PHIL6334/ECON6614-Lecture Notes 3: Hypothesis Testing 1: Basic Elements

Aris Spanos [Spring 2019]

1 Introduction

After estimation (point and interval) we continue the discussion of frequentist statistical inference by focusing on **hypothesis testing.** The discussion will be confined to the basics of frequentist testing by presenting a bare bones but coherent account that brings out the relationship between Fisher and Neyman-Pearson testing as well as a number of foundational problems.

The **main objective** in statistical testing is to use the sample information $\mathbf{x}\in\mathfrak{X}$, as summarized by $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathfrak{X}$, in conjunction with data $\mathbf{x}_0$**:=**$(x_1,x_2,...,x_n)$, to *narrow down* the set of possible values of the unknown parameter $\boldsymbol{\theta}\in\Theta$ using *hypothetical* values of $\boldsymbol{\theta}$. Ideally, the narrowing down reduces $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ to a single point $\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}^*)\}$, $\mathbf{x}\in\mathfrak{X}$. That is, hypothesis testing is all about learning from data $\mathbf{x}_0$ about the 'true' statistical data Generating Mechanism (GM) $\mathcal{M}^*(\mathbf{x})$. Instead of asking the data to pinpoint the value $\boldsymbol{\theta}^*$as in point estimation, one postulates a specific value, say $\boldsymbol{\theta}=\boldsymbol{\theta}_0$, and poses the question to data data $\mathbf{x}_0$ whether $\boldsymbol{\theta}_0$ is 'close enough' to $\boldsymbol{\theta}^*$ or not, using hypothesis testing. That is, testing replaces the *factual* reasoning of estimation with *hypothetical* reasoning. It turns out that hypothetical reasoning enables statistical testing to pose a plethora of sharper questions to the data and elicit more informative answers.

2 Fisher's significance testing

As in the case of estimation, the cornerstone of testing is the notion of a **statistical model**:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathfrak{X}:=\mathbb{R}^n_X,$$

where $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathfrak{X}$, is the (joint) distribution of the sample that encapsulates the probabilistic structure of the sample $\mathbf{X}:=(X_1,...,X_n)$. Frequentist testing takes place within the boundaries of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ and the hypotheses of interest are framed in terms of the model's unknown parameters $\boldsymbol{\theta}\in\Theta$.

Example 1. Consider the **simple (**one parameter**) Normal model**:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\text{: } X_k\backsim\mathsf{NIID}(\mu,\sigma^2),\ x_k\in\mathbb{R},\ \mu\in\mathbb{R},\ \sigma^2>0,\ k\in\mathbb{N}:=(1,2,...,n,...),$$

where for simplicity we assume that σ^2 is **known**.

A typical Fisher type **null hypothesis** takes the form:

$$H_0\text{: } \mu=\mu_0. \tag{1}$$

The question posed by the null hypothesis in (1) is the extent to which data $\mathbf{x}_0$ accords with the pre-specified value μ_0, i.e. $\Theta=(-\infty,\infty):=\mathbb{R}$ is narrowed down to a single value.

The main elements of **a Fisher significance test** $\{d(\mathbf{X}),p(\mathbf{x}_0)\}$.

Table 1: Fisher significance testing - key elements

(a)	A prespecified statistical model: $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, $\mathbf{x}\in\mathfrak{X}:=\mathbb{R}^n_X$,
(b)	a null (H_0: $\theta=\theta_0$) hypothesis,
(c)	a test statistic (distance function) $\tau(\mathbf{X})$,
(d)	the distribution of $d(\mathbf{X})$ under H_0 is known,
(e)	the p-value $\mathbb{P}(d(\mathbf{X})>d(\mathbf{x}_0);H_0)=p(\mathbf{x}_0)$,
(f)	a threshold value c_0 [e.g. .01, .025, .05], such that: $p(\mathbf{x}_0)<c_0\Rightarrow\mathbf{x}_0$ falsifies (rejects) H_0.

For Fisher the null hypothesis is usually a point in the prameter space:

$$H_0\text{: }\mu=\mu_0,\text{ where }\mu_0\text{ is given value.}$$

Fisher used common sense, such as standardizing a best estimator of μ, to construct a **test statistic** that aims to evaluate the discordance between μ_0 and data $\mathbf{x}_0$, based on the distance function. As in the case of a Confidence Interval (CI), an optimal (effective) test begins with an optimal estimator. In Lecture Notes 2 it was shown that $\overline{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$ provides a *best* estimator of μ with a sampling distribution:

$$\overline{X}_n\backsim\mathsf{N}(\mu,\tfrac{\sigma^2}{n}),\tag{2}$$

being *unbiased* ($E(\overline{X}_n)=\mu^*$), *fully efficient* ($Var(\overline{X}_n)=CR(\mu^*)$), Sufficient ($\sum_{i=1}^n X_i$ is a sufficient statistic) and *strongly consistency* ($\mathbb{P}(\lim_{n\to\infty}\overline{X}_n=\mu^*)$), where μ^* denotes the 'true' value of μ.

Digression. Recall that an optimal $(1-\alpha)$ Confidence interval is based on a standardized version of (2), defined in terms of the pivotal quantity:

$$\frac{\sqrt{n}\left(\overline{X}_n-\mu\right)}{\sigma}\overset{\mu=\mu^*}{\backsim}\mathsf{N}(0,1).\tag{3}$$

The question posed by H_0 in (1) amounts to asking the data $\mathbf{x}_0$ whether the distance $(\mu^*-\mu_0)$ is 'large enough' to indicate *discordance* with H_0 or not. In light of the fact that μ^* is unknown, it makes intuitive sense to use its best estimator to define a test statistic in terms of the difference $(\overline{X}_n-\mu_0)$, . Note that under the hypothetical scenario 'what if $\mu=\mu_0$':

$$\overline{X}_n\overset{\mu=\mu_0}{\backsim}\mathsf{N}(\mu_0,\tfrac{\sigma^2}{n}),\tag{4}$$

which after standardization yields the distance function:

$$d(\mathbf{X})=\frac{\sqrt{n}\left(\overline{X}_n-\mu_0\right)}{\sigma}.$$

In view of the fact that σ is known, $d(\mathbf{X})$ constitutes a **statistic**: a function that involves no unknown parameters of the form $d(.)$: $\mathfrak{X} \to \mathbb{R}$. The question now is, 'how does one define whether this distance is large enough?' Since $d(\mathbf{X})$ is a random variable (being a function of the sample $\mathbf{X}$), that question can only be answered in terms of its sampling distribution. **But what is it?**

R.A. Fisher employed *hypothetical reasoning*, *assuming that* H_0 *is true,* we call **under** H_0: $\mu=\mu_0$:

$$d(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma} \overset{H_0}{\backsim} \mathsf{N}(0,1), \tag{5}$$

which renders this test statistic operational since its sampling distribution is known! He used this to define the **p-value**:

$$\mathbb{P}(d(\mathbf{X}) > d(\mathbf{x}_0);\ H_0 \text{ true})=p(\mathbf{x}_0), \tag{6}$$

as an indicator of discordance between data $\mathbf{x}_0$ and H_0; the bigger the value of $d(\mathbf{x}_0)$ the smaller the p-value. More correctly, the p-value is a measure of discordance between μ^* and μ_0, but since μ^* is unknown we use a second best, its best estimator $\overline{X}_n$.

