Summer Seminar: Philosophy of Statistics
Lecture Notes 7: Hypothesis Testing

Aris Spanos [Summer 2019]

1 Introduction

1.1 Difficulties in mastering statistical testing

Statistical testing is arguably the most difficult, confusing and confused chapter in statistical inference for a variety of reasons, including the following.

Inherent difficulties. (a) There is a need to introduce **numerous new notions, concepts and procedures** before one can articulate a coherent picture of what constitutes a statistical test. This makes it very easy for readers to miss the forest for the trees and focus on the superficial elements of testing, such as *the formulae and the probability tables*.

(b) Testing is **conceptually sophisticated** due to the nature and role of the underlying *hypothetical reasoning* that is often insufficiently appreciated.

(c) The **framing of frequentist testing** bequeathed by the three pioneers, Fisher, Neyman and Pearson **was incomplete** in the sense that there was no coherent evidential account pertaining to the results beyond p-values and accept/reject the null rules; see Mayo (1996). As a result, the subsequent discussions of frequentist testing has been beleaguered by **serious foundational problems**. Worse, different applied fields have generated their own secondary literatures attempting to address these foundational problems, but often making matters worse! Indeed, in some fields it has reached a vicious circle where one needs to correct the 'corrections' of those chastising the correctors of the initial problems and their suggested solutions.

Extraneous difficulties. (d) The Fisherian and Neyman-Pearsonian (N-P) variants of frequentist testing are **often misrepresented using oversimplified framings** that neglect the key role of the prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. In addition, the two variants are often presented as essentially incompatible and any attempt to blend them will result in an **inconsistent hybrid** which is "burdened with conceptual confusion" (Gigerenzer 1993, p. 323). Worse, the N-P testing is often viewed as **decision-theoretic** in nature, when in fact the latter distorts the underlying reasoning and misrepresents frequentist testing; see Spanos (2017b).

(e) Statistics spreads across numerous disciplines, providing the framework for empirical modeling and inference, and textbook writers in different disciplines often rely on second hand accounts of past textbooks in their discipline. The resulting narratives often ignore key components of inference, including the underlying statistical model and the reasoning associated with different inference procedures. As a result, their discussion often amounts to an **idiot's guide to statistics** that combins off-the-shelf formulae with statistical tables to yield tabular asterisks ***, **, *.

(f) Unfortunately, there is an crusade by some **Bayesian authors to distort and cannibalize frequentist testing** in a misguided attempt to motivate their own preferred viewpoint on statistical inference. Manifest signs of such a distorted preaching include (i) undue emphasis on 'long-run' frequencies as error probabilities, (ii) loss functions as indispensable tools, (iii) admissibility as a minimal property of estimators, as well as cannibalization of frequetist concepts, such as (iv) using non-regular statistical models, (v) misinterpreting 'error probabilities' as being conditional on θ, (vi) distorting the p-value to resemble posterior probabilities assigned to the null, (vii) misintepreting confidence intervals as Bayesian credible intervals, and (viii) assigning prior probabilities to 'true' nulls and alternatives (whatever that means); reader be aware!

2 Statistical testing before R.A. Fisher

2.1 Francis Edgeworth's testing

Francis Ysidro Edgeworth (Drummond Chair of Political Economy at Oxford) was an economist who made significant contributions to the methods of statistics during the 1880s; see Stigler (1986), Gorroochurn (2016). A typical example of a testing procedure at the end of the 19th century is given by Edgeworth (1885). Viewing his testing procedure *retrospectively* from Fisher's perspective, Edgeworth assumes that the data $\mathbf{x}_0$:=$(x_{11}, x_{12}, \ldots, x_{1n}; x_{21}, x_{22}, ..., x_{2n})$ constitute a realization of the the $2n$-dimensional random (IID) sample $\mathbf{X}$:=$(X_{11}, X_{12}, \ldots, X_{1n}; X_{21}, X_{22}, ..., X_{2n})$ from the *simple (IID) bivariate Normal model*:

$$\mathbf{X}_t \backsim \mathsf{NIID}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ t=1,2,...,n,..., \tag{1}$$

$$\mathbf{X}_t := \begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix},\ \boldsymbol{\mu} := \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},\ \boldsymbol{\Sigma} := \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \tag{2}$$

$$E(X_{1t})=\mu_1,\ E(X_{2t})=\mu_2,\ Var(X_{1t})=Var(X_{2t})=\sigma^2,\ Cov(X_{1t},X_{2t})=0,$$

The *hypothesis of interest* relates to the equality of the two means: (a) $\mu_1=\mu_2$.

Common sense, combined with the statistical knowledge at the time, suggested using the difference between the estimated means:

$$\widehat{\mu}_1-\widehat{\mu}_2, \quad \text{where} \quad \widehat{\mu}_1=\tfrac{1}{n}\textstyle\sum_{t=1}^{n} X_{1t},\ \widehat{\mu}_2=\tfrac{1}{n}\sum_{t=1}^{n} X_{2t},$$

as a basis for deciding whether $\mu_1=\mu_2$. To render the difference $(\widehat{\mu}_1-\widehat{\mu}_2)$ free of the units of measurement, Edgeworth divided it by $\sqrt{Var(\widehat{\mu}_1-\widehat{\mu}_2)}=\sqrt{(\widehat{\sigma}_1^2+\widehat{\sigma}_1^2)}$, where $\widehat{\sigma}_1^2=\frac{1}{n}\sum_{t=1}^{n}(X_{1t}-\widehat{\mu}_1)^2$, $\widehat{\sigma}_2^2=\frac{1}{n}\sum_{t=1}^{n}(X_{2t}-\widehat{\mu}_1)^2$, to define (b) a *distance function*:

$$\xi(\mathbf{X})=\tfrac{|\hat{\mu}_1-\hat{\mu}_2|}{\sqrt{(\hat{\sigma}_1^2+\hat{\sigma}_1^2)}}.$$

To decide whether the observed distance $\xi(\mathbf{x}_0)$ is 'large enough' to infer $\mu_1 \neq \mu_2$, Edgeworth used (c) a *threshold* $2\sqrt{2}$. His argument is that rejection of $(\mu_1-\mu_2)=0$ could *not* have been 'due to pure chance' (accidental) if:

$$\xi(\mathbf{x}_0) > 2\sqrt{2}. \tag{3}$$

Where did the threshold $2\sqrt{2}$ come from? It came from the tail area of $\mathsf{N}(0,1)$ beyond $\pm 2\sqrt{2}$ which is approximately .005. This value was viewed at the time as a 'reasonable' lower bound for the probability of a 'chance' error, i.e. the probability of erroneously inferring a significant discrepancy. But why Normality? At the time, statistical inference did not have the notion of a statistical model (introduced by Fisher 1922a), and thus it relied heavily on large sample size n (asymptotic) results by routinely invoking the Central Limit Theorem. For a detailed discussion of testing the difference between two means in a modern set up see Appendix 13.A.

In summary, *Edgeworth* introduced 3 generic concepts for statistical testing (table 13.1) that have been retained, after being modified by the subsequent literature.

Table 13.1: Edgeworth's testing- key concepts

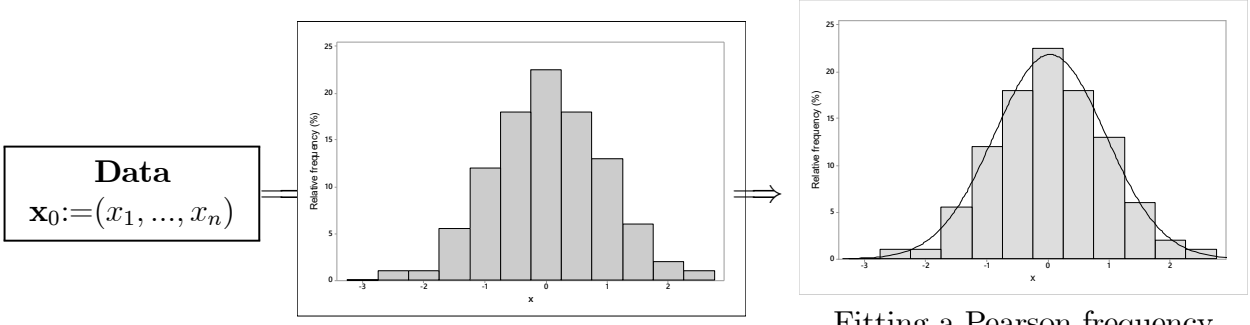
(a) a *hypothesis of interest*: $\mu_1=\mu_2$,

(b) the notion of a *standardized distance*: $\xi(\mathbf{X})$,

(c) a *threshold value* for 'significance': $\xi(\mathbf{x}_0) > 2\sqrt{2}$.

2.2 Karl Pearson's testing

Karl Pearson's approach to statistics was discussed in chapter 12 in some detail. In summary, one would begin with the data $\mathbf{x}_0$ in search of a descriptive model within the Pearson family.



Histogram of the data

Fitting a Pearson frequency curve $f(x;\widehat{\theta}_1,\widehat{\theta}_2,\widehat{\theta}_3,\widehat{\theta}_4)$

Fig. 10.3: The Karl Pearson approach to statistics

Having selected a frequency curve $f_0(x)$ on the basis of $f(x;\widehat{\theta}_1,\widehat{\theta}_2,\widehat{\theta}_3,\widehat{\theta}_4)=f(x;\widehat{\boldsymbol{\theta}})$, Karl Pearson would proceed to assess the appropriateness of $f_0(x)$ by testing the

hypothesis of interest:

$$f_0(x)=f^*(x)\in \mathsf{Pearson}(\theta_1,\theta_2,\theta_3,\theta_4),$$

where $f^*(x)$ denotes the 'true density' in the sense that it could have generated data $\mathbf{x}_0$. To construct a test Pearson proposed the standardized *distance function*:

$$\eta(\mathbf{X})=\sum_{i=1}^{m}\frac{(\widehat{f}_i-f_i)^2}{f_i}=n\sum_{i=1}^{m}\frac{((\widehat{f}_i/n)-(f_i/n))^2}{(f_i/n)}\underset{n\to\infty}{\backsim}\chi^2(m), \tag{4}$$

where $(\widehat{f}_i,\ i=1,2,...,m)$ and $(f_i,\ i=1,2,...,m)$ denote the empirical and assumed [as specified by $f_0(x)$] frequencies. This test statistic compares how close the *observed* are to the *expected* (under $f_0(x)$) relative frequencies. Instead of relying on a simple threshold for the observed distance, like Edgeworth, he went a step further and introduced a primitive version of the *p-value*:

$$\mathbb{P}(\eta(\mathbf{X})>\eta(\mathbf{x}_0))=p(\mathbf{x}_0), \tag{5}$$

as a basis for inferring whether the choice of $f_0(x)$ was adequate or not. The rationale for the p-value was simply based on goodness-of-fit: the smaller the p-value the worse the fit.

Example 13.1. Mendel's cross-breeding experiments (1865-6) were based on *pea-plants* with different *shapes* (round or wrinkled) and colors (yellow or green), which can be framed in terms of two Bernoulli random variables:

$X(\overbrace{\text{round}}^{\text{R}})=0,\ X(\overbrace{\text{wrinkled}}^{\text{W}})=1,\ Y(\overbrace{\text{yellow}}^{\text{Y}})=0, Y(\overbrace{\text{green}}^{\text{G}})=1.$

Mendel's *theory of heredity* was based on two assumptions:

(i) the two random variables X and Y are independent, and

(ii) round (X=0) and yellow (Y=0) are *dominant* traits, but 'wrinkled' (X=1) and 'green' (Y=1) are *recessive* traits with probabilities:

$$\text{dominant: } \mathbb{P}(X=0)=\mathbb{P}(Y=0)=.75,\quad \text{recessive: } \mathbb{P}(X=1)=\mathbb{P}(Y=1)=.25$$

Mendel's *substantive theory* gives rise to the following probabilistic model based on the bivariate Bernoulli distribution below.

Table 13.2: Mendel's model

$x \setminus y$	0	1	$f_x(x)$
0	.5625	.1875	.750
1	.1875	.0625	.250
$f_y(y)$	.750	.250	1.00

Data. The 4 gene-pair experiments Mendel carried out with n=556 gave rise the observed frequencies and relative frequencies given below:

R,Y $(0,0)$	R,G $(0,1)$	W,Y $(1,0)$	W,G $(1,1)$
315	108	101	32

Observed frequencies

$\Rightarrow$

$x \setminus y$	0	1	$\widehat{f}_x(x)$
0	.5666	.1942	.7608
1	.1817	.0576	.2393
$\widehat{f}_y(y)$	.7483	.2518	1.001

Observed relative frequencies

▶ **How adequate is Mendel's theory in light of the data**? Using Pearson's goodness-of-fit chi-square test statistic in (4) yields:

$$\eta(\mathbf{x}_0)=556\left(\tfrac{(.5666-.5625)^2}{.5625}+\tfrac{(.1942-.1875)^2}{.1875}+\tfrac{(.1817-.1875)^2}{.1875}+\tfrac{(.0576-.0625)^2}{.0625}\right)=.463.$$

Given that the tail area of $\chi^2(3)$ is $\mathbb{P}(\eta(\mathbf{X})>.463)=.927$, the p-value indicates excellent goodness-of-fit (no discordance) for Mendel's theory.

Table 13.3: Karl Pearson's testing – key concepts

(a)	Introducing the *Pearson family of distributions* that extended significantly the scope of statistical modeling.
(b)	*Broadening of the scope* of the hypothesis of interest, initiating *Mis-Specification (M-S) testing,* by introducing a goodness-of-fit test in terms of $f_0(x)$ and $f^*(x)$.
(c)	Introducing the notion of a *distance function* whose distribution is asymptotically known (as $n \to \infty$).
(d)	The use of the *tail probability* (p-value) as a basis for evaluating the *goodness-of-fit* of $f_0(x)$ with data $\mathbf{x}_0$.

Karl Pearson's main contributions to statistical testing are listed in table 13.3. As argued next, these features were subsequently re-framed and modified by Fisher.

■ It is very important to highlight the fact that, when viewed from a modern perspective, the hypothesis of interest $f_0(x)=f^*(x)\in\mathsf{Pearson}(\theta_1,\theta_2,\theta_3,\theta_4)$, pertains to the adequacy of the choice of the probability model $f_0(x)$. That is, Pearson's chi-square was the first *misspecification test* for a distribution assumption!

3 Fisher's significance testing

R. A. Fisher founded modern statistics while he was working as a statistician at the *Rothamsted (Agricultural) Experimental Station* in Harpenden (25 miles North of London) from 1919-1933. He was appointed professor of Eugenics at University College, London in 1933, and moved to Cambridge University in 1943 as the Balfour Professor of Genetics. In addition to being the father of modern statistics, he was also a key founder of modern human genetics.

Table 13.4: The simple Normal model

Statistical GM:		$X_t=\mu+u_t,\ t\in\mathbb{N}$,
[1]	Normal:	$X_t \backsim \mathsf{N}(.,.)$,
[2]	Constant mean:	$E(X_t)=\mu$, for all $t\in\mathbb{N}$,
[3]	Constant variance:	$Var(X_t)=\sigma^2$, for all $t\in\mathbb{N}$,
[4]	Independence:	$\{X_t,\ t\in\mathbb{N}\}$ - independent process.

Fisher's recasting of statistical testing was inspired by 'Student' (1908), a paper written by William Gosset, that we encountered in chapter 11. Gosset (1908) introduced the result that for a simple Normal model (table 13.4):

$$\frac{\sqrt{n}(\overline{X}_n-\mu)}{s} \backsim \mathsf{St}(n-1), \tag{6}$$

where $\mathsf{St}(n-1)$ denotes a Student's t distribution with $(n-1)$ degrees of freedom. The crucial importance of this paper stems from the fact that (6) was the *first finite sampling distribution* [valid for any $n>1$] that inspired Fisher to recast statistics into its modern framing. The sampling distribution in (6) inspired Fisher to recast modern statistics by introducing several crucial innovations to statistics.

(a) Fisher (1915) introduced the mathematical framework for formally deriving the finite sampling distribution of the correlation coefficient. In subsequent papers he derived several additional sampling distributions, including those for testing the significance of the regression coefficients as well as partial correlations.

(b) He brought out explicitly the probabilistic assumptions invoked in deriving (6) in the form of the simple Normal (table 13.4):

$$X_k \backsim \mathsf{NIID}(\mu,\sigma^2),\ k=1,2,...,n,... \tag{7}$$

This move introduced the concept of a *statistical model* whose generic form is:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X, \tag{8}$$

where $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}^n_X$ denotes the (joint) distribution of the sample $\mathbf{X}:=(X_1,...,X_n)$ that encapsulates the prespecified probabilistic structure of the underlying stochastic

process $\{X_t,\ t\in\mathbb{N}\}$. The link to the phenomenon of interest comes from viewing data $\mathbf{x}_0{:=}(x_1,x_2,\ldots,x_n)$ as a 'truly typical' realization of the process $\{X_k,\ k\in\mathbb{N}\}$. In addition to the Pearson family (Pearson, 1894), Fisher later enlarged the scope of statistical modeling and inference significantly by introducing the *exponential* and the *transformation* families of distributions; see Lehmann and Romano (2005).

(c) Fisher used the result (6) to construct a *test of significance* for the:

$$\textsf{Null hypothesis:}\quad H_0\text{: } \mu{=}\mu_0, \tag{9}$$

in the context of the simple Normal model (table 13.4) by unlocking the reasoning underlying the estimation result in (6) and modifying it for testing purposes. Let us unpack this claim in detail.

Testing reasoning. What do these sampling distribution claims really mean?

$$d(\mathbf{X};\mu,\sigma^2){=}\tfrac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma}\backsim \mathsf{N}(0,1), \tag{10}$$

$$\tau(\mathbf{X};\mu){=}\tfrac{\sqrt{n}(\overline{X}_n-\mu)}{s}\backsim \mathsf{St}(0,1;n{-}1). \tag{11}$$

Since neither $d(\mathbf{X};\mu,\sigma^2)$ nor $\tau(\mathbf{X};\mu)$ constitute statistics (function of $\mathbf{X}$), how are we to interpret such claims? Fisher (1934) called $d(\mathbf{X};\mu,\sigma^2)$ and $\tau(\mathbf{X};\mu)$ a *pivotal quantities* (or pivots) to stand for a function of both $\mathbf{X}$ and unknown parameters, but their sampling distribution distributions are free of unknown parameters. What is not so obvious is why the unknown parameters disappear from the sampling distributions.

Focusing on (11), the claim is that $E\left[\tau(\mathbf{X};\mu)\right]{=}0$ and $Var\left[\tau(\mathbf{X};\mu)\right]{=}\frac{n-1}{n-3}$. A moment's reflection suggests that $E\left(\frac{\sqrt{n}(\overline{X}_n-\mu)}{s}\right){=}0$ only when $E(\overline{X}_n){=}\mu^*$, i.e. $\overline{X}_n$ is an unbiased estimator of μ, where μ^* denotes the true value of μ. That is, when $\tau(\mathbf{X};\mu)$ is evaluated under $\mu{=}\mu^*$, i.e. using **factual reasoning**:

$$\tau(\mathbf{X};\mu){=}\tfrac{\sqrt{n}(\overline{X}_n-\mu)}{s}\overset{\mu=\mu^*}{\backsim} \mathsf{St}(0,1;n{-}1). \tag{12}$$

Fisher's recasting of frequentist testing was based on transforming (12) into a **test statistic**, by replacing the unknown μ with a hypothesized value μ_0 (H_0), in conjunction with replacing the factual reasoning ($\mu{=}\mu^*$) with **hypothetical**, under $\mu{=}\mu_0$:

$$\tau(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s}\overset{\mu=\mu_0}{\backsim} \mathsf{St}(n{-}1). \tag{13}$$

Note that when $\tau(\mathbf{X})$ is evaluated under $\mu{=}\mu^*$ the result:

$$\tau(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s}\overset{\mu=\mu^*}{\backsim} \mathsf{St}(\delta;n{-}1),\ \delta{=}\tfrac{\sqrt{n}(\mu^*-\mu_0)}{\sigma}, \tag{14}$$

is non-operational since $\mathsf{St}(\delta;n{-}1)$, a non-central Student's t distribution with non-centrality parameter δ, depends on μ^*. But when hypothetical reasoning is extended to another value of μ that belongs to the relevant parameter space, say $\mu_1{\neq}\mu_0$:

$$\tau(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s}\overset{\mu=\mu_1}{\backsim} \mathsf{St}(\delta;n{-}1),\ \delta_1{=}\tfrac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}, \tag{15}$$

this is rendered operational since μ_1, μ_0 are hypothetical values. This plays an important role in the Neyman-Pearson framing.

Indeed, Fisher was very explicit about the *nature of the reasoning* underlying significance testing: "In general, tests of significance are based on hypothetical probabilities calculated from their null hypotheses." (Fisher (1956), p. 47).

(d) Fisher went on to modify/extend Pearson's definition of the *p-value* from a goodness-of-fit indicator into the basic criterion for his finite sample *significance testing*. The key was the realization that the sampling distribution of $\tau(\mathbf{X})$, when evaluated *under* H_0, involves no unknown parameters and yields the tail area:

$$\mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0);\ H_0)=p(\mathbf{x}_0). \tag{16}$$

Fisher reinterpreted the p-value as an *indicator of discordance (contradiction) between data* $\mathbf{x}_0$ and H_0. Formally, the p-value is the probability associated with all possible outcomes $\mathbf{x}\in\mathbb{R}^n_X$ whose test statistic value $d(\mathbf{x})$ is more discordant with H_0 than $d(\mathbf{x}_0)$ is, when evaluated under H_0.

