Excursion 3 Statistical Tests and Scientific Inference

Itinerary

You

3.2 N-P Tests: An Episode in Anglo-Polish Collaboration

> We proceed by setting up a specific hypothesis to test, H_0 in Neyman's and my terminology, the null hypothesis in R. A. Fisher's . . . in choosing the test, we take into account alternatives to H_0 which we believe possible or at any rate consider it most important to be on the look out for . . .Three steps in constructing the test may be defined:
>
> **Step 1.** We must first specify the set of results . . .
>
> **Step 2.** We then divide this set by a system of ordered boundaries . . .
>
> such that as we pass across one boundary and proceed to the next, we come to a class of results which makes us more and more inclined, on the information available, to reject the hypothesis tested in favour of alternatives which differ from it by increasing amounts.
>
> **Step 3.** We then, if possible, associate with each contour level the chance that, if H_0 is true, a result will occur in random sampling lying beyond that level . . .
>
> In our first papers [in 1928] we suggested that the likelihood ratio criterion, λ, was a very useful one . . . Thus Step 2 proceeded Step 3. In later papers [1933–1938] we started with a fixed value for the chance, ε, of Step 3 . . . However, although the mathematical procedure may put Step 3 before 2, we cannot put this into operation before we have decided, under Step 2, on the guiding principle to be used in choosing the contour system. That is why I have numbered the steps in this order. (Egon Pearson 1947, p. 173)

In addition to Pearson's 1947 paper, the museum follows his account in "The Neyman–Pearson Story: 1926–34" (Pearson 1970). The subtitle is "Historical Sidelights on an Episode in Anglo-Polish Collaboration"!

We meet Jerzy Neyman at the point he's sent to have his work sized up by Karl Pearson at University College in 1925/26. Neyman wasn't that impressed:

> Neyman found . . . [K.]Pearson himself surprisingly ignorant of modern mathematics. (The fact that Pearson did not understand the difference between independence and lack of correlation led to a misunderstanding that nearly terminated Neyman's stay . . .) (Lehmann 1988, p. 2)

Thus, instead of spending his second fellowship year in London, Neyman goes to Paris where his wife Olga ("Lola") is pursuing a career in art, and where he could attend lectures in mathematics by Lebesque and Borel. "[W]ere it not for Egon Pearson [whom I had briefly met while in London], I would have probably drifted to my earlier passion for [pure mathematics]" (Neyman quoted in Lehmann 1988, p. 3).

What pulled him back to statistics was Egon Pearson's letter in 1926. E. Pearson had been "suddenly smitten" with doubt about the justification of

tests then in use, and he needed someone with a stronger mathematical background to pursue his concerns. Neyman had just returned from his fellowship years to a hectic and difficult life in Warsaw, working multiple jobs in applied statistics.

[H]is financial situation was always precarious. The bright spot in this difficult period was his work with the younger Pearson. Trying to find a unifying, logical basis which would lead systematically to the various statistical tests that had been proposed by Student and Fisher was a 'big problem' of the kind for which he had hoped . . . (ibid., p. 3)

N-P Tests: Putting Fisherian Tests on a Logical Footing

For the Fisherian simple or "pure" significance test, alternatives to the null "lurk in the undergrowth but are not explicitly formulated probabilistically" (Mayo and Cox 2006, p. 81). Still there are constraints on a Fisherian test statistic. Criteria for the test statistic $d(\boldsymbol{X})$ are

(i) it reduces the data as much as possible;
(ii) the larger $d(\boldsymbol{x}_0)$ the further the outcome from what's expected under H_0, with respect to the particular question;
(iii) the P-value can be computed $p(\boldsymbol{x}_0)=\Pr(d(\boldsymbol{X}) \geq d(\boldsymbol{x}_0); H_0)$.

Fisher, arch falsificationist, sought test statistics that would be *sensitive* to discrepancies from the null. Desiderata (i)–(iii) are related, as emerged clearly from N-P's work.