▲ **Definition (Spanos)**. The p-value is the probability of all sample realizations $\mathbf{x}\in\mathfrak{X}$ such that $d(\mathbf{x})$ accords less well with H_0 than $\mathbf{x}_0$ does, when H_0 is true.

▼ **Traditional definition**. The p-value is defined as the probability, under H_0, of obtaining a result equal to or more extreme than what was actually observed. The clause "equal to or more extreme" is equivocal since it depends on how 'more extreme' is interpreted. This is particularly unclear when the p-value is used in the context of Neyman-Pearson testing where 'more extreme' is invariably related to the alternative hypothesis.

When $p(\mathbf{x}_0)$ is smaller than a certain threshold, say c_0=.025, suggests that data $\mathbf{x}_0$ indicate some discordance with H_0. What about when the p-value is greater than the selected threshold? Fisher preached *strict falsificationism* and rejected any interpretation of that as indicating accordance with H_0.

Numerical Examples

(i) Consider the case where σ=1, n=100, μ_0=12 and data $\mathbf{x}_0$ gave rise to $\overline{x}_n$=12.2. The observed value of the test statistic is $d(\mathbf{x}_0)=\sqrt{100}(12.2-12)=2.0$, which yields:

$$\mathbb{P}(d(\mathbf{X}) > d(\mathbf{x}_0);\ H_0 \text{ true})=.023.$$

This p-value indicates a clear discordance with H_0: μ_0=12, for any threshold $c_0 \geq$ .025.

(ii) Assuming the same values as above except $\overline{x}_n$=12.1 yields $d(\mathbf{x}_0)=\sqrt{100}(12.1-12)=1$, which gives rise to a p-value:

$$\mathbb{P}(d(\mathbf{X}) > d(\mathbf{x}_0);\ H_0 \text{ true})=.159.$$

In this case Fisher would say that data $\mathbf{x}_0$ ($\overline{x}_n$=12.1) indicate no clear discordance with the null, but that should NOT be interpreted as indicating accordance with H_0! Fisher was a strict falsificationist.

Misinterpreting the p-value. It is a *serious error* to interpret the p-value as *conditional* on the null hypotheses H_0 (Cohen, 1994):

$$\mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0)|H_0)=p(\mathbf{x}_0).$$

Conditioning on H_0: $\theta=\theta_0$ or any other value of θ is *meaningless* in frequentist inference since θ is an unknown constant, not a random variable. Conditioning makes sense only when θ is a random variable as in Bayesian statistics. 'H_0 true' deontes an evaluation under a hypothetical scenario H_0: $\theta=\theta_0$ is true. Hence, using the vertical line (|) instead of a semi-colon (;) is incorrect and highly misleading! In light of that, the following interpretations of the p-value are clearly *erroneous*:

▼ (i) assigning probabilities to the null or any other value of θ in Θ is true,

▼ (ii) assigning a probability $(1-p(\mathbf{x}_0))$ that the null is false.

■ It is extremely important to keep in mind that frequentist **error probabilities**, including the p-value, are always attached to the *testing procedure*; they are never attached to θ directly or indirectly! Their aim is to 'calibrate' the capacity of the test to detect discrepancies from θ_0.

Example 2. Consider the **simple Bernoulli model** specified by:

$$\boxed{\mathcal{M}_\theta(\mathbf{x})\text{: } X_k \backsim \mathsf{BerIID}(\theta,\theta(1-\theta)),\ x_k=0,1,\ \theta\in[0,1],\ k=1,2,\dots n,...} \tag{7}$$

Arbuthnot's 1710 conjecture: the ratio of males to females in newborns might *not* be 50-50 ('fair'). This can be tested by framing it as a statistical null hypothesis in terms of θ:

$$H_0\text{: } \theta=\theta_0, \quad \text{where } \theta_0=.5 \text{ denotes 'fair'}$$

in the context of (7) based on the random variable X defined by:

$$\{X=1\}=\{\text{male}\},\ \{X=0\}=\{\text{female}\}.$$

As shown in Lecture Notes (LN) 2 $\overline{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$, is the best estimator of θ with a sampling distribution:

$$\widehat{\theta}_n:=\overline{X}_n \backsim \mathsf{Bin}\left(\theta,\tfrac{\theta(1-\theta)}{n};n\right). \tag{8}$$

Using (8) one can derive the *test statistic*:

$$d(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \overset{H_0}{\backsim} \mathsf{Bin}\left(0,1;n\right), \tag{9}$$

whose Binomial distribution can be approximated accurately using a $\mathsf{N}(0,1)$ distribution for $n>20$; see figure 1.

Data: Let us use Arbuthnot's data of sample size n=14928 newborns during 1710 in London (England), out of which 7640 were boys and 7288 girls. The best estimate of θ is $\widehat{\theta}_n=\frac{7640}{14928}$=.51179, yielding:

$$d(\mathbf{x}_0)=\tfrac{\sqrt{14928}(.51179-.5)}{\sqrt{.5(.5)}}=2.881, \quad \mathbb{P}(d(\mathbf{X})>2.881;\theta=.5)=.0019.$$

The tiny p-value indicates a clear discordance with H_0 since it rejects θ_0=.5 for any threshold $c_0 \geq .002$.

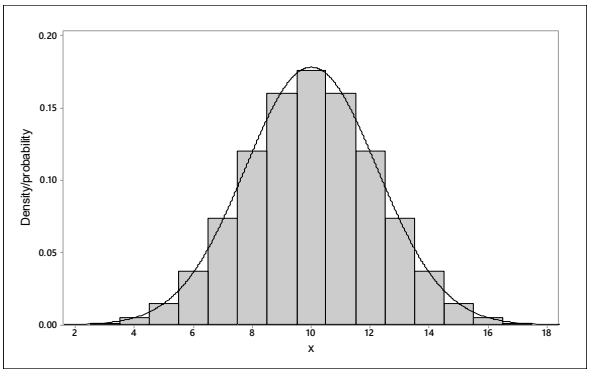

Fig. 1: Normal approximation of the Binomial: $f(y;\theta{=}.5, n{=}20)$

3 Neyman-Pearson (N-P) testing

Neyman and Pearson (1933) pointed out three crucial weaknesses in Fisher's approach to testing:

(i) The ad hoc **choice of a test statistic**. How does one construct an optimal test statistic?

(ii) the **vulnerability** of the **p-value** evaluation to **abuse**. One can evaluate $p(\mathbf{x}_0)$ and then choose a threshold $p(\mathbf{x}_0) \gtrless c_0$ that yields the inference one wants.