To transform the p-value $p(\mathbf{x}_0)$ into an inference pertaining to the 'significance' of H_0 one needs to choose a *threshold* for deciding how small the p-value is 'small enough' to falsify H_0. Fisher suggested several such thresholds, .01, .02, .05, but left the choice to be made on a case by case basis.

As in the case of frequentist estimation, the **primary objective of frequentist testing** is to learn about the 'true' ($\boldsymbol{\theta}=\boldsymbol{\theta}^*$) statistical Data Generating Mechanism (DGM):

$$\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}^*)\},\ \mathbf{x}\in\mathbb{R}^n_X,$$

assumed to have generated data $\mathbf{x}_0$. In light of that, frequentist testing relies on a good estimator of $\boldsymbol{\theta}$ to learn from data about $\boldsymbol{\theta}^*$. In the case of the simple Normal model, $\overline{X}_n$ is the best at pinpointing μ^*, and thus the difference $\left(\overline{X}_n-\mu_0\right)$, defining the test statistic $\tau(\mathbf{X})$, aims to appraise the standardized distance $(\mu^*-\mu_0)$; recall that μ^* has generated the data for the evaluation of $\overline{X}_n$. In contrast to estimation where there is a single factual scenario $\mu=\mu^*$, hypothesis testing relies on hypothetical reasoning and thus one can pose numerous questions to the data based on such hypothetical scenarios using any values $\mu\in\mathbb{R}$.

3.1 A closer look at Fisher's p-value

In light of the above discussion, the p-value $\mathbb{P}(\tau(\mathbf{X})>\tau(\mathbf{x}_0);\ H_0)=p(\mathbf{x}_0)$ aims to evaluate the discordance arising from the difference $(\mu^*-\mu_0)$, or equivalently between:

$$\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x};(\mu^*,\sigma^2)\} \text{ and } \mathcal{M}_0(\mathbf{x})=\{f(\mathbf{x};(\mu_0,\sigma^2)\},\ \mathbf{x}\in\mathbb{R}^n_X,$$

when the sampling distribution of $\tau(\mathbf{X})$ is evaluated under the **hypothetical scenario** $\mu=\mu_0$. In this sense, the p-value is firmly attached to the testing procedure's

capacity to detect discrepancies from the null and cannot (or should not) legitimately be interpreted as the probability assigned to any value of μ in $\mathbb{R}$.

Although it is impossible to pin down Fisher's interpretation of the p-value because his views changed over time and held several renderings at any one time, there is one fixed point among his numerous articulations (Fisher, 1925a, 1955): 'a small p-value can be interpreted as a simple logical disjunction: either an extremely rare event has occurred or H_0 is not true'.

The focus on "a small p-value" needs to be viewed in conjunction with Fisher's *falsificationist stance* about testing in the sense that significance tests can *falsify* but never *verify* hypotheses (Fisher, 1955): "... tests of significance, when used accurately, are capable of rejecting or invalidating hypotheses, in so far as these are contradicted by the data; but that they are never capable of establishing them as certainly true."
Having quoted the above strict falsificationist stance, it is interesting to quote a less stringent one from Fisher (1925): "If p is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis failed to account for the whole of the facts." (p. 80).
The combination of Fisher's logical disjunction, combined with his falsificationist stance, could be interpreted as arguing that a p-value smaller than the designated threshold indicates that H_0 is false in the sense that:

$$\mathcal{M}_0(\mathbf{x})=\{f(\mathbf{x};(\mu_0,\sigma^2)\}\subset\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ \mathbf{x}\in\mathbb{R}^n_X,$$

could not have generated data $\mathbf{x}_0$. What is less apparent is in what sense does this provide evidence against H_0. Fisher did *not* articulate a convincing evidential interpretation that the p-value 'indicates the strength of evidence against the null', despite expressing that aspiration on different occasions:

"The actual value of p ... indicates the strength of evidence against the hypothesis" (Fisher, 1925a, p. 80).

Unfortunately, the p-value could not provide a cogent evidential interpretation for 'rejecting H_0' due primarily to a crucial weakness, known as the large n problem, initially raised by Berkson (1938).

The large n problem. Using Pearson's chi-square test based on:

$$\eta(\mathbf{X})=n\sum_{i=1}^{m}\frac{((\widehat{f}_i/n)-(f_i/n))^2}{(f_i/n)}\underset{n\to\infty}{\backsim}\chi^2(m),$$

Berkson (1938) argued that since $\eta(\mathbf{X})$ usually increases with n, the p-value decreases as n increases. Hence, there is always a large enough n to reject any null hypothesis. His claim needs to be qualified by attaching the clause: when $(f_0(x)-f^*(x))\neq 0$, irrespective of how small this difference is. Taken at face value, Berkson's claim means that a rejection of H_0 based on $p(\mathbf{x}_0)$=.03 and n=50, does not have the same evidential weight for the falsity of H_0 as a rejection with $p(\mathbf{x}_0)$=.03 and n=20000; see Spanos (2014a).

3.2 Significance testing: empirical example

Example 13.4. Arbuthnot's 1710 conjecture: the ratio of males to females in newborns might not be 'fair'. This can be tested using the statistical null hypothesis:

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

in the context of *the simple Bernoulli model* (table 13.8), based on the random variable X defined by: $\{\text{male}\}{=}\{X{=}1\}$, $\{\text{female}\}{=}\{X{=}0\}$.

Table 13.8: The simple Bernoulli model

Statistical GM:	$X_t{=}\theta+u_t,\ t{\in}\mathbb{N}{:=}(1,2,...,n,...)$	
[1]	Bernoulli:	$X_t \backsim \mathsf{Ber}(.,.),\ x_t{=}\{0,1\},$
[2]	Constant mean:	$E(X_t){=}\theta,\ 0{\leq}\theta{\leq}1,$ for all $t{\in}\mathbb{N},$
[3]	Constant variance:	$Var(X_t){=}\theta(1{-}\theta),$ for all $t{\in}\mathbb{N},$
[4]	Independence:	$\{X_t,\ t{\in}\mathbb{N}\}$ - independent process.

Using the MLE of θ, $\hat{\theta}_{ML}{:=}\overline{X}_n{=}\frac{1}{n}\sum_{i=1}^{n}X_i$, where: $\overline{X}_n \backsim \mathsf{Bin}\left(\theta, \frac{\theta(1-\theta)}{n}; n\right)$, one can derive the *test statistic*:

$$d(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n{-}\theta_0)}{\sqrt{\theta_0(1{-}\theta_0)}} \overset{H_0}{\backsim} \mathsf{Bin}\left(0,1;n\right). \tag{17}$$

Note that $\sum_{i=1}^{n}X_i \backsim \mathsf{Bin}\left(n\theta, n\theta(1{-}\theta); n\right)$ can be approximated very accurately using a $\mathsf{N}(n\theta, n\theta(1{-}\theta))$ distribution for $n > 20$, as figure 13.5 attests.

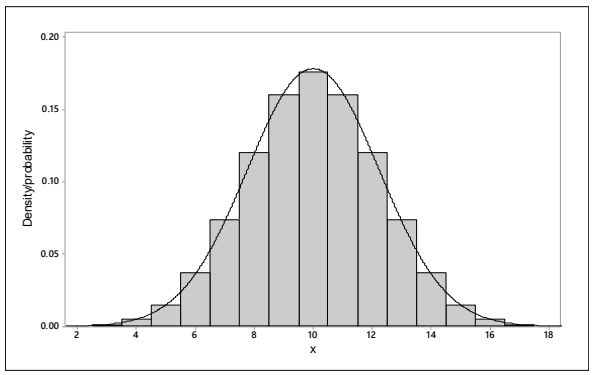


Fig. 13.5: Normal approximation of the Binomial: $f(y;\theta{=}.5, n{=}20)$

Data: $n{=}30862$ newborns during the period 1993-5 in Cyprus, out of which 16029 were boys and 14833 girls. The test statistic takes the form (17) and $\hat{\theta}_n{=}\frac{16029}{30862} = .521$.

$$d(\mathbf{x}_0){=}\frac{\sqrt{30862}(.521{-}.5)}{\sqrt{.5(.5)}}{=}7.366, \quad \mathbb{P}(d(\mathbf{X}){>}7.366;\theta{=}.5){=}.00000017.$$

The tiny p-value indicates strong discordance with H_0.

3.3 Summary of Fisher's significance testing

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

Table 13.11: Fisher's testing - key elements

(a)	a prespecified statistical model: $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$
(b)	a null (H_0: θ=θ_0) hypothesis,
(c)	a test statistic (distance function) $\tau(\mathbf{X})$,
(d)	the distribution of $\tau(\mathbf{X})$ under H_0 is known,
(e)	the p-value $\mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); H_0)$=$p(\mathbf{x}_0)$,
(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.

Example 13.6. Consider a simple (one parameter) Normal model in table 13.12.

Table 13.12: The simple (one parameter) **Normal model**

Statistical GM:		X_t=$\mu + u_t$, $t\in\mathbb{N}$:=$(1, 2, ..., n, ...)$
[1]	Normal:	$X_t \backsim \mathsf{N}(.,.)$, $x_t\in\mathbb{R}$,
[2]	Constant mean:	$E(X_t)$=μ, $\mu\in\mathbb{R}$, $\forall t\in\mathbb{N}$,
[3]	Constant variance:	$Var(X_t)$=σ^2, σ^2-known, $\forall t\in\mathbb{N}$,
[4]	Independence:	$\{X_t,\ t\in\mathbb{N}\}$-independent process.

(a) $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: $X_t \backsim \mathsf{NIID}(\mu, \sigma^2)$ [σ^2 is *known*], $t\in\mathbb{N}$.

(b) Null hypothesis: H_0: μ=μ_0 (e.g. μ_0=0),

(c) a test statistic: $\kappa(\mathbf{X})$=$\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}$,

(d) the distribution of $\kappa(\mathbf{X})$ under H_0: $\kappa(\mathbf{X}) \overset{H_0}{\backsim} \mathsf{N}(0,1)$,

(e) the p-value: $\mathbb{P}(\kappa(\mathbf{X}) > \kappa(\mathbf{x}_0); H_0)$=$p(\mathbf{x}_0)$,

(f) a threshold value for discordance, say c_0=.05.

NOTE that this particular example will be used extensively in the discussion that follows because of the simplicity of the test statistic $\kappa(\mathbf{X})$ and its sampling distributions that happen to be Normal.

4 Neyman-Pearson (N-P) testing

Jerzy Neyman was a Polish mathematician and statistician who spent the first part of his professional career (1921-1934) at various institutions in Warsaw, Poland and then (1934-1938) at University College, London, and the second part (1938-1981) at the University of California, Berkeley, where he founded an influential tradition (school) in statistics. **Egon Pearson** was the son of Karl Pearson, who joined his father's Department of Applied Statistics at University College London (UCL), as a lecturer in 1923. Neyman came to UCL in 1925 to study with Karl Pearson because he was intrigued by his contributions to statistics and his more philosophical book entitled 'The Grammar of Science' published in (1892).

During his visit to UCL Neyman befriended Egon, with whom he began a decade long collaboration. At the time Egon Pearson was going through a soul searching dilemma, follow his father's approach to statistics or that of Fisher's?

"In 1925-6, I was in a state of puzzlement, and realized that, if I was to continue an academic career as a mathematical statistician, I must construct for myself what might be termed statistical philosophy, which would have to combine what I accepted from K. P's [his father] large sample tradition with the newer ideas of Fisher." (Pearson et al. 1990).

Neyman visited his friend at UCL several times during the period 1925-1933 and Egon went to Poland to meet Neyman twice. Their collaborative efforts gave rise to several papers, primarily on hypothesis testing; see Pearson (1966). The highlight of Neyman's and Pearson's professional careers was their collaboration on shaping an optimal theory for hypothesis testing during the years 1925-1935.

4.1 N-P objective: improving Fisher's significance testing

Pearson (1962), reflecting upon the Neyman-Pearson collaborative efforts, described their motivation as follows:

"What Neyman and I experienced, ..., was a dissatisfaction with the logical basis – or lack of it – which seemed to underlie the choice and construction of statistical tests." (p. 395). In particular, their main objective was to ameliorate Fisher's testing by improving what they consider as the weak *features* of that approach.

[a] Fisher's **choice of a test statistics** $d(\mathbf{X})$ on common sense grounds. Fisher would justify his choice of the test statistic as a common sense 'distance function' constructed around the 'best' estimator of θ. Neyman and Pearson question the choice of the distance function as ad hoc and sought objective criteria to define what a 'good' test is. That is, their primary objective was *an optimal theory of testing* analogous to the optimal theory of estimation developed by Fisher in the 1920s.

[b] Fisher's use of a **post-data [$\mathbf{x}_0$ is taken into account] threshold** in conjunction with the p-value to indicate *discordance* with H_0. Neyman and Pearson question the post-data threshold as vulnerable to abuse by practitioners who could evaluate the p-value $p(\mathbf{x}_0)$ and then select a threshold that would give rise to the inference

result they favor.

[c] Fisher's **falsificationist stance** that $d(\mathbf{x}_0)$ and $p(\mathbf{x}_0)$ can only indicate *discordance* but *never accordance* with H_0. Their view was that scientific research is also interested in accordance with H_0.

The culmination of their efforts was the classic Neyman-Pearson (N-P) (1933) paper where they put forward an optimal theory of testing aiming to address the issues [a]-[c], but as Pearson (1962), p. 395, mentioned, their being able to see further was the result of **standing on the shoulders of giants** (primarily Fisher and Student):

"(a) The way of thinking which had found acceptance for a number of years among practicing statisticians, which included the use of tail areas of the distributions of the test statistic.

(b) The classical tradition that, somehow, prior probabilities should be introduced numerically into a solution – a tradition which can certainly be treated in the writings of Karl Pearson and of Student, but to which perhaps only lip service was then being paid.

(c) The tremendous impact of R.A. Fisher. His criticism of Bayes's Theorem and his use of Likelihood.

(d) His geometrical representation in multiple space, out of which readily came the concept of alternative critical regions in the sample space.

(e) His tables of 5 and 1% significance levels, which lend themselves to the idea of choice, in advance of experiment, of the risk of the first kind of error which the experimenter was prepared to take.

(f) His emphasis on the importance of planning an experiment, which led naturally to the examination of the power function, both in choosing the size of the sample so as to enable worthwhile results to be achieved, and in determining the most appropriate test.

(g) Then, too, there were a number of common-sense contributions from that great practicing statistician, Student, some in correspondence, some in personal communication. "

4.2 Modifying Fisher's testing framing: a first view

It is widely accepted that the *key* modification of Fisher's framing by Neyman and Pearson was the introduction of the concept of an **alternative hypothesis** original suggested to Egon Pearson by Gosset; see Pearson (1968), Lehmann (2011). In an attempt to find a more formal way to derive test statistic to replace Fisher's construction based on intuition, Neyman and Pearson (1928) adopted wholeheartedly Fisher's likelihood function and his optimal estimation theory and decided to view hypothesis testing as a comparison of the estimated likelihood functions for the null and alternative hypotheses using their ratio; they called it 'the criterion of likelihood'. To illustrate this '**likelihood criterion**' consider the following example.

Example 13.7. Consider the simple (one parameter) Normal model (table 13.10),

where null and alternative hypotheses are simple:

$$H_0\text{: } \mu=\mu_0 \quad \text{vs.} \quad H_1\text{: } \mu\neq\mu_0. \tag{18}$$

Given that the likelihood function takes the form:

$$L(\mu;\mathbf{x})=(\tfrac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left\{-\tfrac{1}{2\sigma^2}\sum_{k=1}^n(x_k-\mu)^2\right\},$$

the MLE of μ under H_0 is μ_0, yielding the estimated likelihood function:

$$L(\mu_0;\mathbf{x})=(\tfrac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left\{-\tfrac{1}{2\sigma^2}\left[\sum_{k=1}^n(X_k-\overline{X}_n)^2+n(\overline{X}_n-\mu_0)^2\right]\right\}, \tag{19}$$

where (19) follows from the equality:

$$\sum_{k=1}^n(X_k-\mu_0)^2=\sum_{k=1}^n(X_k-\overline{X}_n+\overline{X}_n-\mu_0)^2=\sum_{k=1}^n(X_k-\overline{X}_n)^2+n(\overline{X}_n-\mu_0)^2. \tag{20}$$

The MLE under the alternative is $\widehat{\mu}_{ML}=\overline{X}_n$, giving rise to estimated likelihood function:

$$L(\widehat{\mu}_{ML};\mathbf{x})=(\tfrac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left\{-\tfrac{1}{2\sigma^2}\sum_{k=1}^n(x_k-\overline{x}_n)^2\right\}$$

Hence, the ratio of the two estimated likelihood functions yields:

$$\Lambda(\mathbf{x})=\frac{L(\widehat{\mu}_{ML};\mathbf{x})}{L(\mu_0;\mathbf{x})}=\frac{(\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^n(x_k-\overline{x}_n)^2\right\}}{(\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{k=1}^n(x_k-\overline{x}_n)^2+n(\overline{x}_n-\mu_0)^2\right]\right\}}=\exp\left\{\tfrac{1}{2}\left[\frac{n(\overline{x}_n-\mu_0)^2}{\sigma^2}\right]\right\}. \tag{21}$$

Intuition suggests that one would reject H_0 when $\Lambda(\mathbf{x}_0)$ is large enough, i.e. when $h(\mathbf{x}_0)=\left|\frac{n(\overline{x}_n-\mu_0)^2}{\sigma^2}\right|>c$, which suggests that $\frac{n(\overline{x}_n-\mu_0)^2}{\sigma^2}$ provides a natural distance function. A closer look at $h(\mathbf{x}_0)$ confirms Fisher's choice of a test statistic $\kappa(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}$ as a basis for his significance testing. Indeed, several of the examples used by Neyman and Pearson (1928) appear to suggest that the likelihood ratio affirmed several of Fisher's choices of test statistics; see Gorroochurn (2016). In this sense, the likelihood ratio seemed to suggest **natural test statistics** based on Maximum Likelihood estimators, but one needed a more formal justification for such a choice that is ideally based on some optimal theory of testing analogous to Fisher's optimal estimation theory.

What was not appreciated enough at the time, and continues to this day, is that the concept of an **alternative hypothesis brought into testing the whole of the parameter space** Θ associated with the prespecified statistical model:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X. \tag{22}$$

This is because an optimal theory of testing would require one to compare the null value θ_0 with all other possible values of θ in Θ to establish a 'best' test. Hence, the

alternative hypothesis H_1 should be defined as the *complement* to the null value(s) relative to Θ. That is, the *archetypal* way to specify the null and alternative hypotheses for N-P testing is:

$$H_0\text{: } \theta\in\Theta_0 \text{ vs. } H_1\text{: } \theta\in\Theta_1, \tag{23}$$

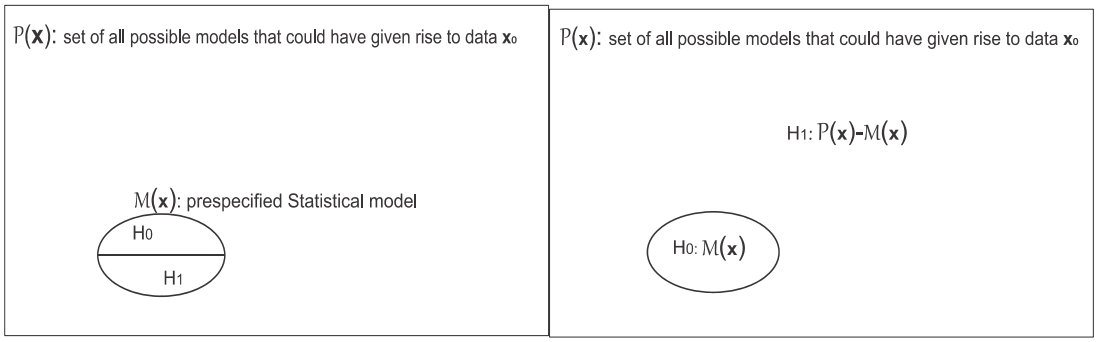
where Θ_0 and Θ_1 constitute a *partition* of Θ: $\Theta_0\cap\Theta_1=\varnothing$, $\Theta_0\cup\Theta_1=\Theta$.

■ Unfortunately, the archetypal specification in (23) has been neglected in a tangle of confusions generated by the subsequent literature, commencing with the misconstrual of the **Neyman-Pearson (N-P) lemma** that was based on a largely **artificial partition** $\Theta{:=}(\theta_0,\theta_1)$.

Learning from data. The reasoning underlying this argument is that N-P tests pose hypothetical questions – framed in terms of $\theta\in\Theta$ – aiming to learn about the true value θ^*, i.e. the generating mechanism $\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}^*)\}$, $\mathbf{x}\in\mathbb{R}^n_X$, requires that the N-P framing covers the whole of Θ and $\mathbb{R}_X$.

