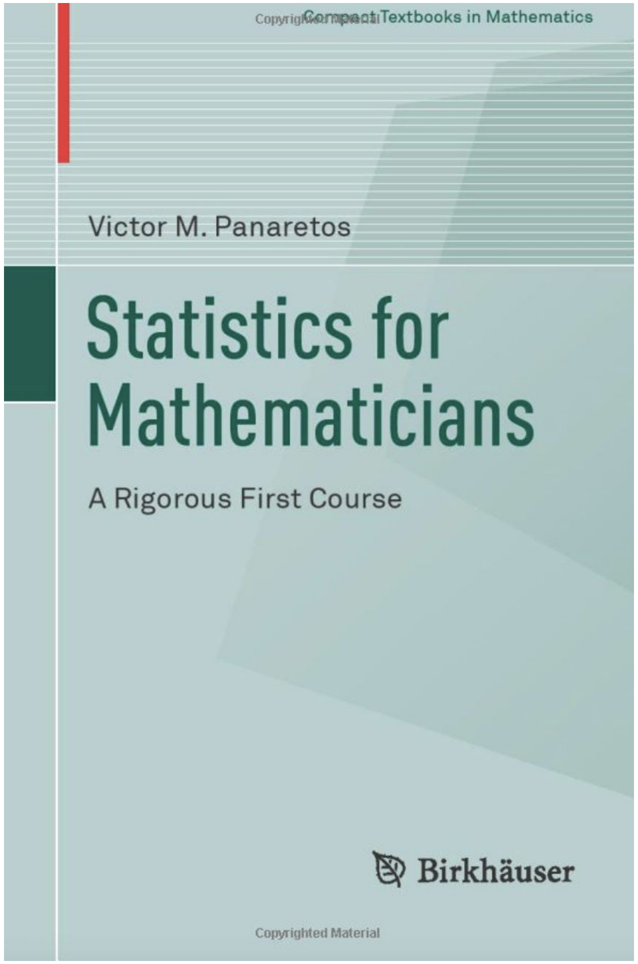
High-Dimensional Probability and Statistics

MATH/STAT/ECE 888: Topics in Mathematical Data Science
Sebastien Roch (Math+Stat)
UW-Madison
Fall 2021

Lecture 24 (11/01/21)

Today’s slides based on Panaretos



https://search.library.wisc.edu/catalog/9912241369602121

Tests and error types
[P, Section 4.1]

Hypothesis testing: motivating example

So far, we have considered the problem of point estimation: given a parametric model $\{F_\theta : \theta \in \Theta\}$, and an iid sample $X_1, \ldots, X_n$ issued from some specific F_θ, estimate the value of θ that generated the sample. There are many contexts, however, where the precise value of the true parameter is not the primary object of our interest. Rather, we are more interested in using the sample to ascertain whether the true value of the parameter belongs to some specific subset of parameter values or not.

Example 4.1 (Coin Tossing)

For a simple example, consider a situation where we wish to ascertain whether a coin is fair, or is biased. We may flip the coin n times and record the outcome of each coin toss. We then wish to use the outcomes in order to decide whether the probability of heads is equal to 1/2 or whether it is different from 1/2. We could formalise this problem by saying that we have $X_1, \ldots, X_n \overset{iid}{\sim} Bern(p)$ and wish to decide whether $p \in \{\frac{1}{2}\}$ or $p \in (0, 1) \setminus \{\frac{1}{2}\}$. $\square$

Hypothesis testing: motivating example cont'd

To make things more concrete, suppose that we know that the parameter has to lie in one of two sets: either in Θ_0 or in Θ_1, where $\Theta_0 \cap \Theta_1 = \emptyset$. We wish to employ the sample $X_1, .., X_n$ that we have at our disposal in order to decide which is the case. This setup arises very often in the sciences, where there are two competing scientific hypotheses. The *null hypothesis* H_0, that states that $\theta \in \Theta_0$,

$$H_0 : \theta \in \Theta_0$$

and the competing *alternative hypothesis* that instead postulates that $\theta \in \Theta_1$,

$$H_1 : \theta \in \Theta_1.$$

Test function

Definition 4.3 (Test Function)

A test function δ is any function $\delta : \mathcal{X}^n \to \{0, 1\}$.

