac{\varepsilon}{\sigma}) \text{ - } \Phi(\text{-}\tfrac{\varepsilon}{\sigma})] \propto \pi_0(\sigma).$$

At this point, Robert (1993) reports that, for large values of σ the previous condition leads to the following:

$$\tfrac{1-\pi_0(\sigma)}{\sigma} \propto \pi_0(\sigma)$$

and to totally specify the proportionality factor, that is the precise expression of the dependence of $\pi_0(\sigma)$, it is suggested to consider 0 having the same weight under both the hypotheses. Formally, this means that the Prior Probability of the Null Hypothesis has to be equal to the Prior Probability of the Alternative times the value the prior density on θ under the Alternative - here assumed to be a Normal with mean 0, that is θ_0 itself, and σ^2 - assumes in 0, that is:

$$\pi_0(\sigma) = [1 \text{ - } \pi_0(\sigma)]g_1(0) = [1 \text{ - } \pi_0(\sigma)]\tfrac{1}{\sigma\sqrt{2\pi}}$$

which implies that 0 is "indifferent" under both the two hypotheses, whatever the value of σ. With some calculation it can be observed that:

$$P(H_0|x_i) = \left[1 + \sqrt{\tfrac{\sigma^2}{1+\sigma^2}}\sqrt{2\pi}\ exp\left\{\tfrac{\sigma^2 x_i^2}{2(\sigma^2+1)}\right\}\right]^{-1}$$

which doesn't suffer the Jeffreys - Lindley's paradox since:

$$\lim_{\sigma^2\to\infty} P(H_0|x_i) = \left[1 + \sqrt{2\pi}\ exp\left\{\tfrac{x_i^2}{2}\right\}\right]^{-1} < 1$$

which can be seen immediately to hold since:

$$\tfrac{\sigma^2}{1+\sigma^2} = 1 \text{ - } \tfrac{1}{1+\sigma^2}.$$

Finally, it is also interesting to oserve that since:

$$\pi_0(\sigma) = [1 \text{ - } \pi_0(\sigma)]\tfrac{1}{\sigma\sqrt{2\pi}} \Leftrightarrow \pi_0(\sigma) = \tfrac{1}{1+\sigma\sqrt{2\pi}}$$

then $\pi_0(\sigma)$ converges to 0 when σ^2 goes to infinity. We will get back later on the Jeffreys - Lindley's paradox when it will be discussed its other interpretation.

So far, it has been seen that Bayesian testing is characterized by both many degrees of subjectivity as well as by some theoretical problems (which can however be fixed by making some assumptions, as how it has been shown, for instance, by Robert (1993)). Anyway, these features should not be considered as an argument in favor of the p - value. For instance, considering the Normal case with $X_i \sim N(\theta, \sigma^2)$, $g_1(\theta)$ being a $N(\theta_0, \sigma^2)$ with σ^2 known, in order to have that to a p - value of 0.05 *actually* corresponds a Posterior Probability of the Null Hypothesis being true of 0.05, π_0 should be choosen to be extremely low. Assuming, for instance, $n = 1$, since it is immediate to note that:

$$n = 1 \Rightarrow B_{0,1}(\boldsymbol{x}) = \sqrt{2}\, e^{-\frac{t^2}{4}}$$

then the Posterior Probability of the Null Hypothesis will actually corresponds to 0.05 when the observed p - value is 0.05 if and only if π_0 will satisfy:

$$0.05 = \left[1 + \frac{1-\pi_0}{\pi_0\sqrt{2}} e^{\frac{(1.96)^2}{4}}\right]^{-1} \Leftrightarrow \pi_0 \simeq 0.089.$$

More in general, among symmetric and unimodal g_1 functions, Berger and Delampady (1987) observe that if the Posterior Probability of the Null Hypothesis is exactly 0.05 and the p - value is 0.05 as well, then π_0 will never be larger than 0.11, for all the sample sizes. Hence, assessing that the Posterior Probability of the Null Hypothesis is 0.05 given that the p - value is also 0.05 would be true if and only if π_0 was chosen to be no more than 0.11, which would clearly make the analysis not objective and strongly biased toward the Alternative. Anyway, despite the poor validity of the conclusions that can be done by observing the p - value, given the presented issues related to the choice of g_1 (exacerbated by how much it can be relevant in determining $P(H_0|\boldsymbol{x})$) it might appear that Bayesian testing doesn't consist

in a valid alternative. A possible solution, widely analyzed by Berger and Sellke (1987) to the not objective choice of g_1 might consist in defining a class $\mathcal{G}$ of possible priors, and then choosing the g_1 belonging to it which minimize both the Bayes Factor and $P(H_0|\boldsymbol{x})$. It will be seen, in the next session, that also these *lower bounds* on the Bayes Factor and $P(H_0|\boldsymbol{x})$ are far way greater than the p - value.

3.2 Lower Bounds on the Bayes Factor and the Posterior Probability of the Null Hypothesis

Let start defining $\mathcal{G}$ as the class of all the prior densities g_1 which are of interest for the researcher as possible choices. We can define, respectively, $\underline{B}_{0,1}(\boldsymbol{x},\, \mathcal{G})$ and $\underline{P}(H_0|\boldsymbol{x},\, \mathcal{G})$ as the lower bounds on the Bayes Factor and the Posterior Probability of the Null Hypothesis within the class $\mathcal{G}$. Formally, they will be given by:

$$\underline{B}_{0,1}(\boldsymbol{x},\, \mathcal{G}) = \inf_{g_1 \in \mathcal{G}} B_{0,1}(\boldsymbol{x}) = f(\boldsymbol{x}|\theta_0) \left\{\sup_{g_1 \in \mathcal{G}} m_{g_1}(\boldsymbol{x})\right\}^{-1}$$

and:

$$\underline{P}(H_0|\boldsymbol{x},\mathcal{G}) = \inf_{g_1 \in \mathcal{G}} P(H_0|\boldsymbol{x}) = \left\{1 + \tfrac{(1-\pi_0)}{\pi_0} \left[\underline{B}_{0,1}(\boldsymbol{x})\right]^{-1}\right\}^{-1}.$$

Berger and Sellke (1987) present the discrepancy between the p - values and the lower bounds on the Posterior Probability of the Null Hypothesis (for $\pi_0 = \frac{1}{2}$) in the Normal case, considering the classes $\mathcal{G}$ of all the possible distributions, of all the symmetric distributions, of all the symmetric and unimodal distribution and of all the Normal distributions.

Lower Bounds for $\mathcal{G}_{\mathcal{A}} = \{All\ densities\}$

The first presented results are those obtained by minimizing the Bayes Factor and the Posterior Probability of H_0 with respect to any possible density g_1. Clearly, these values are

maximally biased toward H_1 and, for this reason, they might appear as not of interest. However, it is worth observing that even in this case $\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{A}})$ and $\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}})$ remain sharply larger than the p - value.

Their computation is based on a result proved by Edwards et al. (1963), who showed that if a *Maximum Likelihood Estimator* $\hat{\theta}(\boldsymbol{x})$ for the parameter θ exists given the observed $\boldsymbol{x}$, then:

$$\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}}) = \frac{f(\boldsymbol{x}|\theta_0)}{f(\boldsymbol{x}|\hat{\theta}(\boldsymbol{x}))} \Leftrightarrow \underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{A}}) = \left\{1 + \frac{(1-\pi_0)f(\boldsymbol{x}|\hat{\theta}(\boldsymbol{x}))}{\pi_0 f(\boldsymbol{x}|\theta_0)}\right\}^{-1}.$$

In the Normal case we are considering, assuming still σ^2 being known, it can be proved that:

$$\hat{\theta}(\boldsymbol{x}) \ = \overline{x} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i$$

is the MLE for θ. Hence, the lower bound on the Bayes Factor becomes:

$$\begin{aligned}
\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}}) &= \tfrac{f(\boldsymbol{x}|\theta_0)}{f(\boldsymbol{x}|\overline{x})} \\
&= \tfrac{(\sigma\sqrt{2\pi})^{-n} exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_0)^2\Big\}}{(\sigma\sqrt{2\pi})^{-n} exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\Big\}} \\
&= exp\Big\{-\tfrac{1}{2\sigma^2}(n\theta_0^2 - 2\theta_0\textstyle\sum_{i=1}^{n} x_i - n\overline{x}^2 + 2\overline{x}\sum_{i=1}^{n} x_i\Big\} \\
&= exp\Big\{-\tfrac{1}{2\sigma^2}(n\theta_0^2 - 2\theta_0 n\overline{x} + n\overline{x}^2\Big\} \\
&= exp\Big\{-\tfrac{1}{2}\big[\tfrac{\sqrt{n}(\overline{x}-\theta_0)}{\sigma}\big]^2\Big\} \\
&= exp\Big\{-\tfrac{t^2}{2}\Big\}
\end{aligned}$$

while the lower bound on the Posterior Probability of the Null Hypothesis is given by:

$$\underline{P}(H_0 \mid \boldsymbol{x},\, \mathcal{G}_{\mathcal{A}}) = \left[1 + \tfrac{1-\pi_0}{\pi_0} e^{\frac{t^2}{2}}\right]^{-1}.$$

In the following table, some values of $\underline{P}(H_0 \mid \boldsymbol{x},\, \mathcal{G}_{\mathcal{A}})$ are compared with their associated p - values.

$p - value$	t	$\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{A}})$
0.10	1.645	0.205
0.05	1.960	0.128
0.01	2.576	0.035
0.001	3.291	0.0044

Table 8: Berger and Sellke (1987), p - values and $\underline{P}(H_0 \mid \boldsymbol{x},\, \mathcal{G}_{\mathcal{A}})$ with $\pi_0 = \frac{1}{2}$

As we have already said, despite minimizing $\underline{P}(H_0 \mid \boldsymbol{x})$ over all the possible densities g_1 is maximally biased in favor of the Alternative, the Posterior Probability of the Null Hypothesis remains fairly high compared to the p - values. For instance, if the observed p - value is 0.05, the Posterior Probability of the Null Hypothesis, computed taking into account all the possible prior densities for θ conditional on $\theta \neq \theta_0$ and its prior probability being $\pi_0 = \frac{1}{2}$, will be at least 0.128. Worth noting is that the values of these lower bounds do not explicitly depend on the size of the sample (but only implicitly through the observed value of t).

Lower Bounds for $\mathcal{G}_{\mathcal{S}} = \{$*All symmetric densities*$\}$

If we look the values of the Posterior Probability of the Null Hypothesis presented in Table 7 (where, we recall, g_1 was a $N(\theta_0,\, \sigma^2)$) and compare them with the lower bounds in Table 8 it can be observed that for a p - value, for instance, of 0.05, the Posterior Probability of the Null Hypothesis is three to five times bigger than its lower bound (these are, respectively, the cases of $n = 1$ and $n = 100$). Intuitively, this provides another valid argument against considering $\mathcal{G}_{\mathcal{A}}$ as possible class for g_1. Hence, at least for the Normal case,

a reasonable subclass of priors g_1 is $\mathcal{G}_{\mathcal{S}} = \{All\ symmetric\ densities\}$. Always in the same work, it is shown how minimizing $\underline{P}(H_0 \mid \boldsymbol{x})$ on $\mathcal{G}_{\mathcal{S}}$ is equivalent to minimizing it on $\mathcal{G}_{\mathcal{TPS}} = \{All\ symmetric\ two\ -\ point\ densities\}$.

$p - value$	t	$\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{S}})$
0.10	1.645	0.340
0.05	1.960	0.227
0.01	2.576	0.068
0.001	3.291	0.0088

Table 9: Berger and Sellke,(1987), p - values and $\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{S}})$ with $\pi_0 = \frac{1}{2}$

In this case, the discrepancy starts becoming more of interests, since the lower bounds of the Posterior Probability of the Null Hypothesis on $\mathcal{G}_{\mathcal{S}}$ are about the double than those on $\mathcal{G}_{\mathcal{A}}$ for all the four considered p - values. However, also with the symmetry restriction, the lower bounds on $\underline{P}(H_0 \mid \boldsymbol{x})$ appear still too far from those computed with g_1 being a $N(\theta_0, \sigma^2)$. Hence, a further reasonable restriction is considered, requiring g_1 to be also unimodal in θ_0.

Lower Bounds for $\mathcal{G}_{\mathcal{US}} = \{Unimodal\ symmetric\ densities\}$

As a first comment, observe that, if the symmetry assumption holds, requiring the prior to be unimodal is equivalent, under a mathematical perspective, to require it being nonincreasing in $\mid \theta$ - $\theta_0 \mid$. As before, the authors state that minimizing $P(H_0 \mid \boldsymbol{x})$ on $\mathcal{G}_{\mathcal{US}}$ is equivalent to minimize it on $\mathcal{G}^*_{\mathcal{US}} = \{All\ symmetric\ uniform\ densities\}$. Once again, without going deeper in the details of the iterative formula they consider, for our aim we can directly present the results (Table 10).

$p-value$	t	$\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{US}})$
0.10	1.645	0.390
0.05	1.960	0.290
0.01	2.576	0.109
0.001	3.291	0.018

Table 10: Berger and Sellke, (1987), p - values and $\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{US}})$ with $\pi_0 = \frac{1}{2}$

In this case, the increase in the lower bound of the Posterior Probability of the Null Hypothesis is only moderate a part from the case in which p - value = 0.001, where it becomes double. This means that minimizing over all the symmetric g_1 already yields some reliable Posterior Probabilities. Finally, we present the lower bounds on $P(H_0 \mid \boldsymbol{x})$ under the assumption that g_1 is a Normal density. It will be of particular interest to compare such lower bounds with the Posterior Probabilities in Table 7.