The above archetypal specification makes it clear that N-P testing takes place *within* the boundaries of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, and all possible values of the parameter space are relevant for statistical purposes, even though only a few might be of substantive interest. This addresses Fisher's criticism of the type II error probability:

"The frequency of the second kind must depend not only on the frequency with which rival hypotheses are in fact true, but also greatly on how closely they resemble the null hypothesis. Such errors are therefore incalculable both in frequency and in magnitude merely from the specification of the null hypothesis, and would never have come into consideration in the theory only of tests of significance, had the logic of such tests not been confused with that of acceptance procedures." (Fisher, 1955, p. 73)



Testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: N-P Testing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: M-S testing

■ This foregrounds the importance of securing the *statistical adequacy* of $\mathcal{M}_\theta(\mathbf{x})$ [validating its probabilistic assumptions] to ensure the reliability of testing inferences. That is, 'learning from data' about θ^* using N-P testing presupposes that the modeler has established that θ^* lies within the boundaries of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ by securing its statistical adequacy.

The second modification of Fisher's framing came in the form of *what constitutes a test*, which is not just a distance function selected on intuitive grounds and a tail

probability! As in the case of an estimator, an N-P *test statistic* is a mapping from the sample space ($\mathbb{R}^n_X$) to the parameter space (Θ):

$$d(.)\text{: } \mathbb{R}^n_X \to \mathbb{R},$$

that *partitions* the sample space ($\mathbb{R}^n_X$) into an *acceptance* C_0 and a *rejection region* C_1:

$$C_0 \cap C_1 = \varnothing,\ C_0 \cup C_1 = \mathbb{R}^n_X,$$

in a way that correspond to Θ_0 and Θ_1, respectively:

$$\mathbb{R}^n_X = \left\{ \begin{array}{c} \boxed{C_0} \leftrightarrow \boxed{\Theta_0} \\ \boxed{C_1} \leftrightarrow \boxed{\Theta_1} \end{array} \right\} = \Theta$$

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*:

$$\text{[i] if } \mathbf{x}_0 \in C_0\text{, accept } H_0, \qquad \text{[ii] if } \mathbf{x}_0 \in C_1\text{, reject } H_0. \tag{24}$$

Table 13.13: 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	$\surd$

This N-P re-framing gave rise to two types of errors (table 13.13), whose *probabilities* (how often a testing procedure errs) are evaluated by:

$$\textbf{type I:} \quad \mathbb{P}(\mathbf{x}_0 \in C_1;\ H_0(\theta)) = \alpha(\theta), \text{ for } \theta \in \Theta_0,\ [\mathbf{x}_0 \in C_1 \iff \text{reject } H_0],$$

$$\textbf{type II:} \quad \mathbb{P}(\mathbf{x}_0 \in C_0;\ H_1(\theta)) = \beta(\theta), \text{ for } \theta \in \Theta_1,\ [\mathbf{x}_0 \in C_1 \iff \text{accept } H].$$

These error probabilities enabled Neyman and Pearson to introduce the third key notion of an 'optimal' test, by fixing $\alpha(\theta)$ to a small value and minimizing $\beta(\theta)$, or maximizing the power $\pi(\theta) = (1 - \beta(\theta))$, for all $\theta \in \Theta_1$.

Optimal test. Select the particular $d(\mathbf{X})$ in conjunction with a rejection region based on a prespecified significance level α:

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

with a view that the combination $(d(\mathbf{X}), C_1(\alpha))$ constitutes a test with the highest *pre-data capacity* to reject H_0 when it is false. This capacity is known as the *power* of the test:

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

The power is a measure of the *generic* (whatever the value $\mathbf{x}\in\mathbb{R}^n_X$) *capacity* of the test to detect discrepancies from H_0 for all θ in Θ_1. That is, the combination of the power $\pi(\theta)$, $\forall\theta_1\in\Theta_1$ and the significance level α calibrate the generic capacities of a particular test to detect discrepancies from H_0.

In summary, the Neyman-Pearson (N-P) proposed modifications to the perceived weaknesses of Fisher's significance testing, have changed Fisher's framework in four important respects.

■ (i) The N-P testing takes place *within* a prespecified $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ by partitioning it into the subsets associated with the null and alternative hypotheses, and any N-P inference involves (directly or indirectly) the whole of the parameter space Θ.

(ii) Fisher's *post-data* p-value has been replaced with a *pre-data* significance level α and the power of the test.

(iii) Combining (i)-(ii), an *optimal test* is defined as $\mathcal{T}_\alpha:=(d(\mathbf{X}),C_1(\alpha))$ where $\mathcal{T}_\alpha$ is chosen so that it maximizes the power $\pi(\theta)$, for all values of θ in Θ_1.

(iv) Fisher's p-value as a measure of *discordance* with H_0, was replaced by the more *behavioristic accept/reject H_0 rules:*

"Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern out behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong." (Neyman and Pearson, 1933, p. 290).

As argued in the sequel, notwithstanding Fisher's (1955) hard-hitting criticisms pertaining to (i)-(iii), the N-P reframing gave rise to **an optimal theory of testing**, analogous to that of estimation framed almost single-handedly by Fisher himself. Where the N-P reframing did not make real progress over Fisher's original formulation was in (iv). Attempts to provide a sound evidential interpretation for the p-value or the accept/reject H_0 results has beleaguered frequentist testing since the 1930s.

4.3 The archetypal N-P testing framing

■ In N-P testing there has been a long debate concerning the proper way to specify the null and alternative hypotheses. It is often thought to be rather arbitrary and subject to abuse. This view stems primarily from inadequate understanding of the role of statistical vs. substantive information. When properly understood in the context of a statistical model $\mathcal{M}_\theta(\mathbf{x})$, N-P testing leaves very little leeway in specifying the H_0 and H_1 hypotheses. This is because the whole of Θ [all possible values of θ] is relevant on *statistical* grounds, irrespective of whether only a small subset might be relevant on *substantive* grounds.

Hence, there is nothing arbitrary about specifying the null and alternative hypotheses in N-P testing. The *default alternative* is always the complement to the null relative to the parameter space of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

Example 13.9. Consider the following hypotheses:

$$H_0\text{: } \mu\leq\mu_0\text{, vs.}\quad H_1\text{: } \mu>\mu_0, \tag{25}$$

in the context of the *simple* (one parameter) *Normal model* (table 13.10):

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\text{: } X_t \backsim \mathsf{N}\left(\mu, \sigma^2\right), \ [\sigma^2 \text{ known}], \ t{=}1, 2, ..., n, ...$$

In light of the results in example 3.7, consider the test $\mathcal{T}_\alpha{:=}\{\kappa(\mathbf{X}),\ C_1(\alpha)\}$:

$$\begin{array}{l} \text{test statistic: } \kappa(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma},\ \overline{X}_n{=}\frac{1}{n}\sum_{k=1}^{n} X_k, \\ \text{rejection region: } C_1(\alpha){=}\{\mathbf{x}\text{: } \kappa(\mathbf{x}) > c_\alpha\}, \end{array} \tag{26}$$

that can be shown to be optimal. To evaluate the error probabilities one needs the distribution of $\kappa(\mathbf{X})$ under H_0 and H_1:

[i] $\kappa(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma} \overset{H_0(\mu_0)}{\backsim} \mathsf{N}(0,1)$,

[ii] $\kappa(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma} \overset{H_1(\mu_1)}{\backsim} \mathsf{N}(\delta_1,1),\ \delta_1{=}\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}{>}0$ for all $\mu_1 > \mu_0$.

These hypothetical sampling distributions are then used to compare H_0 or H_1 via $\kappa(\mathbf{x}_0)$ to the true value $\mu{=}\mu^*$ represented by data $\mathbf{x}_0$ via $\overline{X}_n$, the best estimator of μ. That is the distance $\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}$ is the estimated form of $\frac{\sqrt{n}(\mu^*-\mu_0)}{\sigma}$; recall that by assumption, $\mathcal{M}^*(\mathbf{x})$ has given rise to the data $\mathbf{x}_0$. Both evaluations in [i]-[ii] involve hypothetical reasoning in contrast to:

$$\textsf{Factual:} \quad \text{[iii]} \ \tfrac{\sqrt{n}(\overline{X}_n-\mu^*)}{\sigma} \overset{\mu=\mu^*}{\backsim} \mathsf{N}(0,1),$$

that underlies estimation (point and interval). The evaluation of the type I error probability is based on (i):

$$\alpha{=}\max_{\mu\leq\mu_0} \mathbb{P}(\kappa(\mathbf{X}){>}c_\alpha; H_0(\mu)){=}\mathbb{P}(\kappa(\mathbf{X}){>}c_\alpha; \mu{=}\mu_0).$$

Notice that in cases where the null is for the form H_0: $\mu \leq \mu_0$, the type I error probability is defined as the maximum over all values in the interval $(-\infty, \mu_0]$, which turns out to be $\mu{=}\mu_0$.

The evaluation of type II error probabilities is based on (ii):

$$\beta(\mu_1){=}\mathbb{P}(\kappa(\mathbf{X}) \leq c_\alpha; H_1(\mu_1)) \text{ for all } \mu_1 > \mu_0.$$

The *power* [rejecting the null when false] is equal to $1{-}\beta(\mu_1)$, i.e.

$$\pi(\mu_1){=}\mathbb{P}(\kappa(\mathbf{X}) > c_\alpha; H_1(\mu_1)) \text{ for all } \mu_1 > \mu_0.$$

Figure 13.7 illustrates the type I, II error probabilities, and the power.

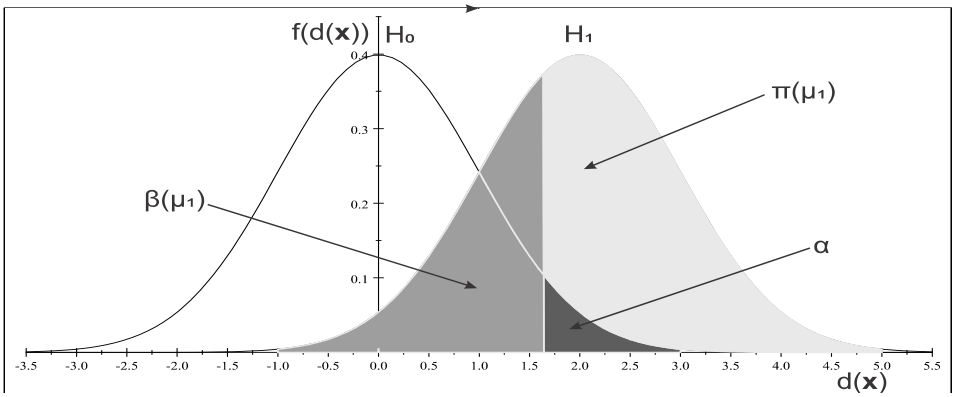


Fig. 13.7: Type I and II error probabilities and the power of the test

Why power? The power $\pi(\mu_1)$ measures the *pre-data* (generic) *capacity* (probativeness) of test $\mathcal{T}_\alpha$:=$\{\kappa(\mathbf{X}),\ C_1(\alpha)\}$ to detect a discrepancy, say γ=$\mu_1-\mu_0$=.1, when present. Hence, when $\pi(\mu_1)$=.35, this test has very low capacity to detect such a discrepancy. If γ=.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.

■ **Being a 'smart alec' with words**. Certain statistics textbooks make a big deal out of the distinction: 'accept H_0' vs. 'fail to reject H_0'.

At a certain level, this is a commendable attempt to bring out the problem of misinterpreting 'accept H_0' as tantamount to 'there is evidence *for* H_0', known as the **fallacy of acceptance**. However, by the same token there is an analogous distinction between 'reject H_0' vs. 'accept H_1', that highlights the problem of misinterpreting 'reject H_0' as tantamount to 'there is evidence *for* H_1', known as the **fallacy of rejection**. What is objectionable about this practice is that **the verbal distinctions by themselves do nothing but perpetuate these fallacies** by just paying lip service instead of addressing them; see section 5.

■ **The zero probability 'paradox'**. The assertion underlying this 'paradox' is that point null hypotheses, H_0: θ=θ_0, are *always false* (not exactly true) in the real world, and thus such testing is pointless. This argument has no merit because a hypothesis or an inferential claim being 'exactly correct' has no place in statistics; 'being a statistician means never having to say a hypothesis or an inferential claim is exactly correct' Senn (2018).

What is assumed is that there is a true θ^* in Θ, and hypothesis testing poses questions to the data seeking to find out whether the hypothesized value θ_0 is 'close enough' to θ^*. Indeed, the whole idea behind frequentist inference is that we can learn from data about θ^* without learning its exact value and such claims are calibrated

using error probabilities. This is the reason why the **post-data severity evaluation** discussed in what follows outputs the *discrepancy* γ from $\theta=\theta_0$ warranted by data $\mathbf{x}_0$ and test $\mathcal{T}_\alpha$.

4.4 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 13.7:

$$\begin{aligned} \mathbb{P}(\textbf{type I error}): \quad & \mathbb{P}(\kappa(\mathbf{X})>c_\alpha;\, \mu=\mu_0)=\alpha, \\ p\textbf{-value}: \quad & \mathbb{P}(\kappa(\mathbf{X})>\kappa(\mathbf{x}_0);\, \mu=\mu_0)=p(\mathbf{x}_0), \end{aligned} \tag{27}$$

it becomes obvious that (a) they share the same test statistic $\kappa(\mathbf{X})$ 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: $\{\mathbf{x}\text{:}\ \kappa(\mathbf{x})>c_\alpha\}$ vs. $\{\mathbf{x}\text{:}\ \kappa(\mathbf{x})>\kappa(\mathbf{x}_0)\}$, $\mathbf{x}\in\mathbb{R}^n_X$, rendering α a *pre-data* and $p(\mathbf{x}_0)$ a *post-data* error probability.

Example 13.10. Consider the test $\mathcal{T}_\alpha$:=$\{\kappa(\mathbf{X}),\ C_1(\alpha)\}$ in example 13.9 for μ_0=10, σ=1, n=100, $\overline{x}_n$=10.4, α=.05 $\Rightarrow c_\alpha$=1.645.

$$\kappa(\mathbf{x}_0)=\tfrac{\sqrt{n}(\overline{x}_n-\mu_0)}{\sigma}=\tfrac{\sqrt{100}(10.175-10)}{1}=1.75>c_\alpha, \quad \text{Reject } H_0.$$

The p-value is: $\mathbb{P}(\kappa(\mathbf{X})>1.75;\ \mu=\mu_0)$=.04.

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*. Hence, it should come as no surprise to learn that the above N-P decision rules (24) could be recast in terms of the p-value:

$$\text{[i]}^*\ \text{if } p(\mathbf{x}_0)>\alpha, \text{ accept } H_0, \quad \text{[ii]}^*\ \text{if } p(\mathbf{x}_0)\leq\alpha, \text{ reject } H_0.$$

Indeed, practitioners often prefer to use the modified rules [i]*-[ii]* because the p-value appears to convey additional information. For instance, rejecting H_0 with $p(\mathbf{x}_0)$=.0001 *seems* more informative than just say H_0 was rejected since $p(\mathbf{x}_0)<\alpha$=.05; see Lehmann and Romano (2005).

Pre-data vs. post-data. Contrary to Gigerenzer (1993), there is no intrinsic inconsistency between Fisher's significance and N-P testing. The key difference between the two types of error probabilities is:

(a) *pre-data* (only n is known), type I, II (power) and coverage, and

(b) *post-data* ($d(\mathbf{x}_0)$ is known), p-value, severity.

The claimed inconsistency between Fisher's significance and N-P testing stems from conflating these two different perspectives of the p-value.

Pre-data perspective. The p-value is traditionally defined as **the probability of obtaining a result 'equal to or more extreme' than the one observed $\mathbf{x}_0$, when H_0 is true**.

The clause 'equal to or more extreme' in the context of a N-P test is invariably interpreted in light of H_1, and their lies the confusion. This has led to the p-value being viewed as the smallest significance level $\alpha_{\min}$ at which H_0 would have been rejected when true; hence the **observed significance level** of a N-P test.

► The same pre-data perspective is also used when asserting that:

$$p(\mathbf{X}) \overset{H_0}{\sim} \mathsf{U}(0,1).$$

Example. For the hypotheses: H_0: $\mu{=}\mu_0$ vs. H_1: $\mu{\neq}\mu_0$, in the context of simple Normal model, the pre-data interpretation of the p-value would give rise to the concept of a **two-sided p-value**:

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

There is no such thing as a two-sided p-value when viewed as a post-data measure of discordance!

Post-data perspective. The post-data p-value is defined as **the probability of all sample realizations $\mathbf{x}{\in}\mathbb{R}^n_X$ for which $d(\mathbf{x})$ accords less well with H_0 than $\mathbf{x}_0$ does, when H_0 is true**.

The real difference between pre-data and post data error probabilities is that the latter uses additional information in the form of **the sign** of $d(\mathbf{x}_0)$ indicating the direction of departure from H_0 suggested by data $\mathbf{x}_0$. This information **renders one of the two tails irrelevant**.

For instance, when $d(\mathbf{x}_0){=}\frac{\sqrt{n}(\overline{x}_n-\mu_0)}{\sigma}{=}3.6>0$, $\mu^*{\notin}(-\infty,\mu_0)$. That is, such a large positive value of $d(\mathbf{x}_0)$ indicates clearly that data $\mathbf{x}_0$ were generated by values of μ^* much larger than $\mu_0{=}.5$; $\mathbf{x}_0$ could not have been generated by $f(\mathbf{x};\mu)$, $\mathbf{x}{\in}\mathfrak{X}$, for $\mu<.5$!

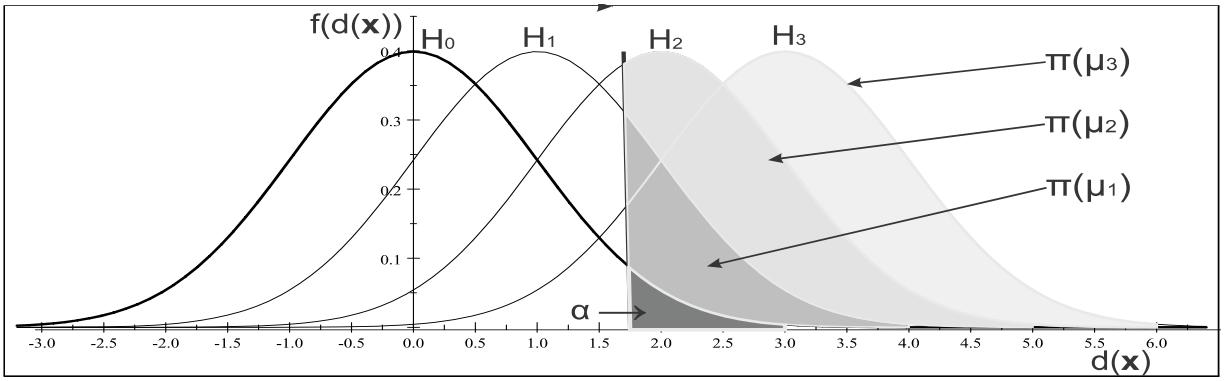


Fig. 4: Power of the test for different discrepancies $\mu_1{<}\mu_2{<}\mu_3$ from μ_0

Third, there is nothing irreconcilable between the pre-data significance level α and the post-data p-value $p(\mathbf{x}_0)$ because the power is relevant for both when seeking

an evidential interpretation of the accept/reject rules. Even Fisher, in unguarded moments, would refer to the power as the 'sensitiveness' of a test: "By increasing the size of the experiment, we can render it more sensitive, meaning by this that it will allow of the detection of a lower degree of sensory discrimination, or in other words, of a quantitatively smaller departure from the null hypothesis." (Fisher, 1934, pp. 21-22).
It is rather unfortunate that to this day the discussion of the crucial weaknesses of the p-value largely ignores the power of the test.

Fourth, 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 the probabilities in (27) are *firmly attached* to the sample realizations $\mathbf{x}\in\mathbb{R}^n_X$. Indeed, attaching probabilities to the unknown constant μ or conditioning on values of μ makes no sense in frequentist statistics. On that issue Fisher (1921) argued: "We may discuss the probability of occurrence of quantities which can be observed or deduced from observations, in relation to any hypotheses which may be suggested to explain these observations. We can know nothing of the probability of hypotheses..." (p. 25).

Fifth, both the p-value and the type I and II error probabilities are NOT *conditional* on H_0 or H_1. Contrary to the widely-held belief **frequentist error probabilities** are *NOT conditional* on H_0 or H_1, since $\theta{=}\theta_0$ or $\theta{=}\theta_1$ being 'true or false' do not constitute *legitimate events* in the context of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$; θ is an *unknown constant*, not a random variable. Hence:

$$\mathbb{P}(d(\mathbf{X})>c_\alpha;\ \theta{=}\theta_0){=}\alpha \text{ and NOT } \mathbb{P}(d(\mathbf{X})>c_\alpha|\ \theta{=}\theta_0){=}\alpha.$$

Instead, 'H_0 or H_1 is true or false' represents a *hypothetical scenarios* under which a tail area of the **sampling distribution** of the test statistic $(d(\mathbf{X}))$ is evaluated. Hence, the claim: "... for all practical purposes in my view, the p value, is indeed a probability conditional or conditioned on an assumption, the null hypothesis." (Schneider, 2018) bespeaks ignorance of basic probability theory; one can condition only on events and random variables, not assumptions and beliefs.

What can go wrong when adopting such a view? (everything!)

