Hypothesis Testing

Contents

There are some problems we meet in statistical practice in which estimation of a parameter is not the primary goal of the analysis; rather, we wish to use our data to decide to trust or not a particular claim about a given population. For example, a drug company may claim that its new drug reduces "bad" cholesterol (or LDL cholesterol) level by more than 20 points (after a certain period of treatment). To try to prove this, the drug company may design a clinical trial and collect data on the level of LDL reduction in selected subjects. The observed data can be analyzed to confirm (or refute) the manufacturer's claim. Properly formulated, such inference problems are called *testing of hypothesis* problems or simply test problems.

1 Basic concepts

Before attempting any inferential procedure, the model and all its underlying assumptions must clearly be defined. This is referred to the ***maintained hypothesis***. These hypothesis define the framework within which the inferential procedure can be applied and interpreted correctly. In the context of the parametric hypothesis testing problem, the maintained hypothesis are simply the assumption made on the distribution of the observed data. In simple situations, this reduces to assuming that we observe an iid sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from a pd $f(x;\theta)$, for some $\theta \in \Theta$ (real or a vector parameter space).

Based on the observed data, we have to decide whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$, where Θ_0, Θ_1 are a partition of the parameter space Θ, i.e. $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$. We formulate this in the following manner:

$$H_0 : \ \theta \in \Theta_0 \ \ vs \ H_1 : \ \theta \in \Theta_1.$$

The hypothesis H_0 is called the ***null hypothesis*** and H_1 is called the ***alternative hypothesis***. In a statistical test procedure, these two hypothesis play an asymmetric role:

- H_0 *represents the* ***current theory***: "null" means statu quo, no change or no effect. Typically, H_0 is also simpler (lower-dimensional) than the alternative hypothesis.
- H_1 *is referred as the* ***researcher's hypothesis***: it is the new claim that one would really like to validate or the question to be answered.

Example 1.1. Take the cholesterol example. Suppose that, in a clinical trial, the observed LDL *decrease* of n subjects are $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 25)$, for some $\mu \in (-\infty, \infty)$. Here, iid and normality (with var. $= 25$) are the maintained hypothesis. The drug company's claim is $\mu > 20$. This claim can be formulated as a test problem with two hypothesis: the null hypothesis $H_0 : \ \mu \leq 20$, and the alternative hypothesis $H_1 : \mu > 20$. □

A hypothesis is called ***simple*** if it completely specifies the underlying distribution; otherwise it is called ***composite***. In our example above, we are testing a composite null against a composite alternative.

In addition, a test is called ***one-tailed*** if the alternative hypothesis is articulated directionally ($H_1 : \theta < \theta_0$ or $H_1 : \theta > \theta_0$, for some given θ_0); otherwise it is called ***two-tailed*** ($H_1 : \theta \neq \theta_0$). In our example above, we have a one-tailed test.

We will use our data to choose between the two hypothesis H_0 and H_1. For that, the basic idea is to try to *"compare" the observed sample with the model under H_0 and under H_1*. The comparison is typically based on a statistic $T \equiv T_n(\boldsymbol{X})$, called the ***test statistic***, which is designed in order to measure the discrepancy between the data and the models under H_0 and under H_1. *Most of the time, a test statistic is taken to be a (function of a)* ***sufficient statistic*** *for the parameter of interest θ, whose distribution (under H_0) is known exactly or asymptotically.*

Assume that an appropriate test statistic T has been chosen and, without loss of generality, *suppose that small values of T support H_0, while large values support H_1*. The next step is to select a number, called the ***critical value***, k and apply the following rule:

> if $T_n(\boldsymbol{x}) \geq k$, reject H_0 in favor of H_1; otherwise, do not reject H_0.

In this way, our *test is nothing but the statistic/function* $\varphi_n(\boldsymbol{X}) = I(T_n(\boldsymbol{X}) \geq k) : \mathbb{R}^n \mapsto \{0, 1\}$, called the ***test function*** or simply test, that takes only values 0 and 1:

- $\varphi_n(\boldsymbol{x}) = 1 \Leftrightarrow$ reject H_0 in favor of H_1.
- $\varphi_n(\boldsymbol{x}) = 0 \Leftrightarrow$ do not reject H_0.

The subset $\mathcal{R}_n = \{\boldsymbol{x} : \ T_n(\boldsymbol{x}) \geq k\}$ of the sample space for which H_0 will be rejected is called the ***rejection region or critical region***. $\mathcal{R}_n^c$, the complement of the rejection region, is called the ***non rejection region***.

Remark The tests used in most practical situations can be expressed as $I(T_n \leq k)$ or $I(T_n \geq k)$. We use $I(T_n \leq k)$ in situations where T_n tends to be small under H_1 and $I(T_n \geq k)$ in situations where T_n tends to be large under H_1. In these notes, we will focus on the last case, but all the methodology can be applied directly to the first situation by using $-T_n$ as the actual test statistic.

Example 1.2. In the cholesterol example, our hypothesis are $H_0 : \ \mu \leq 20$ vs $H_1 : \ \mu > 20$.

As a consistent estimator of μ, $\overline{X}_n$ tends to be larger under H_1 than under H_0. Hence, it is natural to reject H_0 for large values of $\overline{X}_n$ and so to consider, for example, the test function $\varphi_1 = I(\overline{X}_n \geq 21)$. The same reasoning can be applied to the sample median, say M_n, which leads to the test function $\varphi_2 = I(M_n \geq 21)$. The first test φ_1 uses $\overline{X}_n$ as test statistic, while the second test φ_2 uses M_n as test statistic. The rejection region of φ_1 is $\{(x_1 \ldots, x_n) : \ \bar{x}_n \geq 21\}$ and that of φ_2 is $\{(x_1 \ldots, x_n) : \ m_n \geq 21\}$. □

This example raises a couple of questions:

- Why 21 is used as the critical value? How to choose a suitable critical value?
- How to choose between φ_1 and φ_2 ? Is there an "optimal" test? If so, how to find it?

2 Evaluating a test

In a situation where the analyst has to decide between H_0 and H_1, and given that in reality one of these hypothesis is true and the other is false, there are four possible states summarized in the following table:

Decision/Reality	H_0 is true	H_1 is true
Do not reject H_0	Correct	Error of type II
Reject H_0	Error of type I	Correct

For a given test φ_n, the function defined on Θ by

$$\pi_n(\theta) = E_\theta(\varphi_n(\boldsymbol{X})) = P_\theta(\varphi_n(\boldsymbol{X}) = 1) = P_\theta(\text{Reject } H_0)$$

is called the ***power function*** of the test φ_n. This function indicates the probability of rejecting H_0 for every possible value of θ. It plays an important role in evaluating the quality of a test and in comparing different test procedures.

Related to the power function we define:

- ***Probability of type I error***. This is the probability that the test rejects a true null hypothesis; i.e. $P_\theta(\text{Reject } H_0|H_0)$ or, equivalently, $\pi_n(\theta)$, when θ (the true value) $\in \Theta_0$.
- The ***size*** is the largest probability of committing a type I error, i.e. $\max_{\theta\in\Theta_0} \pi_n(\theta)$. If the size of a test is known not to exceed a given $\alpha \in (0,1)$, i.e. $\max_{\theta\in\Theta_0} \pi_n(\theta) \leq \alpha$, then we say that the test is of (significance) ***level*** α.
- ***Power***. This is the probability that the test rejects a false null hypothesis; i.e. $P_\theta(\text{Reject } H_0|H_1)$ or, equivalently, $\pi_n(\theta)$, when $\theta \in \Theta_1$.
- ***Probability of type II error***. This is the probability that the test does not reject a false null hypothesis; i.e. $P_\theta(\text{Do not reject } H_0|H_1) = 1 - \text{Power}$ or, equivalently, $1 - \pi_n(\theta)$, when $\theta \in \Theta_1$.

Example 2.1. Recall our cholesterol example, with $H_0 : \mu \leq 20$ vs $H_1 : \mu > 20$. Assuming that $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 25)$, the power function of the test $\varphi_n = I(\overline{X}_n \geq 21)$ is

$$\pi_n(\mu) = P_\mu(\overline{X}_n \geq 21) = 1 - \Phi\left(\sqrt{n}\frac{21-\mu}{5}\right),$$

where Φ is the cdf of $N(0,1)$.

As this is an increasing function in μ, the size of our test φ_n is $\max_{\mu \leq 20} \pi_n(\mu) = \pi_n(20) = 1 - \Phi(\sqrt{n}/5)$. For example, with $n = 10$, the size is 0.264 and with $n = 100$ the size is 0.023.

Here is the plot of $\pi_n(\mu)$, for different values of the sample size n.