A test function takes the value '0' when we rule in favour of H_0 based on the sample, and it takes the value '1' when we rule in favour of H_1. A test function will typically take the value 0 or 1 depending on whether or not the sample satisfies a certain condition. In other words, test functions are usually constructed by

$$\delta(X_1, \dots, X_n) = \begin{cases} 1, & \text{if } T(X_1, \dots, X_n) \in C, \\ 0, & \text{if } T(X_1, \dots, X_n) \notin C. \end{cases}$$

where T is a statistic called a *test statistic* and C a set in the range of T called the *critical region*. Notice that in compact notation, we may write

$$\delta(X_1, \dots, X_n) = \mathbf{1}\{T(X_1, \dots, X_n) \in C\}.$$

Error types

In hypothesis testing, there are two possible states of nature, and two possible decisions that we can make. Therefore, the "error landscape" is described by the following table:

Decision/Truth	H_0	H_1
0	No error	Type II error
1	Type I error	No error

Definition 4.4 (Error Probabilities)

Let $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ be two competing hypotheses. The Type I error Probability is defined to be the mapping $h : \Theta_0 \to [0, 1]$,

$$h(\theta) = \mathbb{P}_\theta[\delta = 1], \qquad \theta \in \Theta_0.$$

The Type II error Probability is defined to be the mapping $g : \Theta_1 \to [0, 1]$,

$$g(\theta) = \mathbb{P}_\theta[\delta = 0], \qquad \theta \in \Theta_1.$$

► **Remark 4.6 (Warning on Error Probabilities)** Notice that $h(\theta) \neq 1 - g(\theta)$ since the two functions are defined over different domains. It is a common mistake to not realise this.

Type I v. Type II: example

► **Remark 4.7 (Type I vs Type II Error)** It is no coincidence that the two types of error are given two different names, and in fact names that suggest that one kind of error is of primary importance (type I) and the other is secondary (type II). In many practical contexts, the two hypotheses are asymmetric: making one kind of error is far more serious than the other type of error. The more serious type of error is named the Type I error and the other is the Type II error. Therefore, in all practical situations, H_0 is chosen to be the hypothesis whose false rejection is more harmful.

Example 4.8 (Spam Filter)

Suppose we wish an automatic test function to decide whether a new email is spam or not. The new message contains n words $X_1, .., X_n$ and we need a test function in order to decide between two competing hypotheses: "spam" versus "not spam". Notice that marking a message as spam when it is in fact not can have serious consequences (since we will not see it and it could be important). Marking a message as "not spam" when in fact it is spam is annoying, but perhaps not as big of a problem. In this context, it is reasonable to define "H_0 : Message is not spam" and "H_1 : Message is spam". If we do so, the type I error will be precisely the probability to mark a message as spam when it is not. □

Neyman-Pearson
[P, Section 4.2]

A lower bound [taken from Rigollet's notes]

Let ν be a sigma finite measure satisfying $\mathbb{P}_0 \ll \nu$ and $\mathbb{P}_1 \ll \nu$. For example we can take $\nu = \mathbb{P}_0 + \mathbb{P}_1$. It follows from the Radon-Nikodym theorem [Bil95] that both $\mathbb{P}_0$ and $\mathbb{P}_1$ admit probability densities with respect to ν. We denote them by p_0 and p_1 respectively. For any function f, we write for simplicity

$$\int f = \int f(x)\nu(\mathrm{d}x)$$

Lemma 4.3 (Neyman-Pearson). *Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be two probability measures. Then for any test ψ, it holds*

$$\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0) \geq \int \min(p_0, p_1)$$

Moreover, equality holds for the Likelihood Ratio test $\psi^\star = \mathbb{1}(p_1 \geq p_0)$.

A lower bound cont'd [taken from Rigollet's notes]

Proof. Observe first that

$$\begin{aligned}
\mathbb{P}_0(\psi^\star = 1) + \mathbb{P}_1(\psi^\star = 0) &= \int_{\psi^*=1} p_0 + \int_{\psi^*=0} p_1 \\
&= \int_{p_1 \ge p_0} p_0 + \int_{p_1 < p_0} p_1 \\
&= \int_{p_1 \ge p_0} \min(p_0, p_1) + \int_{p_1 < p_0} \min(p_0, p_1) \\
&= \int \min(p_0, p_1)\,.
\end{aligned}$$

A lower bound cont'd [taken from Rigollet's notes]

Next for any test ψ, define its rejection region $R = \{\psi = 1\}$. Let $R^\star = \{p_1 \geq p_0\}$ denote the rejection region of the likelihood ratio test $\psi^\star$. It holds

$$\begin{aligned}
\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0) &= 1 + \mathbb{P}_0(R) - \mathbb{P}_1(R) \\
&= 1 + \int_R p_0 - p_1 \\
&= 1 + \int_{R \cap R^\star} p_0 - p_1 + \int_{R \cap (R^\star)^c} p_0 - p_1 \\
&= 1 - \int_{R \cap R^\star} |p_0 - p_1| + \int_{R \cap (R^\star)^c} |p_0 - p_1| \\
&= 1 + \int |p_0 - p_1| \big[\mathbb{1}(R \cap (R^\star)^c) - \mathbb{1}(R \cap R^\star) \big]
\end{aligned}$$

The above quantity is clearly minimized for $R = R^\star$. $\square$

Aside: total variation distance [taken from Rigollet's notes]

Definition-Proposition 4.4. *The* total variation distance *between two probability measures* $\mathbb{P}_0$ *and* $\mathbb{P}_1$ *on a measurable space* $(\mathcal{X}, \mathcal{A})$ *is defined by*

$$
\begin{aligned}
\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) &= \sup_{R \in \mathcal{A}} |\mathbb{P}_0(R) - \mathbb{P}_1(R)| && (i) \\
&= \sup_{R \in \mathcal{A}} \left| \int_R p_0 - p_1 \right| && (ii) \\
&= \frac{1}{2} \int |p_0 - p_1| && (iii) \\
&= 1 - \int \min(p_0, p_1) && (iv) \\
&= 1 - \inf_{\psi} \left[\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0) \right] && (v)
\end{aligned}
$$

where the infimum above is taken over all tests.