(iii) Fisher's **falsificationist** stance. Scientists would like to know if there is evidence *for* (not just against) a particular substantive claim.

They contemplated their primary objective to be a theory of **optimal testing** – analogous to Fisher's optimal estimation. The first was to find an objective procedure that gives rise to the choice of a test statistic based on optimality criteria, not just Fisher's intuition. . The second was to replace the p-value with pre-data error probabilities that will be less vulnerable to abuse. The third was to find a way to allow for accepting the null. The key to achieving these goals was the concept of an *alternative hypothesis* introduced to supplement Fisher's null hypothesis. How one selects the alternative hypothesis, however, is a source of confusion that lingers on to current discussions ot N-P testing.

3.1 Archetypal Neyman-Pearson (N-P) hypotheses

For a particular null hypothesis H_0: $\theta{\in}\Theta_0{\subset}\Theta$ in the context of a statistical model:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}){=}\{f(\mathbf{x};\theta),\ \theta{\in}\Theta\},\ \mathbf{x}{\in}\mathfrak{X}{:=}\mathbb{R}^n_X,$$

cthe **default alternative hypothesis**, denoted by H_1**,** is always defined as the complement of the null $\theta{\in}\Theta_1{=}\Theta-\Theta_0$ with respect to the particular parameter space Θ: $\theta{\in}\Theta_1{=}\Theta-\Theta_0$.

▲ The **archetypal** way to specify the null and alternative hypotheses for N-P testing is:

$$H_0\text{: } \theta{\in}\Theta_0 \text{ vs. } H_1\text{: } \theta{\in}\Theta_1, \tag{10}$$

where Θ_0 and Θ_1 constitute a **partition** of the *parameter space* Θ; $\Theta_0 \cap \Theta_1 = \varnothing$, $\Theta_0 \cup \Theta_1 = \Theta$. This is because for statistical purposes the whole of the parameter space is relevant, despite the fact that only a few value of the unknown parameter are often of interest for substantive purposes. Indeed, using this partitioning addresses numerous confusions in frequentist testing!

In the case where the sets Θ_0 or Θ_1 contain a single point that determines $f(\mathbf{x};\theta_0)$ completely (no unknowns), the hypothesis is said to be **simple**, otherwise it is **composite**.

Example 1 (continued). In the case of (1) the alternative is by default $\mathbb{R}-\{\mu_0\}$:

$$H_0\text{: } \mu=\mu_0 \text{ vs. } H_1\text{: } \mu\neq\mu_0. \tag{11}$$

Similarly, the one-sided hypotheses:

$$H_0\text{: } \mu\leq\mu_0 \text{ vs. } H_1\text{: } \mu > \mu_0, \tag{12}$$

$$H_0\text{: } \mu\geq\mu_0 \text{ vs. } H_1\text{: } \mu < \mu_0, \tag{13}$$

constitute proper partitions of the parameter space $\mathbb{R}$. The framing of hypotheses:

$$H_0\text{: } \mu=\mu_0 \text{ vs. } H_1\text{: } \mu=\mu_1, \tag{14}$$

makes **no sense** from a statistical perspective because it does not constitute a partition.

N-P testing, in effect, *partitions* Θ into Θ_0 and Θ_1, and poses the question **whether data $\mathbf{x}_0$ accord less well with one or the other subset**. To answer that question the N-P approach uses a test statistic $d(\mathbf{X})$ which maps the partition Θ_0 and Θ_1 into a corresponding partition of the sample space $\mathfrak{X}$, say C_0 and C_1, where $C_0 \cap C_1 = \varnothing$, $C_0 \cup C_1 = \mathfrak{X}$, known as *acceptance* and *rejection regions*, respectively:

$$\mathfrak{X}=\left\{\begin{array}{|c|}\hline C_0 \\ \hline C_1 \\ \hline\end{array} \leftrightarrow \begin{array}{|c|}\hline \Theta_0 \\ \hline \Theta_1 \\ \hline\end{array}\right\}=\Theta$$

Fig. 1 places N-P testing in a broader context where $\mathcal{P}(\mathbf{x})$ denotes the set of all possible models that could have given rise to data $\mathbf{x}_0$.

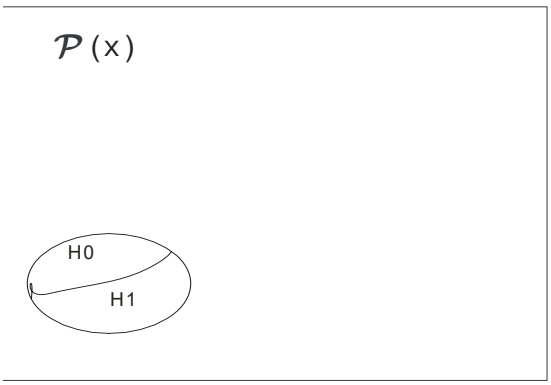

Fig. 1: Neyman-Pearson testing

Table 2: N-P type I and II errors		
N-P rule	H_0 true	H_0 false
Accept H_0	$\surd$	Type II error
Reject H_0	Type I error	$\surd$

The key to applying the N-P approach to testing is to be able to evaluate these two types of error probabilities using hypothetical reasoning associated with H_0 being true or false:

$$\textbf{type I error probability:} \quad \mathbb{P}(\mathbf{x}_0 \in C_1;\ H_0 \text{ true}) = \alpha,$$
$$\textbf{type II error probability:} \quad \mathbb{P}(\mathbf{x}_0 \in C_0;\ H_0 \text{ false}) = \beta,$$
$$\textbf{Power:} \quad \mathbb{P}(\mathbf{x}_0 \in C_1;\ H_0 \text{ false}) = (1-\beta)\,.$$

This depends crucially on knowing the **sampling distribution** of the test statistic $d(\mathbf{X})$ under the scenarios: H_0 true or H_0 false for different values of θ in Θ.

These modifications, in conjunction with the *pre-data* [before $\mathbf{x}_0$ is used for inference] *significance level* (probability of type I error) α, enabled Neyman and Pearson to replace the *post-data* [using $\mathbf{x}_0$] p-value with the *N-P decision rules*:

$$\text{[i] if } \mathbf{x}_0 \in C_0, \text{ accept } H_0, \qquad \text{[ii] if } \mathbf{x}_0 \in C_1, \text{ reject } H_0. \tag{15}$$

Example 1 (continued). Let us consider the hypotheses of interest:

$$H_0\text{: } \mu \leq \mu_0 \text{ vs. } H_1\text{: } \mu > \mu_0. \tag{16}$$

It turns out that the optimal N-P test for (16) concides with that for the hypotheses:

$$H_0\text{: } \mu = \mu_0 \text{ vs. } H_1\text{: } \mu > \mu_0. \tag{17}$$

Provisionally, let us adopt Fisher's test statistic $d(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma}$ as the appropriate distance function. In light of the fact that departures from the null are associated with large values of $d(\mathbf{X})$ a natural **rejection region** is:

$$C_1(\alpha) = \{\mathbf{x}\text{: } d(\mathbf{x}) > c_\alpha\}, \tag{18}$$

where $c_\alpha > 0$ is the threshold rejection value. Given that the sampling distribution of $d(\mathbf{X})$ under H_0 is given in (5), and is completely known, one can evaluate the type I error probability for different rejection values c_α using:

$$\mathbb{P}(d(\mathbf{X}) > c_\alpha;\ H_0 \text{ true}) = \alpha,$$

where α is the type I error; $0 < \alpha < 1$. Similarly, the **acceptance region** is:

$$C_0(\alpha) = \{\mathbf{x}\text{: } d(\mathbf{x}) \leq c_\alpha\},$$

since small values of $d(\mathbf{x})$ indicate accordance with H_0.