(A) **Numerous misinterpretations** of the p-value and the power stem primarily from misinterpreting such error probabilities as conditional on H_0 or H_1.

Cohen (1988): "The power of a statistical test is the probability that it will yield statistically significant results." (p. 1) **NO!**

(B) **The base-rate (**or prosecutor's**) fallacy**; see Achinstein (2010), Howson (2000). In the context of Bayesian reasoning, where θ is viewed as a random variable, this fallacy arises when the conditional probability of a hypothesis H given some evidence E, $P(H|E)$ (the *posterior* probability of H) is conflated with $P(E|H)$ since:

$$P(H|E){=}P(E|H)\left[\tfrac{P(E)}{P(H)}\right], \tag{28}$$

the base rates $[P(H),\ P(E)]$ can be ignored at a Bayesian's peril.

Frequentist error probabilities are immune to this charge since $P(H|E)$ and $P(E|H)$ are meaningless nonsense in **frequentist testing**; see Spanos (2010).

(C) Viewing error probabilities as conditional on H_0 or H_1 lies at the center of the recent replication crises debates thru the **Positive Predictive Value** (PPV) put forward by Ioannidis (2005):

$$\text{PPV}=\Pr(F|R)=\frac{\#\text{ true positive detections}}{\#\text{ positive detections}}=\frac{\Pr(R\cap F)}{\Pr(R)}=\frac{\Pr(R|F)\Pr(F)}{\Pr(R|F)P(F)+\Pr(R|\overline{F})P(\overline{F})},$$

where H_0: no disease, F=H_0 is false, R=test rejects H_0.

▶ The PPV was used by Ioannidis (2005) to make a case that '**most published research findings are false**' because of the wide-spread abuse of *significance testing* (p-hacking, multiple testing, cherry-picking, low power).

Ioannidis (2005) made his case by sitting in his office and making up numbers, say

$$\Pr(F)=.1,\Pr(R|F)=.8,\ \Pr(R|\overline{F})=.15,$$

and then using the apparent (but imaginary) link between these and the error probabilities ($\Pr(R|F)$-power at some discrepancy, $\Pr(R|\overline{F})$-significance level), as well as a prior on 'true nulls' ($\Pr(F)$=.1), concluded that the 'true' rejections rate is:

$$\text{PPV}=\frac{\Pr(R|F)\Pr(F)}{\Pr(R|F)P(F)+\Pr(R|\overline{F})P(\overline{F})}=\frac{(.8)(.1)}{(.8)(.1)+(.15)(.9)}=\frac{.08}{.215}=.372<.5$$

Why $\Pr(F)$=.1? In Economics only around 10% of all null hypotheses are 'true'.

Why $\Pr(R|F)$=.8? One needs .8 or higher for the discrepancy of interest.

Why $\Pr(R|\overline{F})$=.15? The usual threshold is .05, but with all the biases the actual one should be .15. Why not $\Pr(R|\overline{F})$=.21:

$$\text{PPV}=\frac{(.8)(.1)}{(.8)(.1)+(.21)(.9)}=.297$$

This makes a better case! What about the zero probability paradox?

$$\text{PPV}=\frac{(.8)(.0)}{(.8)(0)+(.21)(1)}=0,$$

which cements Ioannidis' case!

▶ Unfortunately, none of his probabilities [$\Pr(F)$, $\Pr(R|F)$,$\Pr(R|\overline{F})$] make sense in frequentist testing: $\Pr(F)$ is just a prior attached to values of θ [a NO, NO in frequentist testing], and $\Pr(R|F)$, $\Pr(R|\overline{F})$ **have nothing to do** with the 'power' and the 'significance level' of a test; Spanos (2010). **Why**?

[a] 'H_0: no disease' and 'F=H_0 is false' are not legitimate events that one can assign probabilities to in frequentist testing; they make sense in Bayesian testing!

[b] All error probabilities (i) are framed in terms of θ and (ii) **depend crucially** on the particular **statistical context**:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ H_0\text{: }\theta\in\Theta_0\text{ vs. }H_1\text{: }\theta\in\Theta_1,\ \mathcal{T}_\alpha:=\{d(\mathbf{X}),C_1(\alpha)\}\text{ and }\mathbf{x}_0. \tag{29}$$

What about the invoked link? The false positive/negative rates for medical devices and procedures have NO statistical context. The manufacturers of medical devices run a very large number (say, 10000) of medical 'tests' with specimens of blood, urine, etc., that are known to be positive or negative.

What is the confusion? Confusing a purely probabilistic deduction with a statistical inference result.

In summary, the PPV constitutes ad hoc posterior measure of *blameworthiness by association* based on a Bayesian meta-model for field-wide inferences. For instance, **all economists** are guilty by **... association** in publishing untrustworthy evidence since:

(i) only 10% of the nulls in economics are true, and

(ii) economists are known to indulge in abusing the p-value.

His conclusion that **"most published research findings are false"** is likely to be accurate, but his explanation of why is totally the wrong!

The fact that the probabilities [$\Pr(F)$, $\Pr(R|F)$,$\Pr(R|\overline{F})$] are meaningful in the context of Bayesian inference, does not render them relevant for **evaluating the reliability of frequentist testing results**. Indeed, when viewed from a frequentist perspective, the PPV has nothing to do with unveiling the untrustworthiness of published empirical evidence, and reflects attention away from certain real sources of untrustworthy evidence, which include:

(i) **statistical misspecification**: invalid probabilistic assumptions imposed on data,

(ii) **uninformed implementation** of inference procedures, and

(iii) **unwarranted evidential interpretations** of their inferential results.

That is, the most crucial source of untrustworthy evidence is the **uninformed, recipe-like implementation of statistical methods** without any real understanding of their assumptions, limitations, proper implementation and unwarranted interpretations of their results. This suggests a refocusing of the proposed strategies for securing the **trustworthiness** of published empirical evidence on **a case by case basis** by appraising whether the study in question has circumvented or dealt with the **potential errors and omissions** that could have undermined the reliability of the particular inferences drawn.

4.5 Optimality of a Neyman-Pearson (N-P) test

How does the introduction of type I and II error probabilities addresses the arbitrariness associated with Fisher's choice of a test statistic? The ideal test is one whose error probabilities are zero, but no such test exist for a given n; an ideal test, like

the ideal estimator, exists only as $n \to \infty$! For $n < \infty$, however, there is a *trade-off* between the above error probabilities: increasing the threshold $c>0$ decreases the type I but increases the type II error probabilities (see fig. 13.8).

Example 13.11. In the case of test $\mathcal{T}_\alpha$:=$\{\kappa(\mathbf{X}),\ C_1(\alpha)\}$ in (26), decreasing the probability of type I error from α=.05 to α=.01, increases the threshold from c_α=1.645 to c_α=2.33, which makes it easier to accept H_0, and this in turn increases the probability of type II error.

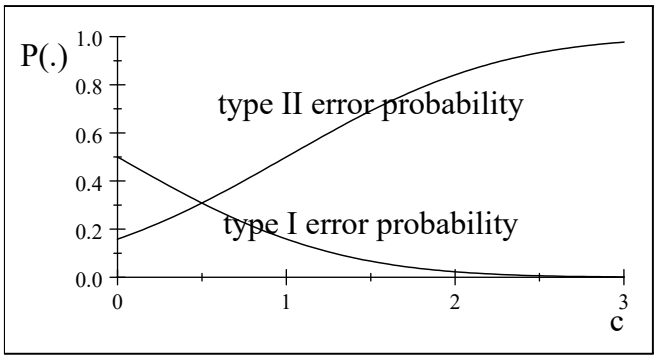


Fig. 13.8: the trade-off: type I vs. II

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.

N-P rationale. In their attempt to justify their twofold strategy, Neyman and Pearson (1933) urged the reader consider the analogy with a criminal offense trial, where the jury are instructed by the judge to find the defendant 'not guilty' unless they have been convinced 'beyond any reasonable doubt' by the evidence:

$$H_0\text{: not guilty},\quad \text{vs.}\quad H_1\text{: guilty.}$$

The clause 'beyond any reasonable doubt' amounts to fixing the type I error to a very small value, to reduce the risk of sending innocent people to the death row. At the same time, one would want the system to minimize the risk of letting guilty people get off scot free".

4.5.1 Optimal properties of N-P tests

The property that sets the gold standard for optimal test is:

[1] **Uniformly Most Powerful (UMP)**: A test $\mathcal{T}_\alpha$:=$\{d(\mathbf{X}),\ C_1(\alpha)\}$ is said to be UMP if it has higher power than any other α-level test $\widetilde{T}_\alpha$, $\forall\theta\in\Theta_1$:

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

That is, an α significance level UMP test $\mathcal{T}_\alpha$ has equal or higher power than any other α significance level test $\widetilde{T}_\alpha$ for all values of θ in Θ_1.

Figure 13.9 illustrates the notion of a UMP by depicting the power of several tests, whose power curves begin at $\pi(\theta_0;\mathcal{T}_\alpha)$=$\alpha$ and monotonically increase as the discrepancy from the null value θ_0 increases. The UMP test corresponds to the power curve in a bold line since it dominates all other power curves for all $\theta\in\Theta_1$; when power curves intersect, no UMP exists.

Example 13.6 (continued). In the context of the simple Normal model (table 13.12) the test $\mathcal{T}_\alpha$:=$\{\kappa(\mathbf{X}),C_1(\alpha)\}$ in (26) is UMP; Lehmann and Romano (2005).

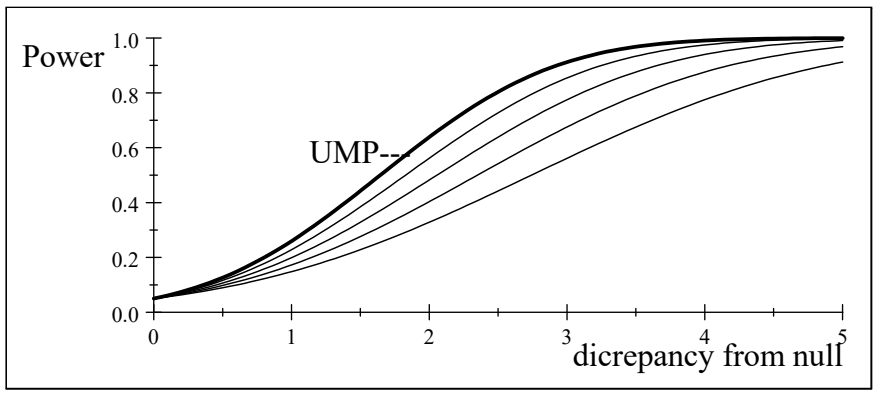


Fig. 13.9: The power of several tests; UMP is the bold line

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 is always greater than that of rejecting H_0 when true, i.e.

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

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 $\theta\in\Theta_1$.

[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 γ= $(\theta_0-\theta_1)$, $\forall\theta_1\in\Theta_1$, however small, as $n\to\infty$, i.e.

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

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!

4.5.2 Evaluating the power of a test

Example 13.12. How does one evaluate the power of test $\mathcal{T}_\alpha$:=$\{\kappa(\mathbf{X}),\ C_1(\alpha)\}$ for different discrepancies γ=$\mu_1-\mu_0$?

Step 1. In light of the fact that for the relevant sampling distribution for the evaluation of $\pi(\mu_1)$ is: [II] $\kappa(\mathbf{X})$=$\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma} \overset{H_1(\mu_1)}{\backsim} \mathsf{N}(\delta_1,1)$, for all $\mu_1>\mu_0$,

the use the $\mathsf{N}(0,1)$ tables requires one to split the test statistic into:

$$\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}=\frac{\sqrt{n}(\overline{X}_n-\mu_1)}{\sigma}+\delta_1,\ \ \delta_1=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma},$$

since under $H_1(\mu_1)$ the distribution of the first component is:

$$\frac{\sqrt{n}(\overline{X}_n-\mu_1)}{\sigma}=(\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}-\delta_1) \overset{H_1(\mu_1)}{\backsim} \mathsf{N}(0,1),\ \text{for } \mu_1>\mu_0.$$

Step 2. Evaluation of the power the test $\mathcal{T}_\alpha$:=$\{\kappa(\mathbf{X}),\ C_1(\alpha)\}$, with σ=1, n=100, c_α=1.645 for different discrepancies $(\mu_1-\mu_0)$ yields the results in table 13.14 (figure 13.10), where Z denotes a generic standard Normal random variable, i.e. $Z\backsim\mathsf{N}(0,1)$.

Table 13.14: Evaluating the power of test $\mathcal{T}_\alpha$

γ=$\mu_1-\mu_0$	δ_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.

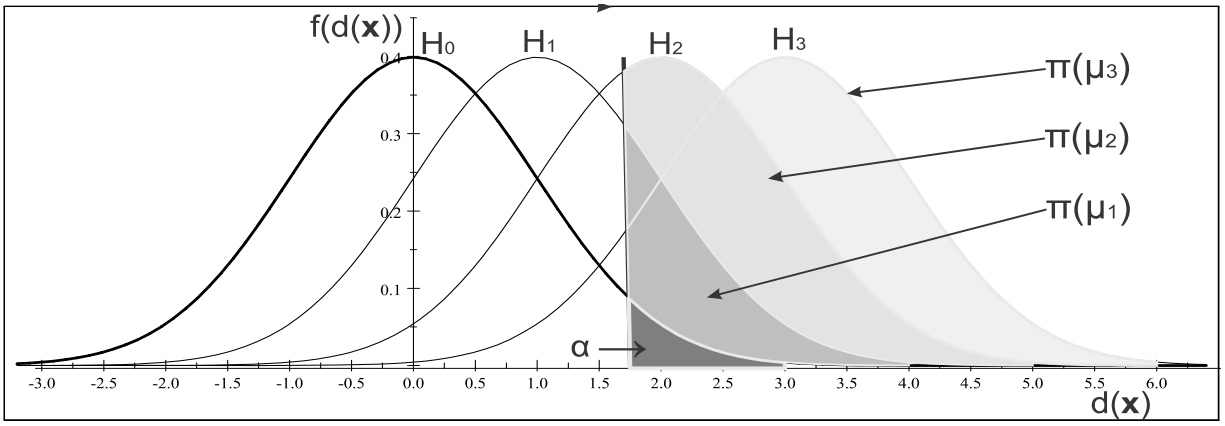


Fig. 13.10: Power of the test for different discrepancies $\mu_1<\mu_2<\mu_3$ from μ_0

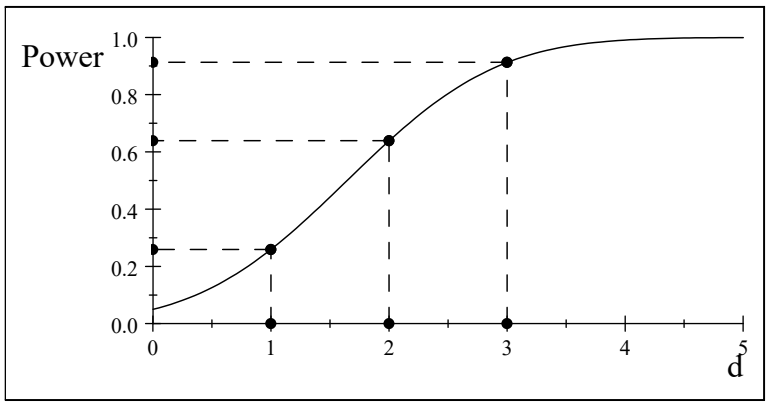


Fig. 13.11: Power curve (table 13.12)

The tail areas associated with the significance level and the power of this test is illustrated in figures 13.10-11.

The power of the test $\mathcal{T}_\alpha$:=$\{\kappa(\mathbf{X}),\ C_1(\alpha)\}$ is typical of an *optimal test* since $\pi(\mu_1)$ increases with the non-zero mean δ_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.

Example 13.13. For test $\mathcal{T}_\alpha$:=$\{\kappa(\mathbf{X}),\ C_1(\alpha)\}$ in (26), with σ=1, α=.05 (c_α=1.645), let us allow the n to increase from 2 to 1000 and evaluate the power for different discrepancies γ, $.075 \leq \gamma \leq .20$, in order to see how the generic capacity (power) of test $\mathcal{T}_\alpha$ increases with n. As can be seen in figure 13.12, the power at n=400 increases with discrepancies γ, but it also increases rapidly with n: $\pi(.02)$=.991, $\pi(.015)$=.912, $\pi(.125)$=.804, $\pi(.125)$=.639, $\pi(.075)$=.442.

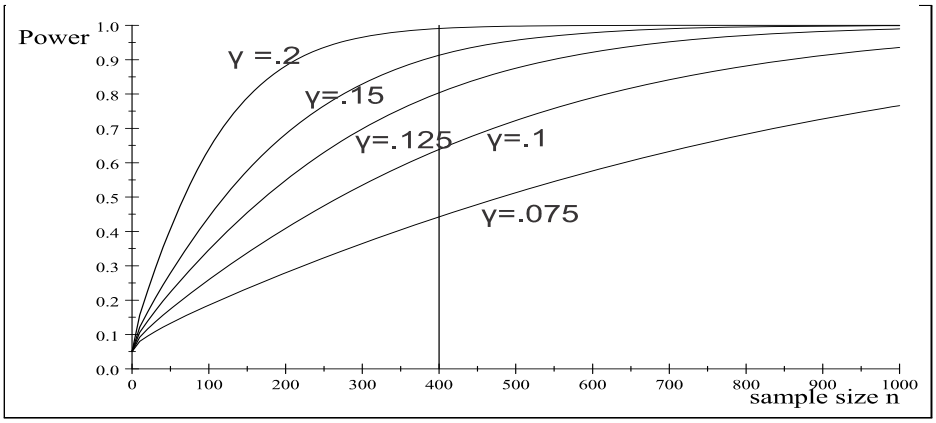


Fig. 13.12: Power for different n and discrepancies $\gamma\sigma$ from the null

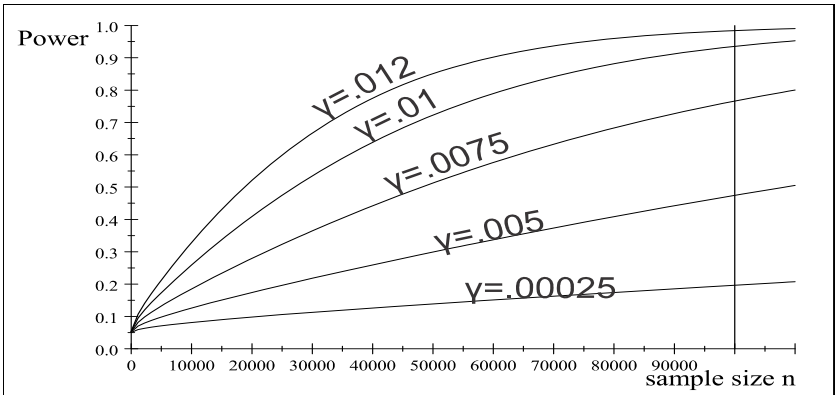


Fig. 13.13: Power for different n and discrepancies $\gamma\sigma$ from the null

When the sample size is increased to n=100,000 (figure 13.13), the power for tiny discrepancies γ from the null indicate significance since the generic capacity of the test increases dramatically: $\pi(.0075)$=.766 30, $\pi(.01)$=.966 64, $\pi(.012)$=.984. Figures 13.12-13 show most clearly that the 'evidential worth' for the presence of a particular $\gamma > 0$ at .05 significance level with n=100 is very different when n=100,000.

■ The figures 13.12-13 indicate most clearly that the current practice of designating statistical significance using asterisks:

$$(.075)^{*}\Leftrightarrow p(\mathbf{x}_0)\leq .05,\ (.075)^{**}\Leftrightarrow p(\mathbf{x}_0)\leq .01,\ (.075)^{***}\Leftrightarrow p(\mathbf{x}_0)\leq .001,$$

is a bad idea since the inference is detached from the context: n, $\mathcal{T}_\alpha$, $\pi(\mu_1)$, $\forall \mu_1 > \mu_0$.

A pre-data role for the power. In light of (a)-(c) one can use the power function pre-data to ensure that before carrying out a study one will be able to detect certain discrepancies of substantive interest with high probability. In practice, it is often difficult to decrease σ and thus the focus for that is usually on the selection of the sample size n.

Example 13.13 (continued). When applying the test $\mathcal{T}_\alpha$:=$\{\kappa(\mathbf{X}),\ C_1(\alpha)\}$ let us assume that the discrepancy of substantive interest is γ=$(\mu_1-\mu_0)$=.2. To ensure that this test has high enough generic capacity (power), say $\pi(.2) \geq .8$, to detect γ=.2 if it exists, one needs to perform certain preliminary calculations to ensure that n is large enough for $\pi(.2) \geq .8$ by solving the equation:

$$\pi(.2)=\mathbb{P}\left(\frac{\sqrt{n}(\overline{X}_n-.2)}{\sigma} > 1.645-\frac{\sqrt{n}(.2)}{\sigma};\mu_1=.2\right)\geq .8,$$

for n. Given that $\mathbb{P}(Z > c)$=.8 $\Rightarrow c$= − .842, $[1.645-\sqrt{n}(.2)]$ = − .841 $\Rightarrow n$=155. Hence, the question 'how large the sample size should be to learn from data about phenomena of interest' is of critical importance in practice. In this case, there is no point in carrying out this study when n=100; one would not learn whether the discrepancy of interest γ=.2 is present or not.