Aside: total variation distance [taken from Rigollet's notes]

Proof. Clearly $(i) = (ii)$ and the Neyman-Pearson Lemma gives $(iv) = (v)$. Moreover, by identifying a test ψ to its rejection region, it is not hard to see that $(i) = (v)$. Therefore it remains only to show that (iii) is equal to any of the other expressions. Hereafter, we show that $(iii) = (iv)$. To that end, observe that

$$\begin{aligned}
\int |p_0 - p_1| &= \int_{p_1 \geq p_0} p_1 - p_0 + \int_{p_1 < p_0} p_0 - p_1 \\
&= \int_{p_1 \geq p_0} p_1 + \int_{p_1 < p_0} p_0 - \int \min(p_0, p_1) \\
&= 1 - \int_{p_1 < p_0} p_1 + 1 - \int_{p_1 \geq p_0} p_0 - \int \min(p_0, p_1) \\
&= 2 - 2\int \min(p_0, p_1)
\end{aligned}$$

$\square$

Back to Neyman-Pearson

Definition 4.9 (Neyman–Pearson Framework)

Let $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ be two competing hypotheses.

1. Fix an $\alpha \in (0, 1)$ and call it the *significance level* or just *level* of the test.
2. Consider only test functions $\delta : \mathcal{X}^n \to \{0, 1\}$ that respect the level, i.e. test functions δ such that

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\delta = 1] \leq \alpha.$$

For ease of reference, we call this class $\mathcal{D}(\Theta_0, \alpha)$. In other words,

$$\mathcal{D}(\Theta_0, \alpha) = \left\{ \delta : \mathcal{X}^n \to \{0, 1\} \,\middle|\, \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\delta = 1] \leq \alpha \right\}.$$

3. Within the class of test functions $\mathcal{D}(\Theta_0, \alpha)$, compare test functions by considering which has lower type II error probability

$$g(\theta) = \mathbb{P}_\theta[\delta = 0], \qquad \theta \in \Theta_1.$$

Equivalently, one can compare test functions by considering which has higher *power*

$$\beta(\theta) = 1 - g(\theta) = \mathbb{P}_\theta[\delta = 1], \qquad \theta \in \Theta_1.$$

Optimal tests

Definition 4.10 (Optimal Tests)

A test function δ of $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ is called optimal at level α (or uniformly most powerful at level α) if the following two hold.

1. $\delta \in \mathcal{D}(\Theta_0, \alpha)$.
2. $\mathbb{P}_{\theta_1}[\psi = 1] \leq \mathbb{P}_{\theta_1}[\delta = 1]$ for all $\theta_1 \in \Theta_1$ and all $\psi \in \mathcal{D}(\Theta_0, \alpha)$.

Neyman-Pearson lemma

Lemma 4.11 (Neyman–Pearson) *Let* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *have joint density (or frequency) function* $f_{\boldsymbol{X}}(\boldsymbol{x};\theta)$ *and suppose we wish to test*

$$H_0 : \theta = \theta_0 \qquad vs \qquad H_1 : \theta = \theta_1.$$

at some level $\alpha \in (0, 1)$, *for* $\theta_0 \neq \theta_1$. *If the random variable*

$$\Lambda(\boldsymbol{X}) = \frac{f_{\boldsymbol{X}}(X_1, \ldots, X_n; \theta_1)}{f_{\boldsymbol{X}}(X_1, \ldots, X_n; \theta_0)} = \frac{L(\theta_1)}{L(\theta_0)}$$

is such that there exists a $Q > 0$ *satisfying*

$$\mathbb{P}_{\theta_0}[\Lambda > Q] = \alpha$$

then the test whose test function is given by

$$\delta(\boldsymbol{X}) = \mathbf{1}\{\Lambda(X) > Q\},$$

is an optimal (most powerful) *test of* H_0 *versus* H_1 *at significance level* α.

Neyman-Pearson lemma cont'd

► **Remark 4.12** A sufficient condition for the existence of a suitable Q for any $\alpha \in (0, 1)$ is that Λ be a continuous random variable under the null hypothesis. If the distribution of Λ under H_0 is discrete or has discontinuities, there may exist $\alpha \in (0, 1)$ such that $\mathbb{P}_{\theta_0}[\Lambda > Q] = \alpha$ cannot be satisfied for any $Q > 0$.