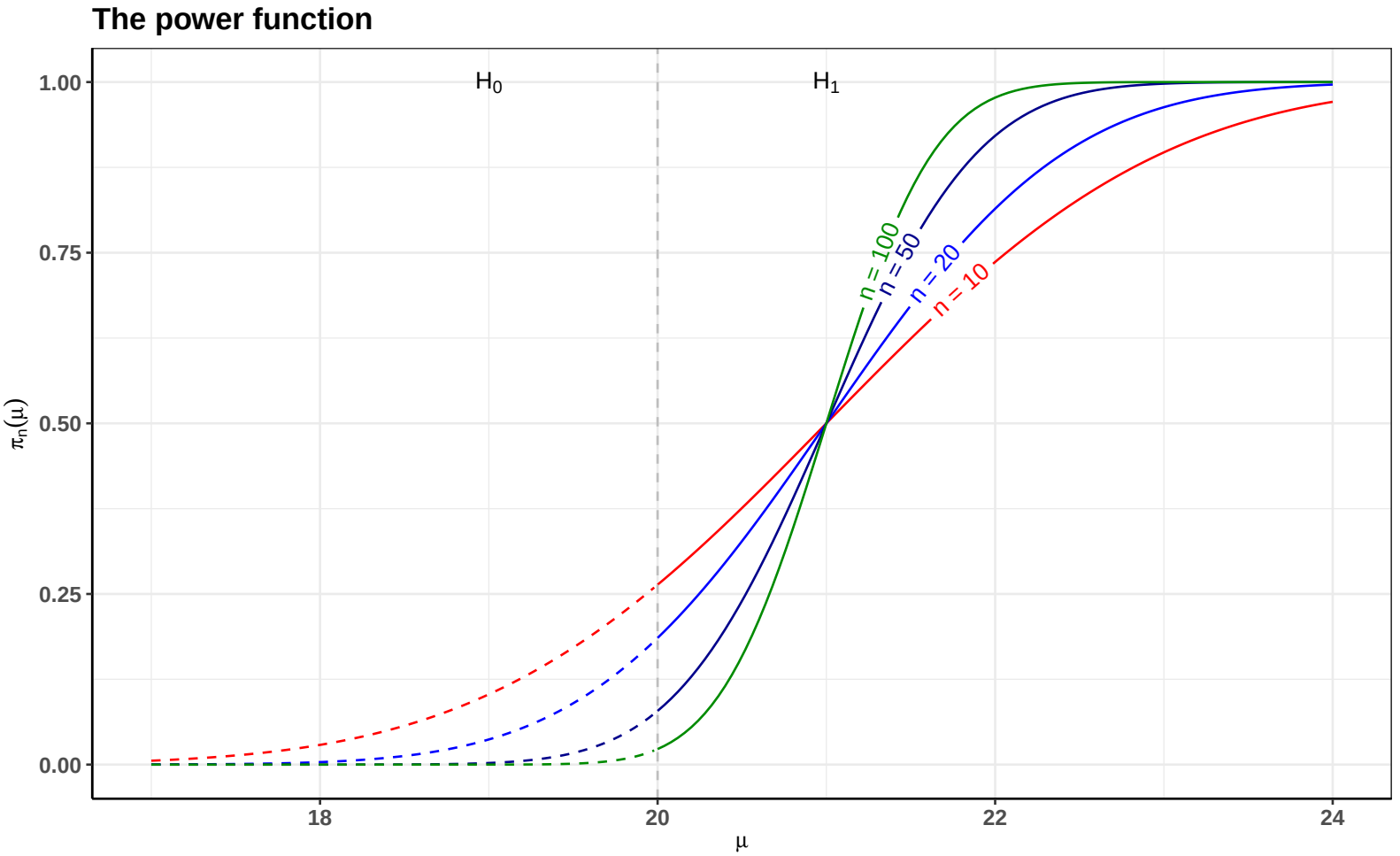


Figure 1: Power (solid line) and Type-I-error probability (dashed line)

For n fixed, as the true μ moves away from 20, the power increases to 1 and the probability of type I error decreases to 0. On the other hand, for fixed μ, as n increases, the power increases (for μ larger than 21) and the probability of type I error decreases. □

Ideally, we would like to have a statistical test with the lowest probability of Type I error (0) and the highest power (1). For a fixed sample size, *it is usually impossible to control both Type I error and power*. In fact, to minimize the probability of type I error, one must not reject H_0 more often and to maximize the power, one must reject H_0 more often: these two goals work against each other. Typically, even in very simple problems, larger power comes at the expense of a larger probability of Type I error. Vice versa, when one tries to reduce the probability of type I error, the power also gets reduced. All that can be done is to control one of these two quantities by adjusting the critical value of the test, or, if possible, by increasing the sample size. Here is an example.

Example 2.2. Let $X_1, \ldots, X_n \overset{iid}{\sim} Ber(\pi)$, where $\pi \in \{0.3, 0.5\}$. Consider testing

$$H_0 : \ \pi = 0.5 \ \ vs \ \ H_1 : \ \pi = 0.3.$$

Put $T_n = \sum_{i=1}^{n} X_i$. It is reasonable to reject the null hypothesis if the total number of successes T_n is "too small". Let's define the test $\varphi_n = I(T_n \leq k)$, where k is the critical value that needs to be fixed. The power function of φ_n is given

by

$$\pi_n(\pi) = P_\pi(T_n \leq k) = \sum_{\ell=0}^{k} C_n^\ell \pi^\ell (1-\pi)^{(n-\ell)}.$$

For example, with $n = 15$ we get (using the `R` function `pbinom(k, size = n, prob =`π`)`)

	Type-I-Error	Power
k	$P(T_n \leq k \vert \pi = 0.5)$	$P(T_n \leq k \vert \pi = 0.3)$
1	4.88×10^{-4}	0.03
4	0.06	0.52
7	0.50	0.95
10	0.94	≈ 1

For example, we can see that the test $\varphi_n = I(T_n \leq 4)$ gives a low type-I-error probability (0.06), but gives also a low power (0.52) and therefore does not provide adequate protection against a type-II-error (probability 0.48). We can increase the power to 0.95 by taking $\varphi_n = I(T_n \leq 7)$. However, by doing this, we increase the type-I-error probability to 0.5.

This demonstrates the conjunction of type-I-error probability and power of a test. The only way to decrease the probability of type I error and to increase the power simultaneously is to increase the sample size and to choose an

appropriate critical value. For example, with $n = 150$ and $k = 60$, we get

k	Type-I-Error	Power
60	0.009	0.996

□

3 Testing strategy: the Neyman-Pearson approach

Given that it is not possible to reduce simultaneously the probabilities of both types of errors, Neyman and Pearson suggested to hold one of the two error probabilities at a pre-specified level that can be considered tolerable, and then minimize the other error probability. More specifically, Neyman-Pearson framework of statistical hypothesis is to

(1) choose a small number $\alpha \in (0, 1)$ ($\alpha = 0.01, 0.05,$ and 0.1 are commonly used in practice). Restrict attention to tests with the pre-chosen level α only,

(2) *among these tests*, try to find a test that has the largest power. When it exists, such a test is called the ***uniformly most powerful test*** (UMP) of level α. More precisely, a test φ^* of level α is UMP if $\forall \varphi$ of level α, $\pi_n^{\varphi^*}(\theta) \geq \pi_n^{\varphi}(\theta)$, $\forall \theta \in \Theta_1$, $\forall n$.

Since reducing type I error also reduces power, to obtain a UMP test (with a pre-chosen α level), one must necessarily stick to tests with the highest possible type I error (not exceeding α).

To apply the Neyman-Pearson approach, two questions must be answered:

(1) how to control the level of a given test and fix it at the pre-chosen α ?
(2) how to find an UMP test (*when it exists*) ?

We will defer this last question to Section 5 and discuss here the first one. Before, observe that the Neyman-Pearson approach makes H_0 **and** H_1 **asymmetric**: we control the type-I-error probability to be at most α; then, we try to make the type-II-error probability *as small as possible*. However, this latter can still be very big even for an UMP test. If we take this approach, the test must be set up so that H_1, the alternative hypothesis, is the one we seek to prove (this is why we refer to it as the *researcher's hypothesis*). By using an α level test, with small α, we guard against saying that the data support the research hypothesis (H_1) when it is false.

Now regarding our first question, without lost of generality and to make things clearer, let's consider the case where our test is given by $\varphi_n = I(T_n \geq k)$, i.e. large values of T_n give evidence against H_0. To fix the level of φ_n to α, we must choose an appropriate critical value k. More precisely, we must choose *the smallest* $k \equiv k(n, \alpha)$ such that

$$\max_{\theta \in \Theta_0} P_\theta(T_n \geq k) \leq \alpha.$$

Finding such a k is greatly simplified if ***the null distribution*** *of the test statistic T_n (i.e. the distribution of T_n under H_0) is known*; otherwise, k can be obtained by *Monte Carlo methods* (not discussed here).

Once the critical value k is chosen, the next step is to calculate $t_n = T_n(\mathbf{x})$, the observed value of T_n, and then to take one of the following decisions:

- If $t_n \geq k$, then, *at level* α, the data **reject** H_0 in favor of H_1 because the event $\{T_n \geq k\}$ is quite rare (it occurs with probability at most α) under H_0. We say that the ***test is significant***.
- If $t_n < k$, then, *at level* α, the data **do not reject** H_0. In this case, either
  - the power is known to be "very large" $\Longrightarrow$ **accept** H_0, i.e reject H_1 in favor of H_0
  - or the power is unknown or known to be "small". In this case, accepting H_0 is a risky decision $\Longrightarrow$ we don't have enough information (i.e. sufficient sample size) to accept or reject H_0. In this case, we say that the ***test is non significant***.