Revisiting the large n problem. Figure 13.12 illustrates how the notion of power can be used to explain the large n problem. Since the power for detecting

any discrepancy $\gamma > 0$ increases with n, there is always a large enough n to reject any null hypothesis $\mu{=}\mu_0$, unless $\mu_0{=}\mu^*$, which can happen only rarely. Hence, even tiny discrepancies from the null, say $(\mu_1{-}\mu_0){=}.000001$, will give rise to the large n problem. It is important to emphasize that there is nothing paradoxical about the power increasing with n; this is what constitutes a consistent test. What is important to take away from this is that both, the p-value and a N-P rejection of H_0, are vulnerable to the large n problem. This renders both inappropriate in providing an evidential interpretation of inference results without being adapted to account for the generic capacity of a test (power).

In summary, it is very important to emphasize three features of an optimal test.

First, an N-P test probes within the boundaries of a particular $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ whose statistical adequacy is presumed. Any departures from the model assumptions are likely to distort the test's error probabilities, resulting in sizeable discrepancies between the nominal (assumed) and actual error probabilities.

Second, an N-P a test is not just a formula associated with particular statistical tables. It is a combination of a test statistic and a rejection region; hence the notation $\mathcal{T}_\alpha{:=}\{d(\mathbf{X}),\ C_1(\alpha)\}$.

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

Example 13.14. In the case of the simple (one parameter) Normal model (table 13.10), consider replacing $\overline{X}_n$ with the second rate unbiased estimator $\widehat{\mu}_2{=}\frac{1}{2}(X_1{+}X_n) \backsim \mathsf{N}(\mu, \frac{\sigma^2}{2})$. The resulting test:

$$\check{T}_\alpha{:=}\{\zeta(\mathbf{X}),\ C_1(\alpha)\},\ \text{for}\ \zeta(\mathbf{X}){=}\tfrac{\sqrt{2}(\widehat{\mu}_2-\mu_0)}{\sigma}\ \text{and}\ C_1(\alpha){=}\{\mathbf{x}{:}\ \zeta(\mathbf{x}){>}c_\alpha\},$$

will *not* be optimal because $\check{T}_\alpha$ is *inconsistent* and its power is much lower than that of $\mathcal{T}_\alpha$! This is because the non-zero mean of the sampling distribution under $\mu{=}\mu_1$ will be $d{=}\frac{\sqrt{2}(\mu_1-\mu_0)}{\sigma}$ and does not change as $n \to \infty$.

Third, it is important to note that by changing the rejection region one can render an optimal N-P test useless! For instance, replacing the rejection region of $\mathcal{T}_\alpha{:=}\{\kappa(\mathbf{X}),\ C_1(\alpha)\}$ with $\overline{C}_1(\alpha){=}\{\mathbf{x}{:}\ \kappa(\mathbf{x}) < c_\alpha\}$, the resulting test $\breve{\mathcal{T}}_\alpha{:=}\{\kappa(\mathbf{X}),\ \overline{C}_1(\alpha)\}$ is practically useless because, in addition to being *inconsistent* (its power does not increase as $n \to \infty$), it is also *biased* (it rejects H_0 when it is true than when it is false) and its power decreases as the discrepancy γ increases.

Example 13.15: Student's t test. In the context of the simple Normal model (table 13.4), consider the 1-sided hypotheses:

$$\begin{aligned} &\text{(1-s>)}: \quad H_0{:}\ \mu{=}\mu_0, \quad \text{vs.} \quad H_1{:}\ \mu > \mu_0 \\ \text{or}\quad &\text{(1-s*)}: \quad H_0{:}\ \mu \leq \mu_0, \quad \text{vs.} \quad H_1{:}\ \mu > \mu_0, \end{aligned} \tag{30}$$

In this case, a UMP (consistent) test $\mathcal{T}_\alpha{:=}\{\tau(\mathbf{X}),\ C_1(\alpha)\}$ is the well-known *Student's*

t test:

$$\begin{array}{l}\text{test statistic: } \tau(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{s},\ s^2=\frac{1}{n-1}\sum_{k=1}^{n}(X_k-\overline{X}_n)^2,\\ \text{rejection region: } C_1(\alpha)=\{\mathbf{x}\text{: } \tau(\mathbf{x})>c_\alpha\},\end{array} \tag{31}$$

where c_α can be evaluated using the Student's t tables; see table 13.5.

To evaluate the two types of error probabilities the distribution of $\tau(\mathbf{X})$ under both the null and alternatives are needed:

$$\begin{array}{l}\text{[I]* } \tau(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}\overset{\mu_0}{\backsim}\mathsf{St}(n-1),\\ \text{[II]* } \tau(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{s}\overset{\mu_1}{\backsim}\mathsf{St}(\delta_1;n-1),\ \text{for all } \mu_1>\mu_0,\end{array} \tag{32}$$

where $\delta_1=\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$ is the *non-centrality parameter*.

Table 13.15 summarizes the main components of a Neyman-Pearson (N-P) test $\{\tau(\mathbf{X}),C_1(\alpha)\}$. When this table is put side-by-side to Fisher's testing table 13.11, it is clear that the elements (a), (c) and (d) are identical, and the changes stem primarily from bringing into the testing the whole of the parameter space Θ and replacing the p-value (post-data) with the significance level (pre-data) error probabilities. As argued in the sequel, the N-P approach to testing can be seen as complementary to Fisher's significance testing, and the two perspectives can be accommodated within the broader error-statistical framework for frequentist testing; see section 5.

Table 13.15: N-P testing - key elements

(a)	a prespecified (parametric) statistical model: $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$,
(b)	a null (H_0: $\theta\in\Theta_0$) and the *alternative* (H_1: $\theta\in\Theta_1$) *within* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$,
(c)	a test statistic (distance function) $\tau(\mathbf{X})$,
(d)	the distribution of $\tau(\mathbf{X})$ under H_0 is known,
(e)	a prespecified significance level α [.01, .025, .05],
(f)	a rejection region $C_1(\alpha)$,
(g)	the distribution of $\tau(\mathbf{X})$ under H_1, i.e. for all $\theta\in\Theta_1$.

4.6 Constructing optimal tests: the N-P lemma

The cornerstone of the N-P approach is the *Neyman-Pearson lemma*. Assume the simple generic statistical model:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\theta)\},\ \theta\in\Theta:=\{\theta_0,\theta_1\}\},\ \mathbf{x}\in\mathbb{R}^n_X, \tag{33}$$

and consider the problem of testing the simple hypotheses:

$$H_0\text{: } \theta=\theta_0 \quad \text{vs.} \quad H_1\text{: } \theta=\theta_1. \tag{34}$$

Existence. There is exists an α-significance level *Uniformly Most Powerful* (UMP) [α-UMP] test, whose generic form is:

$$d(\mathbf{X})=h(\tfrac{f(\mathbf{x};\theta_1)}{f(\mathbf{x};\theta_0)}),\quad C_1(\alpha)=\{\mathbf{x}\colon\, d(\mathbf{x})>c_\alpha\}, \tag{35}$$

where $h(.)$ is a *monotone* function.
Sufficiency. If an α-level test of the form (35) exists, then it is UMP for testing (34).
Necessity. If $\{d(\mathbf{X}),C_1(\alpha)\}$ is α-UMP test, then it will be given by (35).

Note that to implement this result one would need to find a function $h(.)$ yielding a test statistic $d(\mathbf{X})$ whose sampling distribution is known under both H_0 and H_1. The idea of using the ratio $\frac{f(\mathbf{x};\theta_1)}{f(\mathbf{x};\theta_0)}$ as a basis for constructing a test was credited by Egon Pearson (1939) to Gosset in an exchange they had in 1926: "It is the simple suggestion [by Gosset] that the only valid reason for rejecting a statistical hypothesis is that some alternative explains the observed events with a greater degree of probability."

■ At first sight the N-P lemma seems overly simplistic because it assumes a simple statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ whose parameter space has only two points $\Theta{:=}\{\theta_0,\theta_1\}$, which seems totally contrived. Unfortunately, this particular detail is often ignored in some statistics textbook discussions of this lemma. Note, however, that it fits perfectly into the *archetypal* formulation because the two points constitute a partition of Θ. This lemma is often misconstrued as suggesting that for an α-UMP test to exist one needs to confine testing to simple-vs-simple hypotheses even when Θ is uncountable. The truth of the matter is that the construction of an α-UMP test in more realistic cases has nothing to do with simple-vs-simple hypotheses. Instead, it is invariably connected to the *archetypal* N-P testing formulation in (23), and relies primarily on *monotone likelihood ratios* (Karlin and Rubin, 1956) and other features of the prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ for the existence of an α-UMP test.

Example 13.16. To illustrate these comments consider the simple hypotheses:

$$\text{(i) (1-1): } H_0\colon\, \mu{=}\mu_0 \quad \text{vs.} \quad H_1\colon\, \mu{=}\mu_1, \text{ for } \mu_1>\mu_0, \tag{36}$$

in the context of a simple Normal (one parameter) model (table 13.10). This does not satisfy the conditions of the N-P lemma because the parameter space is the whole of the real line; not just two points. Regardless, let us apply the N-P lemma by constructing the ratio using just the two points in (36):

$$\tfrac{f(\mathbf{x};\mu_1)}{f(\mathbf{x};\mu_0)}{=}\exp\left\{\tfrac{n}{\sigma^2}(\mu_1{-}\mu_0)\overline{X}_n{-}\tfrac{n}{2\sigma^2}(\mu_1^2{-}\mu_0^2)\right\}=\exp\left\{\left(\tfrac{n(\mu_1{-}\mu_0)}{\sigma^2}\right)\left(\overline{X}_n{-}\tfrac{\mu_1{+}\mu_0}{2}\right)\right\}. \tag{37}$$

This ratio is clearly *not* a test statistic, as it stands, but it can be transformed into one by noticing that since $(\mu_1{-}\mu_0)>0$ a rejection region $C_1(\alpha){=}\{\mathbf{x}\colon\, \frac{f(\mathbf{x};\mu_1)}{f(\mathbf{x};\mu_0)}{>}c\}$ is equivalent to:

$$\overline{X}_n>c^*{=}\left(\tfrac{\sigma^2}{n(\mu_1{-}\mu_0)}\right)\ln(c)+\left(\tfrac{\mu_1{+}\mu_0}{2}\right). \tag{38}$$

Since c^* is an arbitrary positive constant, it is clear that the relevant test statistic should be a function of $\overline{X}_n$ whose sampling distribution is known under both the null

and alternatives hypotheses. Given that $\overline{X}_n \backsim \mathsf{N}\Big(\mu, \frac{\sigma^2}{n}\Big)$, (37) can be transformed into a familiar test statistic:

$$\kappa(\mathbf{X})=h(\tfrac{f(\mathbf{x};\mu_1)}{f(\mathbf{x};\mu_0)})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma},$$

where the relevant sampling distributions are:

$$\kappa(\mathbf{X}) \overset{\mu=\mu_0}{\backsim} \mathsf{N}(0,1), \quad \kappa(\mathbf{X}) \overset{\mu=\mu_1}{\backsim} \mathsf{N}(\delta_1,1), \ \delta_1=\tfrac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}.$$

These provide the basis for evaluating the type I and II error probabilities.

In this particular case, it turns out that when the ratio $\frac{f(\mathbf{x};\mu_1)}{f(\mathbf{x};\mu_0)}$ is a monotone function of the test statistic $\kappa(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma}$, that can provide the basis for constructing optimal tests even when the null and alternative hypotheses are composite (Lehmann and Romano, 2006). Let us consider certain well-known cases.

[1] For $\mu_1 > \mu_0$, the test $T^{>}_{\alpha}$:=$\{\kappa(\mathbf{X}), C^{>}_1(\alpha)\}$, where
$C^{>}_1(\alpha)$=$\{\mathbf{x}\colon\ \kappa(\mathbf{x}) > c_\alpha\}$, is a α-UMP for the hypotheses:

$$\text{(ii) (1-s}\geq\text{)}: \quad H_0\colon\ \mu \leq \mu_0 \quad \text{vs.} \quad H_1\colon\ \mu > \mu_0,$$

$$\text{(iii) (1-s}>\text{)}: \quad H_0\colon\ \mu=\mu_0 \quad \text{vs.} \quad H_1\colon\ \mu > \mu_0.$$

Despite the difference between (ii) and (iii), the relevant error probabilities coincide. This is because the type I error probability for (ii) is defined as the maximum over all $\mu \leq \mu_0$, which happens to be the end point (μ=μ_0):

$$\alpha=\max_{\mu\leq\mu_0}\mathbb{P}(\kappa(\mathbf{X}) > c_\alpha; H_0(\mu))=\mathbb{P}(\kappa(\mathbf{X}) > c_\alpha; \mu=\mu_0). \tag{39}$$

Similarly, the p-value is the same because:

$$p_{>}(\mathbf{x}_0)=\max_{\mu\leq\mu_0}\mathbb{P}(\kappa(\mathbf{X}) > \kappa(\mathbf{x}_0); H_0(\mu))=\mathbb{P}(\kappa(\mathbf{X}) > \kappa(\mathbf{x}_0); \mu=\mu_0). \tag{40}$$

[2] For $\mu_1 < \mu_0$, the test $T^{<}_{\alpha}$:=$\{\kappa(\mathbf{X}), C^{<}_1(\alpha)\}$, where
$C^{<}_1(\alpha)$=$\{\mathbf{x}\colon\ \kappa(\mathbf{x}) < c_\alpha\}$, is a α-UMP for the hypotheses:

$$\text{(iv) (1-s}\leq\text{)}: \quad H_0\colon\ \mu \geq \mu_0 \quad \text{vs.} \quad H_1\colon\ \mu < \mu_0,$$

$$\text{(v) (1-s}<\text{)}: \quad H_0\colon\ \mu=\mu_0 \quad \text{vs.} \quad H_1\colon\ \mu < \mu_0.$$

Again, the type I error probability and p-value are defined by:

$$\alpha=\max_{\mu\geq\mu_0}\mathbb{P}(\kappa(\mathbf{X}) < c_\alpha; H_0(\mu))=\mathbb{P}(\kappa(\mathbf{X}) < c_\alpha;\ \mu=\mu_0).$$

$$p_{<}(\mathbf{x}_0)=\max_{\mu\geq\mu_0}\mathbb{P}(\kappa(\mathbf{X})<\kappa(\mathbf{x}_0); H_0(\mu))=\mathbb{P}(\kappa(\mathbf{X})<\kappa(\mathbf{x}_0); \mu=\mu_0).$$

4.7 Extending the Neyman-Pearson (N-P) lemma

Given how artificial is the parameter space assumed by the N-P lemma, the question that naturally arises is how one can extend the optimality theory to more realistic cases encountered in practice. The existence of these α-UMP tests can be achieved using two *regularity conditions*.

[A] **Monotonicity**. The ratio (37) is a *monotone function* of the statistic $\overline{X}_n$, in the sense that for any two values $\mu_1>\mu_0$, $\frac{f(\mathbf{x};\mu_1)}{f(\mathbf{x};\mu_0)}$ changes monotonically with $\overline{X}_n$. This implies that $\frac{f(\mathbf{x};\mu_1)}{f(\mathbf{x};\mu_0)}>k$ if and only if $\overline{X}_n>c$ for some c. The general result is that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ has a monotone likelihood ratio of the form:

$$\frac{f(\mathbf{x};\theta_1)}{f(\mathbf{x};\theta_0)}=h(s(\mathbf{x});\theta_0,\theta_1), \tag{41}$$

where for all $\theta_1>\theta_0$, $h(s(\mathbf{x});\theta_0,\theta_1)$ is a non-decreasing function of the statistic $s(\mathbf{X})$ upon which the test statistic will be based; see Karlin and Rubin (1956).

This regularity condition is valid for most statistical models of interest in practice, including the 1-parameter Exponential family of distributions [Normal, Gamma, Beta, Binomial, Negative Binomial, Poisson, etc.], the Uniform, the Exponential, the Logistic, the Hypergeometric etc. It is interesting to note that in his first published paper after the Neyman-Pearson 1933 classic, Fisher (1934) proved the existence of UMP tests in the case of this family as stemming from the property of sufficiency. It was the first and last time Fisher discussed the N-P approach to testing in a positive light.

[B] **Convexity**. The parameter space Θ_1 under H_1 is convex, i.e. for any two values $(\mu_1,\mu_2)\in\Theta_1$, their convex combinations $\lambda\mu_1+(1-\lambda)\mu_2\in\Theta_1$, for any $0\leq\lambda\leq 1$.

When convexity does not hold, like the 2-sided alternative:

$$\text{(vi) (2-s):}\quad H_0\text{: } \mu=\mu_0 \quad \text{vs.} \quad H_1\text{: } \mu\neq\mu_0, \tag{42}$$

[3] the test $T^{\neq}_{\alpha}:=\{\kappa(\mathbf{X}),C_1(\alpha)\}$, $C_1(\alpha)=\{\mathbf{x}\text{: } |\kappa(\mathbf{x})|>c_{\frac{\alpha}{2}}\}$,

is α-UMPU (*Unbiased*); the α-level and p-value are:

$$\alpha=\mathbb{P}(|\kappa(\mathbf{X})|>c_{\frac{\alpha}{2}};\ \mu=\mu_0),\quad p_{\neq}(\mathbf{x}_0)=\mathbb{P}(|\kappa(\mathbf{X})|>|\kappa(\mathbf{x}_0)|;\mu=\mu_0). \tag{43}$$

Example 13.17. Consider the simple Normal model (table 13.10), and the test:

$$T^{>}_{\alpha}:=\{\kappa(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma},\quad C_1(\alpha)=\{\mathbf{x}\text{: } \kappa(\mathbf{x})>c_{\frac{\alpha}{2}}\},$$

to the two-sided hypotheses in (42). Test $\mathcal{T}_{\alpha}$ is not UMP because for negative values of the discrepancy its power is below $\alpha/2$ (fig. 13.13). Indeed, test $T^{>}_{\alpha}$ is biased. When one narrows the group of desirable tests to include unbiasedness, test $T^{>}_{\alpha}$ is

excluded from consideration for being biased in testing the two-sided hypotheses (42). On the other hand, the two-sided version of $T_\alpha^>$:

$$T_\alpha^{\neq}:=\{\kappa(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{\sigma},\quad C_1(\alpha)=\{\mathbf{x}:\ |\kappa(\mathbf{x})|>c_{\frac{\alpha}{2}}\},$$

for testing (42), has lower power for positive discrepancies, but it is unbiased (UMPU); see fig. 13.14.

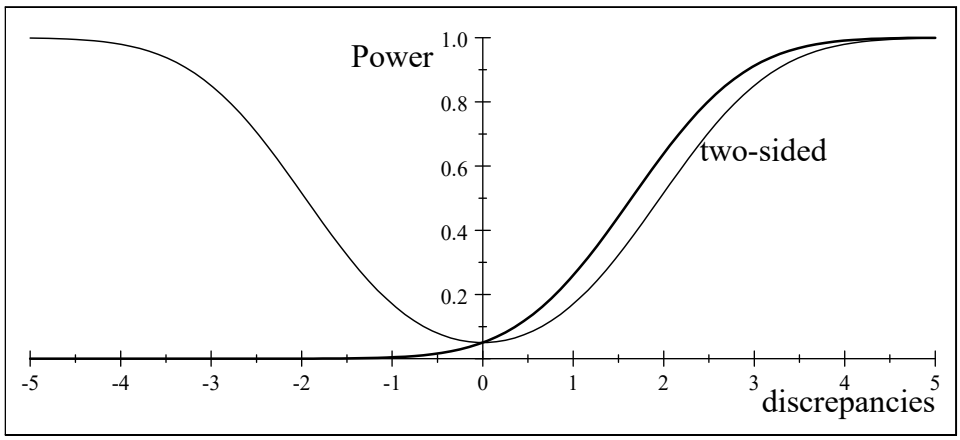


Fig. 13.14: Power of test $T_\alpha^>$ for one-sided (bold) and two-sided (42)

An alternative way to proceed with a view to use the most powerful test and sidestep the bias is to separate the two-sided alternative into two pairs of one-sided hypotheses (Cox and Hinkley, 1974):

$$\text{(i) } H_0\text{: } \mu\leq\mu_0 \text{ vs. } H_1\text{: } \mu>\mu_0,\quad \text{(ii) } H_0\text{: } \mu\geq\mu_0 \text{ vs. } H_1\text{: } \mu<\mu_0,$$

and use the one-sided tests $T_\alpha^>$ and $T_\alpha^<$ to evaluate the relevant p-values, say $p_>(\mathbf{x}_0)$ and $p_<(\mathbf{x}_0)$. The inference result is then based on:

$$p(\mathbf{x}_0)=\min\left(p_>(\mathbf{x}_0),p_<(\mathbf{x}_0)\right).$$

■ **Randomization in testing**? In cases where the test statistic has a *discrete* sampling distribution under H_0, as in (**??**), one might not be able to define α exactly. The traditional way to deal with this problem is to *randomize* (Lehmann and Romano 2006), but this 'solution' raises more problems than it solves. Instead, the best way to address the discreteness issue is to select a value α that is attained for the given n, or approximate the discrete distribution with a continuous one to circumvent the problem. In practice, there are several continuous distributions one can use, depending on whether the original distribution is symmetric or not. In the above case, even for moderately small sizes, the Normal distribution provides a good approximation for the distribution of $(\frac{\overline{X}_n-\theta}{\sqrt{\theta}})$ using the Central Limit Theorem; see chapter 9.