Notice the intuition behind the test: we know that the method of maximum likelihood is a very good estimation method. The higher the likelihood of a parameter, the more plausible this parameter value is as a guess for the true parameter. So, in order to test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, we decide to compare the value of the likelihood function at the two competing parameter values θ_0 and θ_1. If the likelihood of θ_1 is *significantly* higher than the likelihood of θ_0, then we reject H_0 in favour of H_1. How much higher qualifies as *significantly higher*? The theorem tells us that Q-times higher is significantly higher, where Q is a critical value chosen so that the level α be respected.

Neyman-Pearson proof

Proof of Lemma 4.11 We need to verify properties (1) and (2) in Definition 4.10 (p. 103). Since Q is such that $\mathbb{P}_{\theta_0}[\Lambda > Q] = \alpha$, then we immediately have that

$$\mathbb{P}_{\theta_0}[\delta = 1] = \alpha \qquad (\text{since } \mathbb{P}_{\theta_0}[\delta = 1] = \mathbb{P}_{\theta_0}[\Lambda > Q]). \tag{4.1}$$

Therefore $\delta \in \mathcal{D}(\{\theta_0\}, \alpha)$ (i.e. δ indeed respects the level α) which yields (1).

To show (2), let $\psi \in \mathcal{D}(\{\theta_0\}, \alpha)$. For notational ease, write $(X_1, \ldots, X_n)^\top = \boldsymbol{X}$ and $(x_1, \ldots, x_n)^\top = \boldsymbol{x}$. Without loss of generality assume that $f_{\boldsymbol{X}}$ is a density function (otherwise replace any integrals that follow by sums), and observe that

$$f(\boldsymbol{x}; \theta_1) - Q \cdot f(\boldsymbol{x}; \theta_0) > 0 \text{ if } \delta(\boldsymbol{x}) = 1 \quad \& \quad f(\boldsymbol{x}; \theta_1) - Q \cdot f(\boldsymbol{x}; \theta_0) \leq 0 \text{ if } \delta(\boldsymbol{x}) = 0.$$

Therefore, since ψ can only take the values 0 or 1,

$$\psi(\boldsymbol{x})(f(\boldsymbol{x}; \theta_1) - Q \cdot f(\boldsymbol{x}; \theta_0)) \leq \delta(\boldsymbol{x})(f(\boldsymbol{x}; \theta_1) - Q \cdot f(\boldsymbol{x}; \theta_0))$$

$$\int_{\mathcal{X}^n} \psi(\boldsymbol{x})(f(\boldsymbol{x}; \theta_1) - Q \cdot f(\boldsymbol{x}; \theta_0)) d\boldsymbol{x} \leq \int_{\mathcal{X}^n} \delta(\boldsymbol{x})(f(\boldsymbol{x}; \theta_1) - Q \cdot f(\boldsymbol{x}; \theta_0)) d\boldsymbol{x}$$

Neyman-Pearson proof cont'd

Rearranging the terms yields

$$\int_{\mathcal{X}^n} (\psi(\boldsymbol{x}) - \delta(\boldsymbol{x})) f(\boldsymbol{x};\theta_1) d\boldsymbol{x} \le Q \int_{\mathcal{X}^n} (\psi(\boldsymbol{x}) - \delta(\boldsymbol{x})) f(\boldsymbol{x};\theta_0) d\boldsymbol{x}$$

$$\implies \mathbb{E}_{\theta_1}[\psi(\boldsymbol{X})] - \mathbb{E}_{\theta_1}[\delta(\boldsymbol{X})] \le Q\left(\mathbb{E}_{\theta_0}[\psi(\boldsymbol{X})] - \mathbb{E}_{\theta_0}[\delta(\boldsymbol{X})]\right)$$

$$\implies \mathbb{P}_{\theta_1}[\psi(\boldsymbol{X}) = 1] - \mathbb{P}_{\theta_1}[\delta(\boldsymbol{X}) = 1] \le Q\left(\mathbb{P}_{\theta_0}[\psi(\boldsymbol{X}) = 1] - \mathbb{P}_{\theta_0}[\delta(\boldsymbol{X}) = 1]\right)$$

Equation (4.1), combined with the fact that $\psi \in \mathcal{D}(\{\theta_0\}, \alpha)$ and $Q > 0$, implies that the right-hand side is non-positive. This proves (2) in Definition 4.10 (p. 103), and thus completes the proof. $\square$

Example 4.14 (Simple vs Simple Test in Exponential Families)

Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, where $f(x;\theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}$ is a one-parameter exponential family, with η being increasing. Suppose we wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Without loss of generality, assume that $\theta_0 < \theta_1$. The Neyman–Pearson Lemma (Lemma 4.11, p. 103) dictates that we should look for a test statistic of the form

$$\delta = \mathbf{1}\{L(\theta_1)/L(\theta_0) > Q\} = \mathbf{1}\{\log L(\theta_1) - \log L(\theta_0) > \log Q\}.$$

By the exponential family form of $f(x;\theta)$, we obtain that

$$\begin{aligned}\delta &= \mathbf{1}\left\{(\eta(\theta_1) - \eta(\theta_0)\sum_{i=1}^{n} T(X_i) - n(d(\theta_1) - d(\theta_0)) > \log Q\right\}\\ &= \mathbf{1}\left\{\sum_{i=1}^{n} T(X_i) > \frac{\log Q + n(d(\theta_1) - d(\theta_0))}{\eta(\theta_1) - \eta(\theta_0)}\right\}.\end{aligned}$$

Example 4.14 cont'd

Notice that $\eta(\theta_1) - \eta(\theta_0) > 0$ since η is increasing, and $n(d(\theta_1) - d(\theta_0))$ is just a constant. So we can just write

$$\delta = \mathbf{1}\{\tau(X_1, \dots, X_n) > q\},$$

If τ is a continuous random variable, and we want a level α test, then q is going to be the $(1-\alpha)$-quantile of $G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \dots, X_n) \leq t]$, i.e. the $(1-\alpha)$-quantile of the sampling distribution of $\tau(X_1, \dots, X_n)$ when the parameter is taken to be θ_0 (this is called the *null distribution of* τ).

If, on the other hand, we have that η is a decreasing function, then for $\theta_0 < \theta_1$, we have $\eta(\theta_1) - \eta(\theta_0) < 0$. In this case, we can see that the optimal test statistic becomes

$$\delta = \mathbf{1}\{\tau(X_1, \dots, X_n) \leq q\},$$

This time, if we want a level α test, then q must be the α-quantile of $G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \dots, X_n) \leq t]$.

Example 4.14 cont'd

We observe that the form of the test depends on whether η is increasing or decreasing, and on whether $\theta_0 < \theta_1$ or $\theta_0 > \theta_1$. The following table summarises the form of the test statistic for the different cases. In each case, q_s represents the s-quantile of the distribution $G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \dots, X_n) \leq t]$.

	$\theta_0 < \theta_1$	$\theta_0 > \theta_1$
$\eta(\cdot)$ increasing	$\mathbf{1}\{\tau(X_1, \dots, X_n) > q_{1-\alpha}\}$	$\mathbf{1}\{\tau(X_1, \dots, X_n) \leq q_{\alpha}\}$
$\eta(\cdot)$ decreasing	$\mathbf{1}\{\tau(X_1, \dots, X_n) \leq q_{\alpha}\}$	$\mathbf{1}\{\tau(X_1, \dots, X_n) > q_{1-\alpha}\}$

An interesting observation is that the test function does not depend on the precise value of θ_1, but only on whether or not $\theta_1 < \theta_0$ or $\theta_1 > \theta_0$. $\square$

Likelihood ratio test

Definition 4.19 (Likelihood Ratio Test)

Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, yielding a likelihood

$$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta),$$

and let $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ be two competing hypotheses. Define the likelihood ratio as

$$\Lambda(X_1, \ldots, X_n) = \frac{\sup_{\theta\in\Theta_1} L(\theta)}{\sup_{\theta\in\Theta_0} L(\theta)}.$$

The Likelihood Ratio Test (LRT) at level $\alpha \in (0, 1)$ is defined to be the test with test function

$$\delta(X_1, \ldots, X_n) = \mathbf{1}\{\Lambda(X_1, \ldots, X_n) > Q\},$$

where $Q > 0$ is such that $\sup_{\theta\in\Theta_0} \mathbb{P}_\theta[\Lambda(X_1, \ldots, X_n) > Q] = \alpha$, provided such a Q exists.

Likelihood ratio test cont'd

$X_1, \ldots, X_n \overset{iid}{\sim} f(x; \theta, \xi)$, where $\theta \in \mathbb{R}$ and $\xi \in \mathbb{R}^p$ are two unknown parameters. We might be interested in testing

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

at level $\alpha > 0$, for some $\theta_0 \in \mathbb{R}$, without making any reference to (and without caring about) the remaining parameter ξ (a parameter such as ξ is often referred to as a *nuisance parameter*). In this case, the likelihood ratio is formed as

$$\Lambda(X_1, \ldots, X_n) = \frac{\sup_{\theta \in \mathbb{R} \setminus \{\theta_0\}, \xi \in \mathbb{R}^p} L(\theta, \xi)}{\sup_{\theta \in \{\theta_0\}, \xi \in \mathbb{R}^p} L(\theta, \xi)} = \frac{\sup_{\theta \in \mathbb{R}, \xi \in \mathbb{R}^p} L(\theta, \xi)}{\sup_{\xi \in \mathbb{R}^p} L(\theta_0, \xi)} = \frac{L(\hat{\theta}, \hat{\xi})}{\sup_{\xi \in \mathbb{R}^p} L(\theta_0, \xi)},$$