Lower Bounds for $\mathcal{G}_{\mathcal{NOR}} = \{Normal\ densities\}$

We now show which are the values of $\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{NOR}})$, that is the lower bounds on the Posterior Probability of the Null Hypothesis taking into account all those priors g_1 consisting in scale transformations of a (symmetric) Normal distribution. In order to discuss them, another useful result obtained by Edwards et al (1963) is presented. Indeed, they found that:

$$\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{NOR}}) = \sqrt{e}\ t\ e^{-\frac{t^2}{2}} \Leftrightarrow \underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{NOR}}) = \left\{1 + \frac{(1-\pi_0)e^{\frac{t^2}{2}}}{\pi_0 t\sqrt{e}}\right\}^{-1}$$

for all the values of t strictly greater than 1. The results are presented in the following table.

$p - value$	t	$\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{NOR}})$
0.10	1.645	0.412
0.05	1.960	0.321
0.01	2.576	0.133
0.001	3.291	0.0235

Table 11: Berger and Sellke, (1987), p - values and $\underline{P}(H_0 \mid \boldsymbol{x})$

Not surprisingly, the results are very similar to those in Table 7. As a first result, it is very interesting to observe how when $n = 1$ and p - value = 0.10, the Posterior Probability of the Null Hypothesis under $g_1 \sim N(\theta_0, \sigma^2)$ essentially coincides with the above presented lower bound (0.42 the former, 0.412 the latter). Instead, it can be observed that for a p - value = 0.05 the Posterior Probability under $g_1 \sim N(\theta_0, \sigma^2)$ is closer to the lower bound when $n = 5$ being, respectively, 0.33 and 0.321. Finally, on one side, for a p -value = 0.01, once again the Posterior Probability under $g_1 \sim N(\theta_0, \sigma^2)$ is closer to the lower bound when $n = 5$ being, in this case, respectively, 0.13 and 0.133 while, on the other side, for a p - value = 0.001 they become closer when $n = 10$ being, respectively, 0.02 and 0.0235. Hence, we can conclude that the prior $g_1 \sim N(\theta_0, \sigma^2)$ already provides Posterior Probabilities of the Null Hypothesis which are very close to their lower bounds for samples ranging from 1 to 10 observations, while for greater sample sizes the Posterior Probability of the Null Hypothesis starts to significantly diverge with respect to the lower bounds. This last fact consists in the $Jeffreys$ - $Lindley's' Paradox$, which consists - as how it has already been said - in one of the main arguments carried on by those who disagree with the Bayesian testing procedure and results. It will be now described under its other interpretation, which takes into account the behavior of the Null Hypothesis as the sample size n increases. Along with it, is also presented how in the Bayesian testing framework the problem of approximating an Interval Null Hypothesis by a Point one is solved.

3.3 Point Null testing criticisim: Bayesian approximation of an Interval Null Hypothesis and the Jeffreys - Lindley's Paradox

When discussing why, in general, committing the I Type Error is regarded as more severe, it has also been argued why, instead of testing a point Null against a Composite Bilateral Alternative of the kind:

$$H_0^* : \theta = \theta_0 \text{ vs. } H_1^* : \theta \neq \theta_0$$

it would make more sense, in many contexts, to consider:

$$H_0 : \mid \theta \text{ - } \theta_0 \mid \leq \varepsilon \text{ vs. } H_1 : \mid \theta \text{ - } \theta_0 \mid > \varepsilon$$

with ε small. Moreover, we have also observed that the p - value for this test is given by:

$$p \text{ - } value(\boldsymbol{x}, \varepsilon) = \sup_{\theta:|\theta-\theta_0|\leq\varepsilon} P_\theta \ (\mid T(\boldsymbol{X}) \mid \geq \mid t \mid).$$

Moreover, were presented some bounds on $\frac{\varepsilon\sqrt{n}}{\sigma}$ which guaranteed that approximating the p - value observed when testing H_0 vs. H_1 with the p - value observed when testing H_0^* vs. H_1^* would produce no more than a 10% error. Our aim now is to analyze under which conditions the Bayes Factor in favor of H_0^* computed when testing H_0^* vs. H_1^* can approximate the Bayes Factor in favor of H_0 computed in testing H_0 vs. H_1.

As a first point, observe that in this case, the Prior Probability of the Null Hypothesis will no longer be a point probability, but will instead be computed over an interval. Defining the set $\Omega = \{\theta \in \Theta : \mid \theta - \theta_0 \mid \leq \varepsilon\}$ and $\overline{\Omega}$ its complement, it will follow that:

$$\pi_0 = \int_\Omega \pi(\theta) \ d\theta$$

where $\pi(\theta)$ is a continuous prior density and $\Theta = \Omega \cup \overline{\Omega}$. Typically, for ε small, $\pi(\theta)$ will be spiked near θ_0. Defining now $I_\Omega(\theta)$ as the indicator function over the space Ω, the two priors for θ, respectively, under H_0 and H_1 being true will be:

$$g_0(\theta) = \frac{1}{\pi_0}\ \pi(\theta) I_{\Omega}(\theta)$$

$$g_1(\theta) = \frac{1}{1-\pi_0}\ \pi(\theta) I_{\overline{\Omega}}(\theta).$$

The main issues, in this case, are represented by the specifications of ε and g_0. Indeed, Berger and Delampady (1987) observe that specifying π_0 is usually not very problematic, and that it can be easily done by taking into account the nature of the hypothesis itself. Moreover, if the focus is put on the Bayes Factor, its specification becomes unnecessary. At the same time, also defining g_1 is not problematic, given the lower bounds that have been previously defined.

Considering which is the main goal, that is determine when, in a Bayesian context, testing H_0 vs. H_1 can be approximated by testing H_0^* vs. H_1^*, Berger and Delampady (1987) make the following assumptions:

$$g_1^*(\theta) \propto g_1(\theta) \text{ on } \mid \theta \text{ - } \theta_0 \mid \geq \varepsilon \text{ and } \lambda = \textstyle\int_{\Omega} g_1^*(\theta) d\theta \quad \text{ suitably small. (a)}$$

The logic behind such assumptions is that, in order to make testing H_0^* vs. H_1^* the most similar possible to testing H_0 vs. H_1, the prior on θ given H_1^* being true has to be proportional to the prior on θ given H_1 being true. Moreover, a small λ means that under $H_1^* : \theta \neq \theta_0$ the probability that the distance between θ and θ_0 is smaller than ε is suitably small, and therefore closer to 0, where:

$$0 = \textstyle\int_{\Omega} g_1(\theta) d\theta$$

which finally guarantees that $g_1(\theta)$ and $g_1^*(\theta)$ behave similarly also on Ω.

Considering the usual Normal case with σ^2 known, defining $B_{0,1}(\boldsymbol{x})$ the Bayes Factor for testing H_0 vs. H_1, that is:

$$B_{0,1}(\boldsymbol{x}) = \frac{\int_{\Omega} f(\overline{x}|\theta) g_0(\theta) d\theta}{\int_{\overline{\Omega}} f(\overline{x}|\theta) g_1(\theta) d\theta}$$

and $B^*_{0,1}(\boldsymbol{x})$ the Bayes Factor for testing H^*_0 vs. H^*_1, that is:

$$B^*_{0,1}(\boldsymbol{x}) = \frac{f(\overline{x}|\theta)}{m_{g^*_1}(\overline{x})}$$

where, we recall that:

$$m_{g^*_1}(\overline{x}) = \int_\Theta f(\overline{x} \mid \theta) g^*_1(\theta) d\theta$$

it can be proved that if the two conditions in (a) hold, the following Theorem will guarantee that $B_{0,1}(\boldsymbol{x}) \simeq B^*_{0,1}(\boldsymbol{x})$, eluding the problem of choosing ε and g_0.

Theorem 1 Let suppose $\pi(\theta)$ and $g^*_1(\theta)$ being unimodal and symmetric about θ_0 . Recalling that:

$$\varepsilon^* = \varepsilon\sqrt{n}\sigma^{-1}$$

and defining δ as:

$$\delta = [2\varepsilon^*\phi(t)]^{-1}[\Phi(t+\varepsilon^*) - \Phi(t-\varepsilon^*)] \text{ - } 1$$

where ϕ and Φ are, respectively, the standard normal density and the cumulative distribution function of a standardized normal, we will have that if $\mid t \mid \geq 1$, $\varepsilon^* < \mid t \mid$ - 1 and $B^*_{0,1}(\boldsymbol{x}) \leq (1+\delta)^{-1}$, then:

$$B_{0,1}(\boldsymbol{x}) = B^*_{0,1}(\boldsymbol{x})(1 + \frac{\sigma}{\sqrt{n}\tau})$$

where:

$$\text{-}\lambda \leq \lambda[B^*_{0,1}(\boldsymbol{x}) \text{ - } 1][1-\lambda B^*_{0,1}(\boldsymbol{x})]^{-1} \leq \frac{\sigma}{\sqrt{n}\tau}$$

$$\leq \left\{\delta + \lambda(1+\delta)[B^*_{0,1}(\boldsymbol{x}) - 1]\right\}[1 - \lambda B^*_{0,1}(\boldsymbol{x})(1+\delta)]^{-1} \leq \delta$$

where τ is the scale parameter of the Normal prior on θ under H_1. It is of interest to observe that when λ and δ are very small, approximating $B_{0,1}(\boldsymbol{x})$ with $B^*_{0,1}(\boldsymbol{x})$ produces a very small error. Moreover, it can be observed that:

$$\lambda \leq \frac{2\varepsilon^*\sigma}{\sqrt{n}}\ \mathrm{g}_1^*(\theta_0)$$

which means that the lower bound on λ is decreasing in the sample size.

We recall that these results have been shown in order to counterargue against those who dismiss the previously presented problems related to the discrepancy between the p - value and the Posterior Probability of the Null Hypothesis by saying that testing a Point Null against a Composite Bilateral Alternative, in many occasions, makes no sense. Indeed, it has been proved that, making some assumptions, testing a Point Null against a Bilateral Composite Alternative consists in a good approximation of the more natural case in which the Null and the Alternative are interval hypotheses, both in Classic and Bayesian testing. We now present the $Jeffreys - Lindley's\ paradox$ in its interpretation based on the sample size n.

The Jeffreys - Lindley's paradox

The Jeffreys - Linldey's paradox represents one of the strongest arguments supporters of the frequentist approach to testing employ in arguing against the Bayesian approach. For example, Spanos (2013) in comparing the frequentist and the Bayesian approach to testing concludes that while the latter is strongly biased yielding to highly fallacious results, the former can be adjusted in order to avoid producing the Jeffreys - Lindley's paradox. We remind that what we are now going to present consists in a different but mathematically equivalent re-interpretation of what has already been presented in chapter 3 section 1. In

order to describe the paradox under this re-interpretation, we consider the Normal framework we have been considering up to this point.

Definition 7: The Jeffreys - Lindley's paradox Let $\boldsymbol{X} = (X_1,....,X_n)$ be a sample of n i.i.d. random variables $X_i \sim N(\theta,\sigma^2)$, with σ^2 known, and suppose that we are willing to test a point Null Hypothesis $H_0^* : \theta = \theta_0$ vs. a Composite Bilateral Alternative $H_1^* : \theta \neq \theta_0$. Letting the prior on θ under the Alternative Hypothesis being a Normal with mean θ_0 and variance σ^2, the Bayes Factor of the Null against the Alternative has already been said to be:

$$B_{0,1}^*(\boldsymbol{x}) = \sqrt{1+n} \;\; exp\Big\{-\tfrac{t^2}{2(1+n^{-1})}\Big\}$$

where $t = \sqrt{n}(\overline{x}$ - $\theta_0)\sigma^{-1}$. The $Jeffreys - Lindley's\ paradox$ consists in the fact that for a fixed value of t (and, therefore, of the p - value) the Bayes Factor $B_{0,1}^*(\boldsymbol{x})$ goes to infinity as n goes to infinity as well which, in turn, implies that for a fixed value of the p - value, the Posterior Probability of the Null Hypothesis will converge to 1. Formally:

$$\lim_{n\to\infty} B_{0,1}^*(\boldsymbol{x}) = \infty \Leftrightarrow \lim_{n\to\infty} P(H_0^*|\boldsymbol{x}) = \lim_{n\to\infty} \Big\{1 + \tfrac{(1-\pi_0)}{\pi_0}\left[B_{0,1}^*(\boldsymbol{x})\right]^{-1}\Big\}^{-1} = 1.$$

As an example, Robert (2013) observes that assuming $n = 16818$, if $t = 1.96$ - whose associated p - value is 0.05 - the Posterior Probability of the Null Hypothesis will be 0.95. Similarly, always assuming $t = 1.96$, if $n = 164$, the Posterior Probability of the Null Hypothesis will instead be ten times greater than that of the Alternative. Such a discrepancy is regarded as paradoxical since the same dataset might produce strong evidences both in favor and against H_0 only depending on the considered approach.

An argument made by Robert (2013) is that this result is *not* as paradoxically as it might seem. Indeed, according to him, the assumption behind the paradox is that t remains constant in n, which is not of statistical interest: if H_0 is true, then t has a limiting $N(0,1)$

distribution, which has been observed that implies that the p - value is uniformly distributed between 0 and 1, while if H_1 is true, both the p - value and the Bayes Factor will converge to 0, given that t, instead, diverges to infinity.