4.8 Constructing optimal tests: Likelihood Ratio

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 of the 'likelihood' ratio:

$$\lambda_n(\mathbf{X})=\frac{\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta};\mathbf{X})}{\max_{\boldsymbol{\theta}\in\Theta_0} L(\boldsymbol{\theta};\mathbf{X})}=\frac{L(\widehat{\boldsymbol{\theta}};\mathbf{X})}{L(\widetilde{\boldsymbol{\theta}};\mathbf{X})} \tag{44}$$

Note that the max in the numerator is over all $\boldsymbol{\theta}\in\Theta$ [yielding the MLE $\widehat{\boldsymbol{\theta}}$], but that of the denominator is confined to all values under H_0: $\boldsymbol{\theta}\in\Theta_0$ [yielding the constrained MLE $\widetilde{\boldsymbol{\theta}}$]. This is in contrast to the N-P lemma where the numerator is evaluated under the alternative H_1.

(c) The generic rejection region is defined by: $C_1=\{\mathbf{x}\text{: } \lambda_n(\mathbf{x})>c\}$,

but it is almost never the case that the distribution of $\lambda_n(\mathbf{X})$ under H_0 is known. 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:

$$C_1(\alpha)=\{\mathbf{x}\text{: } \tau(\mathbf{X})=h\left(\lambda_n(\mathbf{X})\right)>c_\alpha\}. \tag{45}$$

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

Example 13.21. In the context of a simple Normal model (table 13.4), consider the simple-vs-composite hypotheses:

$$H_0\text{: } \mu=\mu_0 \quad \text{vs.} \quad H_1\text{: } \mu\neq\mu_0, \tag{46}$$

The generic likelihood takes the form: $L(\boldsymbol{\theta};\mathbf{x})=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2\right\}$, giving rise to the numerator of (44):

$$\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta};\mathbf{X})=L(\widehat{\boldsymbol{\theta}};\mathbf{X})=(2\pi\widehat{\sigma}^2)^{-\frac{n}{2}}\exp(-\tfrac{1}{2}n), \tag{47}$$

that denotes the likelihood function evaluated at the unconstrained MLE estimators:

$$\widehat{\mu}=\overline{X}_n, \quad \widehat{\sigma}^2=\tfrac{1}{n}\textstyle\sum_{i=1}^n(x_i-\widehat{\mu})^2. \tag{48}$$

Using (20), the denominator of (44) takes the form:

$$\max_{\boldsymbol{\theta}\in\Theta_0} L(\boldsymbol{\theta};\mathbf{X})=L(\widetilde{\boldsymbol{\theta}};\mathbf{X})=\{2\pi[\widehat{\sigma}^2+(\overline{X}_n-\mu_0)^2]\}^{-\frac{n}{2}}\exp(-\tfrac{1}{2}n). \tag{49}$$

which is evaluated at the *constrained* (under H_0) MLE are:

$$\widetilde{\mu}_0=\mu_0, \quad \widetilde{\sigma}_0^2=\tfrac{1}{n}\textstyle\sum_{i=1}^n(x_i-\mu_0)^2=\widehat{\sigma}^2+(\overline{X}_n-\mu_0)^2. \tag{50}$$

Substituting (47)-(49) into (44) yields:

$$\lambda_n(\mathbf{X})=\frac{\left[(2\pi\widehat{\sigma}^2)^{-\frac{n}{2}}\right]\exp(-\frac{1}{2}n)}{\left[\{2\pi[\widehat{\sigma}^2+(\overline{X}_n-\mu_0)^2]\}^{-\frac{n}{2}}\right]\exp(-\frac{1}{2}n)}=\left\{\frac{\widehat{\sigma}^2}{\widehat{\sigma}^2+(\overline{X}_{nt}-\mu_0)^2}\right\}^{-\frac{n}{2}}=\left\{1+\frac{(\overline{X}_n-\mu_0)^2}{\widehat{\sigma}^2}\right\}^{\frac{n}{2}}\rightarrow$$

$$\rightarrow \lambda_n^{\frac{2}{n}}(\mathbf{X})=\left\{1+\frac{\frac{n}{(n-1)}(\overline{X}_n-\mu_0)^2}{\frac{n}{(n-1)}\widehat{\sigma}^2}\right\}=\left\{1+\frac{n(\overline{X}_n-\mu_0)^2}{(n-1)s^2}\right\}=\left\{1+\frac{\tau(\mathbf{X})^2}{(n-1)}\right\}$$

where $\tau(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{s}$, $s^2=\frac{1}{n-1}\sum_{i=1}^n(x_i-\widehat{\mu})^2$. Hence, rejecting H_0 when $\lambda_n(\mathbf{X})=\left\{1+\frac{(\overline{X}_n-\mu_0)^2}{\widehat{\sigma}^2}\right\}^{\frac{n}{2}}>c$ implies that $\left\{1+\frac{\tau(\mathbf{X})^2}{(n-1)}\right\}>b=c^{\frac{2}{n}}$.

This indicates that $\lambda_n(\mathbf{X})$ is a monotonically increasing function of $\tau(\mathbf{X})^2$, which implies that one can use the known distribution of $\tau(\mathbf{X})$ to define the two-sided t test for (46):

$$\{\tau(\mathbf{X})=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{s},\ C_1(\alpha)=\{\mathbf{x}:\ |\tau(\mathbf{x})|>c_{\frac{\alpha}{2}}\},$$

which can be shown to be UMP Unbiased; see Lehmann and Romano (2005). Note that the sampling distribution of $\tau(\mathbf{X})^2$ is also known, $\tau(\mathbf{X})^2\overset{H_0}{\backsim}\mathsf{F}(1,n-1)$, where $\mathsf{F}(1,n-1)$ denotes the F (for Fisher) distribution with 1 and $(n-1)$ degrees of freedom. This is because when $v\backsim\mathsf{St}(m)$, then $v^2\backsim\mathsf{F}(1,m)$; see Lehmann and Romano (2005). Hence, in principle one could use the F test:

$$\left\{\tau(\mathbf{X})^2=\frac{n(\overline{X}_n-\mu_0)^2}{s^2},\ C_1(\alpha)=\{\mathbf{x}:\ \tau(\mathbf{X})^2>c_\alpha\}\right\},$$

since they will give rise to the same inference result.

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 $\underset{\alpha}{\overset{H_0}{\sim}}$ 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:

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

5 Error-statistical framing of statistical testing

5.1 N-P testing driven by substantively relevant values

The primary aim of this section is to raise several issues when N-P testing is driven by one or more substantive values of interest for the parameters, as a prelude to introducing the error-statistical framing of frequentist testing.

Let us focus on two values of historical interest pertaining the ratio of Boys (B) to Girls (G) in newborns (see Gorroochurn, 2016):

$$\textbf{Arbuthnot}\text{: \#B=\#G}, \quad \textbf{Bernoulli}\text{: 18B to 17G.} \tag{51}$$

These values can be embedded into a simple Bernoulli model (table 13.8), where $\theta{=}\mathbb{P}(X{=}1){=}\mathbb{P}(\text{B})$:

$$\textbf{Arbuthnot}\text{: } \theta_A{=}\tfrac{1}{2}, \qquad \textbf{Bernoulli}\text{: } \theta_B{=}\tfrac{18}{35}. \tag{52}$$

The obvious formulation suggested by substantive arguments is:

$$H_0\text{: } \theta=\theta_A \text{ vs. } H_1\text{: } \theta=\theta_B,$$

which, in light of the above discussion, it is *illegitimate*, because the parameter space of the simple Bernoulli model is not $\Theta{=}\{\theta_A,\theta_B\}$, but $\Theta{=}[0,1]$. Hence, any invoking of the Neyman-Pearson (N-P) lemma to secure a α-UMP test is based on misinterpreting it. What secures a α-UMP test in this case is the monotonic likelihood ratio of the simple Bernoulli model:

$$\tfrac{f(\mathbf{x};\theta_1)}{f(\mathbf{x};\theta_0)}{=}\left(\tfrac{(1-\theta_1)}{(1-\theta_0)}\right)^n\left(\tfrac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum_{k=1}^n x_k}, \text{ where } 0<\theta_0<\theta_1<1,$$

that is a *monotonically increasing function* of the statistic $n\overline{X}_n{=}\sum_{k=1}^n X_k$ since $\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}{>}1$. This ensures the existence of several α-UMP tests depending on the N-P formulation of the hypotheses of interest based on the test statistic $d(\mathbf{X})$ and its sampling distributions:

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

$$d(\mathbf{X}){=} \overset{\theta=\theta_1}{\backsim} \mathsf{Bin}\left(\delta(\theta_1),V(\theta_1);n\right), \text{ for } \theta_1>\theta_0,$$

$$\delta(\theta_1){=}\tfrac{\sqrt{n}(\theta_1-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}}\geq 0,\ V(\theta_1){=}\tfrac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)},\ 0<V(\theta_1)\leq 1.$$

In a N-P setup one is faced with several different legitimate formulations pertaining the two substantive values $\theta_A{=}\frac{1}{2}$ vs. $\theta_B{=}\frac{18}{35}$ defining H_0 and H_1 in conjunction with $\gamma{=}[\theta_B{-}\theta_A]$ being the primary discrepancy of interest, including the ones in table 13.16.The 1-sided tests, for (i), (iii) and (iv), can be shown to be α-level UMP tests,

and the 2-sided tests for (ii) and (iv) can be shown to be α-level UMPU tests; see Lehmann and Romano (2005).

Table 13.16: N-P formulations for θ_A=$\frac{1}{2}$ vs. θ_B=$\frac{18}{35}$

	Null and alternative	Rejection region
(i)	H_0: $\theta \leq \theta_A$ vs. H_1: $\theta > \theta_A$,	$\mathcal{C}_1^{>}(\alpha)$=$\{\mathbf{x}$: $d_A(\mathbf{x}) > c_\alpha\}$
(ii)	H_0: $\theta = \theta_A$ vs. H_1: $\theta \neq \theta_A$,	$\mathcal{C}_1(\alpha)$=$\{\mathbf{x}$: $\lvert d_A(\mathbf{x})\rvert < c_{\frac{\alpha}{2}}\}$
(iii)	H_0: $\theta \geq \theta_B$ vs. H_1: $\theta < \theta_B$,	$\mathcal{C}_1^{<}(\alpha)$=$\{\mathbf{x}$: $d_B(\mathbf{x}) < c_\alpha\}$
(iv)	H_0: $\theta = \theta_B$ vs. H_1: $\theta \neq \theta_B$,	$\mathcal{C}_1(\alpha)$=$\{\mathbf{x}$: $\lvert d_B(\mathbf{x})\rvert < c_{\frac{\alpha}{2}}\}$
(v)	H_0: $\theta \leq \theta_B$ vs. H_1: $\theta > \theta_B$,	$\mathcal{C}_1^{>}(\alpha)$=$\{\mathbf{x}$: $d_B(\mathbf{x}) > c_\alpha\}$

Example 13.25: Arbuthnot's value. Let us test the hypotheses (i) and (ii) the simple Bernoulli model (table 13.8), using the data come in the form of n=30862 newborns during the period 1993-5 in Cyprus, 16029 boys and 14833 girls. In view of the huge sample size it is advisable to choose a smaller significance level, say α=.01 $\Rightarrow$ c_α=2.326, $c_{\frac{\alpha}{2}}$=2.575. The test statistic based on θ_A=θ_0=.5 yields:

$$d_A(\mathbf{x}_0)=\frac{\sqrt{30862}(\frac{16029}{30862}-\frac{1}{2})}{\sqrt{.5(.5)}}=6.808,$$

and thus the p-values take the form:

$$\begin{aligned} &\text{(i)} \quad p_{A>}(\mathbf{x}_0)=\mathbb{P}(d_A(\mathbf{X}) > 6.808;\ H_0)=4.95\times10^{-14} < \alpha=.01, \\ &\text{(ii)} \quad p_{A\neq}(\mathbf{x}_0)=\mathbb{P}(|d_A(\mathbf{X})| >6.808;\ H_0)=9.9\times10^{-12}<\tfrac{\alpha}{2}=.005, \end{aligned} \tag{53}$$

and thus, in both cases H_0: θ=$\frac{1}{2}$ is strongly rejected with tiny p-values.

Example 13.26: **Bernoulli's value**. Testing the Bernoulli value θ_B=θ_0=$\frac{18}{35}$ for the hypotheses (iii)-(v), using the same data as in example 13.24, yields:

$$d_B(\mathbf{x}_0)=\frac{\sqrt{30862}(\frac{16029}{30862}-\frac{18}{35})}{\sqrt{\frac{18}{35}(1-\frac{18}{35})}}=1.789,$$

: 1. 789 4and thus the p-values take the form:

$$\begin{aligned} &\text{(iii)} \quad p_{B<}(\mathbf{x}_0)=\mathbb{P}(d_B(\mathbf{X}) < 1.789;\ H_0)=.963 > \alpha=.01, \\ &\text{(iv)} \quad p_{B\neq}(\mathbf{x}_0)=\mathbb{P}(|d_B(\mathbf{X})| > 1.789;\ H_0)=.073 > \tfrac{\alpha}{2}=.005 \\ &\text{(v)} \quad p_{B>}(\mathbf{x}_0)=\mathbb{P}(d_B(\mathbf{X}) > 1.789;\ H_0)=.036 > \alpha=.01. \end{aligned} \tag{54}$$

The results in (54) seem rather puzzling because for the hypotheses (iii)-(iv), H_0: θ_0=$\frac{18}{35}$ is accepted, but for (v) it is rejected. More puzzling are the three p-values which vary widely! Having used the same data in five different N-P formulations of the null and alternative hypotheses, with each case giving rise to a different p-value,

the first question that arises is which one is the relevant p-value. The above discussion suggests that $p_{B\neq}(\mathbf{x}_0)$, $p_{A\neq}(\mathbf{x}_0)$ and $p_{B<}(\mathbf{x}_0)$ make no post-data sense, but $p_{A>}(\mathbf{x}_0)$ and $p_{B>}(\mathbf{x}_0)$ do; ensure you know why! Having settle that, the second question pertains to whether data $\mathbf{x}_0$ provide evidence for or against the substantive values $\theta_A{=}\frac{1}{2}$ and $\theta_B{=}\frac{18}{35}$. A clear answer to this question will be given in the sequel using the post-data severity evaluation.

5.2 Foundational issues pertaining to statistical testing

5.2.1 Fisher and N-P testing

Let us now appraise how successful the N-P framework was in addressing the perceived weaknesses of Fisher's testing:

[a] Fisher's choice of a test statistics $d(\mathbf{X})$ based on intuitive grounds,

[b] his use of a *post-data* [$d(\mathbf{x}_0)$ is known] threshold in conjunction with the p-value that indicates discordance (reject) with H_0, and

[c] his denial that $d(\mathbf{x}_0)$ can indicate *accordance* with H_0.

Very briefly, the N-P framework partly succeeded in addressing [a] and [c], but did not provide a *coherent evidential account* that answers the basic question (Mayo, 1996): *when do data* $\mathbf{x}_0$ provide evidence for or against a hypothesis H? This is primarily because one could not interpret the N-P results 'Accept H_0' ('Reject H_0') as data $\mathbf{x}_0$ provide evidence *for (against)* H_0 (for H_1). Why?

(i) A particular test $\mathcal{T}_\alpha{:=}\{d(\mathbf{X}), \mathcal{C}_1(\alpha)\}$ could have led to 'Accept H_0' because the power of that test to detect an existing discrepancy γ was very low. This can easily happen when the sample size n is small.

(ii) A particular test $\mathcal{T}_\alpha{:=}\{d(\mathbf{X}), \mathcal{C}_1(\alpha)\}$ could have led to 'Reject H_0' simply because the power of that test was high enough to detect 'trivial' discrepancies from H_0. This can easily happen when the sample size n is very large.

In light of the fact that the N-P framework did not provide a bridge between the accept/reject rules and questions of interest of the scientific inquiry, i.e. when data $\mathbf{x}_0$ provide evidence for or against a hypothesis H, it should come as no surprise to learn that most practitioners sought such answers by trying unsuccessfully (and misleadingly) to distill an evidential interpretation out of Fisher's p-value. In their eyes the pre-designated α is equally vulnerable to manipulation [pre-designation to keep practitioners 'honest' is unenforceable in practice], but the p-value is more informative and data-specific; see Lehmann and Romano (2005).

A crucial problem with the p-value is that a small (large) p-value could not be interpreted as evidence for the presence (absence) of a substantive discrepancy γ for the same reasons as (i)-(ii). The power of the test affects the p-value. For instance, a very small p-value can easily arise in the case of a very large sample size n.

5.2.2 The large n problem

Mayo (2006, p. 809) described the problem as follows: "for any discrepancy from the null, however small, one can find a sample size such as there is a high probability (as high as one likes) that the test will yield a statistically significant result (for any p-value one wishes)." The above comments raise a crucial problem with the current practice in empirical modeling, which ignores the fundamental trade-off between the type I and II error probabilities, by detaching the inference result – accept/reject H_0 or/and small or large p-value – from the particular n and $\mathcal{T}_\alpha$. As shown above, for a particular α an optimal N-P test has power that increases with n. Hence, in practice a rejection at α=.05 with n=25 is very different from a rejection with n=2500 from the evidential perspective. That is, for the same discrepancy from the null (θ_0), say $\gamma_1=\theta_1-\theta_0$, however small, the power of a $\mathcal{T}_\alpha$ test is, say $\pi(\gamma_1)$=.2, with n=25, it could easily become $\pi(\gamma_1)$=1 with n=2500. It turns out that both the Fisher and N-P accounts of testing are vulnerable to this problem, precluding any principled evidential interpretation of their inference results. A .05 statistical significance or a p-value of .04 depends crucially on n, which suggests that such results cannot be detached from the particular context.

■ The dependence of the p-value on the power of the test, and in particular n, calls into question the disputes concerning the relevance of the alternative and the type II error probabilities. When the testing takes place within the boundaries of a prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, the alternative hypothesis is automatically the complement to the null relative to the particular Θ, and any attempts to exorcize the power of the test as irrelevant are misplaced.

To counter the decrease in the p-value as n increases, some textbooks advise practitioners to use rules of thumb based on decreasing α for larger and larger sample sizes; see Lehmann and Romano (2006). Good (1988) proposes to standardize the p-value $p(\mathbf{x}_0)$ to the fixed sample size n=100 using the rule of thumb:

$$p_{100}(\mathbf{x}_0)=\min\left(.5,\left[p(\mathbf{x}_0)\cdot\sqrt{n/100}\right]\right),\ n>10.$$

Example 13.27. $p(\mathbf{x}_0)$=.04 for n=1000 corresponds to $p_{100}(\mathbf{x}_0)$=.126. The severity evaluation discussed below provides a more formal way to take into account the change in the sample size as it affects the power.

5.2.3 Fallacies of acceptance and rejection

The issues with the accept/reject H_0 and p-value results raised above can be formalized into two classic fallacies.

(a) The fallacy of acceptance: *no* evidence against H_0 is misinterpreted as evidence *for* H_0. This fallacy can easily arise in cases where the test in question has low power to detect discrepancies of interest.

(b) The fallacy of rejection: evidence *against* H_0 is misinterpreted as evidence *for* a particular H_1. This fallacy arises in cases where the test in question has high power

to detect substantively minor discrepancies. Since the power of a test increases with the sample size n, this renders N-P rejections, as well as tiny p-values, with large n, highly susceptible to this fallacy.

The above fallacies arise in statistics in a variety of different guises, including the distinction between *statistical and substantive significance*. A few textbooks in statistics warn readers that one should not conflate the two, but there are no principled ways to address the problem head on. In the statistics literature, as well as in the secondary literatures in several applied fields, there have been numerous attempts to circumvent these two fallacies, but none succeeded until recently. These fallacies can be addressed using the post-data severity evaluation of inference results (accept/reject, p-values) by offering an evidential account in the form of the discrepancy from the null warranted by the data; see Mayo and Spanos (2006, 2011), Mayo (2018).

5.3 Post-data severity evaluation: an evidential account

On reflection the above fallacies stem primarily from the fact that there *is* a problem when the p-value and the accept/reject H_0 results are detached from the test itself. That is, the results are viewed as providing the *same evidence* for a particular hypothesis H (H_0 or H_1), regardless of the generic capacity (the power) of the test in question to detect discrepancies from H_0. The intuition behind this reflection is that a small p-value or a rejection of H_0 based on a test with low power (e.g. a small n) for detecting a particular discrepancy γ *provides stronger evidence* for the presence of a particular discrepancy γ than using a test with much higher power (e.g. a large n). Mayo (1996) proposed a frequentist evidential account based on harnessing this intuition in the form of a **post-data severity evaluation** of the accept/reject results. This is based on custom-tailoring the generic capacity of the test to establish the discrepancy γ warranted by data $\mathbf{x}_0$. This evidential account can be used to circumvent the above fallacies, as well as other charges against frequentist testing.

The severity evaluation is a post-data appraisal of the accept/reject and p-value results that revolves around the discrepancy γ from H_0 warranted by data $\mathbf{x}_0$.