Example 3.1. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ is known. For some given μ_0, consider testing

$$H_0 : \ \mu \leq \mu_0 \ \ vs \ \ H_1 : \ \mu > \mu_0.$$

We have seen that it is natural to reject H_0 when $\overline{X}_n$ is "large". For ease of calculation, it is convenient to replace $\overline{X}_n$ by the test statistic $Z_n = \sqrt{n}\dfrac{\overline{X}_n - \mu_0}{\sigma}$. So we consider the test $\varphi_n = I(Z_n \geq k)$. All we need now is to choose k so that the level of this test is α. We have that

$$P_\mu(Z_n \geq k) = P\left(Z \geq k + \sqrt{n}\frac{\mu_0 - \mu}{\sigma}\right), \text{ with } Z \sim N(0,1)$$
$$\Rightarrow \max_{\mu \leq \mu_0} P_\mu(Z_n \geq k) = P_{\mu_0}(Z_n \geq k) = P(Z \geq k).$$

So the smallest k for which $\max_{\mu \leq \mu_0} P_\mu(Z_n \geq k) \leq \alpha$ is obtained by setting $P(Z \geq k) = \alpha$, or $k = z_{1-\alpha}$, i.e. the $(1-\alpha)-$quantile of $N(0,1)$ (in R `qnorm(1-α)`). To conclude, the test

$$\varphi_n = I(Z_n \geq z_{1-\alpha}),$$

or equivalently $\varphi_n = I(\overline{X}_n \geq \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha})$, is of size α.
The power function of this test is given by

$$\pi_n(\mu) = P\left(Z \geq z_{1-\alpha} + \sqrt{n}\frac{\mu_0 - \mu}{\sigma}\right) = 1 - \Phi\left(z_{1-\alpha} + \sqrt{n}\frac{\mu_0 - \mu}{\sigma}\right).$$

Observe that $\pi_n(\mu) \xrightarrow[n\to\infty]{} 1$, $\forall \mu > \mu_0$, such a test is said to be ***consistent***. More precisely, *a test of level α is consistent*

if its power converges to 1 *as n tends to infinity*. This property is often considered necessary for a test to be useful in practice.

Here is the plot of $\pi_n(\mu)$, for $\mu_0 = 20$, $\sigma = 5$ and $\alpha = 5\%$, for different values of the sample size n.

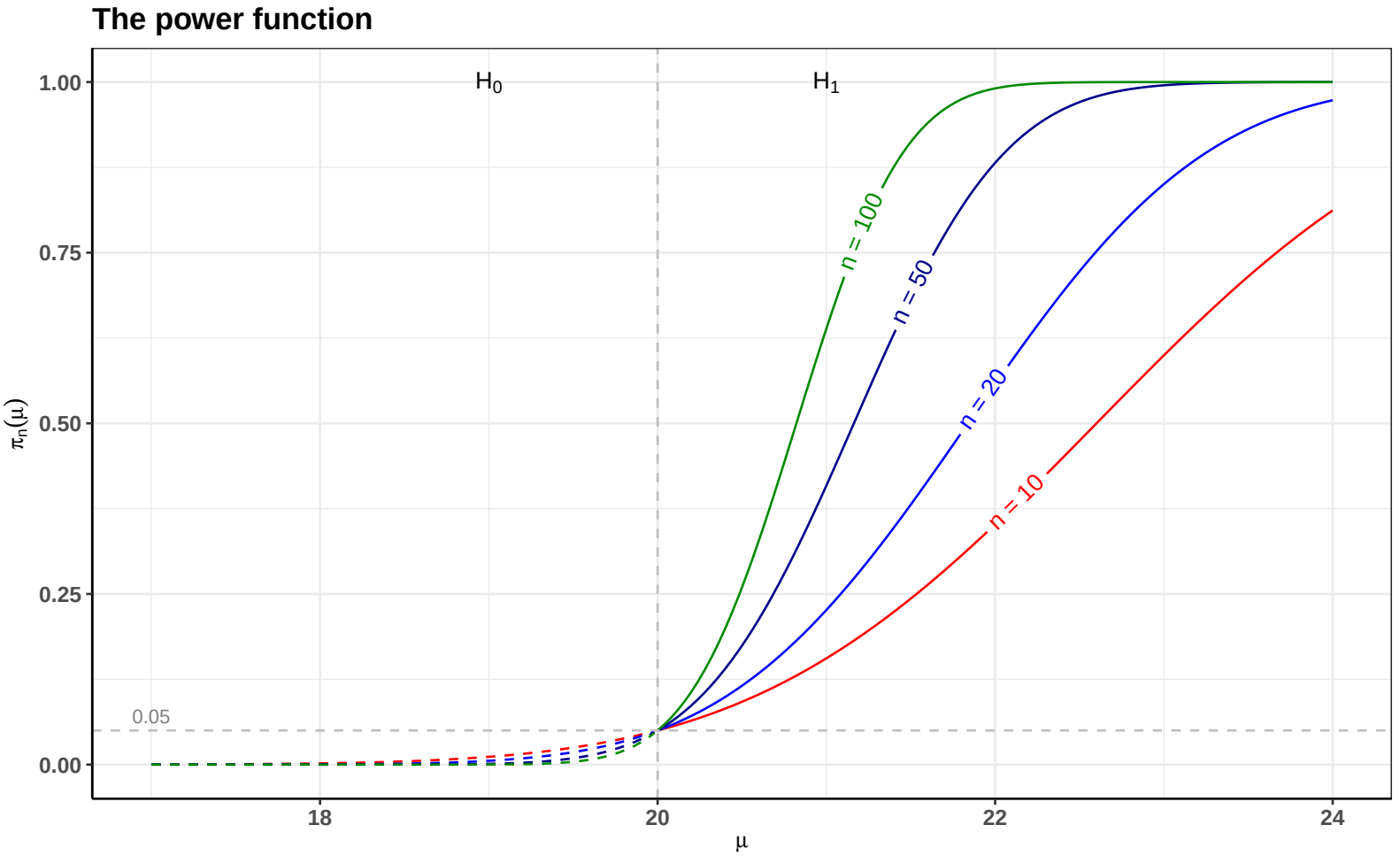


Figure 2: Power (solid line) and Type-I-error probability (dashed line)

□

4 The $p-$Value

The classic approach described above, based on critical values, can be summarized in two main steps:

1. specify the significance level α,
2. calculate the test statistic and compare it to the critical value, then reject/not reject H_0.

Simply reporting that a given hypothesis is rejected is not very informative. This says nothing about the fact that the computed value of the test statistic just barely fell into the rejection region or whether it exceeded the critical value by a large amount. What we need is a relevant way to ***measure the strength of evidence*** against a null hypothesis. And this is exactly what a $p-$value does.

To make things clearer, let's consider the test we studied in the Example 3.1 above. Recall that our decision is to reject H_0 whenever z_n, the observed value of the test statistic Z_n, is greater than $z_{1-\alpha}$. Since $z \mapsto P_{\mu_0}(Z_n \geq z)$ is a decreasing

function, we have that

$$\begin{aligned} z_n \geq z_{1-\alpha} &\Leftrightarrow P_{\mu_0}(Z_n \geq z_n) \leq P_{\mu_0}(Z_n \geq z_{1-\alpha}) \\ &\Leftrightarrow P(Z \geq z_n) \leq \alpha, \ Z \sim N(0,1). \end{aligned}$$

The quantity $P(Z \geq z_n)$, that we can write as $P(Z_n \geq z_n|H_0)$, is called the $p-$value. From the above calculation, we can reformulate our test decision as follows:

- $p-$value $\leq \alpha \ \Rightarrow$ reject H_0 at level α.
- $p-$value $> \alpha \ \Rightarrow$ do not reject H_0 at level α.

A more general definition of the $p-$value can be stated as follows. *Let* $\boldsymbol{T} \equiv T(\boldsymbol{X})$ *be a test statistic such that large values of* T *give evidence against* H_0. Given $t = T(\boldsymbol{x})$, the observed sample value of T, the $p-$value of such a test is

$$\max_{\theta \in \Theta_0} P_\theta(T \geq t).$$

In the particular case where $\Theta_0 = \{\theta_0\}$, the above $p-$value reduces to $P_{\theta_0}(T \geq t) \equiv P(T \geq t|H_0)$ which can be seen as *the probability of observing a test statistic at least as "extreme" (i.e. at the opposite of* H_0 *but along the direction of* H_1*) as the one actually observed, assuming the null hypothesis* H_0 *is true*. This is a simple and an intuitive definition of the $p-$value which is used in many textbooks. *The smaller the* $p-$*value, the more evidence there is in the observed data against the null*

hypothesis and in favor of the alternative hypothesis. Note that, unlike the critical value, the calculation of the $p-$value does not involve the level of the test α.