In line with this last consideration, we now present the Bayesian calibration of the p - value suggested by Bayarri, Berger and Sellke (2001).

3.4 A Bayesian calibration of the p - value

At the start of the this work, we have widely described the problems related with some of the most common misinterpartations of the p - value. In particular, we have discussed why it is wrong to consider it as a frequentist error rate or as the Posterior Probability of the Null Hypothesis. Similarly, we have seen how it can't even be interpreted as a measure of evidence of H_0 againt H_1, since it consists in a measure developed in the Fisherian framework, where the Alternative hypothesis is not considered. However, Bayarri, Berger and Sellke (2001) present an intuitive way to reconcile the Bayesian answers, the p - value and the frequentist results. Indeed, the authors suggest considering the lower bound on the Bayes Factor as a function of the p - value p, with, in particular:

$$\underline{B}_{0,1}(p) = \text{-}e\ p\ log(p)$$

when $p < \frac{1}{e}$, from which the frequentist I Type Error as a function of the p - value, instead, is given by:

$$\alpha(p) = \{1 + [e\ p\ log(p)]^{-1}\}^{-1}.$$

which the authors argue can also be interpreted as the Posterior Probability of the Null Hypothesis arising from using $B_{0,1}(p)$ under the usual assumption that $\pi_0 = \frac{1}{2}$. In this way, the risk associated with the misinterpretation of the frequentist I Type Error probability α as the Posterior Probability of H_0 is avoided, since they coincide. The intuition behind these

calibrations is based on the fact that, under H_0, p is distributed according to a Uniform distribution defined between 0 and 1. Therefore, as Alternative hypothesis, they suggest considering an alternative distribution for p $itself$. Hence, supposing that under H_1 the p - value will be distributed according to a generic density $f(p \mid \zeta)$, ζ being an unknown parameter, the tested hypotheses will be:

$$H_0 : p \sim \mathcal{U}(0,1) \text{ vs. } H_1 : p \sim f(p \mid \zeta).$$

What the authors observe is that if the pivotal quantity $T(\boldsymbol{X})$ is chosen such that extreme values of it consist in strong empirical evidence against H_0, then, conversely, the density of p under H_1 should be *decreasing* in p. In particular, they suggest to consider the class of $be(\zeta,1)$ densities, for $0 < \zeta \leq 1$, so that:

$$f(p \mid \zeta)=\zeta p^{\zeta-1}\ I(0 \leq p \leq 1),\ 0 < \zeta \leq 1.$$

In this case, the Bayes Factor will be given by:

$$B_{0,1}(p) = \frac{f(p|1)}{\int_0^1 f(p|\zeta)g_1(\zeta)d\zeta}$$

from which, the authors state that:

$$\underline{B}_{0,1}(p) = \inf_{all g_1} B_{0,1}(p) = f(p|\ 1)[\sup_{\zeta\in(0,1]} \zeta p^{\zeta-1}]^{-1} = \begin{cases} -e\ p\ log(p) & \text{if } p < \frac{1}{e} \\ 1 & \text{otherwise} \end{cases}$$

where $f(p \mid 1) = 1$. The result is proved as follows:

$$\frac{\delta}{\delta\zeta}\ \zeta\ p^{\zeta-1} = p^{\zeta-1} + \zeta p^{\zeta-1}\ log(p)$$

$$= p^{\zeta-1}[1+\zeta log(p)] = 0$$

$$\Leftrightarrow 1 + \zeta log(p) = 0$$

$$\Leftrightarrow \zeta^* = -\frac{1}{log(p)}$$

which is indeed the maximizer since:

$$\frac{\delta^2}{\delta^2 \zeta}\ \zeta\ p^{\zeta-1} = p^{\zeta-1}log(p) + p^{\zeta-1}log(p) + \zeta p^{\zeta-1}log^2(p)$$

$$= p^{\zeta-1}log(p)[2 + \zeta log(p)]$$

$$\Rightarrow \frac{\delta^2}{\delta^2 \zeta^*}\ \zeta\ p^{\zeta-1} = p^{-\frac{1}{log(p)}-1}\ log(p) < 0$$

being $0 \leq p \leq 1$. Hence, it will follow that:

$$[\sup_{\zeta \in (0,1]}\ \zeta p^{\zeta-1}]^{-1} = [-\frac{1}{log(p)}\ p^{-\frac{1}{log(p)}-1}]^{-1}$$

$$= -log(p)\ p^{\frac{1}{log(p)}+1}$$

$$= -log(p)\ exp\left\{log(p^{\frac{1}{log(p)}+1})\right\}$$

$$= -log(p)\ exp\left\{(\frac{1}{log(p)} + 1)log(p)\right\}$$

$$= -log(p)\ exp\ \{1 + log(p)\}$$

$$= -log(p)\ e\ p$$

which is smaller than 1 if and only if $p < \frac{1}{e}$.

It is interesting to observe how such a lower bound holds for *all* the possible priors g_1 on ζ, so that it becomes an *objective* lower bound of the posterior evidence of H_0 over H_1 for the $be(\zeta,1)$ p - values distributions alternatives. The results are presented in the following table:

$p-value$	0.20	0.10	0.05	0.01	0.005	0.001
$\underline{B}_{0,1}(p)$	0.870	0.625	0.407	0.125	0.072	0.0188
$\alpha(p)$	0.465	0.385	0.289	0.111	0.067	0.0184

Table 12: Bayarri, Berger and Sellke, (2001), P - value calibration as lower bound on Bayes Factor and Posterior Probability of the Null Hypothesis with the $be(\zeta,1)$ prior under the Alternative.

It might be of interest to compare how this proposed calibration performs in providing a measure for the lower bound on the Posterior Probability of H_0 with respect, for instance, with the lower bounds that have been presented in subsection 3.2. In the following table, for different p - values, the lower bounds of the Posterior Probability of H_0 on the restricted classes for g_1 already presented are compared with the proposed calibration.

$p-value$	0.10	0.05	0.01	0.001
-e log(p)	0.6259	0.4072	0.1252	0.0188
$\mathcal{G}_{\mathcal{NOR}}$	0.7007	0.4727	0.1534	0.0240
$\mathcal{G}_{\mathcal{US}}$	0.6393	0.4084	0.1223	0.0183
$\mathcal{G}_{\mathcal{S}}$	0.5151	0.2937	0.0730	0.0088

Table 13: Bayarri, Berger and Sellke (2001), restricted Lower Bounds on the Posterior Probability of H_0 in the Normal setting

A first result of interest is that the Lower Bounds on the Posterior Probability when the prior on θ under the Alternative Hypothesis is restricted to the class of Uniform Symmetric distributions is very similar, for all the four considered p - values, to the proposed calibration. Given that this class is regarded to contain all objective and sensible priors, this suggests the validity of the proposed calibration.

So far, we have widely discussed the most common misinterpretations of the p - value and why they have become a mainstream in the literature, a problem related with the too often overlooked difference between the Neyman - Pearson's and Fisherian approach to testing. Moreover, we have presented some consequences of this misinterpretation, relating in particular with the huge discrepancy existing, in a Bayesian framework, between the p - value and the Posterior Probability of the Null Hypothesis. Also some arguments against both measures have been reported. Finally, a possible rejoinder between the two measures has been presented. Now the discrepancy between the p - value and the Posterior Probability of the Null Hypothesis is analyzed considering a Gamma model with known shape parameter, supposing to be interested in testing a Point Null Hypothesis on the scale parameter.

4 P - Value and Posterior Probability of the Null Hypothesis in the Gamma model

As how it has been said at the end of the previous section, the aim is now to verify whether the already discussed discrepancy between the evidences provided by the p - value and the Posterior Probability of the Null Hypothesis when testing a Point Null Hypothesis vs. a Composite Bilateral Alternative Hypothesis about the location parameter in the Normal case can be observed also in the case in which the two same hypotheses are tested for the scale parameter of a Gamma distribution, assuming the shape parameter to be known. The chosen prior for the scale parameter under the Alternative Hypothesis is a Gamma, since conjugate priors guarantee a posterior distribution which can be expressed in closed form. In subsection 1, after that the problem has been formalized, we derive the Posterior Probability of the Null Hypothesis, proving also the Jeffreys - Lindley paradox to hold in this case. In subsection 2, we describe the computation of the p - value and finally, in subsection 3, for

some fixed values of the involved parameters, p - values and Posterior Probabilities of the Null Hypothesis are compared, also considering its lower bound in the class of all the possible densities for the prior under H_1.

4.1 Posterior Probability of the Null Hypothesis

Let $\boldsymbol{X} = (X_1,...,X_n)$ be a sample of n i.i.d. random variables $X_i \sim ga(\theta, \lambda)$, with θ known, and let suppose that it is of interest to test $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$. Moreover, let still define π_0 as the Prior Probability of the Null Hypothesis and g_1 as the prior on λ under H_1 with, in particular, λ to be distributed according to a Gamma distribution with ζ shape parameter and ω scale parameter, Formalizing the described framework, it follows that:

$$X_i \sim ga(\theta, \lambda) \Rightarrow f(x_i \mid \theta,\lambda) = \tfrac{\lambda^\theta}{\Gamma(\theta)}\ x_i^{\theta-1}\ e^{-\lambda x_i}\ I(x_i > 0)$$

$$\lambda \sim^{H_1} ga(\zeta,\omega) \Rightarrow g_1(\lambda \mid \zeta,\omega) = \tfrac{\omega^\zeta}{\Gamma(\zeta)}\lambda^{\zeta-1}e^{-\omega\lambda}\ I(\lambda > 0).$$

and being the observations i.i.d.:

$$f(\boldsymbol{x}|\theta,\lambda) = (\tfrac{\lambda^\theta}{\Gamma(\theta)})^n \textstyle\prod_{i=1}^n x_i^{\theta-1}\ exp\{-\lambda \sum_{i=1}^n x_i\} I(x_{(1)}>0)$$

where $X_{(1)}$ represents the minimum of the observations. It follows that:

$m_{g_1}(\boldsymbol{x}) = \int_0^{+\infty} f(\boldsymbol{x} \mid \theta, \lambda) g_1(\lambda \mid \zeta, \omega) d\lambda$

$= \int_0^{+\infty} (\tfrac{\lambda^\theta}{\Gamma(\theta)})^n \prod_{i=1}^n x_i^{\theta-1} exp\left\{-\lambda \sum_{i=1}^n x_i\right\} \tfrac{\omega^\zeta}{\Gamma(\zeta)} \lambda^{\zeta-1} e^{-\omega\lambda} I(x_{(1)} > 0) d\lambda$

$= (\tfrac{1}{\Gamma(\theta)})^n \prod_{i=1}^n x_i^{\theta-1}\ \tfrac{\omega^\zeta}{\Gamma(\zeta)}\ I(x_{(1)}>0)\ \int_0^{+\infty} \lambda^{\theta n+\zeta-1} exp\left\{-\lambda(\sum_{i=1}^n x_i + \omega)\right\} d\lambda$

$= (\tfrac{1}{\Gamma(\theta)})^n \prod_{i=1}^n x_i^{\theta-1}\ \tfrac{\omega^\zeta}{\Gamma(\zeta)}\ I(x_{(1)}>0) \tfrac{\Gamma(\theta n+\zeta)}{(\sum_{i=1}^n x_i+\omega)^{\theta n+\zeta}}$

$$\times \underbrace{\int_0^{+\infty} \frac{(\sum_{i=1}^n x_i+\omega)^{\theta n+\zeta}}{\Gamma(\theta n+\zeta)} \lambda^{\theta n+\zeta-1} exp\left\{-\lambda(\sum_{i=1}^n x_i+\omega)\right\} d\lambda}_{=1}$$

$$= [\Gamma(\theta)]^{-n} \prod_{i=1}^n x_i^{\theta-1} \omega^{\zeta} [\Gamma(\zeta)]^{-1} \Gamma(\theta n+\zeta)\ (\textstyle\sum_{i=1}^n x_i+\omega)^{-(\theta n+\zeta)} I(x_{(1)}>0).$$

The associated Bayes Factor in favor of H_0, $B_{0,1}(\boldsymbol{x})$, will be:

$$B_{0,1}(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta,\lambda_0)}{m_{g_1}(\boldsymbol{x})} = \frac{(\frac{\lambda_0^\theta}{\Gamma(\theta)})^n \prod_{i=1}^n x_i^{\theta-1} exp\{-\lambda_0 \sum_{i=1}^n x_i\} I(x_{(1)}>0)}{[\Gamma(\theta)]^{-n} \prod_{i=1}^n x_i^{\theta-1} \omega^{\zeta} [\Gamma(\zeta)]^{-1} \Gamma(\theta n+\zeta) (\sum_{i=1}^n x_i+\omega)^{-(\theta n+\zeta)} I(x_{(1)}>0)}$$

$$= \lambda_0^{\theta n}\ exp\{-\lambda_0 \textstyle\sum_{i=1}^n x_i\} \omega^{-\zeta} \Gamma(\zeta) (\sum_{i=1}^n x_i+\omega)^{\theta n+\zeta} [\Gamma(\theta n+\zeta)]^{-1}.$$

