Negative results: d($\boldsymbol{x}_0$) ≤ c_α:

(**SIST** 339)

A classic fallacy is to construe no evidence against H_0 as evidence of the correctness of H_0.

A canonical example was in the list of slogans opening this book:

> *Ordinary Power Analysis*: If data $\boldsymbol{x}$ are not statistically significantly different from H_0, and the power to detect discrepancy γ is high, then $\boldsymbol{x}$ indicates that the actual discrepancy is no greater than γ

(improves on Cohen a little, he imagines we can identify a "negligible discrepancy)

infer: discrepancy < γ

Probem: Too Coarse

Consider test T+ ($\alpha = .025$): H_0: $\mu = 0$ vs. H_1: $\mu \geq 0$, $\alpha = .025$, $n = 100$, $\sigma = 10$, $\sigma_{\bar{X}} = 1$. Say the cut-off must be $> \bar{x}_{.025} = 2$.

Consider an arbitrary inference $\mu < 1$.

We know POW(T+, $\mu = 1$) = .16 ($1\sigma_{\bar{X}}$ is subtracted from 2). .16 is quite lousy power.

It follows that no statistically insignificant result can warrant $\mu < 1$ for the power analyst.

Suppose, $\bar{x}_0 = -1$. This is $2\sigma_{\bar{X}}$ lower than 1. That should be taken into account.

We do. SEV(T+, $\bar{x}_0$ = -1, μ < 1) = .975.

Z = (-1 -1)/1 = -2
SEV (μ < 1) = Pr (Z > z_0; μ = 1) = .975

It would be even larger for values of μ smaller than 1

μ < 1 is also the upper one-sided CI bound at level .975

But it's not the one-sided CI corresponding to the test:
H_0: μ = 0 vs. H_1: μ ≥ 0

That's the one-sided <u>lower</u> (.975) bound

It matters because some people say you don't need to consider power if you've got CIs (discussed in the readings), since CIs have a duality with tests. (356-8)

Yes, but the test T+ corresponds to forming the one-sided lower bound.

One can look at the upper bound, but there needs to be a rationale for doing so.

(1) $P(d(X) > c_\alpha; \mu = \mu_0 + \gamma)$ Power to detect γ

- Just missing the cut-off c_α is the worst case
- It is more informative to look at the probability of getting a worse fit than you did

(2) $P(d(X) > d(x_0); \mu = \mu_0 + \gamma)$ "attained power"

a measure of the **severity** (or degree of corroboration) for the inference $\mu < \mu_0 + \gamma$

(1) can be low while (2) is high

Not the same as something called "retrospective power" or "ad hoc" power! (There μ is identified with the observed mean)

Shpower and Retrospective Power

"There's a sinister side to statistical power" (**SIST** 354)
I call it *Shpower analysis* because it distorts the logic of ordinary power analysis (from insignificant results).

Because ordinary power analysis is also post data, the criticisms of shpower are wrongly taken to reject both.

Shpower evaluates power with respect to the a hypothesis that the population effect size (discrepancy) equals the observed effect size, e.g., the parameter μ equals the observed mean $\bar{x}_0$, i.e., in *T+* this would be to set μ = $\bar{x}_0$).

The Shpower of test T+: Pr($\bar{X} > \bar{x}_\alpha$;μ = $\bar{x}_0$).

The Shpower of test T+: $\Pr(\bar{X} > \bar{x}_\alpha; \mu = \bar{x}_0)$.

The thinking is since we don't know the value of μ, we might use the observed $\bar{x}_0$ to estimate it, and then compute power in the usual way, except substituting the observed value.

Can't work for the purpose of using power analysis to interpret insignificant results. Why?

Since alternative μ is set = $\bar{x}_0$, and $\bar{x}_0$ is given as statistically insignificant, we are in Case 1 from 5.1 (Exhibit i): the power can never exceed .5.

In other words, since shpower = POW(T+, μ = $\bar{x}_0$), and $\bar{x}_0 < \bar{x}_\alpha$, the power can't exceed .5.

Between H_0 and $\bar{x}_\alpha$ the power goes from α to .5.

a. *The power against H_0 is α.* We can use the power function to define the probability of a Type I error or the significance level of the test:

$$POW(T+, \mu_0) = Pr(\bar{X} > \bar{x}_\alpha; \mu_0), \bar{x}_\alpha = (\mu_0 + z_\alpha \sigma_{\bar{X}}), \sigma_{\bar{X}} = [\sigma/\sqrt{n}])$$

The power at the null is: $Pr(Z > z_\alpha; \mu_0) = \alpha$.

But power analytic reasoning is all about finding an alternative against which the test has *high* capability to have obtained significance. Shpower is always "slim" (to echo Neyman) against such alternatives.

Unsurprisingly, Shpower analytical reasoning has been criticized in the literature: But the critics think they're maligning power analytic reasoning.