Example 4.1. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ is known. For some given μ_0, consider testing

$$H_0: \ \mu = \mu_0 \ \ vs \ \ H_1: \ \mu \neq \mu_0.$$

A natural test function is given by $\varphi_n = I(|Z_n| \geq k)$, where $Z_n = \sqrt{n}\dfrac{\overline{X}_n - \mu_0}{\sigma}$. The test statistic here is $|Z_n|$.
For this test to be of size α, we need to choose k so that $P_{\mu_0}(|Z_n| \geq k) = \alpha$. Thus, $k = z_{1-\alpha/2}$. So, based on this critical value, one should reject H_0, at level α, if $|z_n| \geq z_{1-\alpha/2}$.
An alternative and more convenient approach to do this test is to calculate its $p-$value which, according to the definition above, is given by

$$P_{\mu_0}(|Z_n| \geq |z_n|) = P(|Z| \geq |z_n|) = 2P(Z \geq |z_n|).$$

Based on this, one should reject H_0, at level α, if $2P(Z \geq |z_n|) \leq \alpha$. It is easy to see that

$$|z_n| \geq z_{1-\alpha/2} \Leftrightarrow 2P(Z \geq |z_n|) \leq \alpha. \square$$

Exercise 4.1. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ is known. Find a consistent test of size α for

$$H_0 : \mu \geq \mu_0 \ \ vs \ H_1 : \mu < \mu_0.$$

Calculate its $p-$value and its power function.

5 Likelihood Ratio Test (LRT)

The likelihood ratio test (LRT) is the most general and the most used test in practice. This method provides a unified approach for developing *optimal test procedures*. In fact, it have been shown that *whenever the uniformly most powerful test exists, the LRT procedure leads to it*. In addition to hypothesis testing, the LRT procedure can also be used to construct optimal confidence intervals.

Suppose we observe an iid sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a pd $f(x;\theta)$, for some $\theta \in \Theta$ (Θ can be a vector space). The ***likelihood ratio statistic*** for testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$, with $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$, is defined by

$$\Lambda_n(\boldsymbol{x}) := \frac{\max_{\theta\in\Theta_0} L_n(\theta|\boldsymbol{x})}{\max_{\theta\in\Theta} L_n(\theta|\boldsymbol{x})} = \frac{\max_{H_0} L_n(\theta|\boldsymbol{x})}{\max_{H_0\cup H_1} L_n(\theta|\boldsymbol{x})} = \frac{L_n(\hat{\theta}_0)}{L_n(\hat{\theta})},$$

where $L_n(\theta|\boldsymbol{x}) = \prod_i f(x_i;\theta)$ is the likelihood function of θ based on of the observed sample $\boldsymbol{x}$,
$\hat{\theta} \equiv \hat{\theta}(\boldsymbol{x}) = \arg\max_{\theta\in\Theta} L_n(\theta|\boldsymbol{x})$ is the MLE of θ obtained by maximizing the likelihood over the whole parameter space Θ, and $\hat{\theta}_0 \equiv \hat{\theta}_0(\boldsymbol{x}) = \arg\max_{\theta\in\Theta_0} L_n(\theta|\boldsymbol{x})$ be the ***restricted*** MLE of θ obtained by maximizing the likelihood over the null parameter space $\Theta_0 \subset \Theta$. Notice that, by definition, $0 \leq \Lambda_n \leq 1$.

To understand the reasoning behind the LRT, consider the discrete case where $f(x;\theta)$ is a probability mass function. In this case, given an observed data $\boldsymbol{x}$, Λ_n can be expressed as

$$\frac{P_{\hat{\theta}_0}(\boldsymbol{X}=\boldsymbol{x})}{P_{\hat{\theta}}(\boldsymbol{X}=\boldsymbol{x})} = \frac{\max_{\theta\in\Theta_0} P_\theta(\boldsymbol{X}=\boldsymbol{x})}{\max_{\theta\in\Theta_0\cup\Theta_1} P_\theta(\boldsymbol{X}=\boldsymbol{x})} = \frac{\max Prob(\text{obs. data}|H_0)}{\max Prob(\text{obs. data}|H_0 \cup H_1)}.$$

- A small value of this ratio means that the observed data are less likely to occur under the null hypothesis than in the case where no restrictions (as imposed by the null hypothesis) are applied. This means that there should be a parameter point in Θ_1 for which the observed sample is more likely than for any parameter point in Θ_0. Consequently, H_0 has to be rejected in favor of H_1.
- A value of this ratio close to 1 means that the observed data are nearly as likely to occur under the null hypothesis as without it. H_0 is therefore not binding and there is no reason to reject it.

This motivates the definition of LRT as $I(\Lambda_n \leq c)$, for some $c \in [0,1)$. In this way, the LRT rejects H_0 whenever $\Lambda_n \leq c$. c

must be chosen, in $[0,1)$*, so that the test level is* α*, i.e.* $\max_{\theta\in\Theta_0} P_\theta(\Lambda_n \leq c) \leq \alpha$*.* According to the definition given above, the $p-$value of the LRT is $\max_{\theta\in\Theta_0} P_\theta(\Lambda_n \leq \lambda_n)$, where λ_n is the observed value of Λ_n, i.e. $\lambda_n = \Lambda_n(\boldsymbol{x})$.

To perform LRT, we need to (i) maximize the likelihood function L_n over the full and restricted parameter spaces and (ii) calculate the critical value or the $p-$value as defined above. To do this, we need to know the "null distribution" of Λ_n (i.e. its distribution under H_0) or to express it as a monotonic function of some statistic T_n whose null distribution is known. The following examples illustrate this.

Example 5.1. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The likelihood function of (μ, σ^2) is given by

$$L_n(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right)$$

Assume that $\sigma^2 > 0$ *is known*.

- Let's find the LRT for testing $H_0: \mu = \mu_0 \;\; vs \;\; H_1: \mu \neq \mu_0$.

Since $\max_{\mu=\mu_0} L_n(\mu,\sigma^2) = L_n(\mu_0,\sigma^2)$ and $\max_{\mu} L_n(\mu,\sigma^2) = L_n(\overline{X}_n,\sigma^2)$,

$$\begin{aligned}\Lambda_n = \frac{L_n(\mu_0,\sigma^2)}{L_n(\overline{X}_n,\sigma^2)} &= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left((X_i-\mu_0)^2-(X_i-\overline{X}_n)^2\right)\right)\\ &= \exp\left(-\frac{n}{2\sigma^2}(\overline{X}_n-\mu_0)^2\right) = \exp\left(-\frac{1}{2}Z_n^2\right),\end{aligned}$$

where $Z_n = \sqrt{n}\dfrac{\overline{X}_n-\mu_0}{\sigma}$.

The LRT rejects H_0 if $\Lambda_n \le c$ (with $c \in [0,1)$) $\Leftrightarrow |Z_n| \ge k$, where $k = \sqrt{-2\log(c)} > 0$. This latter is determined so that the size of the test is α, that is $P_{\mu_0}(|Z_n| \ge k) = \alpha$. This is equivalent to say that $P(|Z| \ge k) = \alpha$, thus $k = z_{1-\alpha/2}$. We conclude that the LRT of level α for the above hypothesis is equivalent to $I(|Z_n| \ge z_{1-\alpha/2})$.

The $p-$value of this test is given by

$$\max_{\mu:\mu=\mu_0} P(\Lambda_n \le \lambda_n) = P_{\mu=\mu_0}(\Lambda_n \le \lambda_n) = P_{\mu=\mu_0}\left(|Z_n| \ge \sqrt{2\log(\lambda_n)}\right) = P(|N(0,1)| \ge |z_n|),$$

where z_n (λ_n) is the observed value of Z_n (Λ_n).

- Let's find the LRT for testing $H_0: \mu \le \mu_0$ *vs* $H_1: \mu > \mu_0$.

Since $\mu \mapsto \ell_n(\mu, \sigma^2) = \log L_n(\mu, \sigma^2)$ is strictly concave and has a (unique) maximum at $\overline{X}_n$, we have that

$$\arg\max_{\mu \le \mu_0} L_n(\mu, \sigma^2) = \arg\max_{\mu \le \mu_0} \ell_n(\mu, \sigma^2) = \begin{cases} \overline{X}_n & \text{if } \overline{X}_n \le \mu_0 \\ \mu_0 & \text{if } \overline{X}_n > \mu_0 \end{cases}$$

So,

$$\Lambda_n = \begin{cases} 1 & \text{if } \overline{X}_n \le \mu_0 \\ \frac{L_n(\mu_0, \sigma^2)}{L_n(\overline{X}_n, \sigma^2)} = \exp\left(-\frac{1}{2} Z_n^2\right) & \text{if } \overline{X}_n > \mu_0 \end{cases}$$

The LRT rejects H_0 if $\Lambda_n \le c$ (with $c \in [0,1)$) $\Leftrightarrow |Z_n| \ge k$ **and** $\overline{X}_n > \mu_0 \Leftrightarrow Z_n \ge k$ (with $k > 0$). This latter is determined so that the size of the test is α, that is $\max_{\mu \le \mu_0} P_\mu(Z_n \ge k) = \alpha$. This is equivalent to say that $P(Z \ge k) = \alpha$, thus $k = z_{1-\alpha}$. We conclude that the LRT of level α for the above hypothesis is equivalent to $I(Z_n \ge z_{1-\alpha})$. The $p-$value of this test is

$$\begin{aligned} \max_{\mu : \mu \le \mu_0} P(\Lambda_n \le \lambda_n) &= I(\overline{X} \le \mu_0) + I(\overline{X} > \mu_0) \max_{\mu \le \mu_0} P\left(Z_n \ge \sqrt{2 \log(\lambda_n)}\right) \\ &= I(\overline{X} \le \mu_0) + I(\overline{X} > \mu_0) P(N(0,1) \ge z_n). \end{aligned}$$

Assume now that $\sigma^2 > 0$ *is unknown*.