And the Posterior Probability of the Null Hypothesis will be:

$$P(\lambda_0|\boldsymbol{x}) = \frac{\pi_0 f(\boldsymbol{x}|\theta,\lambda_0)}{\pi_0 f(\boldsymbol{x}|\theta,\lambda_0)\ +\ (1-\pi_0) m_{g_1}(\boldsymbol{x})}$$

$$= \frac{\pi_0 (\frac{\lambda_0^\theta}{\Gamma(\theta)})^n \prod_{i=1}^n x_i^{\theta-1} exp\{-\lambda_0 \sum_{i=1}^n x_i\} I(x_{(1)}>0)}{\pi_0 (\frac{\lambda_0^\theta}{\Gamma(\theta)})^n \prod_{i=1}^n x_i^{\theta-1} exp\{-\lambda_0 \sum_{i=1}^n x_i\} I(x_{(1)}>0)\ +\ (1-\pi_0)[\Gamma(\theta)]^{-n} \prod_{i=1}^n x_i^{\theta-1} \omega^{\zeta} [\Gamma(\zeta)]^{-1} \Gamma(\theta n+\zeta) (\sum_{i=1}^n x_i+\omega)^{-(\theta n+\zeta)} I(x_{(1)}>0)}$$

$$= \frac{\pi_0 [\Gamma(\theta)]^{-n} \prod_{i=1}^n x_i^{\theta-1} \lambda_0^{\theta n} exp\{-\lambda_0 \sum_{i=1}^n x_i\} I(x_{(1)}>0)}{[\Gamma(\theta)]^{-n} \prod_{i=1}^n x_i^{\theta-1} I(x_{(1)}>0) [\pi_0 \lambda_0^{\theta n} exp\{-\lambda_0 \sum_{i=1}^n x_i\}\ +\ (1-\pi_0) \omega^{\zeta} [\Gamma(\zeta)]^{-1} \Gamma(\theta n+\zeta) (\sum_{i=1}^n x_i+\omega)^{-(\theta n+\zeta)}]}$$

$$= \frac{\pi_0 \lambda_0^{\theta n} exp\{-\lambda_0 \sum_{i=1}^n x_i\}}{\pi_0 \lambda_0^{\theta n} exp\{-\lambda_0 \sum_{i=1}^n x_i\}\ +\ (1-\pi_0) \omega^{\zeta} [\Gamma(\zeta)]^{-1} \Gamma(\theta n+\zeta) (\sum_{i=1}^n x_i+\omega)^{-(\theta n+\zeta)}}$$

which can be proved to be equivalent to:

$$P(\lambda_0|\boldsymbol{x}) = \left\{1 + \frac{(1-\pi_0)}{\pi_0} [B_{0,1}(\boldsymbol{x})]^{-1}\right\}^{-1}$$

$$= \left\{1 + \frac{(1-\pi_0)}{\pi_0}\lambda_0^{-\theta n} exp\left\{\lambda_0 \sum_{i=1}^{n} x_i\right\} \omega^{\zeta}[\Gamma(\zeta)]^{-1}(\sum_{i=1}^{n} x_i + \omega)^{-(\theta n+\zeta)}\Gamma(\theta n + \zeta)\right\}^{-1}.$$

Observe that, since:

$$Var(\lambda) = \frac{\zeta}{\omega^2}$$

then, according to the Jeffreys - Lindley Paradox:

$$\lim_{\omega\to 0} P(\lambda_0|\boldsymbol{x}) = \lim_{\zeta\to\infty} P(\lambda_0|\boldsymbol{x}) = 1.$$

The first is immediate to observe since:

$$\lim_{\omega\to 0} \text{(1-}\pi_0)\omega^{\zeta}[\Gamma(\zeta)]^{-1}\Gamma(\theta n + \zeta)(\sum_{i=1}^{n} x_i + \omega)^{-(\theta n+\zeta)} = 0.$$

For the second, considering *Stirling Rule*:

$$n! \sim \sqrt{2\pi}\text{n}^{n+\frac{1}{2}}\ e^{-n} \Leftrightarrow \Gamma(n) = (n\text{ - }1)! \sim \sqrt{2\pi}(\text{n - }1)^{n-\frac{1}{2}}\ e^{-(n-1)} \text{ as } n \to \infty$$

it follows that (letting $\sum_{i=1}^{n} x_i = n\overline{x}$):

$$(1-\pi_0)\omega^{\zeta}[\Gamma(\zeta)]^{-1}\Gamma(\theta n+\zeta)(\sum_{i=1}^{n} x_i+\omega)^{-(\theta n+\zeta)} \sim \text{(1-}\pi_0)\ \omega^{\zeta}\left\{\sqrt{2\pi}(\zeta-1)^{\zeta-\frac{1}{2}}e^{-\zeta+1}\right\}^{-1}\sqrt{2\pi}$$

$$\times\ (\theta n + \zeta - 1)^{\theta n+\zeta-\frac{1}{2}}\ (n\overline{x} + \omega)^{-(\theta n+\zeta)}\ e^{-(\theta n+\zeta-1)}$$

$$= (1\text{ - }\pi_0)\ e^{-\theta n}\ (\zeta-1)^{\frac{1}{2}-\zeta}\ (\theta n + \zeta - 1)^{-\frac{1}{2}+\zeta+\theta n}\ (n\overline{x} + \omega)^{-(\theta n+\zeta)}$$

and:

$$(1\text{ - }\pi_0)\ e^{-\theta n}\ \lim_{\zeta\to\infty}(\zeta-1)^{\frac{1}{2}-\zeta}\ (\theta n + \zeta - 1)^{-\frac{1}{2}+\zeta+\theta n}\ (n\overline{x} + \omega)^{-(\theta n+\zeta)} = 0.$$

4.2 P - Value computation

In order to compute the p - value, at this point, we need to find that statistic which, according to the definition, provides, for extreme values, evidences against the Null Hypothesis. In order to proceed, it has been considered the Likelihood Ratio Test:

$$\text{Reject } H_0 \text{ if } \Lambda = \frac{f(\boldsymbol{x}|\lambda_0)}{f(\boldsymbol{x}|\hat{\lambda}(\boldsymbol{x}))} < k.$$

where $\hat{\lambda}(\boldsymbol{x})$ represents the Maximum Likelihood Estimator for λ. Indeed, once that the ratio has been computed and the Rejection Region has been expressed as function of one statistic (in this case the sum of X_i), it has been investigated for which values it is smaller than k. In other words, it has been studied the behavior of the Likelihood Ratio $\Lambda(t)$ as function of t, where $T = \sum_{i=1}^{n} X_i$: given that it will be found that the $\Lambda(t)$ is not monotone (increases for values smaller than the maximizer t^* and decreases for values greater than it), it will be concluded that both values extremely small and extremely great of T consist in strong evidence against the Null Hypothesis.

The Likelihood Ratio has already been defined as:

$$\Lambda = \frac{f(\boldsymbol{x}|\lambda_0)}{f(\boldsymbol{x}|\hat{\lambda})}.$$

First, we need to find $\hat{\lambda}$:

$$f(\boldsymbol{x} \mid \lambda) = \lambda^{n\theta}[\Gamma(\theta)]^{-n}\textstyle\prod_{i=1}^{n} x_i^{\theta-1}\ e^{-\sum_{i=1}^{n} x_i\lambda}$$

$$\Rightarrow log\ f(\boldsymbol{x} \mid \lambda) = n\theta\ log\ \lambda \text{ - } n\ log\ \Gamma(\theta) + (\theta \text{ - } 1)\textstyle\sum_{i=1}^{n} log\ x_i \text{ - } \lambda \sum_{i=1}^{n} x_i$$

$$\Rightarrow \tfrac{\delta}{\delta\lambda}\ log\ f(\boldsymbol{x} \mid \lambda) = \tfrac{n\theta}{\lambda} \text{ - } \textstyle\sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \hat{\lambda}(\boldsymbol{x}) = \frac{n\theta}{\sum_{i=1}^{n} x_i}$$

which is indeed a maximizer since:

$$\frac{\delta^2}{\delta^2 \lambda}\ logf(\boldsymbol{x} \mid \lambda) = \text{-}\ \frac{n\theta}{\lambda^2} < 0.$$

It follows that the Likelihood Ratio Test becomes:

$$\text{Reject } H_0\ \Lambda = \frac{\lambda_0^{n\theta}[\Gamma(\theta)]^{-n} \prod_{i=1}^{n} x_i^{\theta-1} e^{-\lambda_0 \sum_{i=1}^{n} x_i}}{\hat{\lambda}^{n\theta}[\Gamma(\theta)]^{-n} \prod_{i=1}^{n} x_i^{\theta-1} e^{-n\theta}} < k$$

that is:

$$\text{Reject } H_0 \text{ if } \Lambda = \left(\frac{\lambda_0}{n\theta} \sum_{i=1}^{n} x_i\right)^{n\theta} \exp\{-\lambda_0 \sum_{i=1}^{n} x_i + n\theta)\} < k.$$

At this pont, the aim becomes to study the behavior of $\Lambda \equiv \Lambda(t)$ as function of $t = \sum_{i=1}^{n} x_i$. It follows that:

$$\frac{\delta}{\delta t}\ \Lambda(t) = \frac{\delta}{\delta t}\ \lambda_0^{n\theta}\ (n\theta)^{-n\theta}\ e^{n\theta}\ t^{n\theta}\ e^{-\lambda_0 t}$$

$$= \lambda_0^{n\theta}\ (n\theta)^{-n\theta}\ e^{n\theta}\left[n\theta t^{n\theta-1} e^{-\lambda_0 t} - \lambda_0 t^{n\theta} e^{-\lambda_0 t}\right]$$

$$= \lambda_0^{n\theta}\ (n\theta)^{-n\theta+1}\ e^{n\theta-\lambda_0 t}\ t^{n\theta-1}\ \text{-}\ \lambda_0^{n\theta+1} (n\theta)^{-n\theta}\ e^{n\theta-\lambda_0 t}\ t^{n\theta}$$

$$= \lambda_0^{n\theta}\ (n\theta)^{-n\theta}\ e^{n\theta-\lambda_0 t}\ t^{n\theta}\ \left[\frac{n\theta}{t} - \lambda_0\right]$$

from which it can be observed that:

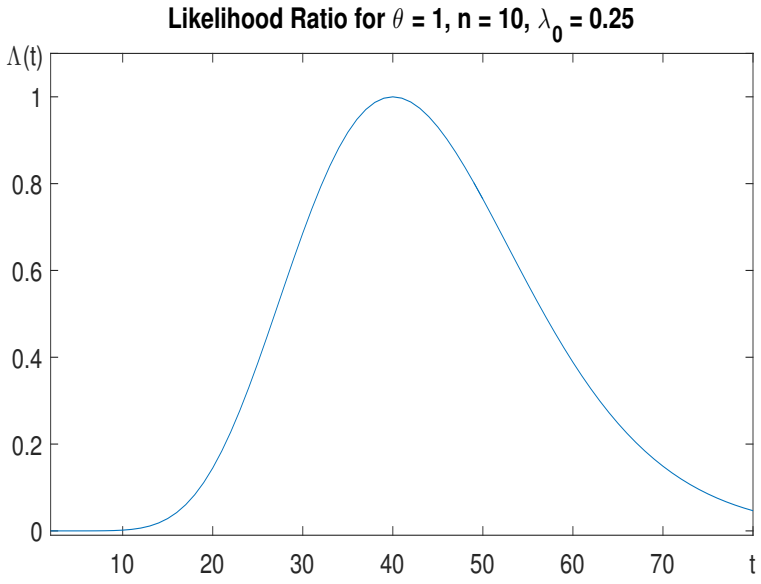
$$(\frac{\delta}{\delta t}\ \Lambda(t) < 0 \Leftrightarrow \frac{n\theta}{t} - \lambda_0 < 0 \Leftrightarrow t > \frac{n\theta}{\lambda_0}) \wedge (\frac{\delta}{\delta t}\ \Lambda(t) > 0 \Leftrightarrow \frac{n\theta}{t} - \lambda_0 > 0 \Leftrightarrow t < \frac{n\theta}{\lambda_0}).$$

Hence, we can conclude that the Likelihood Ratio $\Lambda(t)$, as function of the sum of the observations, is increasing for values smaller of $t^* = n\theta(\lambda_0)^{-1}$ and decreasing otherwise. Moreover,

it can be observed that $\Lambda(t)$ is defined only for positive values ($t = \sum_{i=1}^{n} x_i$, $x_i \geq 0$ for all i) and that can assume only values between 0 and 1. Finally:

$$\lim_{t \to +\infty} \Lambda(t) = \lim_{t \to 0} \Lambda(t) = 0.$$

Letting, for instance $\theta = 1$, $n = 10$, $\lambda_0 = 0.25$ (choices made only to have a well defined graph), graphically $\Lambda(t)$ appears to be:



from which it will follow that:

$$\Lambda(t) < k \Leftrightarrow t < t_{k,1} \vee t > t_{k,2}$$

and therefore:

$$R = \{\boldsymbol{x} \in \boldsymbol{X} : t < t_{k,1} \vee t > t_{k,2}\}.$$

At this point, if our goal was to find those values of $t_{k,1}$ and $t_{k,2}$ such that the *ex ante fixed* Probability of committing the I Type Error was α, that is:

$$\alpha = P(T < t_{k,1} \mid \lambda = \lambda_0) + P(T > t_{k,2} \mid \lambda = \lambda_0)$$

or, equivalently:

$$\alpha = P(\chi^2_{2n\theta} < 2\lambda_0 t_{k,1} \mid \lambda = \lambda_0) + P(\chi^2_{2n\theta} > 2\lambda_0\ t_{k,2} \mid \lambda = \lambda_0)$$

we would have that a possible, simplifying solution was considering those values such that the probability of observing more extreme values is $\frac{\alpha}{2}$. Defining, once again, $\chi^2_{n,\alpha}$ as that value such that:

$$P(\chi^2_n > \chi^2_{n,\alpha}) = \alpha$$

it will follow that:

$$2\lambda_0 t_{k,1} = \chi^2_{2n\theta,1-\frac{\alpha}{2}}$$
$$2\lambda_0 t_{k,2} = \chi^2_{2n\theta,\frac{\alpha}{2}}.$$