Fisher introduced the idea of a parametric statistical model, which may be written $M_\theta(\boldsymbol{x})$. Karl Pearson and others had been prone to mixing up a parameter θ, say the mean of a population, with a sample mean $\overline{x}$. As a result, concepts that make sense for statistic $\overline{\boldsymbol{X}}$, like having a distribution, were willy-nilly placed on a fixed parameter θ. Neyman and Pearson [N-P] gave mathematical rigor to the components of Fisher's tests and estimation. The model can be represented as a pair (S, Θ) where S denotes the set of all possible values of the *sample* $\boldsymbol{X} = (X_1, \ldots, X_n)$ – one such value being the data $\boldsymbol{x}_0 = (x_1, \ldots, x_n)$ – and Θ denotes the set of all possible values of the unknown *parameter(s)* θ. In hypothesis testing, Θ is used as shorthand for the family of probability distributions or, in continuous cases, densities *indexed* by θ. Without the abbreviation, we'd write the full model as

$$M_\theta(\boldsymbol{x}) := \{f(\boldsymbol{x};\ \theta),\ \theta \in \Theta\},$$

where $f(\boldsymbol{x};\theta)$, for all $\boldsymbol{x} \in$ S, is the distribution (or density) of the sample. We don't test all features of the model at once; it's part of the test specification

to indicate which features (parameters) of the model are under test. The *generic form* of *null* and *alternative* hypotheses is

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

where (Θ_0, Θ_1) constitute subsets of Θ that partition Θ. Together, Θ_0 and Θ_1 exhaust the parameter space. N-P called H_0 the *test hypothesis*, which is preferable to null hypothesis, since for them it's on par with alternative H_1; but for brevity and familiarity, I mostly call H_0 the null. I follow A. Spanos' treatment.

Lambda Criterion

What were Neyman and Pearson looking for in their joint work from 1928? They sought a criterion for choosing, as well as generating, sensible test statistics. Working purely on intuition, which they later imbued with a justification, N-P employ the likelihood ratio. Pearson found the spark of the idea from correspondence with Gosset, known as Student, but we will see that generating good tests requires much more than considering alternatives.

How can we consider the likelihood ratio of hypotheses when one or both can contain multiple values of the parameter? They consider the maximum values that the likelihood could take over ranges of the parameter space. In particular, they take the maximum likelihood over all possible values of θ in the entire parameter space Θ (not Θ_1), and compare it to the maximum over the restricted range of values in Θ_0, to form the ratio

$$\Lambda(\boldsymbol{X}) = \frac{\max_{\theta\in\Theta} \mathrm{L}(\boldsymbol{X};\theta)}{\max_{\theta\in\Theta_0} \mathrm{L}(\boldsymbol{X};\theta)}.$$

Let's look at this. The numerator is the value of θ that makes the data $\boldsymbol{x}$ most probable over the entire parameter space. It is the *maximum likelihood estimator* for θ. Write it as $\hat{\theta}$. The denominator is the value of θ that maximizes the probability of $\boldsymbol{x}$ restricted just to the members of the null Θ_0. It may be called the *restricted* likelihood. Write it as $\widetilde{\theta}$:

$$\Lambda(\boldsymbol{X}) = \frac{\mathrm{L}(\hat{\theta}\text{-unrestricted})}{\mathrm{L}(\widetilde{\theta}\text{-restricted})}.$$

Suppose that looking through the entire parameter space Θ we cannot find a θ value that makes the data more probable than if we restrict ourselves to the parameter values in Θ_0. Then the restricted likelihood in the

denominator is large, making the ratio $\Lambda(\boldsymbol{X})$ small. Thus, a small $\Lambda(\boldsymbol{X})$ corresponds to H_0 being in accordance with the data (Wilks 1962, p. 404). It's a matter of convenience which way one writes the ratio. In the one we've chosen, following Aris Spanos (1986, 1999), the larger the $\Lambda(\boldsymbol{X})$, the more discordant the data are from H_0. This suggests the null would be rejected whenever

$$\Lambda(\boldsymbol{X}) \geq k_\alpha$$

for some value of k_α.