To evaluate the **type II error probability** one needs to know the sampling distribution of $d(\mathbf{X})$ when H_0 is false. However, since H_0 is false refers to H_1: $\mu > \mu_0$, this evaluation will involve all values of μ greater than μ_0 (i.e. $\mu_1 > \mu_0$):

$$\beta(\mu_1)=\mathbb{P}(d(\mathbf{X}) \leq c_\alpha;\ H_0 \text{ false})=\mathbb{P}(d(\mathbf{X}) \leq c_\alpha;\ \mu=\mu_1),\ \forall(\mu_1>\mu_0)$$

The relevant sampling distribution takes the form:

$$d(\mathbf{X})=\frac{\sqrt{n}\left(\overline{X}_n-\mu_0\right)}{\sigma} \overset{\mu=\mu_1}{\backsim} \mathsf{N}(\delta,1),\ \delta=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma},\ \text{for all } \mu_1>\mu_0. \tag{19}$$

That is, under H_1 the sampling distribution of $d(\mathbf{X})$ is Normal, but its mean is non-zero and changes with μ_1. Given that $d(\mathbf{X})=d_1(\mathbf{X})+\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$, where $d_1(\mathbf{X})=\frac{\sqrt{n}\left(\overline{X}_n-\mu_1\right)}{\sigma} \overset{\mu=\mu_1}{\backsim} \mathsf{N}(0,1)$, one can use:

$$d_1(\mathbf{X})=d(\mathbf{X})-\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma} \overset{\mu=\mu_1}{\backsim} \mathsf{N}(0,1), \tag{20}$$

the relevant tail areas. Figure 2 illustrates the type I, II error probabilities, and the power.

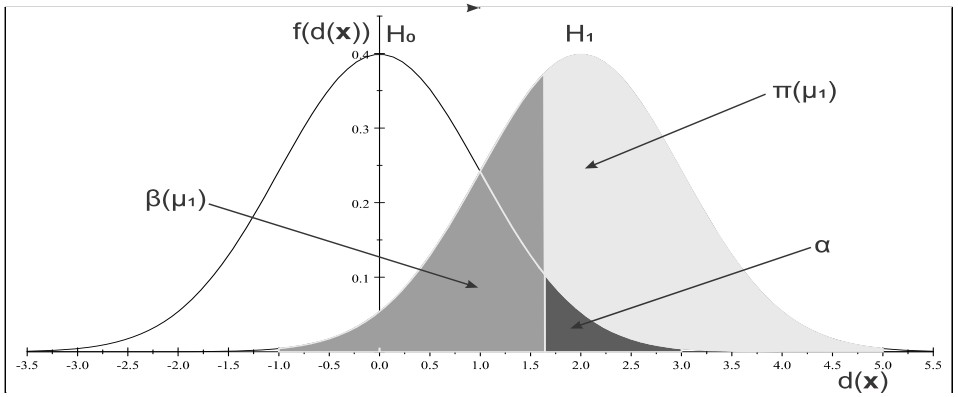

Fig. 2: Type I and II error probabilities and the power of the test

Why power? The power $\pi(\mu_1)$ measures the *pre-data* (generic) *capacity* (probativeness) of test $\mathcal{T}_\alpha$:=$\{d(\mathbf{X}),\ C_1(\alpha)\}$ to detect a discrepancy, say $\gamma=\mu_1-\mu_0=.1$, when present. Hence, when $\pi(\mu_1)=.35$, this test has very low capacity to detect such a discrepancy. If $\gamma=.1$ is the discrepancy of substantive interest, this test is practically useless for that purposes because we know beforehand that this test does not enough capacity (probativeness) to detect γ even if present! What can one do in such a case? The power of $\mathcal{T}_\alpha$ is monotonically increasing with $\delta_1=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$, and thus, increasing the sample size n or decreasing σ increases the power.

Example 1 (continued). In the case of test $\mathcal{T}_\alpha$:=$\{d(\mathbf{X}),\ C_1(\alpha)\}$ in (**??**), decreasing the probability of type I error from $\alpha=.05$ to $\alpha=.01$, increases the threshold from $c_\alpha=1.645$ to $c_\alpha=2.33$, which makes it easier to accept H_0, and this in turn increases the probability of type II error.

Example 1 (continued). Evaluation of the power the test $\mathcal{T}_\alpha$:=$\{d(\mathbf{X}),\ C_1(\alpha)\}$, with σ=1, n=100, c_α=1.645 for different discrepancies $(\mu_1-\mu_0)$ yields the results in table 3, where Z denotes a generic standard Normal random variable, i.e. $Z\backsim\mathsf{N}(0,1)$.

Table 3: Evaluating the power of test $\mathcal{T}_\alpha$		
γ=$\mu_1-\mu_0$	$\delta_1=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$	$\pi(\mu_1)$=$\mathbb{P}(\frac{\sqrt{n}(\overline{X}_n-\mu_1)}{\sigma}>c_\alpha-\delta_1;\mu_1)$
γ=.1	δ=1,	$\pi(10.1)$=$\mathbb{P}(Z>1.645-1)$=.259,
γ=.2	δ=2,	$\pi(10.2)$=$\mathbb{P}(Z>1.645-2)$=.639,
γ=.3	δ=3,	$\pi(10.3)$=$\mathbb{P}(Z>1.645-3)$=.913.

The tail areas associated with the significance level and the power of this test is illustrated in figures 3-4.