For what, instead, concerns the computation of the p - value, we have to take into account for the asymmetry of the statistic $T = \sum_{i=1}^n X_i$ (we recall that $T \sim ga(n\theta,\ \lambda)$). Many authors (Rohatgi and Saleh (2001), Kulinskaya (2007), Pace and Salvan (1996) among the others) suggest the following: defining t^* as the observed value of the statistic $T = \sum_{i=1}^n X_i$, the p - value will be given by:

$$p - value = 2min\{P(T < t^* \mid \lambda = \lambda_0), P(T > t^* \mid \lambda = \lambda_0)\}$$

or, once again, equivalently:

$$p - value = 2min\{P(\chi^2_{2n\theta} < 2\lambda_0 t^* \mid \lambda = \lambda_0), P(\chi^2_{2n\theta} > 2\lambda_0 t^* \mid \lambda = \lambda_0)\}.$$

Hence, given which is our purpose - that is to make a comparison between the evidences provided by the p - value against H_0 and its Posterior Probability associated with that p - value - we can obtain the values of t^* which yield that p - value, and evaluating the Posterior Probability of H_0 in them. The described procedure is presented for the case in which p -

value = 0.001: the other cases are identical.

$$0.001 = 2min\{P(\chi^2_{2n\theta} < 2\lambda_0 t^* \mid \lambda = \lambda_0), P(\chi^2_{2n\theta} > 2\lambda_0 t^* \mid \lambda = \lambda_0)\}$$

$$\Leftrightarrow 0.0005 = min\{P(\chi^2_{2n\theta} < 2\lambda_0 t^* \mid \lambda = \lambda_0), P(\chi^2_{2n\theta} > 2\lambda_0 t^* \mid \lambda = \lambda_0)\}.$$

From which, considering the two separate cases:

$$1)\ P(\chi^2_{2n\theta} < 2\lambda_0 t^* \mid \lambda = \lambda_0) < P(\chi^2_{2n\theta} > 2\lambda_0 t^* \mid \lambda = \lambda_0)$$

$$\Rightarrow 0.0005 = P(\chi^2_{2n\theta} < 2\lambda_0 t^* \mid \lambda = \lambda_0) \Rightarrow 0.9995 = P(\chi^2_{2n\theta} > 2\lambda_0 t^* \mid \lambda = \lambda_0)$$

$$\Rightarrow 2\lambda_0 t^* = \chi^2_{2n\theta,0.9995} \Rightarrow t_1^* = \tfrac{1}{2\lambda_0}\ \chi^2_{2n\theta,0.9995}.$$

$$2)\ P(\chi^2_{2n\theta} < 2\lambda_0 t^* \mid \lambda = \lambda_0) > P(\chi^2_{2n\theta} > 2\lambda_0 t^* \mid \lambda = \lambda_0)$$

$$\Rightarrow 0.0005 = P(\chi^2_{2n\theta} > 2\lambda_0 t^* \mid \lambda = \lambda_0) \Rightarrow 2\lambda_0 t^* = \chi^2_{2n\theta,0.0005} \Rightarrow t_2^* = \tfrac{1}{2\lambda_0}\chi^2_{2n\theta,0.0005}.$$

We can therefore conclude that if the observed p - value is 0.001, then *one* of these two values of t has been observed:

$$t_1^* = \tfrac{1}{2\lambda_0}\ \chi^2_{2n\theta,0.9995} \vee t_2^* = \tfrac{1}{2\lambda_0}\chi^2_{2n\theta,0.0005}.$$

With the same procedure that has been shown for this case, it can be found that, for the other three levels of the observed p - values we are considering (defining, once again, t^* as the observed values of t):

$$p - value = 0.01 \Leftrightarrow (\ t^* = t_1^* = \tfrac{1}{2\lambda_0}\ \chi^2_{2n\theta,0.995} \vee t^* = t_2^* = \tfrac{1}{2\lambda_0}\chi^2_{2n\theta,0.005})$$

$$p - value = 0.05 \Leftrightarrow (\ t^* = t_1^* = \tfrac{1}{2\lambda_0}\ \chi^2_{2n\theta,0.975} \vee t^* = t_2^* = \tfrac{1}{2\lambda_0}\chi^2_{2n\theta,0.025})$$

$$p - value = 0.10 \Leftrightarrow (\ t^* = t_1^* = \tfrac{1}{2\lambda_0}\ \chi^2_{2n\theta,0.95} \vee t^* = t_2^* = \tfrac{1}{2\lambda_0}\chi^2_{2n\theta,0.05}).$$

4.3 P - Value and Posterior Probability comparison

At this point, in order to provide a numerical comparison between the p - value and Posterior Probability of the Null Hypothesis, it is needed to assume some values for the parameters ω, ζ, θ and λ_0, as well as for the sample size n. For the purpose, we consider the case in which we are testing:

$$H_0 : \lambda = 1 \text{ vs. } H_1 : \lambda \neq 1.$$

Moreover, letting, for instance, $\theta = \zeta = \omega = 1$, it follows that:

$$X_i \sim ga(1,\, \lambda) \Rightarrow f(x_i \mid \lambda) = \lambda\ e^{-\lambda x_i}\ I(x_i > 0)$$

$$\lambda \sim^{H_1} ga(1,\, 1) \Rightarrow g_1(\lambda) = e^{-\lambda}\ I(\lambda > 0).$$

Observe that:

$$\forall\ \alpha > 0,\ \lambda \sim^{H_1} ga(\alpha,\, \alpha) \Rightarrow E(\lambda) = 1,\ Var(\lambda) = \tfrac{1}{\alpha}$$

In other words, if $\lambda_0 = 1$, then all the priors on λ under H_1 with same value for both the shape and scale parameter will have as expected value λ_0 but a different value for the variance, from which it will follow a different behavior of the Posterior Probability of H_0.

Finally, as possible sample sizes, we refer to those considered by Berger and Delampady (1987), that is $n = 1, 5, 10, 20, 50, 100$. Considering $\pi_0 = \frac{1}{2}$, the Posterior Probability of the Null Hypothesis becomes:

$$P(\lambda_0|\boldsymbol{x}) = \frac{\frac{1}{2}exp\{-t\}}{\frac{1}{2}exp\{-t\} + \frac{1}{2}\Gamma(n+1)(t+1)^{-(n+1)}}$$

where $t = \sum_{i=1}^{n} x_i$. Now, in six different tables (one for each sample size), are presented the values of the Posterior Probability of H_0 - and its associated Bayes Factor - corresponding to each p - value. Out of the parentheses there are the values corresponding to the case in which the p - value has been computed on the left tail. That is, for instance, considering the case of p - value = 0.001, since it has been assumed $\lambda_0 = 1$:

$$t_1^* = \frac{1}{2}\chi^2_{2n\theta,0.9995} \vee t_2^* = \frac{1}{2}\chi^2_{2n\theta,0.0005}$$

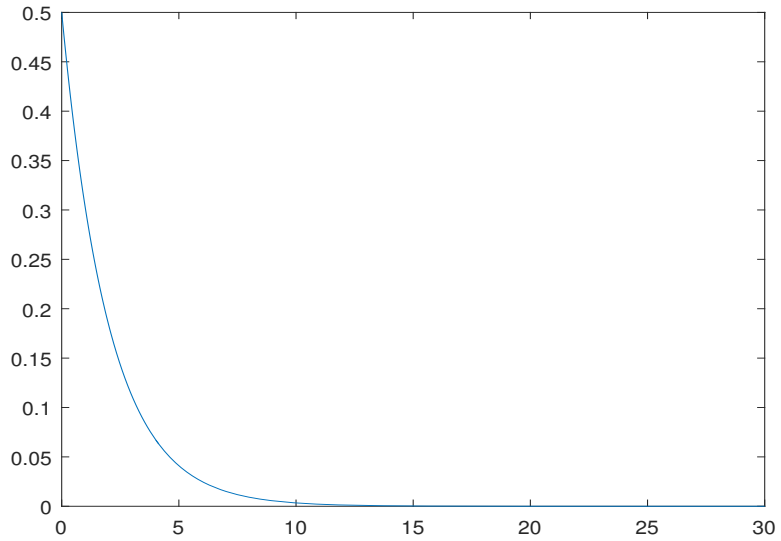
where, since it has also been assumed that $\theta = 1$, considering the case $n = 1$, according to how it has been previously defined the quantity $\chi^2_{2n\theta,\alpha}$, it will follow that:

$$0.9995 = P(\chi_2^2 > \chi^2_{2,0.9995}) \Leftrightarrow \chi^2_{2,0.9995} = 0.001 \Leftrightarrow t_1^* = 0.0005$$

and

$$0.0005 = P(\chi_2^2 > \chi^2_{2,0.0005}) \Leftrightarrow \chi^2_{2,0.0005} = 15.20 \Leftrightarrow t_2^* = 7.60$$

(All the values in the following tables have been computed similarly).

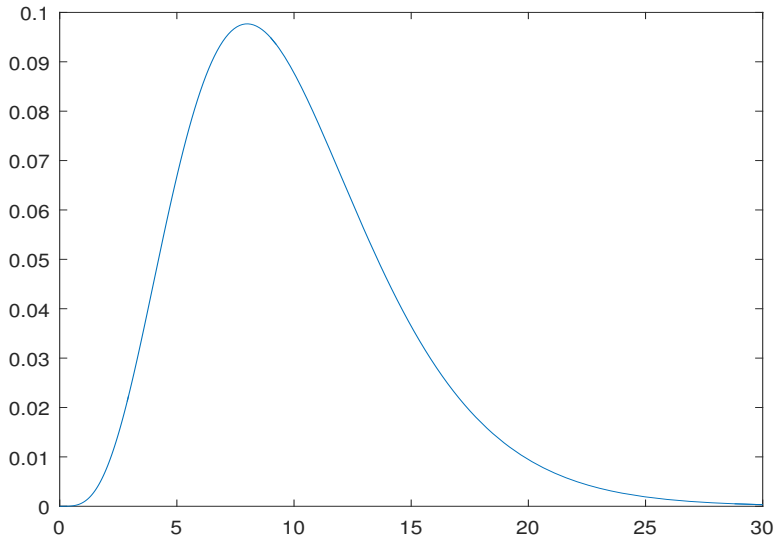


(Graph of the density of a χ^2_{2n}, $n = 1$)

Data evidences		*Bayesian measures*	
t_1^* (t_2^*)	*p - value*	$P(\lambda_0\|\boldsymbol{x}, t_1^*)$ ($P(\lambda_0\|\boldsymbol{x}, t_2^*)$)	$B_{0,1}(\boldsymbol{x}, t_1^*)$ ($B_{0,1}(\boldsymbol{x}, t_2^*)$)
0.0005 (7.600)	0.001	0.5001 (0.0357)	1.0004 (0.0370)
0.005 (5.295)	0.01	0.5012 (0.1658)	1.0048 (0.1988)
0.0255 (3.690)	0.05	0.5062 (0.3545)	1.0251 (0.5492)
0.0515 (2.995)	0.10	0.5122 (0.4440)	1.0500 (0.7986)

Table 14: P - value, Posterior Probability of H_0 and Bayes Factor when $n = 1$ and $\pi_0 = \frac{1}{2}$

As a first result, it is of interest to observe that, even for the same p - value, the Posterior Probability of the Null Hypothesis is different if we consider the left or the right tail, which is due to the shape of the χ^2_{2n} distribution. Moreover, it is of interest to compare the results in this table with those found by Berger and Delampady (1987) that have been presented in Table 7 when testing a Point Null Hypothesis on the location parameter of a Normal distribution. In this case, the Posterior Probability of H_0 remains essentially constant when the p - value is computed on the left tail, while, when the p - value is computed on the right tail, it follows a pattern similar to that of the Normal case (the values reported by Berger and Delampady were, respectively 0.09, 0.21, 0.35 and 0.42)

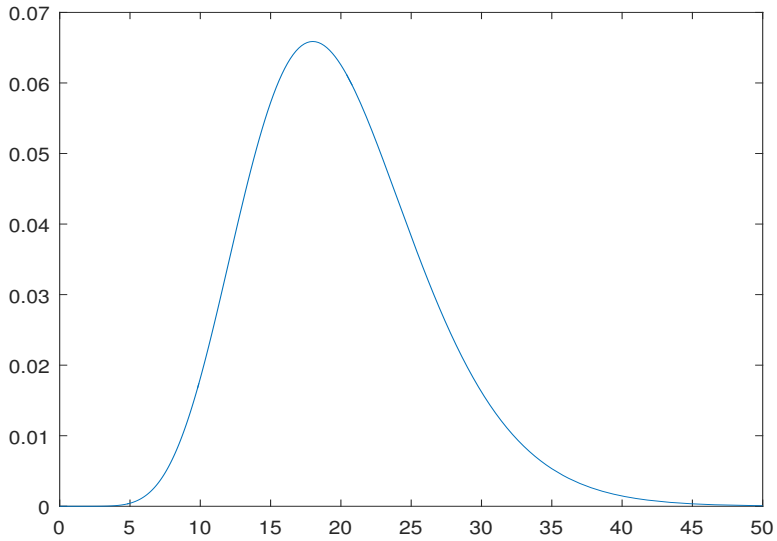


(Graph of the density of a χ^2_{2n}, $n = 5$)

Data evidences		*Bayesian measures*	
t_1^* (t_2^*)	*p - value*	$P(\lambda_0 \vert \boldsymbol{x}, t_1^*)$ ($P(\lambda_0 \vert \boldsymbol{x}, t_2^*)$)	$B_{0,1}(\boldsymbol{x}, t_1^*)$ ($B_{0,1}(\boldsymbol{x}, t_2^*)$)
0.64 (15.71)	0.001	0.0778 (0.0266)	0.0844 (0.0273)
1.08 (12.60)	0.01	0.1864 (0.1513)	0.2291 (0.1783)
1.63 (10.24)	0.05	0.3493 (0.3751)	0.5368 (0.6003)
1.97 (9.16)	0.10	0.4437 (0.4913)	0.7976 (0.9658)

Table 15: P - value, Posterior Probability of H_0 and Bayes Factor when $n = 5$ and $\pi_0 = \frac{1}{2}$

With $n = 5$ - as with all the following values of n - instead, a clear increasing pattern in the Posterior Probability of H_0 can be observed also when the p - value is computed on the left tail. As in the Normal case, the Posterior Probability of H_0 associated with $n = 5$ is smaller than that with $n = 1$ for p - value = 0.01 and p - value = 0.001 on both tails, while, only on the right tail, it is greater for the other two considered p - values. A similar behavior can be observed a in the Normal model.