So far all of this was to form the distance measure $\Lambda(\boldsymbol{X})$. It's looking somewhat the same as the Likelihoodist account. Yet we know that the additional step 3 that error statistics demands is to compute the probability of $\Lambda(\boldsymbol{X})$ under different hypotheses. Merely reporting likelihood ratios does not produce meaningful control of errors; nor do likelihood ratios mean the same thing in different contexts. So N-P consider the probability distribution of $\Lambda(\boldsymbol{X})$, and they want to ensure the probability of the event $\{\Lambda(\boldsymbol{X}) \geq k_\alpha\}$ is sufficiently small under H_0. They set k_α so that

$$\Pr(\Lambda(\boldsymbol{X}) \geq k_\alpha; H_0) = \alpha$$

for small α. Equivalently, they want to ensure high probability of accordance with H_0 just when it adequately describes the data generation process. Note the complement:

$$\Pr(\Lambda(\boldsymbol{X}) < k_\alpha; H_0) = (1 - \alpha).$$

The event statement to the left of ";" does not reverse positions with H_0 when you form the complement, H_0 stays where it is.

The set of data points leading to $(\Lambda(\boldsymbol{X}) \geq k_\alpha)$ is what N-P call the *critical region* or *rejection region* of the test $\{\boldsymbol{x}: \Lambda(\boldsymbol{X}) \geq k_\alpha\}$ – the set of outcomes that will be taken to reject H_0 or, in our terms, to infer a discrepancy from H_0 in the direction of H_1. Specifying the test procedure, in other words, boils down to specifying the rejection (of H_0) region.

Monotonicity. Following Fisher's goal of maximizing sensitivity, N-P seek to maximize the capability of detecting discrepancies from H_0 when they exist. We need the sampling distribution of $\Lambda(\boldsymbol{X})$, but in practice, $\Lambda(\boldsymbol{X})$ is rarely in a form that one could easily derive this. $\Lambda(\boldsymbol{X})$ has to be transformed in clever ways to yield a test statistic $d(\boldsymbol{X})$, a function of the sample that has a known distribution under H_0. A general trick to finding a suitable test statistic $d(\boldsymbol{X})$ is to find a function $h(\cdot)$ of $\Lambda(\boldsymbol{X})$ that is *monotonic* with respect to a statistic $d(\boldsymbol{X})$. The greater $d(\boldsymbol{X})$ is,

the greater the likelihood ratio; the smaller d($\boldsymbol{X}$) is, the smaller the likelihood ratio. Having transformed $\Lambda(\boldsymbol{X})$ into the test statistic d($\boldsymbol{X}$), the rejection region becomes

$$\text{Rejection Region, RR} \coloneqq \{\boldsymbol{x}\text{: } d(\boldsymbol{x}) \geq c_\alpha\},$$

the set of data points where $d(\boldsymbol{x}) \geq c_\alpha$. All other data points belong to the "non-rejection" or "acceptance" region, NR. At first Neyman and Pearson introduced an "undecided" region, but tests are most commonly given such that the RR and NR regions exhaust the entire sample space S. The term "acceptance," Neyman tells us, was merely shorthand: "The phrase 'do not reject *H*' is longish and cumbersome . . . My own preferred substitute for 'do not reject *H*' is 'no evidence against *H* is found'" (Neyman 1976, p. 749). That is the interpretation that should be used.

The use of the $\Lambda(\cdot)$ criterion began as E. Pearson's intuition. Neyman was initially skeptical. Only later did he show it could be the basis for good and even optimal tests.

> Having established the usefulness of the Λ-criterion, we realized that it was essential to explore more fully the sense in which it led to tests which were likely to be effective in detecting departures from the null hypothesis. So far we could only say that it seemed to appeal to intuitive requirements for a good test. (E. Pearson 1970 p. 470, I replace λ with Λ)

Many other desiderata for good tests present themselves.

We want a higher and higher value for $\Pr(d(\boldsymbol{X}) \geq c_\alpha; \theta_1)$ as the discrepancy $(\theta_1 - \theta_0)$ increases. That is, the larger the discrepancy, the easier (more probable) it should be to detect it. This came to be known as the *power function*. Likewise, the power should increase as the sample size increases, and as the variability decreases. The point is that Neyman and Pearson did not start out with a conception of optimality. They groped for criteria that intuitively made sense and that reflected Fisher's tests and theory of estimation. There are some early papers in 1928, but the N-P classic result isn't until the paper in 1933.