where $(\hat{\theta}, \hat{\xi})$ is an MLE of (θ, ξ). The Likelihood Ratio Test at level $\alpha \in (0, 1)$ will be defined again as the test with test function

$$\delta(X_1, \ldots, X_n) = \mathbf{1}\{\Lambda(X_1, \ldots, X_n) > Q\},$$

where $Q > 0$ is such that $\sup_{\xi \in \mathbb{R}^p} \mathbb{P}_{\theta_0, \xi}[\Lambda(X_1, \ldots, X_n) > Q] = \alpha$, provided such a

Example 4.22 (Bilateral Test for Means of Gaussian Distributions)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where μ and σ^2 are unknown. Suppose we wish to test the hypothesis pair

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu \neq \mu_0$$

at level $\alpha > 0$, for some fixed $\mu_0 \in \mathbb{R}$. Let us use the Likelihood Ratio method in order to derive a suitable test. We notice that we have two parameters, but are only interested in one of them. Following the reasoning presented above, we need to determine

$$\Lambda(X_1, \ldots, X_n) = \frac{L(\hat{\mu}, \hat{\sigma}^2)}{\sup_{\sigma^2 > 0} L(\mu_0, \sigma^2)}, \tag{4.2}$$

where $(\hat{\mu}, \hat{\sigma}^2)$ is the MLE of (μ, σ^2). For the numerator, one may calculate that

$$\frac{\partial}{\partial \sigma^2} \ell(\mu_0, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (X_i - \mu_0)^2.$$

Example 4.22 (Bilateral Test for Means of Gaussian Distributions) cont'd

Following the same steps as in Exercise 3.16 (p. 71), we conclude that

$$\arg\sup_{\sigma^2>0} L(\mu_0, \sigma^2) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2.$$

In other words, the supremum in the numerator in Eq. (4.2) satisfies

$$\sup_{\sigma^2>0} L(\mu_0, \sigma^2) = L\left(\mu_0\,,\ \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2\right),$$

and so the numerator is equal to

$$\begin{aligned}\sup_{\sigma^2>0} L(\mu_0, \sigma^2) &= \left[\frac{1}{2\pi(1/n)\sum_{i=1}^{n}(X_i - \mu_0)^2}\right]^{n/2} \exp\left\{-\frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{(2/n)\sum_{i=1}^{n}(X_i - \mu_0)^2}\right\}\\ &= \left[\frac{ne^{-1}}{2\pi\sum_{i=1}^{n}(X_i - \mu_0)^2}\right]^{n/2}.\end{aligned}$$

Example 4.22 (Bilateral Test for Means of Gaussian Distributions) cont'd

Next, we turn to the denominator in Eq. (4.2). Recalling Example 3.16 (p. 71), we have that the MLE of (μ, σ^2) is given by the pair:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}, \qquad \& \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

It follows that

$$\begin{aligned} L(\hat{\mu}, \hat{\sigma}^2) &= \left[\frac{1}{2\pi(1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]^{n/2} \exp\left\{-\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{(2/n)\sum_{i=1}^{n}(X_i - \overline{X})^2}\right\} \\ &= \left[\frac{ne^{-1}}{2\pi\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]^{n/2}. \end{aligned}$$

Consequently, the likelihood ratio is

$$\Lambda(X_1, \ldots, X_n) = \frac{L(\hat{\mu}, \hat{\sigma}^2)}{\sup_{\sigma^2 > 0} L(\mu_0, \sigma^2)} = \left[\frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]^{n/2}.$$

Example 4.22 (Bilateral Test for Means of Gaussian Distributions) cont'd

This can be further simplified by recalling that

$$\sum_{i=1}^{n}(X_i - \mu_0)^2 = \sum_{i=1}^{n}(X_i - \overline{X} + \overline{X} - \mu_0)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - \mu_0)^2,$$

since the cross-terms vanish. Using this fact, we may write

$$\Lambda(X_1, \ldots, X_n) = \left[\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - \mu_0)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]^{n/2} = \left\{1 + \frac{n(\overline{X} - \mu_0)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right\}^{n/2}.$$