(Graph of the density of a χ^2_{2n}, $n = 10$. As how can be seen, the greater n, the more the distribution becomes flatter: when $n = 5$ is spiked in about 0.10,while when $n = 10$ is spiked in about 0.065)

Data evidences		*Bayesian measures*	
t_1^* (t_2^*)	*p - value*	$P(\lambda_0\|\boldsymbol{x}, t_1^*)$ ($P(\lambda_0\|\boldsymbol{x}, t_2^*)$)	$B_{0,1}(\boldsymbol{x}, t_1^*)$ ($B_{0,1}(\boldsymbol{x}, t_2^*)$)
2.70 (23.75)	0.001	0.0319 (0.0277)	0.0330 (0.0285)
3.72 (20.00)	0.01	0.1466 (0.1659)	0.1718 (0.1989)
4.80 (17.09)	0.05	0.3607 (0.4150)	0.5642 (0.7094)
5.43 (15.71)	0.10	0.4832 (0.5408)	0.9350 (1.1777)

Table 16: P - value, Posterior Probability of H_0 and Bayes Factor when $n = 10$ and $\pi_0 = \frac{1}{2}$

When $n = 10$ the Posterior Probability of H_0 remains quite similar to the case of $n = 5$, with just a slight increase that ranges from 0.001 and 0.05 for all the p - values on both tails (a part from the left - tail of the p - value 0.001, where it decreases). In comparison with the Normal example, instead, the Posteriors Probabilities of λ_0 are slightly greater than the Posteriors Probabilities of θ_0 for all the p - values, in particular when the p - value is computed on the right tail.

Data evidences		*Bayesian measures*	
t_1^* (t_2^*)	*p - value*	$P(\lambda_0\|\boldsymbol{x},\, t_1^*)$ ($P(\lambda_0\|\boldsymbol{x},\, t_2^*)$)	$B_{0,1}(\boldsymbol{x},\, t_1^*)$ ($B_{0,1}(\boldsymbol{x},\, t_2^*)$)
8.46 (38.05)	0.001	0.0263 (0.0316)	0.0270 (0.0326)
10.36 (33.39)	0.01	0.1587 (0.1930)	0.1886 (0.2392)
12.22 (29.67)	0.05	0.4153 (0.4709)	0.7103 (0.8900)
13.26 (27.88)	0.10	0.5520 (0.6012)	1.2320 (1.5075)

Table 17: P - value, Posterior Probability of H_0 and Bayes Factor when $n = 20$ and $\pi_0 = \frac{1}{2}$

For $n = 20$, with respect to the case of $n = 10$, an increase in the Posterior Probability of H_0 - ranging from 0.05 to 0.07 - can be observed on both tails, in particular for p - value = 0.05 and p - value = 0.10. Nonetheless, it remains small its increase for the other two p - values while, once again, the behavior of the Posterior of λ_0 is essentially the same of the Posterior of θ_0 in the Normal case.

Data evidences		*Bayesian measures*	
t_1^* (t_2^*)	*p - value*	$P(\lambda_0\|\boldsymbol{x},\, t_1^*)$ ($P(\lambda_0\|\boldsymbol{x},\, t_2^*)$)	$B_{0,1}(\boldsymbol{x},\, t_1^*)$ ($B_{0,1}(\boldsymbol{x},\, t_2^*)$)
29.95 (76.59)	0.001	0.0330 (0.0414)	0.0341 (0.0432)
33.65 (70.09)	0.01	0.2112 (0.2488)	0.2678 (0.3312)
37.11 (64.78)	0.05	0.5191 (0.5608)	1.0794 (1.2769)
38.97 (62.17)	0.10	0.6560 (0.6877)	1.9070 (2.2021)

Table 18: P - value, Posterior Probability of H_0 and Bayes Factor when $n = 50$ and $\pi_0 = \frac{1}{2}$

When $n = 50$, given the increase in the Posterior Probability of λ_0 - ranging from 0.06 to 0.10 for all the p - values considered on both tails, a part from 0.001 - it starts to be visible its clear convergence to 1, which is in line with the Jeffreys - Lindley paradox in its interpretation based on the sample size.

Data evidences		*Bayesian measures*	
t_1^* (t_2^*)	*p - value*	$P(\lambda_0\|\boldsymbol{x},\, t_1^*)$ ($P(\lambda_0\|\boldsymbol{x},\, t_2^*)$)	$B_{0,1}(\boldsymbol{x},\, t_1^*)$ ($B_{0,1}(\boldsymbol{x},\, t_2^*)$)
70.33 (136.21)	0.001	0.0444 (0.0533)	0.0465 (0.0563)
76.12 (127.63)	0.01	0.2733 (0.3061)	0.3761 (0.4411)
81.37 (120.53)	0.05	0.6042 (0.6334)	1.5265 (1.7278)
84.14 (117.00)	0.10	0.7298 (0.7504)	2.7010 (3.0064)

Table 19: P - value, Posterior Probability of H_0 and Bayes Factor when $n = 100$ and $\pi_0 = \frac{1}{2}$

Finally, for $n = 100$, the Posterior Probabilities of λ_0 become more similar on both the tails, and assume, values very close to those of θ_0 in the Normal case, in particular for the p - values computed on the left tail. Also the increases with respect to the case of $n = 50$ are significant.

Lower Bounds for the Posterior Probability of λ_0 on $\mathcal{G}_{\mathcal{A}} = \{All\ densities\}$

When presenting the Lower Bounds on the Posterior Probability in the Normal model, considering as class for the prior under the Alternative Hypothesis all the possible densities, has been presented a result proved by Edwards et al. (1963), that is:

$$\underline{B}_{0,1}(\boldsymbol{x},\, \mathcal{G}_{\mathcal{A}}) = \tfrac{f(\boldsymbol{x}|\theta_0)}{f(\boldsymbol{x}|\hat{\theta}(\boldsymbol{x}))} \Leftrightarrow \underline{P}(H_0 \mid \boldsymbol{x},\, \mathcal{G}_{\mathcal{A}}) = \left\{1 + \tfrac{(1-\pi_0) f(\boldsymbol{x}|\hat{\theta}(\boldsymbol{x}))}{\pi_0 f(\boldsymbol{x}|\theta_0)}\right\}^{-1}.$$

In the Gamma model we are considering, given that the MLE estimator for λ as already been observed to be:

$$\hat{\lambda}(\boldsymbol{x}) = \tfrac{n\theta}{t}$$

it can be found that:

$$\underline{B}_{0,1}(\boldsymbol{x},\, \mathcal{G}_{\mathcal{A}}) = (\tfrac{t}{n\theta}\lambda_0)^{n\theta} exp\big\{-t(\lambda_0 - \tfrac{n\theta}{t})\big\}$$

from which:

$$\underline{P}(H_0 \mid \boldsymbol{x},\, \mathcal{G}_{\mathcal{A}}) = \left[1 + \tfrac{(1-\pi_0)}{\pi_0}(\tfrac{t}{n\theta}\lambda_0)^{-n\theta} exp\left\{t(\lambda_0 - \tfrac{n\theta}{t})\right\}\right]^{-1}.$$

Considering the values we have assumed for θ, λ_0 and π_0, it follows that:

$$\underline{P}(H_0 \mid \boldsymbol{x},\, \mathcal{G}_{\mathcal{A}}) = \left[1 + (\tfrac{t}{n})^{-n} exp\left\{t(1 - \tfrac{n}{t})\right\}\right]^{-1}.$$

The values, for the six different sample sizes we have considered, are presented in the following tables:

Data evidences		*Bayesian measures*	
t_1^* (t_2^*)	*p - value*	$\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_1^*)$ ($\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_2^*)$)	$\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_1^*)$ ($\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_2^*)$)
0.0005 (7.600)	0.001	0.0014 (0.0102)	0.0014 (0.0103)
0.005 (5.295)	0.01	0.0133 (0.0673)	0.0135 (0.0722)
0.0255 (3.690)	0.05	0.0633 (0.2003)	0.0678 (0.2505)
0.0515 (2.995)	0.10	0.1174 (0.2895)	0.1330 (0.4074)

Table 20: P - value and Lower Bounds for the Posterior Probability of H_0 and Bayes Factor when $n = 1$ and $\pi_0 = \frac{1}{2}$

It is of interest to compare these lower bounds with the values reported in Table 14. First, is worth noting that, for all the considered p - values, the lower bounds on the Posterior Probability of λ_0, minimized over all the possible priors g_1, is much smaller (on both tails) than the Posterior computed under g_1 distributed according to gamma. In particular on the left tail, while the lower bound ranges from 0.0014 to 0.1174 (resp. p - value = 0.001 and p - value = 0.1), we have seen that for $n = 1$, the Posterior Probability of λ_0 was about 0.50 for all the four considered p - values. Once again, this is due to the shape of the χ^2_{2n} distribution.

Data evidences		*Bayesian measures*	
t_1^* (t_2^*)	*p - value*	$\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_1^*)$ ($\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_2^*)$)	$\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_1^*)$ ($\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_2^*)$)
0.64 (15.71)	0.001	0.0027 (0.0068)	0.0027 (0.0068)
1.08 (12.60)	0.01	0.0232 (0.0484)	0.0237 (0.0509)
1.63 (10.24)	0.05	0.0967 (0.1603)	0.1071 (0.1910)
1.97 (9.16)	0.10	0.1642 (0.2436)	0.1965 (0.3221)

Table 21: P - value and Lower Bounds for the Posterior Probability of H_0 and Bayes Factor when $n = 5$ and $\pi_0 = \frac{1}{2}$

When $n = 5$, the difference between the lower bounds and the Posterior Probabilities presented in Table 15 is less evident than with n =1. For instance, for p - value = 0.05, while the Posterior Probabilities presented in Table 15 were 0.3493 and 0.3751 on the two tails (left and right respectively), the lower bounds are 0.0967 and 0.1603.

Data evidences		*Bayesian measures*	
t_1^* (t_2^*)	*p - value*	$\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_1^*)$ ($\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_2^*)$)	$\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_1^*)$ ($\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_2^*)$)
2.70 (23.75)	0.001	0.0031 (0.0061)	0.0330 (0.0061)
3.72 (20.00)	0.01	0.0264 (0.0444)	0.0271 (0.0465)
4.80 (17.09)	0.05	0.1053 (0.1505)	0.1177 (0.1771)
5.43 (15.71)	0.10	0.1771 (0.2328)	0.2151 (0.3034)

Table 22: P - value and Lower Bounds for the Posterior Probability of H_0 and Bayes Factor when $n = 10$ and $\pi_0 = \frac{1}{2}$

When $n = 10$, can be done considerations similar to the case of n =5. It is interesting to observe a slight increase in the lower bounds for all the sample sizes on the left tail and, conversely, a decrease on the right tail, which is due to the fact that the χ^2_{2n} is becoming flatter.

Data evidences		*Bayesian measures*	
t_1^* (t_2^*)	*p - value*	$\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_1^*)$ ($\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_2^*)$)	$\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_1^*)$ ($\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_2^*)$)
8.46 (38.05)	0.001	0.0034 (0.0056)	0.0035 (0.0056)
10.36 (33.39)	0.01	0.0289 (0.0415)	0.0297 (0.0433)
12.22 (29.67)	0.05	0.1117 (0.1441)	0.1258 (0.1683)
13.26 (27.88)	0.10	0.1809 (0.2251)	0.2277 (0.2904)

Table 23: P - value and Lower Bounds for the Posterior Probability of H_0 and Bayes Factor when $n = 20$ and $\pi_0 = \frac{1}{2}$

With $n = 20$, the discrepancy between the lower bounds and the Posterior Probabilities in Table 17 increase on both tails. For instance, when p - value = 0.10, the Posterior Probabilities of λ_0 are 0.5520 and 0.6012 (respectively on the right and left tail), while the lower bounds are 0.1809 and 0.2251. Moreover, it is worth noting how the lower bounds have remained almost equal, with a difference of no more than 0.05.