The severe tester uses attained power $\Pr(d(\mathbf{X}) > d(\boldsymbol{x}_0); \mu')$ to evaluate severity, but to address criticisms of power analysis, we have to stick to ordinary power (**SIST** 355).

Ordinary Power POW (μ'): $\Pr(d(\mathbf{X}) > c_\alpha; \mu')$
Shpower: Observed or retro-power: $\Pr(d(\mathbf{X}) > c_\alpha; \mu = \bar{x}_0)$

An article by Hoenig and Heisey (2001) ("The Abuse of Power") calls power analysis abusive. Is it? Aris Spanos and I say no (in a 2002 note),

Power-analytic reasoning: High power to get significance when μ = μ', together with your *not getting significance* indicates μ < μ'

But if μ' replace μ' by $\bar{x}_0$,it will never be high.

Exhibit (vii) (SIST, p. 359): Gelman and Carlin (2014) appear to be at odds with the upshot of quiz on p. 323, start of Tour I.

From our mountains out of molehill fallacies, if POW(μ') is high then a just significant result is *poor* evidence that μ > μ'; while if POW(μ') is low it's good evidence that μ > μ'.

A way to make sense of their view is to see them as saying if the observed mean is so out of whack with what's known, that we suspect the assumptions of the test are questionable or invalid.

You have grounds to question the low power computation because you question the reported error probabilities, be it due to selective reporting, publication bias, or violated statistical assumptions. (See **SIST** pp. 360-1)

5.6 Positive Predictive Value: Fine for Luggage (SIST 361)

To understand how the *diagnostic screening* criticism tests really took off, go back to a paper by John Ioannidis (2005).

> Several methodologists have pointed out that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles should be interpreted based only on p-values. …

> It can be proven that most claimed research findings are false (p. 0696).

However absurd such behavior sounds, 70 years after Fisher exhorted us never to rely on "isolated results," let's suppose Ioannidis is right.

But it gets worse. Even the single significant result is very often the result of the cherry-picking, and barn-hunting we are all too familiar with.

> Commercially available 'data mining' packages actually are proud of their ability *to yield statistically significant results through data dredging* (ibid., p. 0699).

The DS criticism of tests shows that if:

1. you publish upon getting a single P-value < .05,
2. you dichotomize tests into "up-down" outputs rather than report discrepancies and magnitudes of effect,
3. you data dredge, and cherry-pick and/or
4. there is a sufficiently low prevalence of genuine effects in your field

then the proportion of true nulls among those found statistically significant– (FFR)–differs from and can be much greater than the Type I error set by the test.

For the severe tester, committing #3 alone is suspect, unless we adjustment to get proper error probabilities

High prevalence of true hypotheses in your field should not atone for this sin.

Diagnostic Screening

- *If we imagine randomly selecting a hypothesis from an urn of nulls 90% of which are true*
- *Consider just 2 possibilities H_0: no effect, H_1: meaningful effect, all else ignored*
- *Take the prevalence of 90% as Pr(H_0 you picked)= .9, Pr(H_1) = .1*
- *Rejecting H_0 with a single (just .05) significant result, cherry picking to boot*

The unsurprising result is that most "findings" are false:
$Pr(H_0|$ findings with a P-value of .05) ≠ Pr (reject at level .05; H_0

Only the second one is a Type 1 error probability)

Positive Predictive Value (PPV) (1 – FFR). To get the (PPV) we are to apply Bayes' rule using the given relative frequencies (or prevalences):

$$\text{PPV: } \Pr(\text{D}|+) = \frac{\Pr(+|\text{D})\,\Pr(\text{D})}{[\Pr(+|\text{D})\,\Pr(\text{D}) + \Pr(+|\sim\text{D})\,\Pr(\sim\text{D})]}$$

$$= \frac{1}{(1+\text{B})}$$

$$\text{where} \qquad \text{B} = \frac{\Pr(+|\sim\text{D})\,\Pr(\sim\text{D})}{\Pr(+|\text{D})\,\Pr(\text{D})}$$

Sensitivity

SENS: Pr(+|D).

H_1: D: Dangerous bag

(~ power)

H_0: ~D: no danger

Specificity

SPEC: Pr(–|~D);

(1 – α)

Even with Pr(D) = .5, with Pr(+|~D) = .05 and Pr(+|D)= .8, we still get a rather high

$$\text{PPV} = \frac{1}{\left[\frac{1+Pr(+|\sim D)}{Pr(+|D)}\right]}$$

1/ (1 + 1/16) = 16/17

With Pr(D) = .5, all we need for a PPV greater than .5 is for Pr(+|~D) to be less than Pr(+|D).

With a small prevalence Pr(D) e.g., < Pr(+|~D) (< α)

We get PPV < .5 even with a maximal sensitivity Pr(+|D) of 1. In

There is still a boost from the prior prevalence.

Recall absolute vs relative confirmation (B – boost)

Chart SIST 365

What is prevalence? (bott 366)

Probabilistic instantiation fallacy *(367)* The outcome may be ***X*** = 1 or 0 according to whether the hypothesis we've selected is true.