Powerful Tests. Pearson describes the days when he and Neyman are struggling to compare various different test statistics – Neyman is in Poland, he is in England. Pearson found himself simulating power for different test statistics and tabling the results. He calls them "empirical power functions." Equivalently, he made tables of the complement to the empirical power function: "what was tabled was the percentage of samples for which a test at 5 percent level failed to establish significance, as the true mean shifted from μ_0 by steps of $\sigma/\sqrt{n}$ (ibid. p. 471). He's construing the test's capabilities in terms

of percentage of samples. The formal probability distributions serve as shortcuts to cranking out the percentages. "While the results were crude, they show that our thoughts were turning towards the justification of tests in terms of power"(ibid.).

While Pearson is busy experimenting with simulated power functions, Neyman writes to him in 1931 of difficulties he is having in more complicated cases, saying: I found a test in which, paradoxically, "*the true hypothesis will be rejected more often than some of the false ones*. I told Lola [his wife] that we had invented such a test. She said: 'good boys!'" (ibid. p. 472). A test should have a higher probability of leading to a rejection of H_0 when H_1: $\theta \in \Theta_1$ than when H_0: $\theta \in \Theta_0$. After Lola's crack, pretty clearly, they would insist on *unbiased tests*: the probability of rejecting H_0 when it's true or adequate is always less than that of rejecting it when it's false or inadequate. There are direct parallels with properties of good estimators of θ (although we won't have time to venture into that).

Tests that violate unbiasedness are sometimes called "worse than useless" (Hacking 1965, p. 99), but when you read for example in Gigerenzer and Marewski (2015) that N-P found Fisherian tests "worse than useless" (p. 427), there is a danger of misinterpretation. N-P aren't bad-mouthing Fisher. They know he wouldn't condone this, but want to show that without making restrictions explicit, it's possible to end up with such unpalatable tests. In the case of two-sided tests, the additional criterion of unbiasedness led to uniformly most powerful (UMP) unbiased tests.

Consistent Tests. Unbiasedness by itself isn't a sufficient property for a good test; it needs to be supplemented with the property of *consistency*. This requires that, as the sample size n increases without limit, the probability of detecting any discrepancy from the null hypothesis (the power) should approach 1. Let's consider a test statistic that is unbiased yet inconsistent. Suppose we are testing the mean of a Normal distribution with σ known. The test statistic to which the Λ gives rise is

$$\mathrm{d}(\boldsymbol{X}) = \sqrt{n}(\overline{x} - \theta_0)/\sigma.$$

Say that, rather than using the sample mean $\overline{x}$, we use the average of the first and last values. This is to estimate the mean θ as $\hat{\theta} = 0.5(X_1 + X_n)$. The test statistic is then $\sqrt{2}(\hat{\theta} - \theta_0)/\sigma$. This is an unbiased estimator of θ. The distribution of $\hat{\theta}$ is $\mathrm{N}(\theta, \sigma^2/2)$. Even though this is unbiased and enables control of the Type I error, it is inconsistent. The result of looking only at two outcomes is that the power does not increase as n increases. The power of

this test is much lower than a test using the sample mean for any $n > 2$. If you come across a criticism of tests, make sure *consistency* is not being violated.

Historical Sidelight. Except for short visits and holidays, their work proceeded by mail. When Pearson visited Neyman in 1929, he was shocked at the conditions in which Neyman and other academics lived and worked in Poland. Numerous letters from Neyman describe the precarious position in his statistics lab: "You may have heard that we have in Poland a terrific crisis in everything" [1931] (C. Reid 1998, p. 99). In 1932, "I simply cannot work; the crisis and the struggle for existence takes all my time and energy" (Lehmann 2011, p. 40). Yet he managed to produce quite a lot. While at the start, the initiative for the joint work was from Pearson, it soon turned in the other direction with Neyman leading the way.