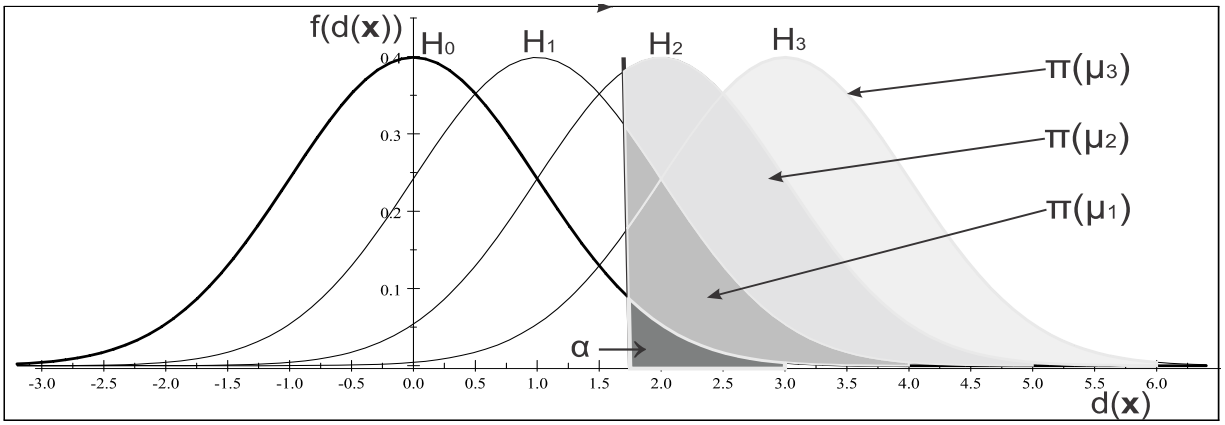

Fig. 3: Power of the test for different discrepancies $\mu_1<\mu_2<\mu_3$ from μ_0

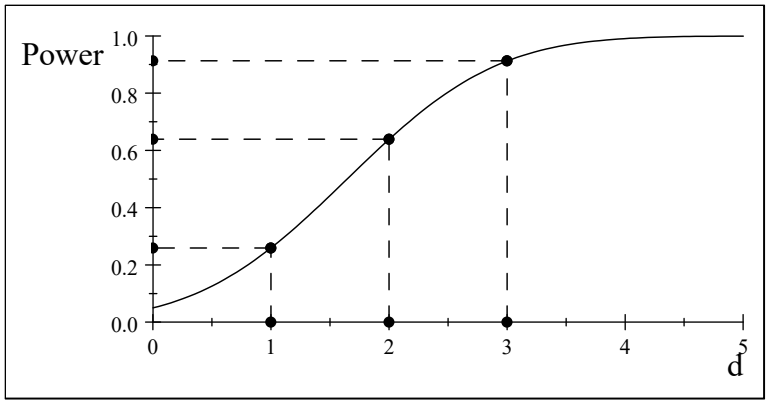

Fig. 4: Power curve (table 2)

The power of the test $\mathcal{T}_\alpha$:=$\{d(\mathbf{X}),\ C_1(\alpha)\}$ is typical of an *optimal test* since $\pi(\mu_1)$ increases with the non-zero mean $\delta_1=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$, and thus the power: (a) increases as the sample size n increases, (b) increases as the discrepancy γ=$(\mu_1-\mu_0)$ increases, and (c) decreases as σ increases.

The power of this test is *typical* **of an optimal (UMP) test** since $P(\mu_1)$ increases with the non-zero mean $\delta_1=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$:

(a) the power increases with the sample size n,

(b) the power increases with the discrepancy $\gamma=(\mu_1-\mu_0)$,

(c) the power decreases with σ.

The features (a)-(b) are often used to decide pre-data on how large n should be to detect departures $(\mu_1-\mu_0)$ of interest as part of the pre-data design of the study.

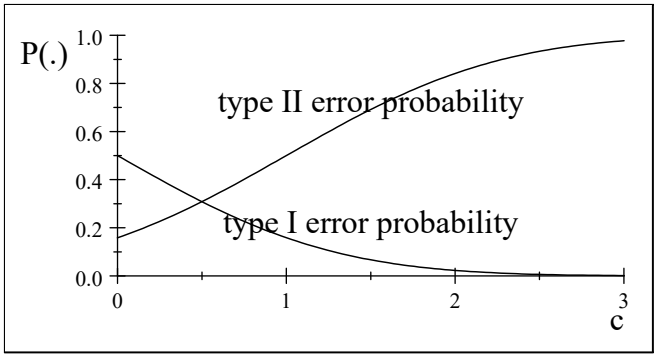

Fig. 5: the trade-off: type I vs. II

To address this trade-off Neyman and Pearson (1933) proposed a twofold strategy.

Defining an optimal N-P test. An *optimal N-P test* is based on: (a) fixing an *upper bound* α for the type I error probability:

$$\mathbb{P}(\mathbf{x}\in C_1;\ H_0(\theta) \text{ true}) \leq \alpha, \text{for all } \theta\in\Theta_0,$$

and then (b) select $\{d(\mathbf{X}),\ C_1(\alpha)\}$ that *minimizes* the type II error probability, or equivalently, *maximizes* the *power*:

$$\pi(\theta)=\mathbb{P}(\mathbf{x}_0\in C_1;\ H_1(\theta) \text{ true})=1-\beta(\theta),\ \text{for all } \theta\in\Theta_1.$$

The general rule is that in selecting an *optimal* N-P test the whole of the parameter space Θ is relevant. This is why partitioning both Θ and the sample space $\mathbb{R}^n_X$ using a test statistic $d(\mathbf{X})$ provides the key to N-P testing.

How can one address this trade-off?

Neyman and Pearson (1933) suggested that a natural way to address this trade-off is as follows:

(a) Specify the null and alternative hypothesis in such a way so as to render the type I error the more serious of the two potential errors.

(b) Fix the type I error probability to a small number, say α=.05 or α=.01:

$$\mathbb{P}(d(\mathbf{X}) > c_\alpha;\ H_0 \text{ true})=\alpha,$$

and call α the **significance level** of the test, where the choice of α depend on the particular circumstances.

(c) For a given α, choose the optimal test $\{d(\mathbf{X}), C_1(\alpha)\}$ that *minimizes* the type II error probability for all values $\mu_1>\mu_0$.

The last step is often replaced with an equivalent step:

(c)* for a given α, choose a test $\{d(\mathbf{X}), C_1(\alpha)\}$ that *maximizes* the **power** of the test in question for all values $\mu_1 > \mu_0$:

$$P(\mu_1) = \mathbb{P}(d(\mathbf{X}) > c_\alpha;\ H_1(\mu_1) \text{ true}),\ \text{for all } \mu_1 > \mu_0. \tag{21}$$