- Let's find the LRT for testing $H_0: \mu = \mu_0 \ \ vs \ \ H_1: \mu \neq \mu_0$.
  We know that
  $$\max_{\mu=\mu_0,\sigma^2} L_n(\mu,\sigma^2) = L_n(\mu_0,\tilde{\sigma}_0^2) \text{ and } \max_{\mu,\sigma^2} L_n(\mu,\sigma^2) = L_n(\overline{X}_n,\hat{\sigma}_n^2),$$
  where $\tilde{\sigma}_0^2 = n^{-1}\sum_{i=1}^n (X_i-\mu_0)^2$ and $\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (X_i-\overline{X}_n)^2$. So,
  $$\begin{aligned}
  \Lambda_n &= \frac{L_n(\mu_0,\tilde{\sigma}_0^2)}{L_n(\overline{X}_n,\hat{\sigma}_n^2)} = \left(\frac{\tilde{\sigma}_0^2}{\hat{\sigma}_n^2}\right)^{-n/2}, \text{ with} \\
  \frac{\tilde{\sigma}_0^2}{\hat{\sigma}_n^2} &= \frac{\sum_i (X_i-\mu_0)^2}{\sum_i (X_i-\overline{X}_n)^2} = \frac{\sum_i (X_i-\overline{X}_n)^2 + n(\overline{X}_n-\mu_0)^2}{\sum_i (X_i-\overline{X}_n)^2} = 1 + \frac{n(\overline{X}_n-\mu_0)^2}{\sum_i (X_i-\overline{X}_n)^2} \\
  &= 1 + \frac{n(\overline{X}_n-\mu_0)^2}{(n-1)S_n^2} = 1 + \frac{T_n^2}{n-1},
  \end{aligned}$$
  where $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X}_n)^2$ and $T_n = \sqrt{n}\frac{\overline{X}_n-\mu_0}{S_n}$. We see that $\Lambda_n \leq c \iff |T_n| \geq k$. And since under H_0, $T_n \sim t_{n-1}$, where t_{n-1} is the Student's t-distribution with $n-1$ degrees of freedom, the LRT of size α for the above hypothesis is equivalent to $I\left(|T_n| \geq t_{n-1;1-\frac{\alpha}{2}}\right)$.
- Let's find the LRT for testing $H_0: \mu \leq \mu_0 \ \ vs \ \ H_1: \mu > \mu_0$.

Under H_0, the MLE of (μ, σ^2) is

$$\begin{cases} (\overline{X}_n, \hat{\sigma}_n^2) & \text{if } \overline{X}_n \leq \mu_0 \\ (\mu_0, \tilde{\sigma}_0^2) & \text{if } \overline{X}_n > \mu_0 \end{cases}$$

So,

$$\Lambda_n = \begin{cases} 1 & \text{if } \overline{X}_n \leq \mu_0 \\ \frac{L_n(\mu_0, \tilde{\sigma}_0^2)}{L_n(\overline{X}_n, \hat{\sigma}_n^2)} = \left(1 + \frac{T_n^2}{n-1}\right)^{-n/2} & \text{if } \overline{X}_n > \mu_0 \end{cases}$$

The LRT rejects H_0 if $\Lambda_n \leq c$ (with $c \in [0,1)$) $\Leftrightarrow |T_n| \geq k$ **and** $\overline{X}_n > \mu_0 \Leftrightarrow T_n \geq k$. By the same reasoning as before (see the case of known σ), we conclude that the LRT of level α is equivalent to $I\left(T_n \geq t_{n-1;1-\alpha}\right)$. □

Exercise 5.1. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Find the LRT of level α for testing $H_0 : \mu \geq \mu_0$ *vs* $H_1 : \mu < \mu_0$, assuming first that σ is known and then without this assumption.

6 Asymptotic Tests

LRT construction, as discussed in the previous section, is a complicated process that requires a case-by-case analysis. The main difficulty is to find the ***exact*** distribution, under H_0, of Λ_n (or some function of it), so that the appropriate

critical value can be determined.

A way to overcome this difficulty is to rely on asymptotic theory, which, as we will see, allows the development of unified test procedures that can be applied to a wide range of problems. Unlike exact tests, asymptotic tests (those based on asymptotic theory) can, by definition, only be applied in situations where ***the sample size is large enough***. We say that a test with power function $\pi_n(\theta)$ is of ***asymptotic level*** α if

$$\lim_{n\to\infty} \pi_n(\theta) \leq \alpha, \ \forall \theta \in \Theta_0.$$

There are three classical asymptotic test methods based on the likelihood function: the ***Wald*** tests, the ***Score*** (or ***Rao***) tests, and the (log-)***likelihood ratio*** tests. Under appropriate regularity conditions, these tests are asymptotically equivalent with the best possible power (asymptotically).

Let $\hat{\boldsymbol{\theta}}$ denote the MLE of $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$. In what follows, the *regularity assumptions* required to ensure that $\sqrt{\boldsymbol{I}_n(\hat{\boldsymbol{\theta}})}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_d(\mathbf{0}, \mathbb{1})$ are supposed to be fulfilled; see previous sections.

6.1 Simple null hypothesis

To aid intuition, we first consider the simplest case of one parameter ($d = 1$). Let

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

The three test statistics of interest are:

$$W_n \equiv W_n(\theta_0) := I_n(\hat{\theta})(\hat{\theta} - \theta_0)^2 \qquad \text{(Wald)}$$

$$R_n \equiv R_n(\theta_0) := \frac{S_n^2(\theta_0)}{I_n(\theta_0)} \qquad \text{(Score)}$$

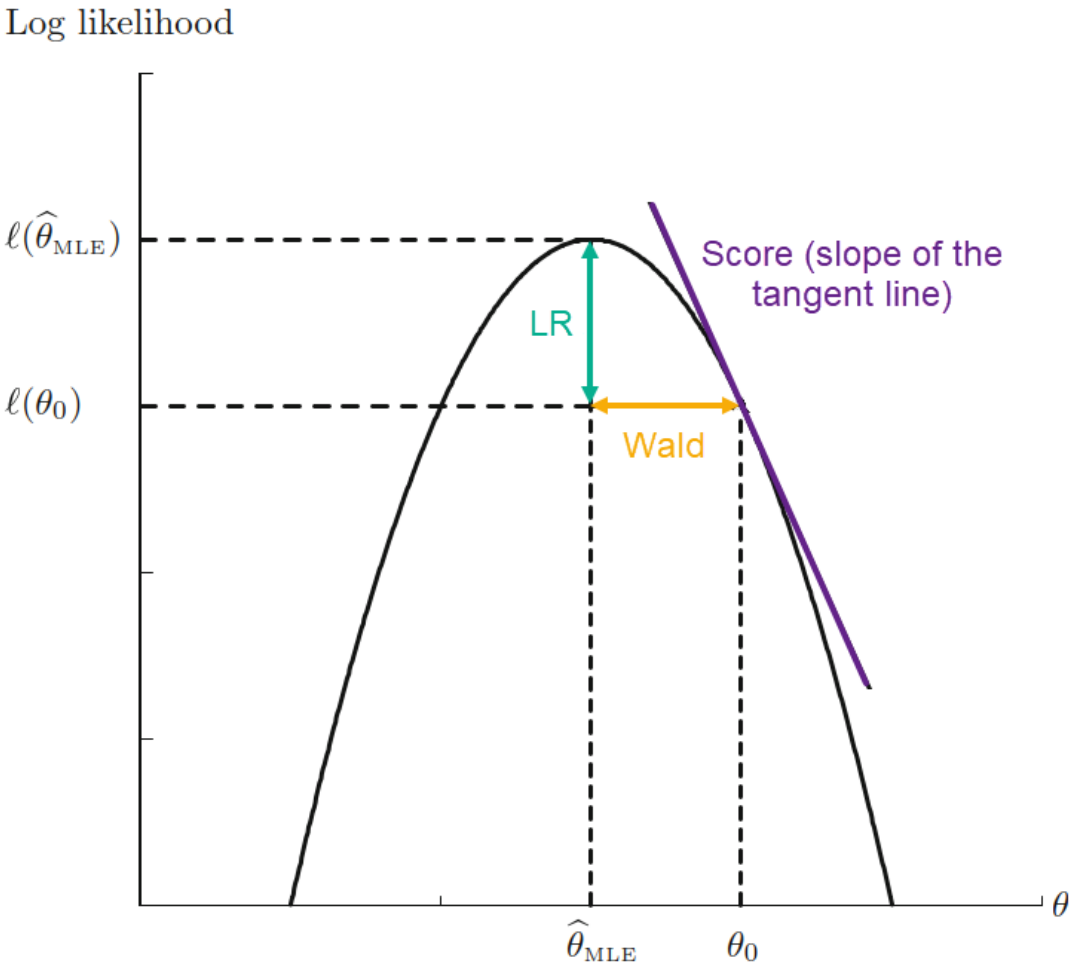
$$T_n \equiv T_n(\theta_0) := 2\left(\ell_n(\hat{\theta}) - \ell_n(\theta_0)\right) \qquad \text{(Likelihood ratio)}$$

In the above formulas $\ell_n(\theta) = \sum_i \log f(X_i, \theta)$ is the log-likelihood and S_n is the corresponding Score function, i.e. $S_n(\theta) = \ell_n'(\theta|\boldsymbol{X}) = \sum_i \partial_\theta \log f(X_i, \theta)$.