By comparison, Egon Pearson's greatest troubles at the time were personal: He had fallen in love "at first sight" with a woman engaged to his cousin George Sharpe, and she with him. She returned the ring the very next day, but Egon still gave his cousin two years to win her back (C. Reid 1998, p. 86). In 1929, buoyed by his work with Neyman, Egon finally declares his love and they are set to be married, but he let himself be intimidated by his father, Karl, deciding "that I could not go against my family's opinion that I had stolen my cousin's fiancée . . . at any rate my courage failed" (ibid., p. 94). Whenever Pearson says he was "suddenly smitten" with doubts about the justification of tests while gazing on the fruit station that his cousin directed, I can't help thinking he's also referring to this woman (ibid., p. 60). He was lovelorn for years, but refused to tell Neyman what was bothering him.

N-P Tests in Their Usual Formulation: Type I and II Error Probabilities and Power

Whether we accept or reject or remain in doubt, say N-P (1933, p. 146), it must be recognized that we can be wrong. By choosing a distance measure $d(\boldsymbol{X})$ wherein the probability of different distances may be computed, if the source of the data is H_0, we can determine the probability of an erroneous rejection of H_0 – a Type I error.

The test specification that dovetailed with the Fisherian tests in use began by ensuring the probability of a Type I error – an erroneous rejection of the null – is fixed at some small number, α, the *significance level* of the test:

$$\textbf{Type I error probability} = \Pr(d(\boldsymbol{X}) \geq c_\alpha; H_0) \leq \alpha.$$

Compare the Type I error probability and the *P*-value:

***P*-value**: $\Pr(d(\boldsymbol{X}) \geq d(\boldsymbol{x}_0); H_0) = p(\boldsymbol{x}_0)$.

So the N-P test could easily be given in terms of the P-value:

Reject H_0 iff $p(\boldsymbol{x}_0) \leq \alpha$.

Equivalently, the rejection (of H_0) region consists of those outcomes whose *P*-value is less than or equal to α. Reflecting the tests commonly used, N-P suggest the Type I error be viewed as the "more important" of the two. Let the relevant hypotheses be H_0: $\theta = \theta_0$ vs. H_1: $\theta > \theta_0$.

The Type II error is failing to reject the null when it is false to some degree. The test leads you to declare "no evidence of discrepancy from H_0" when H_0 is false, and a discrepancy exists. The alternative hypothesis H_1 contains more than a single value of the parameter, it is *composite*. So, abbreviate by $ß(\theta_1)$: the Type II error probability assuming $\theta = \theta_1$, for θ_1 values in the alternative region H_1:

Type II error probability (at θ_1) $= \Pr(d(\boldsymbol{X}) < c_\alpha; \theta_1) = ß(\theta_1)$, for $\theta_1 \in \Theta_1$.

In Figure 3.2, this is the area to the left of c_α, the vertical dotted line, under the H_1 curve. The shaded area, the complement of the Type II error probability (at θ_1), is the *power* of the test (at θ_1):

Power of the test (POW) (at θ_1) $= \Pr(d(\boldsymbol{X}) \geq c_\alpha; \theta_1)$.

This is the area to the right of the vertical dotted line, under the H_1 curve, in Figure 3.2. Note $d(\boldsymbol{x}_0)$ and c_α are always approximations expressed as decimals. For continuous cases, Pr is the probability density.

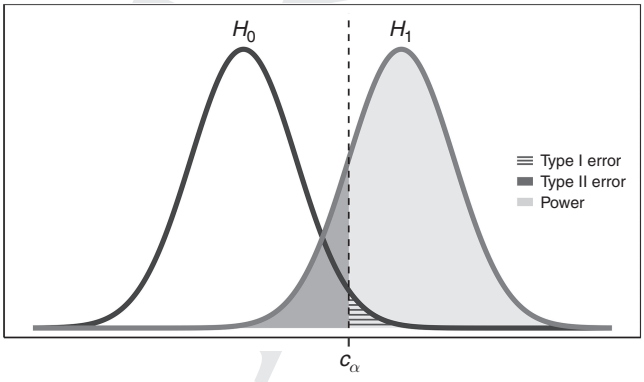

Figure 3.2 Type II error and power.