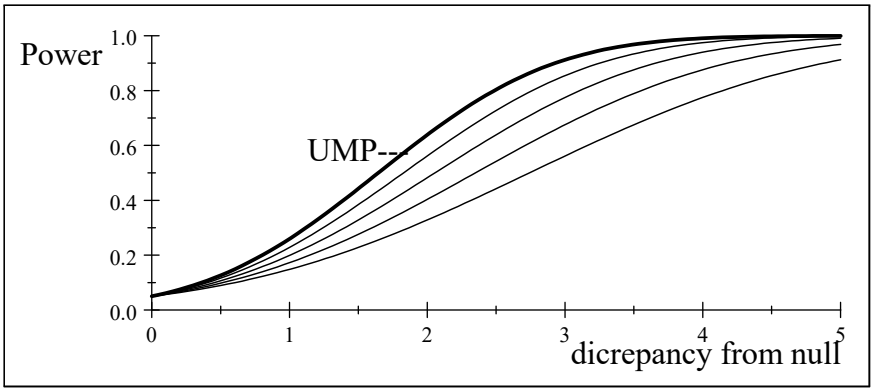

Fig. 6: The power of several tests; UMP is the bold line

Uniformly Most Powerful (UMP). A test $T_\alpha := \{d(\mathbf{X}),\ C_1(\alpha)\}$ is said to be *UMP* if it has higher power than any other α-level test $\widetilde{T}_\alpha$, *for all* values $\theta \in \Theta_1$. In symbols:

$$P(\theta; T_\alpha) \geq P(\theta; \widetilde{T}_\alpha) \text{ for all } \theta \in \Theta_1.$$

That is, a UMP N-P test is one whose power curve dominates that of every other possible test in the sense that for all $\mu_1 > \mu_0$ its power is greater than or equal to that of the other tests; see fig. 6.

Additional properties of N-P tests

[2] **Unbiasedness:** A test $\mathcal{T}_\alpha := \{d(\mathbf{X}),\ C_1(\alpha)\}$ is said to be *unbiased* if the probability of rejecting H_0 when false (power of the test) is always greater than that of rejecting H_0 when true (the type I error probability). In symbols:

$$\max_{\theta \in \Theta_0} \mathbb{P}(\mathbf{x}_0 \in C_1;\ H_0(\theta)) \leq \alpha < \mathbb{P}(\mathbf{x}_0 \in C_1;\ H_1(\theta)) \text{ for } \theta \in \Theta_1.$$

That is, a test rejects H_0 more often when H_0 is false than when H_0 is true! In contrast, a biased test would reject H_0 more often when its is true and when it is false, i.e. the significance level α is higher than the power $\pi(\theta)$ for all $\theta \in \Theta_1$.

Example 1 (continued). The question that one might naturally raise at this stage is whether the same test statistic $d(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma}$ can be used to specify a UMP for testing the two-sided hypotheses:

$$H_0\text{: } \mu = \mu_0,\ \text{vs. } H_1\text{: } \mu \neq \mu_0. \tag{22}$$

The rejection region in this case should naturally allow for discrepancies on either side of μ_0 and would take the form:

$$C_1^{\neq}(\alpha) = \{\mathbf{x}\text{: } |d(\mathbf{x})| > c_{\frac{\alpha}{2}}\}. \tag{23}$$

It turns out that the test defined by $\{d(\mathbf{X}), C_1^{\neq}(\alpha)\}$ is not UMP, but it is UMP Unbiased; see Spanos (1999).

[3] **Consistency**: A test $\mathcal{T}_\alpha$:=$\{d(\mathbf{X}),\ C_1(\alpha)\}$ is said to be *consistent* if its power goes to one for all discrepancies $\gamma=(\theta_0-\theta_1)\neq 0$, $\forall\theta_1\in\Theta_1$, however small, as $n\to\infty$. In symbols:

$$\lim_{n\to\infty}\pi_n(\theta)=\mathbb{P}(\mathbf{x}_0\in C_1;\ H_1(\theta)\text{ true})=1\text{ for all }\theta\in\Theta_1.$$

As in estimation, consistency is a *minimal* (necessary but not sufficient) property for a test. A 'decent' test should be capable to detect any discrepancy from the null when an infinite number of observations is available!

Returning to the original intentions by Neyman and Pearson (1933) to improve upon Fisher's significance testing, we can see that by bringing into the set up the notion of an *alternative hypothesis* defined as the compliment to the null, they defined an **optimal test** in terms of notion of an α significance level UMP test. The notion of optimality renders the choice of the test statistic and the associated rejection region a matter of mathematical optimization, replacing Fisher's intuition what test statistic makes sense to use in different cases. It turned out that in most cases Fisher's initial intuition coincided with the notion of an optimal test.

The main components of a N-P test are given in table 4.

Table 4: N-P testing - key elements

(i)	a statistical model: $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathfrak{X}$,
(ii)	a null (H_0) and an alternative (H_1) hypothesis within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$,
(iii)	a test statistic $d(\mathbf{X})$,
(iv)	the distribution of $d(\mathbf{X})$ under H_0 [ascertainable],
(v)	the significance level (or size) α,
(vi)	the rejection region $C_1(\alpha)=\{\mathbf{x}:\ d(\mathbf{x})>c_\alpha\}$,
(vii)	the distribution of $d(\mathbf{X})$ under H_1 [ascertainable].

Remarks. (i) It is very important to remember that a N-P test is not just a test statistic! It is at least a pair T_α:=$\{d(\mathbf{X}),\ C_1(\alpha)\}$ that defines a procedure with particular capacities, as calibrated by the error probabilities.

(ii) The optimality of an N-P test is inextricably bound up with the optimality of the estimator the test statistic is based on. Hence, it is no accident that most optimal N-P tests are based on consistent, fully efficient and sufficient estimators.

(iii) Also, by changing the rejection region one can render an optimal N-P test useless! For instance, replacing the rejection region of $\{d(\mathbf{X}),\ C_1(\alpha)\}$ with: $\overline{C}_1(\alpha)=\{\mathbf{x}:\ \varphi(\mathbf{x})<c_\alpha\}$, the resulting test $\overline{T}_\alpha$:=$\{d(\mathbf{X}),\ \overline{C}_1(\alpha)\}$ is practically useless because it is biased and its power decreases as the discrepancy γ increases.

3.2 Significance level α vs. the p-value

It is important to note that there is a mathematical relationship between the type I error probability (significance level) and the p-value. Placing them side by side in the case of example 1:

$$\begin{aligned} \mathbb{P}(\textbf{type I error}): \quad & \mathbb{P}(d(\mathbf{X})>c_\alpha; \mu=\mu_0)=\alpha, \\ p\textbf{-value}: \quad & \mathbb{P}(d(\mathbf{X})>d(\mathbf{x}_0); \mu=\mu_0)=p(\mathbf{x}_0), \end{aligned} \tag{24}$$

it becomes obvious that:

(a) they share the same test statistic $d(\mathbf{X})$ and are both evaluated using the tail of the sampling distribution under H_0, but

(b) differ in terms of their tail areas of interest: $\{\mathbf{x}\text{:}\ d(\mathbf{x})>c_\alpha\}$ vs. $\{\mathbf{x}\text{:}\ d(\mathbf{x})>d(\mathbf{x}_0)\}$, $\mathbf{x}\in\mathbb{R}^n_X$, rendering α a *pre-data* and $p(\mathbf{x}_0)$ a *post-data* error probability.