Note that $T_n = -2\log \dfrac{L_n(\theta_0)}{L_n(\hat{\theta})} = -2\log \Lambda_n \in [0, \infty)$. This is the so-called *log-likelihood ratio statistic*, but for simplicity's sake we'll just call it the likelihood ratio (LR) statistic.

Although the above formulas defining the three statistics are very different, they are based on the same principle of

assessing the distance between the MLE $\hat{\theta}$ and the null value θ_0: the greater the difference between these two, the stronger the evidence against the null hypothesis H_0. Wald uses $(\hat{\theta} - \theta_0)^2$, which we can describe as a "horizontal" difference (see the figure below). LRT uses $\big(\ell_n(\hat{\theta}) - \ell_n(\theta_0)\big)$, i.e. the "vertical" or likelihood difference. Score uses $\big(S_n(\hat{\theta}) - S_n(\theta_0)\big)^2 = S_n^2(\theta_0)$, i.e. the "slope" (of the likelihood) difference. In any case, once the "distance" has been chosen, H_0 must be rejected if it is found to be "too far" away from the data (i.e. the test statistic is "too large").


Log likelihood
$\ell(\widehat{\theta}_{\text{MLE}})$
$\ell(\theta_0)$
LR
Wald
Score (slope of the tangent line)
θ
$\widehat{\theta}_{\text{MLE}}$
θ_0

In the Wald statistic, $(\hat{\theta} - \theta_0)^2$ is multiplied by the Fisher information I_n to obtain a "standardized distance" with a fixed/known distribution. In fact, we have seen that $\sqrt{I_n(\hat{\theta})}(\hat{\theta} - \theta) \xrightarrow{d} N(0,1)$, so, by Slutsky's theorem, $W_n(\theta) \xrightarrow{d} \chi^2_1$ (*here θ refers to the true parameter value*). Thus, under H_0, $W_n \xrightarrow{d} \chi^2_1$.
Note that the Wald statistic can be expressed as $W_n = Z_n^2$, with

$$Z_n \equiv Z_n(\theta_0) := \frac{\hat{\theta} - \theta_0}{\sqrt{\widehat{Avar}(\hat{\theta})}},$$

with $\widehat{Avar}(\hat{\theta}) = I_n^{-1}(\hat{\theta})$. $I_n(\hat{\theta})$ can be replaced by the observed Fisher information $J_n(\hat{\theta}) = -\ell_n''(\hat{\theta})$, or by any asymptotically equivalent quantity (like $I_n(\theta_0)$ or $J_n(\theta_0)$), without altering the asymptotic distribution of W_n.

As for the LRT, observe that, by second-order Taylor's expansion of $\theta \mapsto \ell_n(\theta)$ around $\hat{\theta}$,

$$\begin{aligned} \ell_n(\theta) - \ell_n(\hat{\theta}) &\approx \frac{1}{2}(\theta - \hat{\theta})^2 \ell_n''(\hat{\theta}) \\ \Rightarrow 2\big(\ell_n(\hat{\theta}) - \ell_n(\theta)\big) &\approx \frac{-n^{-1}\ell_n''(\hat{\theta})}{I(\theta)} \left(\sqrt{nI(\theta)}(\hat{\theta} - \theta)\right)^2 \end{aligned}$$

Since, $\sqrt{I_n(\theta)}(\hat{\theta} - \theta) \xrightarrow{d} N(0,1)$ and $-n^{-1}\ell_n''(\hat{\theta}) \xrightarrow{p} I(\theta)$, we conclude that $T_n(\theta) \xrightarrow{d} \chi^2_1$. Thus, under H_0, $T_n \xrightarrow{d} \chi^2_1$.

As for the Score test, using the first-order Taylor expansion of $\theta \mapsto S_n(\theta)$ around θ_0, and following the same line of reasoning as above shows that, under H_0, $S_n \xrightarrow{d} \chi^2_1$. Similarly to the Wald statistic, in the Score statistic, one could replace $I_n(\theta_0)$ by any consistent estimator of it like for example $J_n(\theta_0)$. Note also that the Score statistic can be expressed as $\left(\frac{S_n(\theta_0)}{\sqrt{Var(S_n(\theta_0))}}\right)^2$.

To conclude, we can say that, *under H_0, the three statistics (Wald, Score, LR) share the same asymptotic distribution, namely the χ^2_1*. Based on this result, the Wald test, the asymptotic LRT and the Score test reject H_0, in favor of H_1, if the observed value of their statistics (i.e. W_n for Wald, R_n for the Score and T_n for LR) is $\geq \chi^2_{1;1-\alpha}$, where $\chi^2_{1;1-\alpha}$ is the $(1-\alpha)-$quantile of χ^2_1. These tests have an ***asymptotic size*** equal to α. We can check this easily for the LRT, for example, by observing that its type I error probability is $P_{\theta_0}(T_n \geq \chi^2_{1;1-\alpha}) \to P(\chi^2_1 \geq \chi^2_{1;1-\alpha}) = \alpha$.

In practice, it is common to use the ***asymptotic $p-$value*** $P(\chi^2_1 \geq t_n)$ (t_n is the observed value of T_n) and to reject H_0 whenever this quantity is less than α. The exact same considerations applies to the other two statistics (Wald and Score).

Wald, Score, and LR tests are asymptotically equivalent: they reach the same decision with probability approaching 1 as $n \to \infty$. However, their performance can be *quite different for a finite sample size*. Each method has its strengths and limitations:

- Based on the statistic $Z_n = \sqrt{I_n(\hat{\theta})}(\hat{\theta} - \theta_0)$, it is straightforward to create one-sided Wald tests (e.g. tests of $H_0 : \theta = \theta_0$ vs $H_1 : \theta < \theta_0$ or $H_1 : \theta > \theta_0$), but this is more difficult with Score and likelihood ratio statistics.
- The Wald statistic is by far the simplest and the easiest to interpret. It yields immediate confidence intervals and it is largely available in standard computing packages.
- The Wald test is not limited to MLE estimation, one just need to know the asymptotic distribution of the estimator under studied.
- The Score test does not require the MLE $\hat{\theta}$ whereas the other two tests do. The Score test also tends to give the best Type I error rates for small sample sizes.
- The Score test and likelihood ratio test are invariant under reparameterization, whereas the Wald test is not. For example, the Wald test about a scale parameter σ depends on whether the null hypothesis is expressed as $H_0 : \sigma = \sigma_0$ or $H_0 : \sigma^2 = \sigma_0^2$.
- In general, the likelihood ratio is more difficult to compute than Wald or Score tests, but it tends to have the greatest power.

Example 6.1. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Consider $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$. Recall the log-likelihood function

$$\ell_n(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

Assume that $\sigma^2 > 0$ *is known*. Also recall that

$$S_n(\mu) = \partial_\mu \ell_n(\mu, \sigma^2) = \frac{n}{\sigma^2}(\overline{X}_n - \mu),$$
$$I_n(\mu) = -E\left(S_n'(\mu)\right) = \frac{n}{\sigma^2},$$

and $\overline{X}_n$ is the MLE of μ. It follows that

$$W_n = \frac{n}{\sigma^2}(\overline{X}_n - \mu_0)^2 = \left(\sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma}\right)^2,$$
$$R_n = \frac{\left(\frac{n}{\sigma^2}(\overline{X}_n - \mu_0)\right)^2}{\frac{n}{\sigma^2}} = \frac{n}{\sigma^2}(\overline{X}_n - \mu_0)^2,$$
$$T_n = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left((X_i - \mu_0)^2 - (X_i - \overline{X}_n)^2\right) = \frac{n}{\sigma^2}(\overline{X}_n - \mu_0)^2.$$

The three test statistics are identical in this model. □

Example 6.2. Let $X_1, \ldots, X_n \overset{iid}{\sim} Ber(\pi)$. Consider $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$. Recall that

$$\begin{aligned}
\ell_n(\pi) &= n\hat{\pi}\log(\pi) + n(1-\hat{\pi})\log(1-\pi), \\
S_n(\pi) &= n\frac{\hat{\pi}-\pi}{\pi(1-\pi)}, \\
I_n(\pi) &= \frac{n}{\pi(1-\pi)},
\end{aligned}$$

where $\hat{\pi} = n^{-1}\sum_i X_i$ is the MLE of π. It follows that

$$\begin{aligned}
W_n &= n\frac{(\hat{\pi}-\pi_0)^2}{\hat{\pi}(1-\hat{\pi})} = \left(\sqrt{n}\frac{\hat{\pi}-\pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})}}\right)^2, \\
R_n &= \frac{\left(n\frac{\hat{\pi}-\pi_0}{\pi_0(1-\pi_0)}\right)^2}{\frac{n}{\pi_0(1-\pi_0)}} = n\frac{(\hat{\pi}-\pi_0)^2}{\pi_0(1-\pi_0)}, \\
T_n &= 2\left\{n\hat{\pi}\log\frac{\hat{\pi}}{\pi_0} + n(1-\hat{\pi})\log\frac{1-\hat{\pi}}{1-\pi_0}\right\}.\square
\end{aligned}$$