Data evidences		*Bayesian measures*	
t_1^* (t_2^*)	*p - value*	$\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_1^*)$ ($\underline{P}(H_0 \mid \boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_2^*)$)	$\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_1^*)$ ($\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_2^*)$)
29.95 (76.59)	0.001	0.0038 (0.0051)	0.0038 (0.0052)
33.65 (70.09)	0.01	0.0308 (0.0391)	0.0317 (0.0407)
37.11 (64.78)	0.05	0.1175 (0.1381)	0.1331 (0.1602)
38.97 (62.17)	0.10	0.1929 (0.2180)	0.2389 (0.2787)

Table 24: P - value and Lower Bounds for the Posterior Probability of H_0 and Bayes Factor when $n = 50$ and $\pi_0 = \frac{1}{2}$

When $n = 50$, the probabilities on the two tails become even closer, with a difference of no more than 0.02, and an increase of no more than 0.01 with respect to the lower bounds computed for $n = 20$, which means that the difference in the sample size does not affect anymore significantly the lower bounds. On the other side, a sharp increase was instead observed on the Posterior of λ_0 under the gamma prior in Table 18. In particular, an increase of 0.10 (from 0.5520 to 0.6560 and from 0.4153 to 0.5191) was observed on the left tail when p - value = 0.05 and p - value = 0.10.

Data evidences		*Bayesian measures*	
t_1^* (t_2^*)	*p - value*		$\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_1^*)$ ($\underline{B}_{0,1}(\boldsymbol{x}, \mathcal{G}_{\mathcal{A}}, t_2^*)$)
70.33 (136.21)	0.001	0.0040 (0.0049)	0.0040 (0.0050)
76.12 (127.63)	0.01	0.0321 (0.0380)	0.0332 (0.0394)
81.37 (120.53)	0.05	0.1206 (0.1350)	0.1372 (0.1561)
84.14 (117.00)	0.10	0.1964 (0.2142)	0.2444 (0.2726)

Table 25: P - value and Lower Bounds for the Posterior Probability of H_0 and Bayes Factor when $n = 100$ and $\pi_0 = \frac{1}{2}$

Finally, with $n = 100$, there are almost no difference in the lower bounds with respect to $n = 50$. while another increase of about 0.10 was observed in the Posterior Probability of λ_0. Hence, as the sample size increases, while the lower bounds remain essentially constant, the Posterior Probability of λ_0, considering the Gamma prior under H_1, increases and converges to 1, which has already been said consisting in the Jeffreys - Lindley's paradox.

The main conclusion that can be drawn is that the problem of the discrepancy between

the Posterior Probability of the Null Hypothesis in the Gamma model (when considering a Point Null Hypothesis on the shape parameter) appears to be relevant as well as in the Normal case. However, further investigation are needed for different values of λ_0, as well as for different priors g_1, even for the same $\lambda_0 = 1$. Indeed, also within the class of all the g_1 distributed according to a gamma distribution, given that different parameters can yield the same expected value - assumed to be λ_0 - but different values of the variance (as how it has been already said for $\lambda_0 = 1$), it also should be investigated how the variance of the gamma priors affect the Posterior Probability of H_0 keeping constant the expected value equal to λ_0, so that also the effect of the Jeffreys - Lindley paradox can be better analyzed.

For what it concerns the analysis of the lower bounds of the Posterior Probability of λ_0 computed minimizing with respect to all the possible prior densities under H_1, the huge discrepancy between such lower bounds and the Posterior Probabilities of λ_0 needs to be further investigated. However, when the lower bounds computed by Berger and Sellke (1987) in the Normal model were presented, it was already discussed that considering all the possible densities for the prior on θ under H_1 consisted in a too stringent minimization, even tough interesting for several reasons.

5 P - value and reproducibility issues: a case study in psychology

In 2016, Professors Valen E.Johnson, Richard D.Payne, Tianying Wang, Alex Asher and Soutrik Mandal performed a re - analysis of a study originally conducted by the Open Science Collaboration (OSC) whose goal was to assess the reproducibility of statistical results in psychology. In the original analysis performed by the OSC, 100 studies were chosen from some of the most important psychology journals, and their conclusions were that despite 97% of the

considered studies presented statistically significant results (where a result was deemed to be significant if presented a $p - value \leq 0.052$), only the 36% of the reproduced tests presented again a statistically significant result, and the effects (always in the reproduced tests) were on average half the magnitude of the original ones. One of the main reasons behind these low rates of reproducibility, according to the authors, was the so called *Publication Bias*, which essentially consists in the selective choice of the reported p - values. To better understand how the Publication Bias can affect the validity of the p - value based conclusions, consider the extreme case in which every researcher, after having conducted many experiments, reported only those yielding a p - value below a fixed threshold. Not only would be impossible to properly interpret these reported p - values , but it would also follow an artificial proliferation of statistically significant results, as it will be seen appears to be the case in psychology.

One of the conclusions the authors draw, working on the OSC dataset, is that once considering (estimating) how many statistical analyses have been conducted and how many tests have been performed, taking into account also the publication bias, in psychology the fraction of the Null Hypotheses which are true can be assumed to be the 90%. In other words, according to the OSC data, the probability that, in a published test with p - value = 0.05, the Null Hypothesis of absence of relation (or, at most, of the presence of only a negligible effect) is true is 0.9.

Despite the OSC considered 100 experiments in their replication, most of their findings were based on a subset of them called the *Meta - Analytic* (MA) subset, encompassing 73 studies. The particular choice of these studies is due to a particular statistical feature they share: indeed, in such studies, it is possible to transform the observed effect sizes to the correlation scale. The main advantages of such transformations are related with the fact that correlation coefficients can be easily interpreted, and that considering a z - transformation the standard errors of the transformed coefficients depend only on the sample size of the considered study. Defining as r the sample correlation coefficient based on a generic bivariate normal sample

of size n and ρ the population correlation coefficient, Fisher (1915) proved that:

$$Z = \frac{1}{2}log(\frac{1+r}{1-r}) \sim N(\frac{1}{2}log(\frac{1+\rho}{1-\rho}),\ \frac{1}{n-3})$$

(to be more precise, Z is approximately distributed as the above defined Normal). For these reasons, also the authors rely, for their analysis, on the MA subset: such an approximation, despite a possible loss in efficiency, they will find that does not produce any immediate distortion in the fraction of the tested hypotheses which are true. Moreover, in the MA subset, the 96% of the included studies (70 out of 73) reported significant results, which is almost in line with the overall sample of the OSC study, where we have already said that statistically significant results were reported in the 97% of the studies. Already from such percentages, it can be deducted the presence of a severe Publication Bias. Hence, in order to proceed with the estimation of the quantities of interest, the authors describe the statistical model they have considered: we provide a brief description of it as well.

5.1 A statistical model for the MA subset

Let define $\{z_{i,1}\}_{i=1}^{M}$ as the collection of all the z coefficients for tests $i \in \{1,,,M\}$ observed in the original tests, and $\{z_{i,2}\}_{i=1}^{M}$ similarly for the replicated tests. Moreover, let $\{n_{i,1}\}_{i=1}^{M}$ and $\{n_{i,2}\}_{i=1}^{M}$ be the collections of the associated sample sizes. The first assumption the authors make is that the 73 z coefficients in the MA subset represent just a part of the larger M population of tests - which includes also those tests that have not been published because lacking significativity or simply because not of interest - and whose size is also of interest to be estimated. Another reasonable assumption they make is that a result statistically significant is always published. To formalize these two assumptions, they define an M dimensional vector $\boldsymbol{\zeta} = \{\zeta_i\}_{i=1}^{M}$ including the M z - transformed population correlation coefficients for each test, and another vector $\boldsymbol{I} = \{\boldsymbol{I}_i\}_{i=1}^{M}$ of indicators function assuming value 1 if the test statistic i was published and 0 otherwise. It follows that:

$$I_i = 1 \; \forall \; i \in \{1, , , 73\} \wedge I_i = 0 \; \forall \; i \in \{74, , , M\}.$$

Moreover, they define another vector of indicators $\boldsymbol{J} = \{\boldsymbol{J}_i\}_{i=1}^{M}$ assuming value 1 if the original study yielded a statistically significant result (that is yielded a p - value < 0.052) and 0 otherwise. Another assumption they make is that the probabilty that a test statistic is published given that it is not significant is $0 < \alpha < 1$, while the probability that a test statistic is published given that it is significant is 1. Formally:

$$P(I_i = 1 \mid J_i = 0) = \alpha \wedge P(I_i = 1 \mid J_i = 1) = 1.$$

Another quite strong assumption they make is the independence between the test statistics obtained from the different tests.

Moreover, they assume that the p - value observed when testing a Point Null Hypothesis is adequate in approximating the p - value observed when testing an Interval Null Hypothesis (such approximation has been described in section 1.3). Finally, they consider two different possible priors for the mean of the z - transformed correlation coefficients under H_1 (that is rejected Null Hypothesis of absence of relation; non - zero effect). One is the *normal moment function*, indexed in τ:

$$f(\zeta_i \mid \tau) = \frac{\zeta_i^2}{\tau\sqrt{2\pi\tau}} exp\Big\{-\frac{\zeta_i^2}{2\tau}\Big\}$$

while the other is a Normal with mean 0 and variance τ:

$$f(\zeta_i \mid \tau) = \frac{1}{\sqrt{2\pi\tau}} exp\Big\{-\frac{\zeta_i^2}{2\tau}\Big\}.$$

However, using a Bayesian chi - squared test for model fit, they find that only the first prior yields an adequate model for the transformed population coefficients. Let finally define the variance of $z_{i,j}$ as $\sigma_{i,j}^2 = (n_{i,j} - 3)^{-1}$. It follows that the joint sampling density of the z - transformed correlation coefficient of an original, published and statistically significant result

is given by:

$$f(z_{i,1},\, J_i = 1,\, I_i = 1 \mid \zeta) = P(I_i = 1 \mid J_i = 1)\phi(z_{i,1} \mid \zeta_i,\ \sigma^2_{i,j})\ I(|z_{i,1}| \geq \sigma_{i,j}\ q_{0.974})$$

$$= \phi(z_{i,1} \mid \zeta_i,\ \sigma^2_{i,j})\ I(|z_{i,1}| \geq \sigma_{i,j}\ q_{0.974})$$

where $\phi(z_{i,j} \mid \zeta_i,\ \sigma^2_{i,j})$ is the marginal sampling distribution of the z - transformed correlation coefficient for the study i repetition j ($j = 1$ original study, $j = 2$ repetition). Instead, $q_{0.974}$ is the 0.974 quantile from a standard normal distribution. Similarly, the joint sampling density of the z - transformed correlation coefficient of an original, published but not statistically significant result is given by:

$$f(z_{i,1},\, J_i = 0,\, I_i = 1 \mid \zeta) = P(I_i = 1 \mid J_i = 0)\phi(z_{i,1} \mid \zeta_i,\ \sigma^2_{i,j})\ I(|z_{i,1}| \leq \sigma_{i,j}\ q_{0.974})$$

$$= \alpha\ \phi(z_{i,1} \mid \zeta_i,\ \sigma^2_{i,j})\ I(|z_{i,1}| \leq \sigma_{i,j}\ q_{0.974})$$

and finally, the joint sampling density of the z - transformed correlation coefficient of an original, unpublished and not statistically significant result is given by:

$$f(z_{i,1},\, J_i = 0,\, I_i = 0 \mid \zeta) = P(I_i = 0 \mid J_i = 0)\phi(z_{i,1} \mid \zeta_i,\ \sigma^2_{i,j})\ I(|z_{i,1}| \leq \sigma_{i,j}\ q_{0.974})$$

$$= (1 - \alpha)\ \phi(z_{i,1} \mid \zeta_i\ ,\ \sigma^2_{i,j})\ I(|z_{i,1}| \leq \sigma_{i,j}\ q_{0.974})$$

where we recall that it has been assumed that $P(I_i = 0 \mid J_i = 1) = 0$. Observe that, for $i > 73$, the authors consider the associated $z_{i,1}$ (that is the z - transformed coefficients for all the M - 73 tests who have not produced statistically significant results and that have

not been published) as missing data. For what it concerns the replicated studies, the density of the $z_{i,2}$ is clearly independent from I_i and J_i, and therefore it is given by:

$$f(z_{i,2} \mid \zeta_i) = \phi(z_{i,2} \mid \zeta_i \ , \sigma^2_{i,j}).$$

At this point, let finally determine another M dimensional vector $\boldsymbol{Z} = \{Z_i\}_{i=1}^{M}$, where each Z_i consists in a random variable distributed according to a Bi(1,1 - π_0), which assumes value 0 if $\zeta_i = 0$ (with probability π_0), and 1 otherwise.

$$W_i = \begin{cases} 0 & \text{if } \zeta_i = 0, \quad prob = \pi_0 \\ 1 & \text{if } \zeta_i \neq 0, \quad prob = 1 - \pi_0 \end{cases}.$$

Hence, for all the M tests (70 published with statistically significant results, 3 published with not statistically significant results and M - 73 not published), the prior density on the mean of the z - transformed coefficients, given the value of the scale parameter τ and whether ζ_i itself is equal or different from 0 is given by, considering the previously defined moment prior density:

$$f(\zeta_i \mid \tau, W_i) = (1 - W_i)\delta_0 + W_i \frac{\zeta_i^2}{\tau\sqrt{2\pi\tau}} exp\left\{-\frac{\zeta_i^2}{2\tau}\right\}$$

where δ_0 indicates a unit mass at 0. Assuming a Jeffreys prior for α (probability that a not statistically significant result is published) and π_0 (prior probabilty of the Null Hypothesis of absence of effect), a prior proportional to $\frac{1}{\tau}$ for τ and a prior for M proportional to $\frac{1}{M^2}$, the joint posterior distribution they finally present is:

$$f(\boldsymbol{z}, \boldsymbol{\zeta}, \boldsymbol{W}, M, \pi_0, \tau, \alpha \mid D) \propto \binom{M}{70\ 3} \prod_{j=1}^{2} f(z_{i,j},\ I_i = 1,\ J_i = 1 \mid \zeta_i)$$

$$\times \prod_{j=1}^{2} f(z_{i,j},\ I_i = 0,\ J_i = 0 \mid \zeta_i)$$

$$\times \prod_{j=1}^{2} f(z_{i,j},\ I_i = 1,\ J_i = 0 \mid \zeta_i)$$

$$\times \prod_{i=1}^{M} f(\zeta_i \mid \tau, W_i)\pi_0^{1-W_i}(1-\pi_0)^{W_i}$$

$$\times \frac{1}{\tau M^2}\ (\pi_0)^{-\frac{1}{2}}\ (1-\pi_0)^{-\frac{1}{2}}\ (\alpha_0)^{-\frac{1}{2}}\ (1-\alpha_0)^{-\frac{1}{2}}$$

where D represents $\{z_{i,j}\}$, $I_i = 1$ and J_i for $i \in \{1, ..., 73\}$. Note that $I_i = J_i = 0$ for $i > 73$ (given the assumption that all the statistically significant results have been published). In explaining the expression, the authors consider the first combinatorial term as necessary to define in how many ways the 70 published studies with statistically significant results, the 3 published studies with not significant results and the remaining M - 73 unpublished results with not statistically significant results could have occurred in the M performed studies. Moreover, while the authors precise that the density of ζ_i can either be a moment or a normal density, since later, in the paper, they find that only the moment density fits the model, we immediately consider it to be only this one.