A *uniformly most powerful* (UMP) N-P test of a hypothesis at level α is one that minimizes ß(θ_1), or, equivalently, maximizes the power for all $\theta > \theta_0$. One reason alternatives are often not made explicit is the property of being a best test for any alternative. We'll explore power, an often-misunderstood creature, further in Excursion 5.

Although the manipulations needed to derive a test statistic using a monotonic mapping of the likelihood ratio can be messy, it's exhilarating to deduce them. Wilks (1938) derived a general asymptotic result, which does not require such manipulations. He showed that, under certain regularity conditions, as n goes to infinity one can define the asymptotic test, where "~" denotes "is distributed as".

$$2\ln\Lambda(\mathbf{X}) \sim \chi^2(r), \text{ under } H_0, \text{ with rejection region RR} \coloneqq \{\boldsymbol{x}: 2\ln\Lambda(\boldsymbol{x}) \geq c_\alpha\},$$

where $\chi^2(r)$ denotes the chi-square distribution with r degrees of freedom determined by the restrictions imposed by H_0.[4] The monotonicity of the likelihood ratio condition holds for familiar models including one-parameter variants of the Normal, Gamma, Beta, Binomial, Negative Binomial, Poisson (the Exponential family), the Uniform, Logistic, and others (Lehmann 1986). In a wide variety of tests, the Λ principle gives tests with all of the intuitively desirable test properties (see Spanos 2018, chapter 13).

Performance versus Severity Construals of Tests

"The work [of N-P] quite literally transformed mathematical statistics" (C. Reid 1998, p. 104). The idea that appraising statistical methods revolves around optimality (of some sort) goes viral. Some compared it "to the effect of the theory of relativity upon physics" (ibid.). Even when the optimal tests were absent, the optimal properties served as benchmarks against which the performance of methods could be gauged. They had established a new pattern for appraising methods, paving the way for Abraham Wald's decision theory, and the seminal texts by Lehmann and others. The rigorous program overshadowed the more informal Fisherian tests. This came to irk Fisher. Famous feuds between Fisher and Neyman erupted as to whose paradigm would reign supreme. Those who sided with Fisher erected examples to show that tests could satisfy predesignated criteria and long-run error control while leading to counterintuitive tests in specific cases. That was Barnard's point on the eclipse

[4] The general likelihood ratio $\Lambda(\boldsymbol{X})$ should be contrasted with the simple likelihood ratio associated with the well-known Neyman–Pearson (N-P) lemma, which assumes that the parameter space Θ includes only two values, i.e., $\Theta \coloneqq (\theta_0, \theta_1)$. In such a case no estimation is needed because one can take the simple likelihood ratio. Even though the famous lemma for UMP tests uses the highly artificial case of point against point hypotheses (θ_0, θ_1), it is erroneous to suppose the recommended tests are intended for this case. A UMP test, after all, alludes to all the possible parameter values, so just picking two and ignoring the others would not be UMP.

experiments (Section 3.1): no one would consider the class of repetitions as referring to the hoped-for 12 photos, when in fact only some smaller number were usable. We'll meet up with other classic chestnuts as we proceed.

N-P tests began to be couched as formal mapping rules taking data into "reject H_0" or "do not reject H_0" so as to ensure the probabilities of erroneous rejection and erroneous acceptance are controlled at small values, independent of the true hypothesis and regardless of prior probabilities of parameters. Lost in this *behavioristic* formulation was how the test criteria naturally grew out of the requirements of probative tests, rather than good long-run performance. Pearson underscores this in his paper (1947) in the epigraph of Section 3.2: Step 2 comes before Step 3. You must first have a sensible distance measure. Since tests that pass muster on performance grounds can simultaneously serve as probative tests, the severe tester breaks out of the behavioristic prison. Neither Neyman nor Pearson, in their applied work, was wedded to it. Where performance and probativeness conflict, probativeness takes precedent. Two decades after Fisher allegedly threw Neyman's wood models to the floor (Section 5.8), Pearson (1955) tells Fisher: "From the start we shared Professor Fisher's view that in scientific enquiry, a statistical test is 'a means of learning'" (p. 206):

> . . . it was not till after the main lines of this theory had taken shape with its necessary formalization in terms of critical regions, the class of admissible hypotheses, the two sources of error, the power function, etc., that the fact that there was a remarkable parallelism of ideas in the field of acceptance sampling became apparent. Abraham Wald's contributions to decision theory of ten to fifteen years later were perhaps strongly influenced by acceptance sampling problems, but that is another story. (ibid., pp. 204–5)

In fact, the tests as developed by Neyman–Pearson began as an attempt to obtain tests that Fisher deemed intuitively plausible, and this goal is easily interpreted as that of computing and controlling the severity with which claims are inferred.