■ A hypothesis H passes a *severe test* $\mathcal{T}_\alpha$ with data $\mathbf{x}_0$ if:

(S-1) $\mathbf{x}_0$ accords with H, and

(S-2) with very high probability, test $\mathcal{T}_\alpha$ would have produced a result that accords less well with H than $\mathbf{x}_0$ does, if H were false.

Recall that the *post-data definition* of **the p-value** is the probability of all sample realizations $\mathbf{x}\in\mathbb{R}^n_X$ for which $d(\mathbf{x})$ *accords less well with* H_0 *than* $\mathbf{x}_0$ *does*, if H_0 were true.

Severity can be viewed as an feature of a test $\mathcal{T}_\alpha$ as it relates to a particular data $\mathbf{x}_0$ *and* a specific claim H being considered. Hence, the severity function has three arguments, $SEV(\mathcal{T}_\alpha, \mathbf{x}_0, H)$, denoting the severity with which H passes $\mathcal{T}_\alpha$ with $\mathbf{x}_0$; see Mayo and Spanos (2006), Mayo (2018).

To explain how the above severity evaluation can be applied in practice, let us return to the problem of assessing the two substantive values of interest $\theta_A{=}\frac{1}{2}$ and $\theta_B{=}\frac{18}{35}$. When these values are viewed in the context of the error statistical perspective, it becomes clear that the way to frame the probing is to choose one of the values as the null hypothesis, and let the difference between them $(\theta_B{-}\theta_A)$ represent the discrepancy of substantive interest.

Example 13.28. Probing the Arbuthnot value θ_A using the hypotheses:

$$\text{(i)} \quad H_0\text{: } \theta \leq \theta_A \text{ vs. } H_1\text{: } \theta > \theta_A,$$
$$\text{(ii)} \quad H_0\text{: } \theta = \theta_A \text{ vs. } H_1\text{: } \theta \neq \theta_A,$$

in example 13.25 gave rise to a *rejection* of H_0 at $\alpha{=}.01$ ($c_\alpha{=}2.326$) since:

$$d_A(\mathbf{x}_0){=}\frac{\sqrt{30762}(\frac{16029}{30762}-\frac{1}{2})}{\sqrt{.5(.5)}}{=}7.389,$$

$$\text{(i) } p_{A>}(\mathbf{x}_0){=}\mathbb{P}(d_A(\mathbf{X}) > 7.389;\ H_0){=}7.4{\times}10^{-14} < \alpha{=}.01,$$
$$\text{(ii) } p_{A\neq}(\mathbf{x}_0){=}\mathbb{P}(|d_A(\mathbf{X})| > 7.389;\ H_0){=}1.48{\times}10^{-13}{<}\tfrac{\alpha}{2}{=}.005.$$

An important feature of the severity evaluation is that it is *post-data*, and thus the sign of the observed test statistic $d(\mathbf{x}_0)$ provides information that indicates the *directional* inferential claims that 'passed'. In relation to the above example, the severity 'accordance' condition (S-1) implies that the rejection of $\theta_0{=}\frac{1}{2}$ with $d(\mathbf{x}_0){=}7.389{>}0$, indicates that the form of the inferential claim that 'passed' is of the generic form:

$$\theta > \theta_1{=}\theta_0{+}\gamma, \quad \text{for some } \gamma \geq 0. \tag{55}$$

The directional feature of the severity evaluation is very important in addressing several criticisms of N-P testing, including:

[a] switching between one-sided, two-sided or simple-vs-simple hypotheses,

[b] interchanging the null and alternative hypotheses, and

[c] manipulating the level of significance in an attempt to get the desired result.

To establish the particular discrepancy γ warranted by data $\mathbf{x}_0$, the severity post-data 'discordance' condition (S-2) calls for evaluating the probability of the event: "outcomes $\mathbf{x}$ that accord less well with $\theta{>}\theta_1$ than $\mathbf{x}_0$ does", i.e. $[\mathbf{x}\text{: } d(\mathbf{x}) \leq d(\mathbf{x}_0)]$:

$$SEV(T_\alpha^{>};\theta > \theta_1){=}\mathbb{P}(\mathbf{x}\text{: } d_A(\mathbf{x}) \leq d_A(\mathbf{x}_0); \theta > \theta_1 \text{ is false}). \tag{56}$$

The scenario $\theta > \theta_1$ is false allows for the possibility that any such inferential claim for a particular θ_1 is false, and to counter that one needs to evaluate the severity for all θ_1 in the direction indicated by the data. Hence, in practice $\mathsf{Sev}(T_\alpha^{>};\mathbf{x}_0;\theta{>}\theta_1)$ is evaluated at $\theta{=}\theta_1$ for all possible values of θ_1:

$$SEV(T_\alpha^{>};\theta > \theta_1){=}\mathbb{P}(\mathbf{x}\text{: } d_A(\mathbf{x}) \leq d_A(\mathbf{x}_0);\ \theta{=}\theta_1),\ \text{for } \theta_1{=}\theta_0{+}\gamma, \text{ for all } \gamma \geq 0. \tag{57}$$

The evaluation of SEV for the above hypotheses in (i)-(ii) is based on:

$$\begin{array}{l} d(\mathbf{X})=\frac{\sqrt{n}(\hat{\theta}_n-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \overset{\theta=\theta_1}{\backsim} \mathsf{Bin}\left(\delta(\theta_1), V(\theta_1); n\right), \text{ for } \theta_1>\theta_0, \\ \delta(\theta_1)=\frac{\sqrt{n}(\theta_1-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \geq 0, \ V(\theta_1)=\frac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)}, \ 0<V(\theta_1)\leq 1. \end{array} \tag{58}$$

Since the hypothesis that 'passed' is of the form $\theta>\theta_1=\theta_0+\gamma$, the objective of $\text{SEV}(T_\alpha^>;\theta>\theta_1)$ is to determine the *largest* discrepancy $\gamma\geq 0$ warranted by data $\mathbf{x}_0$.

Example 13.28 (continued). For the observed test statistic $d_A(\mathbf{x}_0)=7.389$, table 13.15 evaluates $\text{SEV}(\mathcal{T}_\alpha;\theta>\theta_1)$ for different values of γ, with the evaluations based on:

$$\frac{[d_A(\mathbf{X})-\delta(\theta_1)]}{\sqrt{V(\theta_1)}} \overset{\theta=\theta_1}{\backsim} \mathsf{Bin}\left(0,1;n\right) \simeq \mathsf{N}(0,1). \tag{59}$$

Note that for the above data, the scaling $\sqrt{V(\theta_1)}=.9996\simeq 1$ can be ignored.

The evaluation of $SEV(T_\alpha^>;\gamma>\gamma_1)$, like that of the power, is based on the distribution of the test statistic under the alternative, but unlike power SEV uses $d_A(\mathbf{x}_0)$ as the threshold instead of c_α. Figure 13.15 plots the severity curve evaluated for several alternatives in table 13.17. For $\gamma:=(\theta_1-\theta_0)=.015$, the evaluation of severity components yields:

$$d_A(\mathbf{x}_0)=7.389, \quad \delta(\theta_1)=\frac{\sqrt{30762}(.515-.5)}{\sqrt{.5(.5)}}=5.262, \quad d_A(\mathbf{x}_0)-\delta(\theta_1)=7.389-5.262=2.127.$$

Hence, the evaluation of $SEV(T_\alpha^>;\gamma>\gamma_1)$ using the $\mathsf{N}(0,1)$ tables yields:

$$SEV(T_\alpha^>;\gamma>\gamma_1)=\mathbb{P}(d_A(\mathbf{X})\leq 7.389;\ \theta_1=.515)=.983,$$

since $\Phi(z\leq 2.127)=.983$, where $\Phi(.)$ is the cdf of $\mathsf{N}(0,1)$.

Similarly, for $\gamma=.0142$: $\frac{\sqrt{n}(\hat{\theta}_n-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}}-\delta(\theta_1)=7.389-\frac{\sqrt{30762}(.5142-.5)}{\sqrt{.5(.5)}}=2.408,$

$$SEV(T_\alpha^>;\gamma>\gamma_1)=\mathbb{P}(d_A(\mathbf{X})\leq 7.389;\theta_1=.5142)=.992.$$

for $\gamma=.01$, $7.389-\frac{\sqrt{30762}(.51-.5)}{\sqrt{.5(.5)}}=3.881$, $SEV(T_\alpha^>;\gamma>\gamma_1=.51)=.999995$,

for $\gamma=.015$, $7.389-\frac{\sqrt{30762}(.515-.5)}{\sqrt{.5(.5)}}=2.127$, $SEV(T_\alpha^>;\gamma\leq\gamma_1=.515)=.983$,

for $\gamma=.017$, $7.389-\frac{\sqrt{30762}(.517-.5}{\sqrt{.5(.5)}}=1.426$, $SEV(T_\alpha^>;\gamma\leq\gamma_1=.517)=.923$.

Table 13.17: Severity of 'Reject H_0: $\theta\leq .5$ vs. H_1: $\theta>.5$' with $(T_\alpha^>;\mathbf{x}_0)$

$\theta>\theta_1=\theta_0+\gamma$, $\gamma=$	.01	.013	.0143	.015	.016	.017	.018	.02	.021	.025
$\mathsf{Sev}(\theta>\theta_1)=$	.999	.997	.991	.983	.962	.923	.859	.645	.500	.084

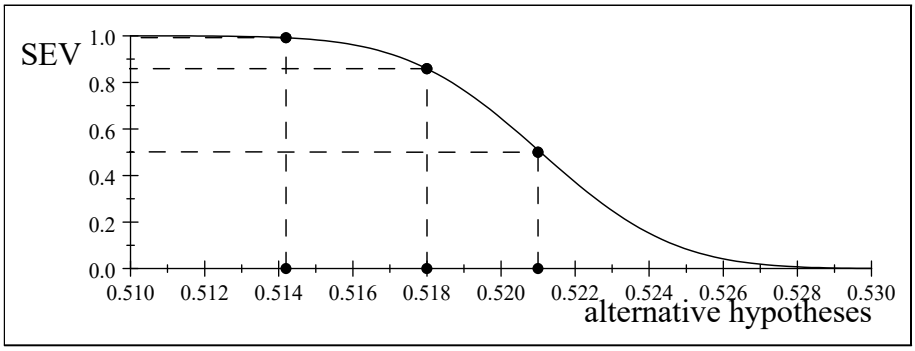


Fig. 13.15: The severity curve (table 13.17)

In light of the above example, the main features of the severity evaluation are:

(a) SEV evaluation is error probabilistic in nature.

(b) The underlying reasoning is hypothetical.

(c) The evaluation is post-data in the sense that the direction of departure is indicated by $d(\mathbf{x}_0)$ and not by the specified althernative hypothesis.

(d) The inferential claim associated with SEV comes in the form of the discrepancy γ^* from the null warranted for the particular test and data $\mathbf{x}_0$ associated with different severity threholds.

Severity and evidential interpretation. Taking a very high probability, say .95, as a threshold, the largest discrepancy from the null warranted by this data is:

$$\gamma \leq .01637, \text{ since } SEV(T_\alpha^{>}; \theta > .51637)=.95.$$

That is, the post-data severity evaluation uses the information in the particular **statistical context**:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ H_0\text{: } \theta\in\Theta_0 \text{ vs. } H_1\text{: } \theta\in\Theta_1,\ \mathcal{T}_\alpha:=\{d(\mathbf{X}), C_1(\alpha)\} \text{ and } \mathbf{x}_0, \tag{60}$$

to propose an inferential interpretation of the accept/reject results that replace the broadly dichotomous inferences with the warranted discrepancy from the null.

Is this discrepancy substantively significant? In general, to answer this question one needs to appeal to substantive subject matter information to assess the warranted discrepancy on substantive grounds. In human biology it is commonly accepted that the sex ratio at birth is approximately $\theta^\star$=.5122; see Hardy (2002). In light of the fact that the discrepancy .0122 is well within the range of the SEV inferential claim $\gamma \leq .01637$, the statistical significance entails *substantive significance*.

Returning to the original substantive discrepancy of interest between θ_B and θ_A:

$$\gamma^*=(18/35) - .5=.0142857,$$

one can use SEV to conclude that:

$$SEV(T_\alpha^{>}; \theta > .5143) = .991,$$

indicating that there is *excellent evidence* for the claim $\theta > \theta_1 = \theta_0 + \gamma^*$.

In terms of the ultimate objective of statistical inference, that of learning from data, this evidential interpretation seems highly effective because it narrows down the original parameter space from $\theta^* \in [0, 1]$ to a very small subset $\theta^* \in (.5, .5164]$!

5.4 Revisiting issues bedeviling frequentist testing

The above post-data evidential interpretation based on the severity assessment can also be used to shed light on a number of issues raised in the previous sections.

5.4.1 Addressing the large n problem

The post-data severity evaluation of the accept/reject H_0 result, addresses the large n problem by taking into consideration the generic capacity (power) of the test in evaluating the warranted discrepancy γ^* from H_0.

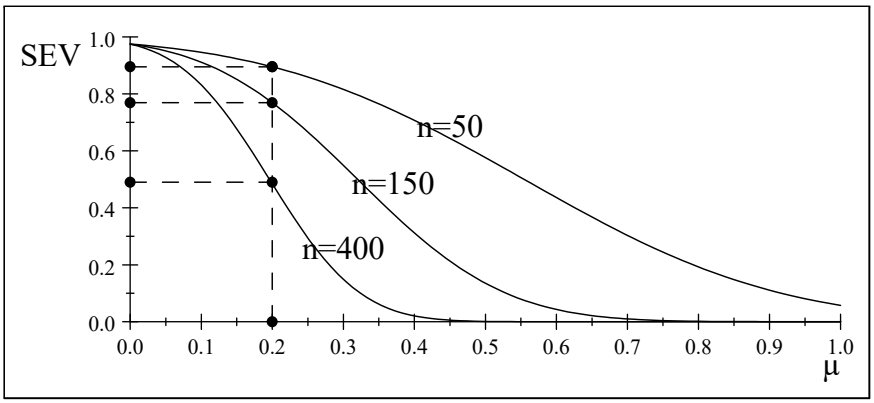


Fig. 13.16: Severity for $\mu > .2$ with different sample sizes

Example 13.29. In the context of the simple Normal model (table 13.10), consider the hypotheses: H_0: $\mu \leq \mu_0$ vs. H_1: $\mu > \mu_0$ for $\mu_0 = 0$, $\alpha = .025, \sigma = 2$. The severity curves shown below are associated with test $\mathcal{T}_\alpha$ and are based on the same outcome $\kappa(\mathbf{x}_0) = 1.96$ but different sample sizes ($n = 25, n = 100, n = 400$), indicating that the severity for inferring $\mu > .2$ decreases as n increases: (i) $n = 50$, $SEV(\mu > 0.2) = .895$, (i) $n = 150$, $SEV(\mu > 0.2) = .769$, (iii) $n = 400$, $SEV(\mu > 0.2) = .49$; see figure 13.16.

5.4.2 The arbitrariness of the N-P specification of hypotheses

The question that naturally arises at this stage is whether having good evidence for inferring $\theta > \frac{18}{35}$ depends on the particular way one has chosen to specify the hypotheses for the N-P test. Intuitively, one would expect that the evidence should not be dependent on the specification of H_0 and H_1 as such, but on test $\mathcal{T}_\alpha$, data $\mathbf{x}_0$ and the statistical model $\mathcal{M}_\theta(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n_X$.

Example 13.30. To explore that issue let us focus on probing the Bernoulli value: H_0: $\theta_0=\frac{18}{35}$, where the test statistic in example 13.24 yields:

$$d_B(\mathbf{x}_0)=\frac{\sqrt{30762}(\frac{16029}{30762}-\frac{18}{35})}{\sqrt{\frac{18}{35}(1-\frac{18}{35})}}=2.379. \tag{61}$$

This value leads to *rejecting* H_0 at α=.01 when one uses the formulation in (v), but accepts H_0 when formulations (iii) and (iv) are used. Which result should one believe?

Irrespective of these conflicting results, the severity 'accordance' condition (S-1) implies that, in light of the observed test statistic $d_B(\mathbf{x}_0)$=2.379 > 0, the *directional claim* that 'passed' on the basis of (61) is of the form:

$$\theta > \theta_1=\theta_0+\gamma.$$

To evaluate the particular discrepancy γ warranted by data $\mathbf{x}_0$, condition (S-2) calls for the evaluation of the probability of the event: 'outcomes $\mathbf{x}$ that accord less well with $\theta > \theta_1$ than $\mathbf{x}_0$ does', i.e. [$\mathbf{x}$: $d_B(\mathbf{x}) \leq d_B(\mathbf{x}_0)$], giving rise to (56), but with a different θ_0. Table 13.18 lists several such evaluations of:

$$SEV(T_\alpha^{>};\theta>\theta_1)=\mathbb{P}(d_B(\mathbf{X}) \leq 2.379;\ \theta_1=.5143+\gamma),$$

for different values of γ. NOTE that these evaluations are based on (58). A number of negative values of γ are included in order to address the original substantive hypotheses of interest, as well as bring out the fact that the results of table 13.17 and 13.18 are identical when viewed as inferences pertaining to θ; they simply have a different null values θ_0, and thus different discrepancies, but identical θ_1 values.

Table 13.18: Severity of Accept H_0: $\theta=\frac{18}{35}$ **vs.** H_1: $\theta>\frac{18}{35}$ with $(T_\alpha^{>};\mathbf{x}_0)$

$\theta>\theta_0+\gamma$, $\gamma=$	-.0043	-.0013	.000	.0007	.0017	.0027	.0037	.0057	.0067	.0107
$\mathsf{Sev}(\theta>\theta_1)=$	.999	.997	.991	.983	.962	.923	.859	.645	.500	.084

The severity evaluations in tables 13.15 and 13.16, not only render the choice between (i)-(v) irrelevant, they also scotch the widely used argument pertaining to the *asymmetry* between the null and alternative hypotheses. The N-P convention of selecting a small α is generally viewed as reflecting a strong bias against rejection of the null. On the other hand, Bayesian critics of frequentist testing argue the exact opposite: N-P testing is biased against accepting the null; see Lindley (1957). The evaluations in tables 13.15 and 13.16 demonstrate that the evidential interpretation of frequentist testing based on severe testing addresses such asymmetries.

5.4.3 Addressing the fallacy of rejection

The potential arbitrariness of the N-P specification of the null and alternative hypotheses and the associated p-values is brought out in probing the Bernoulli value θ_B=$\frac{18}{35}$ using the different formulations of the hypotheses in (54). Can the severity evaluation explain away these conflicting and confusing results?

The choice (iii) where H_1: $\theta < \frac{18}{35}$ was driven solely by substantive information relating to θ_A=$\frac{1}{2}$, which is often a bad idea because it ignores the statistical dimension of inference. The choice (iv) where H_1: $\theta < \frac{18}{35}$ was based on lack of information about the direction of departure, which makes sense pre-data, but not post-data. The choice (v) where H_1: $\theta > \frac{18}{35}$ reflects the *post-data* direction of departure indicated by $d_B(\mathbf{x}_0)$=2.379 > 0. In light of that, the severity evaluation confirms that $p_{B>}(\mathbf{x}_0)$=.009 is the only relevant p-value.

From the severity perspective a pertinent definition of the p-value is: the p-value is the probability of all possible outcomes $\mathbf{x}\in\mathbb{R}^n_X$ that accord less well with H_0 than $\mathbf{x}_0$ does, when H_0 is true.

This perspective brings out the vulnerability of both the p-value and the N-P reject H_0 rule to the fallacy of rejection in cases where n is large. Viewing the above relevant p-value ($p_{B>}(\mathbf{x}_0)$=.009) from the severity vantage point, it is directly related to H_1 passing a severe test. This is because the probability that test $\mathcal{T}_\alpha$ would have produced a result that accords less well with H_1 than $\mathbf{x}_0$ does ($\mathbf{x}$**:** $d_B(\mathbf{x})<d_B(\mathbf{x}_0)$), if H_1 were false (H_0 true) is very high since:

$$\mathsf{Sev}(T^{>}_\alpha;\mathbf{x}_0;\theta>\theta_0) \quad =\mathbb{P}(d_B(\mathbf{X})<d(\mathbf{x}_0);\theta\leq\theta_0)=1-\mathbb{P}(d_B(\mathbf{X})>d_B(\mathbf{x}_0);\theta=\theta_0)=.991$$

■ This suggests that the crucial weakness of the p-value is that is establishes the existence of *some* discrepancy $\gamma \geq 0$, but provides no information concerning the magnitude warranted by $\mathbf{x}_0$. The severity evaluation remedies that by relating $\mathsf{Sev}(T^{>}_\alpha;\mathbf{x}_0;\theta > \theta_0)$=.991 to the discrepancy γ warranted by data $\mathbf{x}_0$ that revolves around the inferential claim $\theta > \theta_0 + \gamma$. Given that the p-value is evaluated at θ=θ_0, the implicit discrepancy associated with the p-value is γ=0. This ignores the generic capacity of the test that gave rise to the rejection of H_0.