The problem the authors recognize with the obtained density is that, given that it is of interest to make inference on M, π_0, α and τ, the parameters $\boldsymbol{z}$, $\boldsymbol{\zeta}$ and $\boldsymbol{W}$ can be considered as *nuisance* parameters, and therefore it is necessary to marginalize with respect to them. The final expression they obtain is the following:

$$f(M,\, \alpha_0,\, \pi_0,\, \tau \mid D) \propto \binom{M}{70\ 3} \prod_{i=1}^{73} A_i(\alpha, \pi_0, \tau) \prod_{i=74}^{M} B_i(\alpha, \pi_0, \tau)$$

$$\times \frac{1}{\tau M^2}\ (\pi_0)^{-\frac{1}{2}}\ (1-\pi_0)^{-\frac{1}{2}}\ (\alpha)^{-\frac{1}{2}}\ (1-\alpha)^{-\frac{1}{2}}$$

where $A_i(\alpha,\, \pi_0,\, \tau)$ is defined as:

$$A_i(\alpha,\, \pi_0,\, \tau) = \int f(\zeta_i \mid \tau, W_i)\pi_0^{1-W_i}(1-\pi_0)^{W_i} \prod_{j=1}^{2} f(z_{i,j},\ I_i,\ J_i \mid \zeta_i) d\zeta_i dW_i$$

for all $i \leq 73$ and a given M, while $B_i(\alpha,\, \pi_0,\, \tau)$ si defined as:

$$B_i(\alpha,\, \pi_0,\, \tau) = \int f(\zeta_i \mid \tau, W_i)\pi_0^{1-W_i}(1-\pi_0)^{W_i} \prod_{j=1}^{2} f(z_{i,j},\ I_i,\ J_i \mid \zeta_i) dz_{i,1} dz_{i,2} d\zeta_i dW_i$$

for $i > 73$ and a given M. The main problem related with $f(M,\, \alpha_0,\, \pi_0,\, \tau \mid D)$ is that it does not allow any explicit computation of the marginal posterior density of the four parameters: however, MCMC algorithm can be implemented to perform the analysis. The results they report are presented in the following table:

Model	Parameter	π_0	τ	α
Moment prior	Posterior Mean	0.930	0.0877	0.00569
Moment prior	95% credible interval	(0.884, 0.961)	(0.0604, 0.1252)	(0.00129, 0.0138)

Table 26: Johnson et al. (2016), estimates of the re - analysis based on the MA subset

and the estimate of M is 706, with the 95% confidence interval being (495,956). A first result is that, on 706 estimated total number of publications (which means an estimated 633 unpublished not statistically significant results) the estimated Prior Probability that the Null is true (that is $\zeta_i = 0$, absence of relation between the two variables) is 0.93: this result is in line with the fact that the estimated value of the probability that a non statistical significant result is published is 0.5 %. This means that if statistically significant results are considered those whose p - value is below the 0.052 threshold, if 700 tests were performed, the false positives - that is the cases in which the absence of correlation was wrongly rejected - would have been on average $700 \times 0.93 \times 0.052 = 34$. What the authors also observe is that,

starting from the fact that given that in the original OSC study the researchers observed that the average power (probability of rejecting a false Null Hypothesis, that is rejecting the absence of relation when indeed there is) for those tests whose statistical significance was 5% is 0.75, from which on 700 tests on average they would have been observed 700 $\times$ 0.07 $\times$ 0.75 = 37 true positives (where 0.07 = 1 - 0.93 is the estimated Prior Probability that the Null Hypothesis is false, that is that $\zeta_i \neq 0$). Hence, according to the estimates, they would have been observed 34 + 37 = 71 positive findings: it is indeed recalled that in the original MA subset, containing 73 studies, 70 reported statistically significant results.

Moreover, what the authors observe is that of this hypothetical population of 700 studies, among the false positives (cases in which the Null Hypothesis of absence of relation has been wrongly rejected) would replicate only 34 $\times$ 0.052 = 2 while, among the 37 true positive, they would replicate 37 $\times$ 0.75 = 28. Hence, according to the authors estimates, of the 73 studies, they would be expected to replicate 2 + 28 = 30, which is essentially in line with what is observed in the MA subset, where of the 70 statistically significant results 28 replicated.

To finally produce a comparison with Bayesian estimates, the authors derive the Bayes Factor for each experiment. They find that, letting n_i being the sample size of the considered experiment i, the Bayes Factor in favor of H_1 (expressed as function of the sample correlation coefficient r) is:

$$B_{1,0}(r) = \frac{1}{\tau\sqrt{\tau d_1}}(\frac{1}{d_1} + d_2^2)exp\Big\{d_1\frac{d_2^2}{2}\Big\}$$

where:

$$d_1 = n \text{ - } 3 + \tau^{-1};\ \ d_2 = \frac{z_i(n-3)}{d_1}$$

and that the Posterior Probability of the Null Hypothesis is given by:

$$P(H_0 \mid r) = \Big\{1 + \frac{(1-\pi_0)}{\pi_0}B_{1,0}(r)\Big\}^{-1}$$

where for π_0 and τ they can be considered the previously presented estimates. What the authors observe is that, for instance, when $n = 10$ and p - value $= 0.05$, the Posterior Probability of H_0 is 0.842, which means that the Null Hypothesis is rejected at a 0.05 significance level being instead estimated to be true with a probability of 0.842. A similar conclusion can be made for $n = 30$, and with $n = 100$ the Posterior Probability of H_0 becomes even greater (close to 1). Another interesting result is that for both $n = 30$ and $n = 10$, to a p - value of 0.005 it corresponds an estimated Posterior Probability of the Null Hypothesis of about 0.6, and for p - value $= 5 \times 10^{-5}$ when $n = 10$ the Posterior Probability of H_0 is still between 0.15 and 0.20, while only with $n = 30$ is close to 0.

5.2 Final remarks

The main conclusion that can be done is that - at least considering the OSC dataset (and, in particular, its MA subset) - the misuse of the p - value is behind a strong Publication Bias, which leads to low reproducibility rates in psychology.

Moreover, considering the comparison made between the p - value and the Posterior Probability of the Null Hypothesis of absence of relation (or negligible effect), it appears that the discrepancy between the two measures is quite strong.

It has been discussed that the problems arising with the use of the p - value are not due to the measure itself, but by both its misinterpretations as a frequentist I Type Error probability or as the Posterior Probability of the Null Hypothesis. Indeed, the p - value consists in just a measure to summarize the evidences provided by the observed data against a proposed model for them (Wasserstein 2016). For this reason, in particular in those context where a binary decision has to be taken, it can't be considered alone.

Moreover, it has been observed that computing the Posterior Probability of the Null Hypoth-

esis consists in a reliable way to assess the consequences of the misuse of the p - value fallacy. It has been shown that, in some cases, to a p - value of 0.05 it is associated a Probability of the Null Hypothesis even greater than 0.5. However, also this measure has its limitation, as, for instance, the Jeffreys - Linldey paradox, or the determination of the Prior Probabilities of the two Hypotheses (customary considered as equally probable). For this reason, the Posterior Probability of the Null Hypothesis should not be considered as a substitute of the p - value, but as a complementary result which enriches the study and provides more validity to the conclusions.

The suggestion is that further investigations are necessary, both in more theoretical as applied research. The discrepancy that has been found between the Posterior Probability of the Null Hypothesis and the p - value in testing a Point Null Hypothesis on the scale of parameter in a Gamma model with known shape parameter is, in the very end, not surprising. However, it is needed to precisely assess the magnitude of this discrepancy, in particular considering other values of λ_0, and taking into account different priors under the Alternative Hypothesis. A more general extension appears necessary to all the distributions.

For what it concerns the study of the reproducibility rates, it would be of interest to observe the results that would arise in other fields using the same statistical model considered by Johnson et al. (2016).

Bibliography

[1] Aitkin M., (1991) "Posterior Bayes Factor", *Journal of the Royal Statistical Society, Series B (Methodological)*, *Vol.*53, *No.*1, *pp.*111 - 142.

[2] Asher A., Johnson Valen E., Mandal S., Payne Richard D., Wang T., (2016) "On the reproducibility of psychological science", *Journal of the American Statistical Association, Accepted Manuscript.*

[3] Bayarri M.J., Berger James O. , Sellke T., (2001) "Calibration of p - value for testing Precise Null Hypotheses", *The American Statistician*, *Vol.*55, *No.*1 *pp.*62 − 71.

[4] Bayarri M.J., Hubbard R., (2003) "Confusion Over Measures of Evidence (p's) versus Errors (α's) in Classical Statistical Testing", *The American Statistician*, *Vol.*57, *No.*3, *pp.*171 − 178.

[5] Berger James O., (2003) "Could Fisher, Jeffreys and Neyman have agreed on Testing?", *Statistical Science*, *Vol.*18, *No.*1, *pp.*1 − 32.

[6] Berger James O., (1985) "*Statistical decision theory and Bayesian analysis : with* 23 *illustrations*, 2. *ed*", New York, Springer.

[7] Berger James O., Delampady M., (1987) "Testing Precise Hypotheses", *Statistical Science*, *Vol.*2, *No.*3, *pp.*317-335.

[8] Berger James O., Sellke Thomas, (1987) "Testing a Point Null Hypothesis : The Irreconcilability of P - Values and Evidence", *Journal of the American Statistical Association*, *Vol.*82, *No.*397, *pp.*112-122.

[9] Bernardo Jose' M., Smith Adrian F.M. (2000) "*Bayesian Theory*", Cichester, Wiley.

[10] Chopin N., Robert Christian P., Rosseau J., (2009) "Harold Jeffreyss Theory of Probability Revisited", *Statistical Science*, *Vol.*24, *No.*2, *pp.* 141 - 172.

[11] Cifarelli Donato M., Muliere P., (1989) " *Statistica Bayesiana*", Pavia, Iuculano Editore.

[12] Dickhaus T., (2014) "*Simultaneous Statistical Inference with applications in the Life Sciences*", New York, Springer.

[13] Edwards Ward, Lindman Harold, Savage Leonard J. (1963) "Bayesian Statistical Inference for Psychological Research", *Psychological Review*, *Vol.*70, *No.*3, *pp.* 193 - 242.

[14] Fisher Ronald A., (1971) "*The Design of Experiments*", New York, Hafner Publishing Company.

[15] Fisher Ronald A., (1934) "*Statistical Methods for Research Workers*", F. A. E. Crew, Edinburgh D. Ward Cutler, Rothamsted, Central Agricultural Library.

[16] Johnson Valen E., (2004) "A Bayesian χ^2 Test for Goodness - of - fit", *The Annals of Statistics*, *Vol.*32, *No.*6, *pp.*2361 - 2384.

[17] Kulinskaya E., (2007) "On two - sided p - values for non - symmetric distributions", *arXiv* : 0810.2124.

[18] Lazar Nicole A., Wasserstein Ronald L., (2016) "The ASA's Statement on p - Values: Context, Process, and Purpose", *The American Statistician*, $Vol.70$, $No.2$, *pp.* 129 - 133.

[19] Lehmann E.L. , Romano Joseph P., (2005) "*Testing statistical hypotheses*", New York, Springer.

[20] Mukhopadhyay Nitis, (2000) "*Probability and Statistical* $Inference$", New York, Marcel Dekker.

[21] Robert Christian P., (1993) "A note on the Jeffreys - Lindley Paradox", *Statistica Sinica*, $Vol.3$, $No.2$, *pp.*601 - 608.

[22] Robert Christian P., (2013) "On the Jeffreys Lindleys paradox", $arXiv$: 1303.5973.

[23] Rohatgi Vijay K., Saleh A.K. Md Ehsanes, (2001) "*An introduction to probability and statistics*", New York, Wiley.