Example 4.22 (Bilateral Test for Means of Gaussian Distributions) cont'd

Observe now that

$$\Lambda > Q \iff \underbrace{\frac{n(\overline{X}-\mu_0)^2}{\sum_{i=1}^n (X_i-\overline{X})^2/(n-1)}}_{T^2} > \underbrace{(n-1)(Q^{2/n}-1)}_{:=C} \iff \underbrace{\left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right|}_{|T|} > \sqrt{C},$$

so the likelihood ratio test is

$$\delta(X_1,\ldots,X_n) = \mathbf{1}\{\Lambda > Q\} = \mathbf{1}\left\{\left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right| > \sqrt{C}\right\},$$

and $\sqrt{C}$ needs to be selected so that $\mathbb{P}_{H_0}\left[\left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right| > \sqrt{C}\right] = \alpha$. But, when H_0 is true, we have that $T \sim \mathsf{t}_{n-1}$, the latter denoting Student's distribution with $n-1$ degrees of freedom (see Theorem 2.9, p. 48). It follows that $\sqrt{C} = t_{n-1,1-\alpha/2}$, where the latter is the $(1-\alpha/2)$-quantile of the t_{n-1} distribution. In conclusion, the LRT is

$$\delta = \mathbf{1}\left\{|\overline{X}-\mu_0| > t_{n-1,1-\alpha/2} S/\sqrt{n}\right\}.$$

The p-value
[P, Section 4.4]

p-value

1. It is not always clear a priori what the "right" significance level is. Should we take $\alpha = 0.05$, or should we take $\alpha = 0.04$? It is the scientist who should suggest what the "right" significance level is, and then the mathematician gives the test function. But what if the scientist does not really know what the precise level should be, or if two different scientists suggest two different levels? This can be an issue because it might be that, for the same data, picking $\alpha = 0.05$ could result in H_0 being rejected, while picking $\alpha = 0.04$ could result in H_0 not being rejected.
2. Suppose we are somehow able to pick a precise level α, so that we have bypassed the problem stated above. Once the level is set, we use the optimal test (if available), and then for our given data set we make a decision. Suppose we reject H_0 at the level α. The problem now is that we have no clear indication of how comfortable or how marginal our decision was. For instance, would our decision have been different, had we selected a slightly smaller α?

p-value cont'd

Definition 4.28 (p-Value)

Let $X_1, \ldots, X_n \overset{iid}{\sim} f(\cdot; \theta)$ and $H_0 : \theta \in \Theta_0$ be a null hypothesis that is of one of the three following forms:

$$\{H_0 : \theta = \theta_0\} \qquad \text{or} \qquad \{H_0 : \theta \leq \theta_0\} \qquad \text{or} \qquad \{H_0 : \theta \geq \theta_0\}.$$

Let δ_α be a test function for H_0, of one of the two following forms:

$$\delta_\alpha(X_1, \ldots, X_n) := \mathbf{1}\{T(X_1, \ldots, X_n) > q_{1-\alpha}\} \quad \text{or} \quad \delta_\alpha(X_1, \ldots, X_n) := \mathbf{1}\{T(X_1, \ldots, X_n) \leq q_\alpha\},$$

where T is some test statistic, and q_z is the z-quantile of the distribution $G_0(t) = \mathbb{P}_{\theta_0}[T(X_1, \ldots, X_n) \leq t]$. Then, we define

$$p(X_1, \ldots, X_n) := \inf\{\alpha \in (0, 1) : \delta_\alpha(X_1, \ldots, X_n) = 1\}.$$

to be the p-value.

p-value cont’d

In other words, the p-value is a random variable that tells us which is the smallest significance level α for which our testing method would reject the null hypothesis H_0 on the basis of the sample $X_1, \ldots, X_n$. Why does this quantity have any relevance? Because it gives us a measure of how stable our decision is under perturbations of a given level α: if the p-value is very small, then this means that we reject H_0 even if we are very strict and impose a rather small α (i.e. very small probability of type I error). If the p-value is relatively large, this means that we would only have rejected H_0 if we were willing to tolerate a high probability of type I error. How small should the p-value be in order to decide that we have rejected ? The answer is left up to the scientist, who can decide depending on his/her deeper knowledge of the experiment at hand. Notice that this approach gives a solution to the problems (1) and (2) outlined above.

p-value cont'd

Lemma 4.30 (Calculation of p-Values) *In the setup given in Definition 4.28, we have:*

1. *If δ_α is of the form $\delta_\alpha(X_1, \ldots, X_n) := \mathbf{1}\{T(X_1, \ldots, X_n) > q_{1-\alpha}\}$, then*

$$p(X_1, \ldots, X_n) = 1 - G_0(T(X_1, \ldots, X_n))$$

2. *If δ_α is of the form $\delta_\alpha(X_1, \ldots, X_n) := \mathbf{1}\{T(X_1, \ldots, X_n) \leq q_\alpha\}$, then*

$$p(X_1, \ldots, X_n) = G_0(T(X_1, \ldots, X_n))$$

Proof of Lemma 4.30 It suffices to prove (1), as (2) is proven directly analogously. In the setting (1), we can use the fact that G_0 is non-decreasing to write:

$$\delta_\alpha(X_1, \ldots, X_n) = 1 \implies T(X_1, \ldots, X_n) > q_{1-\alpha} \implies G_0(T(X_1, \ldots, X_n)) \geq G_0(q_{1-\alpha})$$
$$\implies G_0(T(X_1, \ldots, X_n)) \geq 1 - \alpha \implies \alpha \geq 1 - G_0(T(X_1, \ldots, X_n)).$$