Example. Consider test $\mathcal{T}_\alpha$:=$\{d(\mathbf{X}),\ C_1(\alpha)\}$ for μ_0=10, σ=1, n=100, $\overline{x}_n$=10.2, α=.025 $\Rightarrow c_\alpha$=1.96 :

$$d(\mathbf{x}_0)=\tfrac{\sqrt{100}(10.2-10)}{1}=2.0 > c_\alpha, \quad \text{Reject } H_0.$$

The p-value is: $\mathbb{P}(d(\mathbf{X})>2.0;\ \mu=\mu_0)$=.023.

Their common features and differences bring out several issues.

First, the p-value can be viewed as the smallest significance level α at which H_0 would have been rejected with data $\mathbf{x}_0$. For that reason the p-value is often referred to as the *observed significance level*. For a qualification, see the fourth issue below.

Second, both the p-value and the type I and II error probabilities are NOT *conditional* on H_0 or H_1; the sampling distribution of $d(\mathbf{X})$ is evaluated under different hypothetical scenarios pertaining to the values of θ in Θ. Hence, neither the significance level α nor the p-value can be interpreted as probabilities *attached* to particular values of μ, associated with H_0 or H_1, since all error probabilities are *firmly attached* to the sample realizations $\mathbf{x}\in\mathbb{R}^n_X$.

Third, there is nothing irreconcilable between the significance level α and the p-value. α is a *pre-data* error probability defining the generic capacity of the test in question, and $p(\mathbf{x}_0)$ is a *post-data* error-probability evaluating the discordance between H_0 and data $\mathbf{x}_0$, based on the same generic capacity (power).

Fourth, **there is no such thing as a two-sided p-value**! The real difference between pre-data and post data error probabilities is that the latter use *additional information* in the form of the sign of $d(\mathbf{x}_0)$, which points out the direction of departure from H_0 indicated by data $\mathbf{x}_0$. This information renders one of the two tails irrelevant, and calls into question the concept of a *two-sided* p-value $p(\mathbf{x}_0)$.

Example 1 (continued). Consider testing the two-sided hypotheses:

$$H_0\text{:}\ \mu=\mu_0,\ \text{vs.}\ H_1\text{:}\ \mu\neq\mu_0, \tag{25}$$

using the test $T_\alpha^{\neq}$:=$\{d(\mathbf{X}), C_1^{\neq}(\alpha)\}$, where the rejection region is $C_1^{\neq}(\alpha)$=$\{\mathbf{x}$: $|d(\mathbf{x})| > c_{\frac{\alpha}{2}}\}$. Using μ_0=10, σ=1, n=100, $\overline{x}_n$=10.175, α=.05 $\Rightarrow$ $c_{\frac{\alpha}{2}}$=1.96.

$$d(\mathbf{x}_0)=\frac{\sqrt{100}(10.175-10)}{1}=1.75 < c_\alpha, \quad \text{Accept } H_0.$$

In light of the fact that $d(\mathbf{x}_0)$=1.75 > 0, the p-value is:

$$p(\mathbf{x}_0)=\mathbb{P}(d(\mathbf{X}) > d(\mathbf{x}_0); \mu=\mu_0) = .038,$$

which is smaller than α=.05.

Recall the **definition**: The p-value is the probability of all sample realizations $\mathbf{x}\in\mathfrak{X}$ such that $d(\mathbf{x})$ accords less well with H_0 than $\mathbf{x}_0$ does, when H_0 is true.

(a) ($\mathbf{x}$: $d(\mathbf{x}) > d(\mathbf{x}_0)$) defines the set of all $\mathbf{x}\in\mathfrak{X}$ such that $d(\mathbf{x})$ accords less well with H_0 than $\mathbf{x}_0$ does.

(b) ($\mathbf{x}$: $d(\mathbf{x}) < d(\mathbf{x}_0)$) defines the set of all $\mathbf{x}\in\mathfrak{X}$ such that $d(\mathbf{x})$ accords better with H_0 than $\mathbf{x}_0$ does.

In contrast, the clause "equal to or more extreme than $d(\mathbf{x}_0)$" can include $\{\mathbf{x}$: $|d(\mathbf{x})| \geq c_{\frac{\alpha}{2}}\}$!

3.3 Constructing optimal tests: Likelihood Ratio (LR) test

The likelihood ratio test procedure can be viewed as a generalization/extension of the *Neyman-Pearson lemma* to more realistic cases where the null and/or the alternative might be composite hypotheses. Its general formulation in the context of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ takes the following form.

(a) The hypotheses of interest are specified by: H_0: $\boldsymbol{\theta}\in\Theta_0$ vs. H_1: $\boldsymbol{\theta}\in\Theta_1$,

where Θ_0 and Θ_1 constitute a partition of Θ.

(b) The test statistic is a function that stems from the 'likelihood' ratio:

$$\lambda_n(\mathbf{x})=\frac{\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta};\mathbf{x})}{\max_{\boldsymbol{\theta}\in\Theta_0} L(\boldsymbol{\theta};\mathbf{x})}=\frac{L(\widehat{\boldsymbol{\theta}};\mathbf{x})}{L(\widetilde{\boldsymbol{\theta}};\mathbf{x})}, \tag{26}$$

when $\lambda_n(\mathbf{X})$ viewed as a random variable defined by $\lambda_n(.)$: $\mathbb{R}_X \longrightarrow [1,\infty)$.

Note that the max in the numerator is dervied over all values of $\boldsymbol{\theta}\in\Theta$ [yielding the MLE $\widehat{\boldsymbol{\theta}}$], but that of the max in the denominator is confined to all values under H_0: $\boldsymbol{\theta}\in\Theta_0$ [yielding the constrained MLE $\widetilde{\boldsymbol{\theta}}$].

CAUTION: the *likelihood ratio* $\lambda_n(\mathbf{X})$=$\frac{L(\widehat{\boldsymbol{\theta}};\mathbf{X})}{L(\widetilde{\boldsymbol{\theta}};\mathbf{X})}$ differs from the **likelihoodist ratio** $\ell(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\mathbf{x}_0)$=$\frac{L(\boldsymbol{\theta}_1;\mathbf{x}_0)}{L(\boldsymbol{\theta}_0;\mathbf{x}_0)}$ in one key respect: $\lambda_n(\mathbf{X})$ is a random variable with its own sampling distribution, but $\ell(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1;\mathbf{x}_0)$ denotes a numerical fraction of two likelihood values.

(c) The generic rejection region based on $\lambda_n(\mathbf{X})$ is defined by:

$$C_1=\{\mathbf{x}\text{: } \lambda_n(\mathbf{x}) > c\},$$

but it is rarely the case that the distribution of $\lambda_n(\mathbf{X})$ under H_0 is ascertainable. More often than not, one needs to use a transformation $h(.)$ to ensure that $h\left(\lambda_n(\mathbf{X})\right)$ has a *known* sampling distribution under H_0. This can then be used to define the rejection region:

$$C_1(\alpha)=\{\mathbf{x}\text{: } d(\mathbf{X})=h\left(\lambda_n(\mathbf{X})\right)>c_\alpha\}. \tag{27}$$

In terms of constructing optimal tests, the LR procedure can be shown to yield several well-known optimal tests; Lehmann and Romano (2005).