Let's check the Type I error and the power of these tests using some simulations.

```r
p.value <- function(pi0, pi, n) {
  x <- rbinom(n, size = 1, prob = pi)
  hat.pi <- mean(x)
  Wald <- n * ((hat.pi - pi0)^2 / (hat.pi * (1 - hat.pi)))
  Score <- n * ((hat.pi - pi0)^2 / (pi0 * (1 - pi0)))
  LR <- 2 * (n * hat.pi * ifelse(hat.pi == 0, 0, log(hat.pi / pi0)) +
               n * (1 - hat.pi) * ifelse(hat.pi == 1, 0, log((1 - hat.pi) / (1 - pi0))))
  p.valueW <- pchisq(Wald, df = 1, lower.tail = FALSE)
  p.valueS <- pchisq(Score, df = 1, lower.tail = FALSE)
  p.valueL <- pchisq(LR, df = 1, lower.tail = FALSE)
  return(c(wald = p.valueW, Score = p.valueS, LR = p.valueL))
}
# Type I error (pi.0 = true.pi = 0.5); n = 10, 100, 1000
rbind((replicate(5000, p.value(0.5, 0.5, 10)) <= 0.05) |> rowMeans(),
(replicate(5000, p.value(0.5, 0.5, 100)) <= 0.05) |> rowMeans(),
(replicate(5000, p.value(0.5, 0.5, 1000)) <= 0.05) |> rowMeans())
```

```
      wald  Score    LR
```

```
[1,] 0.115 0.0188 0.115
[2,] 0.059 0.0590 0.059
[3,] 0.051 0.0510 0.051
```

```r
# Power (pi.0 = 0.4 and true.pi = 0.5); n = 10, 100, 1000
rbind((replicate(5000, p.value(0.4, 0.5, 10)) <= 0.05) |> rowMeans(),
(replicate(5000, p.value(0.4, 0.5, 100)) <= 0.05) |> rowMeans(),
(replicate(5000, p.value(0.4, 0.5, 1000)) <= 0.05) |> rowMeans())
```

```
      wald  Score     LR
[1,] 0.183 0.0546 0.0678
[2,] 0.546 0.5460 0.5460
[3,] 1.000 1.0000 1.0000
```

The generalization of the three tests to the ***multiparameter case*** is not very difficult. Suppose we wish to test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, where $\boldsymbol{\theta} \in \mathbb{R}^d$. Then

$$W_n \equiv W_n(\boldsymbol{\theta}_0) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^t \boldsymbol{I}_n(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \quad \text{(Wald)}$$

$$R_n \equiv R_n(\boldsymbol{\theta}_0) = \boldsymbol{S}_n^t(\boldsymbol{\theta}_0)\boldsymbol{I}_n^{-1}(\boldsymbol{\theta}_0)\boldsymbol{S}_n(\boldsymbol{\theta}_0) \quad \text{(Score)}$$

$$T_n \equiv T_n(\boldsymbol{\theta}_0) = 2\left(\ell_n(\hat{\boldsymbol{\theta}}) - \ell_n(\boldsymbol{\theta}_0)\right) \quad \text{(Likelihood ratio)}$$

Under H_0, these three statistics converge to χ^2_d, i.e. chi-squared distribution with d degrees of freedom. As consequence, the likelihood ratio test, for example, rejects H_0 when $T_n \geq \chi^2_{d;1-\alpha}$, or, equivalently, when $P(\chi^2_d \geq t_n) \leq \alpha$. The same applies to the other two statistics.

Example 6.3. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where μ and σ are unknown. Consider testing

$$H_0: \ \mu = \mu_0 \text{ and } \sigma^2 = \sigma_0^2 \ \ vs \ \ H_1: \ \mu \neq \mu_0 \text{ or } \sigma^2 \neq \sigma_0^2.$$

We know that

$$\ell_n(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2,$$

$$\boldsymbol{S}_n(\boldsymbol{\theta}) = \left(n\frac{\overline{X}_n - \mu}{\sigma^2},\ -\frac{n}{2\sigma^2} + \frac{\sum_i (X_i - \mu)^2}{2\sigma^4}\right)^t,$$

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix},$$

and that $\left(\overline{X}_n, \hat{\sigma}_n^2\right)$, with $\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$, is the MLE of (μ, σ^2). Put $\tilde{\sigma}_0^2 = n^{-1}\sum_{i=1}^{n}(X_i - \mu_0)^2$.
The Wald statistic is given by

$$\left(\overline{X}_n - \mu_0 \quad \hat{\sigma}_n^2 - \sigma_0^2\right) \begin{pmatrix} \frac{n}{\hat{\sigma}_n^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}_n^4} \end{pmatrix} \begin{pmatrix} \overline{X}_n - \mu_0 \\ \hat{\sigma}_n^2 - \sigma_0^2 \end{pmatrix} = n\frac{(\overline{X}_n - \mu_0)^2}{\hat{\sigma}_n^2} + \frac{n}{2}\left(1 - \frac{\sigma_0^2}{\hat{\sigma}_n^2}\right)^2 \xrightarrow{d} \chi_2^2, \text{ if } H_0 \text{ is true.}$$

The Score statistic is given by

$$\begin{aligned}
&\begin{pmatrix} n\frac{\overline{X}_n-\mu_0}{\sigma_0^2} & -\frac{n}{2\sigma_0^2}+\frac{\sum_i(X_i-\mu_0)^2}{2\sigma_0^4} \end{pmatrix} \begin{pmatrix} \frac{\sigma_0^2}{n} & 0 \\ 0 & \frac{2\sigma_0^4}{n} \end{pmatrix} \begin{pmatrix} n\frac{\overline{X}_n-\mu_0}{\sigma_0^2} \\ -\frac{n}{2\sigma_0^2}+\frac{\sum_i(X_i-\mu_0)^2}{2\sigma_0^4} \end{pmatrix} \\
&= n\frac{(\overline{X}_n-\mu_0)^2}{\sigma_0^2}+\frac{n}{2}\left(1-\frac{\tilde{\sigma}_0^2}{\sigma_0^2}\right)^2 \xrightarrow{d} \chi_2^2, \text{ if } H_0 \text{ is true.}
\end{aligned}$$

The LRT statistic is given by

$$\begin{aligned}
T_n &= 2(\ell_n(\bar{X}_n,\hat{\sigma}_n^2)-\ell_n(\mu_0,\sigma_0^2)) \\
&= n\frac{(\bar{X}_n-\mu_0)^2}{\sigma_0^2}-n\log\left(\frac{\hat{\sigma}_n^2}{\sigma_0^2}\right)+n\left(\frac{\hat{\sigma}_n^2}{\sigma_0^2}-1\right) \xrightarrow{d} \chi_2^2, \text{ if } H_0 \text{ is true.}\square
\end{aligned}$$

6.2 Composite null hypothesis

In practice, we are rarely interested in a simple null hypothesis in which all the d components of the vector parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ are assigned, as done previously. Often, we're only interested in a subset of $\boldsymbol{\theta}$, or in some specific constraints

on its components. This can be formulated, in most cases, as follows

$$H_0 : \boldsymbol{A\theta} = \boldsymbol{a} \text{ vs } H_1 : \boldsymbol{A\theta} \neq \boldsymbol{a},$$

where $\boldsymbol{A}$ is an $r \times d$ *full rank matrix* ($r \leq d$ and the r rows of $\boldsymbol{A}$ are linearly independent) and $\boldsymbol{a}$ is an r-vector.
For example, suppose that $d = 5$, then

- $\theta_1 = \ldots = \theta_5 = 0 \Leftrightarrow \boldsymbol{A\theta} = \boldsymbol{0}$, with $\boldsymbol{A} = \mathbb{1}$ (5×5 identity matrix).
- $\theta_1 = 1, \theta_2 = 2, \ldots, \theta_5 = 5 \Leftrightarrow \boldsymbol{A\theta} = \boldsymbol{a}$, with $\boldsymbol{A} = \mathbb{1}$ and $\boldsymbol{a} = (1, 2, \ldots, 5)^t$.
- $\theta_2 = \theta_4 = 0 \Leftrightarrow \boldsymbol{A\theta} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, with $\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$
- $\theta_2 + \theta_4 = 1 \Leftrightarrow \boldsymbol{A\theta} = 1$, with $\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \end{pmatrix}$
- $\theta_1 = 2\theta_2$ and $\theta_3 = \theta_4 \Leftrightarrow \boldsymbol{A\theta} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, with $\boldsymbol{A} = \begin{pmatrix} 1 & -2 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{pmatrix}$

The generalized Wald test statistic, the generalized Score test statistic and the generalized LRT statistic are given,

respectively, by

$$
\begin{aligned}
W_n &= (\boldsymbol{A}\hat{\boldsymbol{\theta}} - \boldsymbol{a})^t (\boldsymbol{A}\boldsymbol{I}_n^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{A}^t)^{-1} (\boldsymbol{A}\hat{\boldsymbol{\theta}} - \boldsymbol{a}), \\
R_n &= \boldsymbol{S}_n^t(\tilde{\boldsymbol{\theta}}_{\boldsymbol{a}})\boldsymbol{I}_n^{-1}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{a}})\boldsymbol{S}_n(\tilde{\boldsymbol{\theta}}_{\boldsymbol{a}}), \\
T_n &= 2\left(\ell_n(\hat{\boldsymbol{\theta}}) - \ell_n(\tilde{\boldsymbol{\theta}}_{\boldsymbol{a}})\right),
\end{aligned}
$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, and $\tilde{\boldsymbol{\theta}}_{\boldsymbol{a}}$ is the *restricted MLE of* $\boldsymbol{\theta}$ *obtained by maximizing the (log-)likelihood under* H_0*, i.e. under the constraint that* $\boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{a}$.