Not only did Fisher reply encouragingly to Neyman's letters during the development of their results, it was Fisher who first informed Neyman of the split of K. Pearson's duties between himself and Egon, opening up the possibility of Neyman's leaving his difficult life in Poland and gaining a position at University College in London. Guess what else? Fisher was a referee for the all-important N–P 1933 paper, and approved of it.

> To Neyman it has always been a source of satisfaction and amusement that his and Egon's fundamental paper was presented to the Royal Society by Karl Pearson, who was hostile and skeptical of its contents, and favorably refereed by the formidable Fisher,

who was later to be highly critical of much of the Neyman–Pearson theory. (C. Reid 1998, p. 103)

Souvenir J: UMP Tests

Here are some familiar Uniformly Most Powerful (UMP) unbiased tests that fall out of the Λ criterion (letting μ be the mean):

(1) One-sided Normal test. Each X_i is NIID, N(μ, σ^2), with σ known: H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$.

$$\mathrm{d}(\boldsymbol{X}) = \sqrt{n}(\overline{X} - \mu_0)/\sigma,\ \mathrm{RR}(\alpha) = \{\boldsymbol{x}\colon \mathrm{d}(\boldsymbol{x}) \geq c_\alpha\}.$$

Evaluating the Type I error probability requires the distribution of d($\boldsymbol{X}$) under H_0: d($\boldsymbol{X}$) ~ N(0,1).

Evaluating the Type II error probability (and power) requires the distribution of d($\boldsymbol{X}$) under $H_1[\mu = \mu_1]$:

$$\mathrm{d}(\boldsymbol{X}) \sim \mathrm{N}(\delta_1,1), \text{ where } \delta_1 = \sqrt{n}(\mu_1 - \mu_0)/\sigma.$$

(2) One-sided Student's t test. Each X_i is NIID, N(μ, σ^2), σ unknown: H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$:

$$\mathrm{d}(\boldsymbol{X}) = \sqrt{n}(\overline{X} - \mu_0)/s, \quad \mathrm{RR}(\alpha) = \{\boldsymbol{x}\colon \mathrm{d}(\boldsymbol{x}) \geq c_\alpha\},$$

$$s^2 = \left[\frac{1}{(n-1)}\right]\sum(X_i - \overline{X})^2.$$

Two-sided Normal test of the mean H_0: $\mu = \mu_0$ against H_1: $\mu \neq \mu_0$:

$$\mathrm{d}(\boldsymbol{X}) = \sqrt{n}(\overline{X} - \mu_0)/s, \quad \mathrm{RR}(\alpha) = \{\boldsymbol{x}\colon |\mathrm{d}(\boldsymbol{x})| \geq c_\alpha\}.$$

Evaluating the Type I error probability requires the distribution of d($\boldsymbol{X}$) under H_0: d($\boldsymbol{X}$) ~ St(n − 1), the Student's t distribution with (n − 1) degrees of freedom (df).

Evaluating the Type II error probability (and power) requires the distribution of d($\boldsymbol{X}$) under $H_1[\mu = \mu_1]$: d($\boldsymbol{X}$) ~ St(δ_1, n − 1), where $\delta_1 = \sqrt{n}(\mu_1 - \mu_0)/\sigma$ is the non-centrality parameter.

This is the UMP, unbiased test.

(3) The difference between two means (where it is assumed the variances are equal):

H_0: $\gamma := \mu_1 - \mu_2 = \gamma_0$ against H_1: $\gamma_1 \neq \gamma_0$.

A Uniformly Most Powerful Unbiased (UMPU) test is defined by