5.4.4 Addressing the fallacy of acceptance

Example 13.31. **Arbuthnot's 1710 conjecture reparameterized.**

The equality of males and females can be tested in the context of a simple Bernoulli model (table 13.8) but now {X=1}={female}, {X=0}={male}, using the hypotheses:

$$\text{(i)} \quad H_0\text{: } \varphi \leq \varphi_A \text{ vs. } H_1\text{: } \varphi > \varphi_A,\ \varphi_A\text{=.5},$$

$$\text{(ii)} \quad H_0\text{: } \varphi \geq \varphi_A \text{ vs. } H_1\text{: } \varphi < \varphi_A,$$

in terms of φ=$\mathbb{P}(X$=1)=$E(X)$; NOTE that φ=1−θ in terms of the notation in table

13.8. The best (UMP) tests for (i)-(ii) take the form:

$$\text{(i)} \quad T_{\alpha}^{>}:=\{d_A(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\varphi_0)}{\sqrt{\varphi_0(1-\varphi_0)}},\ C_1^{>}(\alpha)=\{\mathbf{x}:d(\mathbf{X})>c_\alpha\}\},$$

$$\text{(ii)} \quad T_{\alpha}^{<}:=\{d_A(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\varphi_0)}{\sqrt{\varphi_0(1-\varphi_0)}},\ C_1^{<}(\alpha)=\{\mathbf{x}:d(\mathbf{X})<-c_\alpha\}\},$$

where $\widehat{\varphi}_n:=\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$ as the best estimator of φ and:

$$d_A(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\varphi_0)}{\sqrt{\varphi_0(1-\varphi_0)}} \overset{H_0}{\backsim} \mathsf{Bin}\left(0,1;n\right). \tag{62}$$

Using the data with n=30762 newborns during the period 1993-5 in Cyprus, 16029 boys and 14733 girls, α=.01 $\Rightarrow c_\alpha=\pm 2.326$, yields $\widehat{\varphi}_n(\mathbf{x}_0)=\frac{14733}{30762}=.4789$:

$$\text{(i)} \quad d_A(\mathbf{x}_0)=\frac{\sqrt{30762}(\frac{14733}{30762}-.5)}{\sqrt{.5(.5)}}=-7.389<2.326,$$

$$\text{(ii)} \quad d_A(\mathbf{x}_0)=\frac{\sqrt{30762}(\frac{14733}{30762}-.5)}{\sqrt{.5(.5)}}=-7.389<-2.326,$$

where (i) indicates accepting of H_0, but (ii) indicates rejecting H_0, confirmed by the associated p-values:

$$\text{(i)} \quad p_{A>}(\mathbf{x}_0)=\mathbb{P}(d_A(\mathbf{X})>-7.389;\ H_0)=1.0,$$

$$\text{(ii)} \quad p_{A<}(\mathbf{x}_0)=\mathbb{P}(d_A(\mathbf{X})<-7.389;\ H_0)=2.065\times 10^{-10}<\alpha=.01.$$

At the *coarse accept/reject* H_0 level, there is nothing contradictory about the above N-P results of accepting $\varphi\leq\varphi_A$ and rejecting $\varphi>\varphi_A$. They agree that, with very high probability, the test procedure suggests that the true value of φ, $\varphi^*\in[0,.5)$ or equivalently $\varphi^*\notin[5,1]$.

The problem is that the coarseness of these results renders them largely uninformative with respect to the main objective of testing which is to learn from $\mathbf{x}_0$ about φ^*. This is the issue the post-data severity evaluation aims to address. It is achieved by supplementing the coarse accept/reject rules with effective probing of the null value φ_A=.5 with a view to output the discrepancy warranted by data $\mathbf{x}_0$ in the direction indicated by the sign of $d_A(\mathbf{x}_0)=-7.389$. Hence, as a post-data error probability, the severity evaluation is guided by $d_A(\mathbf{x}_0)\gtrless 0$ and not by the accept/reject H_0 results as such because the latter is driven by its pre-data specification; see Spanos (2013).

As argued above, the sign of $d_A(\mathbf{x}_0)=-7.389$ indicates the relevant direction of departure from φ_A=.5, and thus the p-value that makes sense as a post-data error probability is the one for case (ii) since the values of $\mathbf{x}\in\{0,1\}^n$ that accord less well with φ_A=.5 lie to the left of that value. Similarly, for evaluating the post-data severity the relevant inferential claim is:

$$\varphi\leq\varphi_1=\varphi_0+\gamma, \quad \text{for some } \gamma\leq 0,$$

$$SEV(T_\alpha^<, \varphi \leq \varphi_1)=\mathbb{P}(\mathbf{x}\text{: } d_A(\mathbf{x}) > d_A(\mathbf{x}_0); \varphi_1=\varphi_0+\gamma). \quad (63)$$

where the evaluation of SEV is based on (58); note that $\sqrt{V(\varphi_1)} \simeq 1$. In light of $\varphi \leq \varphi_1=\varphi_0+\gamma,\ \gamma\leq 0$, the objective of SEV$(T_\alpha^>; \varphi \leq \varphi_1)$ is to determine the *largest* discrepancy $\gamma\leq 0$ warranted by data $\mathbf{x}_0$.

Table 13.19: Severity of N-P 'Accept H_0': $\varphi\leq .5$ vs. H_1: $\varphi>.5$' $(T_\alpha^>; d_A(\mathbf{x}_0)<0)$

$\gamma=$	-.01	-.014	-.015	-.01636	-.017	-.019	-.0211	-.022	-.024	-.025
$\varphi_1=\varphi_0+\gamma,$	.49	.486	.485	.48364	.483	.481	.4789	.478	.476	.475
Sev$(\varphi>\varphi_1)=$	1.0	.993	.983	.951	.923	.766	.500	.371	.152	.084

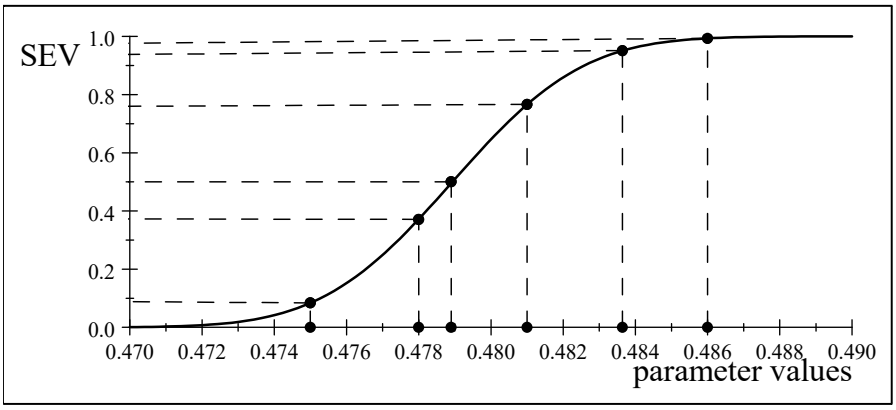


Fig. 13.17: The severity curve (table 13.19)

The post-data severity evaluations are as follows:

for $\gamma= -.01$, -7.389$-\frac{\sqrt{30762}(.49-.5)}{\sqrt{.5(.5)}} = -3.881,\ SEV(T_\alpha^>; \varphi\leq\varphi_1=.49)=.999995,$

for $\gamma= -.014$, -7.389$-\frac{\sqrt{30762}(.486-.5)}{\sqrt{.5(.5)}} = -2.478,\ SEV(T_\alpha^>; \varphi\leq\varphi_1=.486)=.993$

for $\gamma = -.015$, -7.389$-\frac{\sqrt{30762}(.485-.5)}{\sqrt{.5(.5)}} = -2.127,\ SEV(T_\alpha^>; \varphi\leq\varphi_1=.485)=.983,$

for $\gamma= -.01636$, -7.389$-\frac{\sqrt{30762}(.48364-.5}{\sqrt{.5(.5)}} = -1.65,\ SEV(T_\alpha^>; \varphi\leq\varphi_1=.48364)=.951,$

for $\gamma= -.017$, -7.389$-\frac{\sqrt{30762}(.483-.5}{\sqrt{.5(.5)}} = -1.426,\ SEV(T_\alpha^>; \varphi\leq\varphi_1=.48364)=.923,$

for $\gamma = -.019$, -7.389$-\frac{\sqrt{30762}(.481-.5}{\sqrt{.5(.5)}} = -.724,\ SEV(T_\alpha^>; \varphi\leq\varphi_1=.48364)=.766,$

for $\gamma = -.0211$, -7.389$-\frac{\sqrt{30762}(.4789-.5)}{\sqrt{.5(.5)}} = 0,\ SEV(T_\alpha^>; \varphi\leq\varphi_1=.4789)=.5,$

for $\gamma = -.022$, -7.389$-\frac{\sqrt{30762}(.478-.5)}{\sqrt{.5(.5)}} =.328,\ SEV(T_\alpha^>; \varphi\leq\varphi_1=.478)=.371,$

for $\gamma = -.024$, -7.389$-\frac{\sqrt{30762}(.476-.5)}{\sqrt{.5(.5)}} =1.03,\ SEV(T_\alpha^>; \varphi\leq\varphi_1=.476)=.152,$

for $\gamma = -.025$, -7.389$-\frac{\sqrt{30762}(.475-.5)}{\sqrt{.5(.5)}}$=1.381, $SEV(T_\alpha^>; \varphi \leq \varphi_1 = .475)$=.084.

Table 4 reports the severity curve for various discrepancies and plotted in fig. 5. Using the severity threshold of .95, the warranted discrepancy from φ_A=.5 can be derived using the equation:

$$-7.389-\frac{\sqrt{30762}(\gamma)}{\sqrt{.5(.5)}} = -1.645 \rightarrow \gamma^* = -.01636,$$

where the value -1.645 is chosen because $1-\Phi(-1.645)$=.95; Φ denotes the cumulative distribution function (cdf) of $\mathsf{N}(0,1)$. The inferential claim warranted by data $\mathbf{x}_0$ is: $\varphi \leq .48364$.

Similarly, the severity evaluation of the discrepancy associated with the substantive value of φ, $\varphi^\bigstar$=.4878, has $SEV(T_\alpha^>; \varphi \leq .4878)$=.999, which clearly exceeds the .95 threshold. This indicates that the test result based on $d_A(\mathbf{x}_0)$=-7.389 is both statistically and substantively significant with data $\mathbf{x}_0$.

Finally, it is important to note that the post-data severity evaluations in table 13.19 are in complete inferential agreement with those in tables 13.17 and 13.18.

A comparison of Figures 13.17 and 13.15 indicates that their severity curves differ with respect to their slope since their evaluations pertain to the two different tails of the sampling distribution of $d_A(\mathbf{X})$ under H_1.

5.4.5 The arbitrariness of the significance level

It is often argued by critics of frequentist testing that the choice of the significance level in the context of the N-P approach, or the threshold for the p-value in the case of Fisherian significance testing, is totally arbitrary and vulnerable to manipulation.

To see how the severity evaluations can address this problem, let us try to 'manipulate' the original significance level (α=.01) to a different one that would alter the accept/reject results for H_0: $\theta=\frac{18}{35}$.

Example 13.32. Looking back at results in examples 13.25-26 for Bernoulli's conjectured value using the data from Cyprus, it is easy to see that choosing a larger significance level, say α=.02 $\Rightarrow$ c_α=2.053, would lead to *rejecting* H_0 for both N-P formulations:

$$\text{(2-s) } (H_1\text{: } \theta \neq \tfrac{18}{35}) \text{ and (1-s>) } (H_1\text{: } \theta > \tfrac{18}{35})$$

since $d(\mathbf{x}_0)$=2.379>2.053, and the corresponding p-values are:

$$\text{(2-s): } p(\mathbf{x}_0)=\mathbb{P}(|d(\mathbf{X})| > 2.379; H_0)=.0174,$$
$$\text{(1-s>): } p_>(\mathbf{x}_0)=\mathbb{P}(d(\mathbf{X}) > 2.379; H_0)=.009.$$

How does this change the original severity evaluations? The short answer is: *it doesn't!* The severity evaluations remain invariant to any changes to α because they depend on the direction of departure indicated by the sign of $d(\mathbf{x}_0)$=2.379>0, and *not* by the direction of the alternative or the value of c_α. From the severity perspective,

the relevant direction of departure in light of $d(\mathbf{x}_0) > 0$ is $\theta > \theta_1 = \theta_0 + \gamma$, indicating the generic form of the hypothesis that 'passed', and thus, the same evaluations given in tables 13.14 apply.

5.4.6 Warranted discrepancy vs. estimation-based effect sizes

The warranted discrepancy γ^* should be contrasted with *estimation-based* effect size estimates whose primary aim is to get a more appropriate measure of the 'magnitude of the scientific effect'. Although there is no agreement in the literature about the most appropriate effect size estimate, Cohen's (1988) proposed measure is:

$$g = (.5176 - .5) = .0176,$$

which is much larger than $\gamma^* = .01254$ and $SEV(T_\alpha; \theta > .5176) = .50$.
This reveals the arbitrariness of grading such effects sizes as small, medium and large. Using Cohen's benchmarks, $g = .0176$ is tiny, since the benchmark for small is $g_s = .05$, despite the fact that $\gamma^* \leq .01254$ implies substantive significance.

As a general rule, it is never a good idea to view the point estimate, say $\widehat{\theta}(\mathbf{x}_0) = \overline{x}_n$, as (approximately) coinciding with θ^*, since it represents just a single value from the sampling distribution of $\widehat{\theta}(\mathbf{X})$. That is why point estimation does *not* output an inferential claim that $\overline{x}_n$ approximates θ^* sufficiently close.

5.4.7 Replication for observational data?

For data generated by phenomena that are invariant to time and location, replicability amounts to similar inferential claims in the form of the post-data severity.

Example. The **data** refer to $n = 10514$ newborns, 5442 boys, and 5072 girls in **Cyprus** during 1993. assuming $\alpha = .01 \Rightarrow c_\alpha = 2.326$, and ask the question whether the severity results replicate with similar sample size data for London in 1706, $n = 15369$, 7952 boys and 7417 girls, reported by Arthbunot (1712).

For SEV$(T_\alpha; \theta > \theta_1) \geq .85$, the post-data severity yields:

$$\text{London 1706:} \quad \gamma^* \leq .0132, \text{ since } SEV(T_\alpha; \theta > .5132) = .85, \tag{64}$$

which is very similar to the warranted discrepancy for the Cyprus 1993 data:

$$\text{Cyprus 1993:} \quad \gamma^* \leq .01254, \text{ since } SEV(T_\alpha; \theta > .51254) = .85. \tag{65}$$

Notice that in both cases, the **substantive discrepancy** of .0122 is within the above upper bounds, and thus statistical signicance also implies substantive significance.

5.5 Observed Confidence Intervals and Severity

In addition to addressing the fallacies of acceptance and rejection, the post-data severity evaluation can be used to address the issue of degenerate post-data error probabilities and the inability to distinguish between different values of μ within an observed CI; Mayo and Spanos (2006). This is achieved after two key changes:

(a) Replace the *factual* reasoning underlying CIs, which becomes degenerate post-data, with the *hypothetical* reasoning underlying the post-data severity evaluation.

(b) Replace any claims pertaining to overlaying the true μ^* with severity-based inferential claims of the form:

$$\mu > \mu_1{=}\mu_0{+}\gamma, \quad \mu \leq \mu_1{=}\mu_0{+}\gamma, \quad \text{for some } \gamma \geq 0. \tag{66}$$

This can be achieved by relating the observed bounds:

$$\overline{x}_n \pm c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}}) \tag{67}$$

to particular values of μ_1 associated with (66), say $\mu_1{=}\overline{x}_n{-}c_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})$, and evaluating the post-data severity of the inferential claim:

$$\mu > \mu_1{=}\overline{x}_n{-}c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}}).$$

A moment's reflection, however, suggests that the connection between establishing the warranted discrepancy γ from μ_0 and the observed CI (67) is more apparent than real. The severity assessment of $\mu > \mu_1$ is based on post-data hypothetical reasoning and does not pertain directly to μ_1 or μ^*, but to the warranted discrepancy $\gamma^*{=}\mu_1{-}\mu_0$ in light of $\mathbf{x}_0$. The severity evaluation probability relates to the inferential claim (66), and has nothing to do with the coverage probability. The equality of the tail areas stems from the mathematical duality, but that does not imply inferential duality.

6 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 $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. Hypothesis testing gives rise to learning from data by partitioning $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$:

$$\mathcal{M}_0(\mathbf{x}){=}\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}{\in}\Theta_0\},\ \text{or}\ \mathcal{M}_1(\mathbf{x}){=}\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}{\in}\Theta_1\},\ \mathbf{x}{\in}\mathbb{R}^n_X? \tag{68}$$

and framing the hypotheses in terms of $\boldsymbol{\theta}$, H_0: $\theta{\in}\Theta_0$ vs. H_1: $\theta{\in}\Theta_1$.
A test $T_\alpha{=}\{d(\mathbf{X}),\ C_1(\alpha)\}$ is defined in terms of a test statistic and a rejection region, and its *optimality* (effectiveness) is calibrated in terms of the relevant type I and II error probabilities evaluated using *hypothetical reasoning.* These error probabilities specify how often these procedures lead to erroneous inferences, and thus determine the power of the test:

$$\pi(\theta_1){=}\mathbb{P}(\mathbf{x}_0{\in}C_1;\ H_1(\theta_1)\ \text{true}){=}\beta(\theta_1),\ \forall\theta_1{\in}\Theta_1.$$

An inference is reached by an inductive procedure which, with high probability, will reach true conclusions from valid premises $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. Hence, the *trustworthiness* of

frequentist inference depends on two pre-conditions: (a) *optimal* inference procedures, (b) after securing the *adequacy* of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

Addressing foundational issues. The *first* such issue concerns the presumption that the true $\mathcal{M}^*(\mathbf{x})$ lies within the boundaries of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. This can be addressed by securing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, vis-a-vis data $\mathbf{x}_0$, using trenchant *Mis-Specification (M-S) testing*, before applying frequentist testing; see chapter 15. As shown above, when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is misspecified, the *nominal* and *actual error probabilities* can be very different, undermining the reliability of the test in question. The *second* issue concerns the form and nature of the evidence $\mathbf{x}_0$ can provide *for* $\theta\in\Theta_0$ or $\theta\in\Theta_1$. Neither the p-value, nor the N-P accept/reject rules provide an evidential interpretation, primarily because they are highly vulnerable to the fallacies of acceptance and rejection. These fallacies, however, can be circumvented using a post-data severity evaluation of p-value and accept/reject results to output the discrepancy γ from the null warranted by data $\mathbf{x}_0$. This warranted inferential claim for a particular test T_α and data $\mathbf{x}_0$, gives rise to learning from data. The post-data severity evaluation was shown to address, not only the classic fallacies of acceptance and rejection, but several other foundational problems bedeviling frequentist testing since the 1930s.

Crucial distinctions

Factual vs. hypothetical reasoning, simple vs. composite hypotheses, testing within vs. testing outside a statistical model, significance level vs. p-value, pre-data vs. post-data error probabilities, fallacy of acceptance vs. fallacy of rejection, statistical vs. substantive significance, hypothesis testing vs. Confidence Intervals.

Essential ideas

- Hypothesis testing, based on hypothetical reasoning, constitutes the most powerful and flexible inference procedure in learning from data about the stochastic generating mechanism $\mathcal{M}^*(\mathbf{x})$, $\mathbf{x}\in\mathbb{R}^n_X$ that gave rise to data $\mathbf{x}_0$. Pre-data error probabilities calibrate the generic capacity of a test to achieve that.

- The p-value is the probability of all possible outcomes $\mathbf{x}\in\mathbb{R}^n_X$ that accord less well with H_0 than $\mathbf{x}_0$ does, when H_0 is true. Any attempt to assign it an evidential interpretation is doomed because the concept ignores the generic capacity of the test; the large n problem highlighted that weakness. Since the p-value is a post-data error probability, it cannot be two-sided; the observed $d(\mathbf{x}_0)$ designates the only relevant side.

- The N-P reframing recasts Fisher's testing by partitioning the parameter and sample spaces, rendering N-P testing strictly within the boundaries of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. In contrast, Mis-Specification (M-S) testing constitutes probing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

- In N-P testing the null and alternative hypotheses should constitute a partition of the parameter space of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. For statistical purposes all possible values

$\theta \in \Theta$ are relevant for N-P testing, irrespective of whether only one or more values are of substantive interest.

- There can be no legitimate evidential interpretation of the accept/reject results and the p-value, that can address the fallacies of acceptance and rejection, without accounting for the generic capacity (power) of the test applied to the particular data $\mathbf{x}_0$.

- The post-data severity evaluation provides an evidential interpretation of the statistical significance/insignificance by outputting the warranted discrepancy from the null, $\gamma_1=\theta_0-\theta_1$, after taking into account the generic capacity of the particular test $\mathcal{T}_\alpha$ and data $\mathbf{x}_0$.

- Egregious and misleading claims in frequentist hypothesis testing:

(a) The type I, II error and the p-value constitute conditional probabilities.
(b) The type I, II error probabilities and the p-value can be assigned to θ.
(c) Confidence intervals are more reliable than p-values.
(d) The p-value provides a stand alone measure of evidence against the null.
(e) The N-P lemma yields a UMP for any two simple hypotheses, regardless of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X$.
(f) Statistical significance can be detached from $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ and data $\mathbf{x}_0$, hence ***, **, *. Rejecting the null at $\alpha=.05$ with $n=40$ or $n=10000$ are very different on evidential grounds.