Similar to the simple null hypothesis case, ***these three statistics converge, Under*** H_0***, to*** χ^2_r, i.e. chi-squared distribution with r degrees of freedom. This latter corresponds to the number of restrictions imposed by H_0.

Before moving on and looking at some examples, let's focus on the LRT and consider the case of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$ vs $H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{10}$, for some given $\boldsymbol{\theta}_{10}$.

In the remainder of this section, in order to simplify calculations and make it easier to follow-up, we'll concentrate solely on the Wald and the LR tests.

Example 6.4. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where μ and σ are unknown. Consider testing

$$H_0: \ \mu = \mu_0 \ \ vs \ \ H_1: \ \mu \neq \mu_0.$$

The Wald statistic is given by

$$\begin{aligned} W_n &= \left(\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \overline{X}_n \\ \hat{\sigma}_n^2 \end{pmatrix} - \mu_0 \right)^t \left(\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \frac{\hat{\sigma}_n^2}{n} & 0 \\ 0 & \frac{2\hat{\sigma}_n^4}{n} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right)^{-1} \left(\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \overline{X}_n \\ \hat{\sigma}_n^2 \end{pmatrix} - \mu_0 \right) \\ &= n \frac{(\overline{X}_n - \mu_0)^2}{\hat{\sigma}_n^2} \xrightarrow{d} \chi_1^2, \text{ if } H_0 \text{ is true.} \end{aligned}$$

Since the MLE of (μ, σ^2) under H_0 is $(\mu_0, \tilde{\sigma}_0^2)$, where $\tilde{\sigma}_0^2 = n^{-1} \sum_i (X_i - \mu_0)^2$, the LRT statistic is given by

$$\begin{aligned} T_n &= 2(\ell_n(\bar{X}_n, \hat{\sigma}_n^2) - \ell_n(\mu_0, \tilde{\sigma}_0^2)) \\ &= n \log \left(\frac{\tilde{\sigma}_0^2}{\hat{\sigma}_n^2} \right) = n \log \left(1 + \frac{(\overline{X}_n - \mu_0)^2}{\hat{\sigma}_n^2} \right) \xrightarrow{d} \chi_1^2, \text{ if } H_0 \text{ is true.} \square \end{aligned}$$

Example 6.5. Let $X_1, \ldots, X_n \overset{iid}{\sim} Ber(\pi_1)$ and $Y_1, \ldots, Y_n \overset{iid}{\sim} Ber(\pi_2)$ be two independent samples. Consider testing

$$H_0: \ \pi_1 = \pi_2 \ \ vs \ \ H_1: \ \pi_1 \neq \pi_2.$$

Based on the sample (X_i, Y_i), $i = 1, \ldots, n$, it's easy to see that the Log-likelihood for (π_1, π_2) is

$$\ell_n(\pi_1, \pi_2) = n\hat{\pi}_1 \log(\pi_1) + n(1 - \hat{\pi}_1) \log(1 - \pi_1) + n\hat{\pi}_2 \log(\pi_2) + n(1 - \hat{\pi}_2) \log(1 - \pi_2),$$

where $(\hat{\pi}_1, \hat{\pi}_2) := \left(n^{-1} \sum_i X_i, n^{-1} \sum_i Y_i\right)$ is the MLE of (π_1, π_2). The FI matrix is

$$\boldsymbol{I}_n(\pi_1, \pi_2) = \begin{pmatrix} \frac{n}{\pi_1(1-\pi_1)} & 0 \\ 0 & \frac{n}{\pi_2(1-\pi_2)} \end{pmatrix}.$$

The Wald statistic is given by

$$\begin{aligned} W_n &= \left(\begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix} - 0 \right)^t \left(\begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n} & 0 \\ 0 & \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right)^{-1} \left(\begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix} - 0 \right) \\ &= n \frac{(\hat{\pi}_1 - \hat{\pi}_2)^2}{\hat{\pi}_1(1 - \hat{\pi}_1) + \hat{\pi}_2(1 - \hat{\pi}_2)} \xrightarrow{d} \chi^2_1, \text{ if } H_0 \text{ is true.} \end{aligned}$$

Under H_0, the log-likelihood reduces to

$$\ell_n(\pi_1, \pi_1) = n(\hat{\pi}_1 + \hat{\pi}_2)\log(\pi_1) + n(2 - \hat{\pi}_1 - \hat{\pi}_2)\log(1 - \pi_1).$$

It's easy to see that, in this case, the MLE of π_1 $(= \pi_2)$ is $\tilde{\pi}_0 = \frac{\sum_{i=1}^{n}(X_i + Y_i)}{2n}$. As a result, the LRT statistic is given by

$$\begin{aligned} T_n &= 2\left(\ell_n(\hat{\pi}_1, \hat{\pi}_2) - \ell_n(\tilde{\pi}_0, \tilde{\pi}_0)\right) \\ &= 2n\sum_{j=1}^{2}\left(\hat{\pi}_j \log\left(\frac{\hat{\pi}_j}{\tilde{\pi}_0}\right) + (1 - \hat{\pi}_j)\log\left(\frac{1 - \hat{\pi}_j}{1 - \tilde{\pi}_0}\right)\right) \xrightarrow{d} \chi^2_1, \text{ if } H_0 \text{ is true.}\square \end{aligned}$$

We now turn to a special, but very useful, case of the general formulation of the LR statistic presented above. Consider the case of a parametric model indexed by (θ, η), where θ and η are unknown. Let's say that θ is our parameter of interest and η is a nuisance parameter (θ and η can be vectors). In this context, our objective is to test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, for some given θ_0. Let L denote the likelihood function of the model under study, $(\hat{\theta}, \hat{\eta}) = \arg\max_{\theta,\eta} L(\theta, \eta)$ be the MLE of (θ, η) and $\hat{\eta}_0 \equiv \hat{\eta}(\theta_0) := \arg\max_{\eta} L(\theta_0, \eta)$ be the restricted MLE of η under H_0. According to the general definition given above, the LRT statistic is

$$T_n \equiv T_n(\theta_0) := -2\log\frac{L(\theta_0, \hat{\eta}_0)}{L(\hat{\theta}, \hat{\eta})} = -2\log\frac{L_p(\theta_0)}{L_p(\hat{\theta})},$$

where $L_p(\theta) = \max_\eta L(\theta, \eta)$ is the profile likelihood for θ, i.e. $L_p(\theta) = L(\theta, \hat{\eta}(\theta))$, with $\hat{\eta}(\theta) = \arg\max_\eta L(\theta, \eta)$. In this case, $T_n \xrightarrow{d} \chi^2_r$, if H_0 is true, where r is the length of θ. T_n is sometimes referred to as the profile likelihood ratio statistic.

Example 6.6. Let's consider the case of the linear model $Y = \beta_0 + \beta_1 x + \epsilon$, where, for example, $\beta_0 = 10$, $\beta_1 = 20$, $x = 1, \ldots, 10$ and $\epsilon \sim N(0, 10^2)$. Let's say we observe $n = 10$ observations from (Y, x). In this context, we'd like to test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. Under the linear model assumption, the LRT statistic is

$$2(\ell_n(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2) - \ell_n(\hat{\beta}_{00}, 0, \hat{\sigma}_0^2)),$$

where $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ are the unrestricted MLEs (i.e. those we saw in Section 7.3) and $\hat{\beta}_{00}$ and $\hat{\sigma}_0^2$ are the restricted MLEs (i.e. those that maximize likelihood assuming $\beta_1 = 0$).

```r
library(stats4)


set.seed(5)
x <- 1:10
y <- 10 + 20 * x + rnorm(10, sd = 10)
```

```r
neglogLiReg <- function(a, b, s, y) {
  -sum(dnorm(y, mean = a + b * x, sd = s, log = TRUE))
}

loglik <- mle(\(a, b, s) neglogLiReg(a, b, s, y = y),
              start = list(a = mean(y), b = 0, s = sd(y))) |> logLik()
loglik0 <- mle(\(a, s) neglogLiReg(a, b = 0, s, y = y),
               start = list(a = mean(y), s = sd(y))) |> logLik()

LogLR <- (2 * (loglik - loglik0))[1] |> print()
pchisq(LogLR, df = 1, lower.tail = FALSE)
```

```
[1] 37.1
[1] 1.14e-09
```

A simple way to perform such a test in R is to use the `drop1()` function.

```r
glm(y ~ x) |> drop1(test = "LRT")
```

```
Single term deletions

Model:
y ~ x
       Df Deviance   AIC scaled dev. Pr(>Chi)
<none>          812  78.3
x       1    33072 113.4        37.1  1.1e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```