It follows that $\inf\{\alpha \in (0, 1) : \delta_\alpha(X_1, \ldots, X_n) = 1\} = 1 - G_0(T(X_1, \ldots, X_n))$, and the proof is complete. $\square$

p-value cont’d

► **Remark 4.31 (Interpreting p-Values)** The Lemma gives us a further way of understanding p-values. Let’s concentrate on case (1), where we reject for large values of T. Notice that $1 - G_0(T(X_1, \ldots, X_n))$ equals the probability of observing something as large, or even larger than what we observed, when H_0 is true. Therefore, when the p-value is small, we have in fact observed something that would be very improbable/unusual if H_0 were indeed true. So we expect that H_0 is false. A common mistake is to interpret the p-value as the *probability that H_0 is true*. This is wrong, and in fact does not even make sense, because the parameter θ is not a random variable.

Distribution of p-values

Exercise 57 Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$. Suppose we wish to test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ using the test function δ_α of the form

$$\delta_\alpha(T(X_1, \ldots, X_n)) = \mathbf{1}\{T(X_1, \ldots, X_n) > q_{1-\alpha}\} \text{ or } \delta_\alpha(T(X_1, \ldots, X_n)) = \mathbf{1}\{T(X_1, \ldots, X_n) \leq q_\alpha\},$$

where q_α is the α-quantile of G_0, the CDF of $T(X_1, \ldots, X_n)$ when $\theta = \theta_0$. Assuming that G_0 is continuous, show that, under H_0, the p-value is uniformly distributed on $[0, 1]$.

Example 4.32

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$ and consider the hypothesis pair:

$$H_0 : \mu = 0 \qquad \text{vs} \qquad H_1 : \mu \neq 0$$

We recall (see Example 4.21, p. 115) that the likelihood ratio test for this pair is given by:

$$\delta(X_1, \ldots, X_n) = \mathbf{1}\left\{ \left(\frac{\bar{X}}{1/\sqrt{n}} \right)^2 > \chi^2_{1,1-\alpha} \right\},$$

where $\chi^2_{1,1-\alpha}$ is the $1 - \alpha$ quantile of the χ^2_1 distribution. Notice, therefore, that this test statistic conforms to the setup given in Definition 4.28. We may thus define the corresponding p-value as

$$p(X_1, \ldots, X_n) = 1 - G_{\chi^2_1}\left(n\bar{X}^2\right),$$

where $G_{\chi^2_1}$ denotes the CDF of the χ^2_1 distribution. Observe that when $\bar{X}$ is at a large distance from 0, then the p-value will be small. In fact, the p-value is monotonically decreasing in $\bar{X}$ (note that $G_{\chi^2_1}$ is a monotonically increasing function from $(0, \infty)$ to $(0, 1)$ because the density of a χ^2_1 is strictly positive over the entire interval $(0, \infty)$—see Definition 1.16, p. 13). $\square$

p-value v. Neyman-Pearson

One might finally ask: is there any link between Fisher's and Neyman & Pearson's approach to hypothesis tests? In the case where $G_0(t)$ is strictly monotonic,[2] there is a particularly simple and elegant connection:

Corollary 4.33 *In the setup given in Definition 4.28, let $\alpha_0 \in (0, 1)$ and assume that G_0 is continuous and strictly increasing. If we define a test function*

$$\psi(X_1, \dots, X_n) := \mathbf{1}\{p(X_1, \dots, X_n) \leq \alpha_0\},$$

then $\psi(X_1, \dots, X_n) = \delta_{\alpha_0}(X_1, \dots, X_n)$. In other words, if we reject the null whenever the p-value is smaller than α_0, then our test reduces to δ_{α_0}.

p-value v. Neyman-Pearson cont'd

Proof Without loss of generality, we assume that we are in the setup where the p-value corresponds to a statistic of the form $\delta_\alpha(X_1, \ldots, X_n) := \mathbf{1}\{T(X_1, \ldots, X_n) > q_{1-\alpha}\}$. Now, observe that, using Lemma 4.30, and we have:

$$p(X_1, \ldots, X_n) < \alpha_0 \iff 1 - G_0(T(X_1, \ldots, X_n)) < \alpha_0 \iff G_0(T(X_1, \ldots, X_n)) > 1 - \alpha_0.$$

Under our assumptions, G_0^{-1} exists and is strictly increasing. Applying it to both sides of the last inequality yields:

$$p(X_1, \ldots, X_n) < \alpha_0 \iff T(X_1, \ldots, X_n) > \underbrace{G_0^{-1}(1 - \alpha_0)}_{=q_{1-\alpha_0}} \iff \delta(X_1, \ldots, X_n) = 1.$$