Example 1 (continued). Consider the hypotheses:

$$H_0\text{: } \mu\leq\mu_0 \quad \text{vs.} \quad H_1\text{: } \mu>\mu_0, \tag{28}$$

in the context of the simple Normal model, whose likelihood function is:

$$L(\boldsymbol{\theta};\mathbf{x})=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^n(x_i-\mu)^2\right\}.$$

Note that $\max\limits_{\boldsymbol{\theta}\in\Theta}L(\theta;\mathbf{x})=L(\widehat{\theta};\mathbf{X})$ where $\widehat{\theta}=\overline{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$ is the ML estimator of μ, yielding:

$$L(\widehat{\boldsymbol{\theta}};\mathbf{x})=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^n(x_i-\overline{x}_n)^2\right\}. \tag{29}$$

On the other hand, $\max\limits_{\boldsymbol{\theta}\in\Theta_0}L(\theta;\mathbf{x})=\max\limits_{\mu\leq\mu_0}L(\theta;\mathbf{x})=L(\widetilde{\theta};\mathbf{X})$ where:

$$\widetilde{\theta}=\begin{cases}\mu_0 & \text{if } \overline{x}_n\geq\mu_0\\ \overline{x}_n & \text{if } \overline{x}_n<\mu_0\end{cases},$$

and thus, under H_0: $\mu\leq\mu_0$,the LF is:

$$L(\widetilde{\boldsymbol{\theta}};\mathbf{x})=\begin{cases}(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu_0)^2\right\} & \text{for } \overline{x}_n\geq\mu_0,\\ (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\overline{x}_n)^2\right\} & \text{for } \overline{x}_n<\mu_0.\end{cases}$$

The case $\overline{x}_n<\mu_0$ as irrelevant since there is no test: $L(\widehat{\boldsymbol{\theta}};\mathbf{x})=L(\widetilde{\boldsymbol{\theta}};\mathbf{x})\to\lambda_n(\mathbf{X})=1$.

On the other hand, for $\overline{x}_n\geq\mu_0$ the likelihood ratio yields a statistic:

$$\begin{aligned}\lambda_n(\mathbf{X}) &=\tfrac{L(\widehat{\boldsymbol{\theta}};\mathbf{x})}{L(\widetilde{\boldsymbol{\theta}};\mathbf{x})}=\exp\left\{-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^n(x_i-\overline{x}_n)^2+\tfrac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu_0)^2\right\}=\\ &=\exp\left\{\tfrac{n}{2\sigma^2}\left(\overline{x}_n-\mu_0\right)^2\right\}\to 2\ln\lambda_n(\mathbf{X})=\tfrac{n}{\sigma^2}\left(\overline{x}_n-\mu_0\right)^2.\end{aligned}$$

Since $\frac{n}{\sigma^2}\left(\overline{x}_n-\mu_0\right)^2>c\to|\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}|>c_\alpha$, the rejection region $C_1=\{\mathbf{x}\text{: } 2\ln\lambda_n(\mathbf{x})>c\}$, is equivalent to the N-P test:

$$\{d(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma},\ C_1(\alpha)=\{\mathbf{x}\text{: } \tau(\mathbf{X})>c_\alpha\},$$

which can be shown to be UMP.

Asymptotic Likelihood Ratio Test. One of the most crucial advantages of the likelihood ratio test in practice is that even when one cannot find a transformation

$h(.)$ of $\lambda_n(\mathbf{X})$ that will yield a test statistic whose finite sample distributions are known, one can use the asymptotic distribution. Wilks (1938) proved that *under certain restrictions*:

$$2\ln\lambda_n(\mathbf{X})=2\left(\ln L(\widehat{\boldsymbol{\theta}};\mathbf{X})-\ln L(\widetilde{\boldsymbol{\theta}};\mathbf{X})\right)\underset{n\to\infty}{\overset{H_0}{\backsim}}\chi^2(r),$$

where $\overset{H_0}{\underset{\alpha}{\sim}}$ reads "under H_0 is asymptotically distributed as" and r denotes *the number of restrictions* involved in defining Θ_0. That is, under certain regularity restrictions on the underlying statistical model, when $\mathbf{X}$ is an IID sample, the asymptotic distribution (as $n\to\infty$) of $2\ln\lambda_n(\mathbf{X})$ is chi-square with as many degrees of freedom as there are restrictions, irrespective of the distributional assumption. This result can be used to define the asymptotic likelihood ratio test:

$$\{2\ln\lambda_n(\mathbf{X}),\ C_1(\alpha)=\{\mathbf{x}:\ 2\ln\lambda_n(\mathbf{x})>c_\alpha\},\ \textstyle\int_{c_\alpha}^{\infty}\psi(x)dx=\alpha.$$

4 Summary and conclusions

In frequentist inference **learning from data $\mathbf{x}_0$** about the stochastic phenomenon of interest is accomplished by applying *optimal* inference procedures with ascertainable error probabilities in the context of a statistical model:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X. \tag{30}$$

Hypothesis testing gives rise to learning from data $\mathbf{x}_0$ by partitioning $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ into two subsets framed in terms of the parameter(s):

$$H_0\text{: }\theta\in\Theta_0,\quad \text{vs.}\quad H_0\text{: }\theta\in\Theta_1, \tag{31}$$

and use $\mathbf{x}_0$ to ask (using hypothetical reasoning) whether $\boldsymbol{\theta}^*\in\Theta_0$ or $\boldsymbol{\theta}^*\in\Theta_1$?

A test $T_\alpha:=\{d(\mathbf{X}),\ C_1(\alpha)\}$ is defined in terms of a test statistic $(d(\mathbf{X}))$ and a rejection region $(C_1(\alpha))$, and its optimality is calibrated in terms of the relevant **error probabilities**:

$$\textbf{type I:}\quad \mathbb{P}(\mathbf{x}_0\in C_1;\ H_0(\theta)\text{ true})\leq\alpha(\theta),\text{ for }\theta\in\Theta_0,$$
$$\textbf{type II:}\quad \mathbb{P}(\mathbf{x}_0\in C_0;\ H_1(\theta_1)\text{ true})=\beta(\theta_1),\text{ for }\theta_1\in\Theta_1.$$

These error probabilities specify how often these procedures lead to erroneous inferences. For a given significance level α the optimal N-P test is the one whose pre-data capacity (power):

$$\mathcal{P}(\theta_1)=\mathbb{P}(\mathbf{x}_0\in C_1;\ \theta=\theta_1),\ \text{for all }\theta_1\in\Theta_1,$$

is equal or greater than any other α-significance level test for all $\theta_1\in\Theta_1$; UMP.

The existence of a UMP test depends crucially on the prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, and the framing of the hypotheses. It is not as rare as some statistics textbooks would have us believe; see Lehmann and Romano (2005).