The probability of ***X*** = 1 is .5, it does not follow that a specific hypothesis we might choose–say, your blood pressure drug is effective–has a probability of .5 of being true, for a frequentist-

Other problems arise is using the terms from significance tests for FFR or PPV assessments: Pr(+|D) and Pr(+|~D) in the DS criticism.

The DS model of tests considers just two possibilities “no effect” and “real effect”.

H_0: 0 effect ($\mu = 0$),
H_1: the discrepancy against which the test has power $(1 - \beta)$.

It is assumed the probability for finding any effect, regardless of size, is the same.

$[\alpha/(1 - \beta)]$ used as the likelihood ratio to get a posterior of H_1

If the H_1 for which $(1 - \beta)$ is high, they take it as high likelihood for H_1

That’s why this is on a chapter on power.

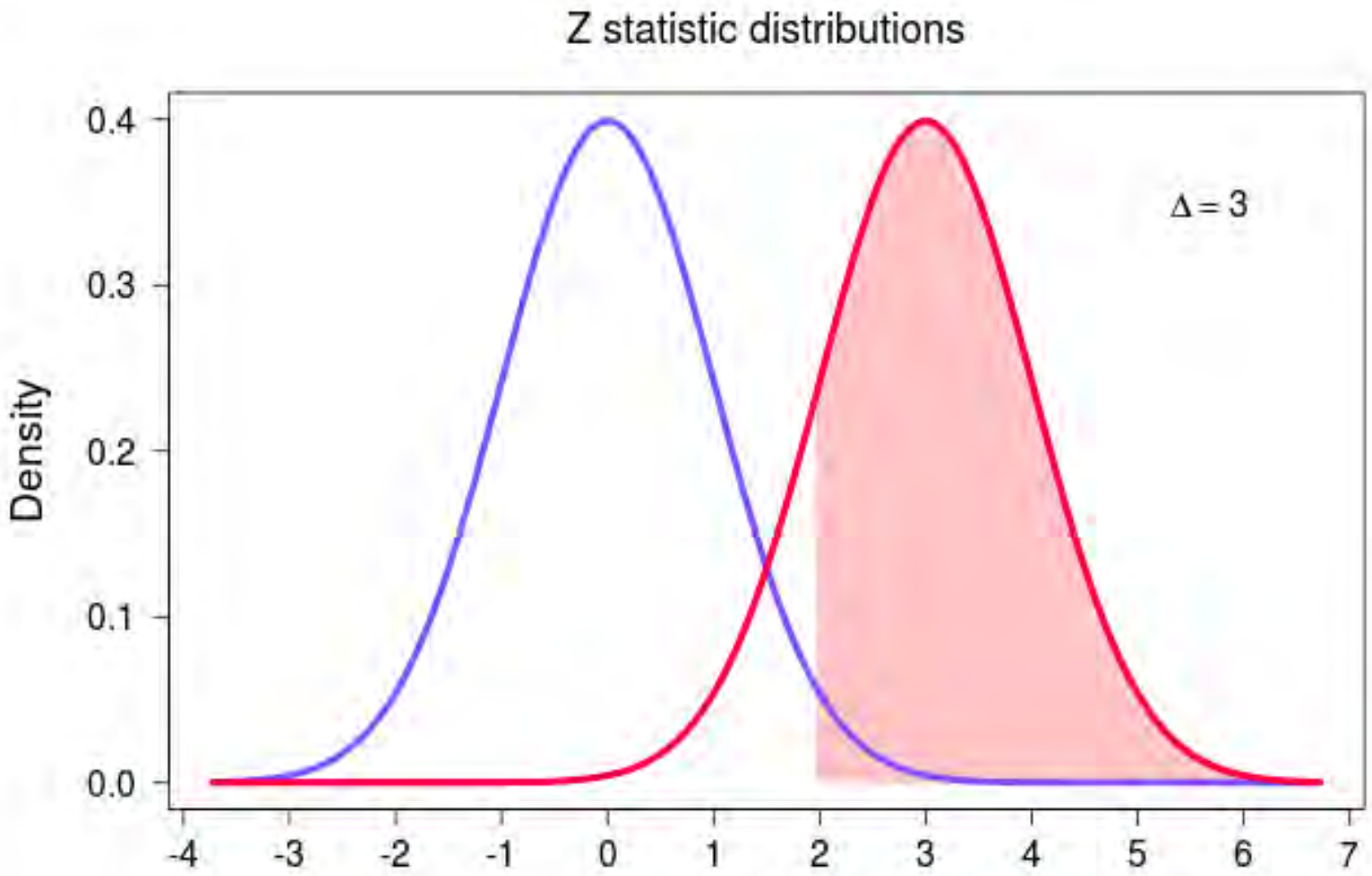
For an H_1 where (1 – β) is high, take our H_1

H_1: $\mu \geq \mu_{.84}$

$\mu_{.84}$ is the alternative against which the test has .84 power.

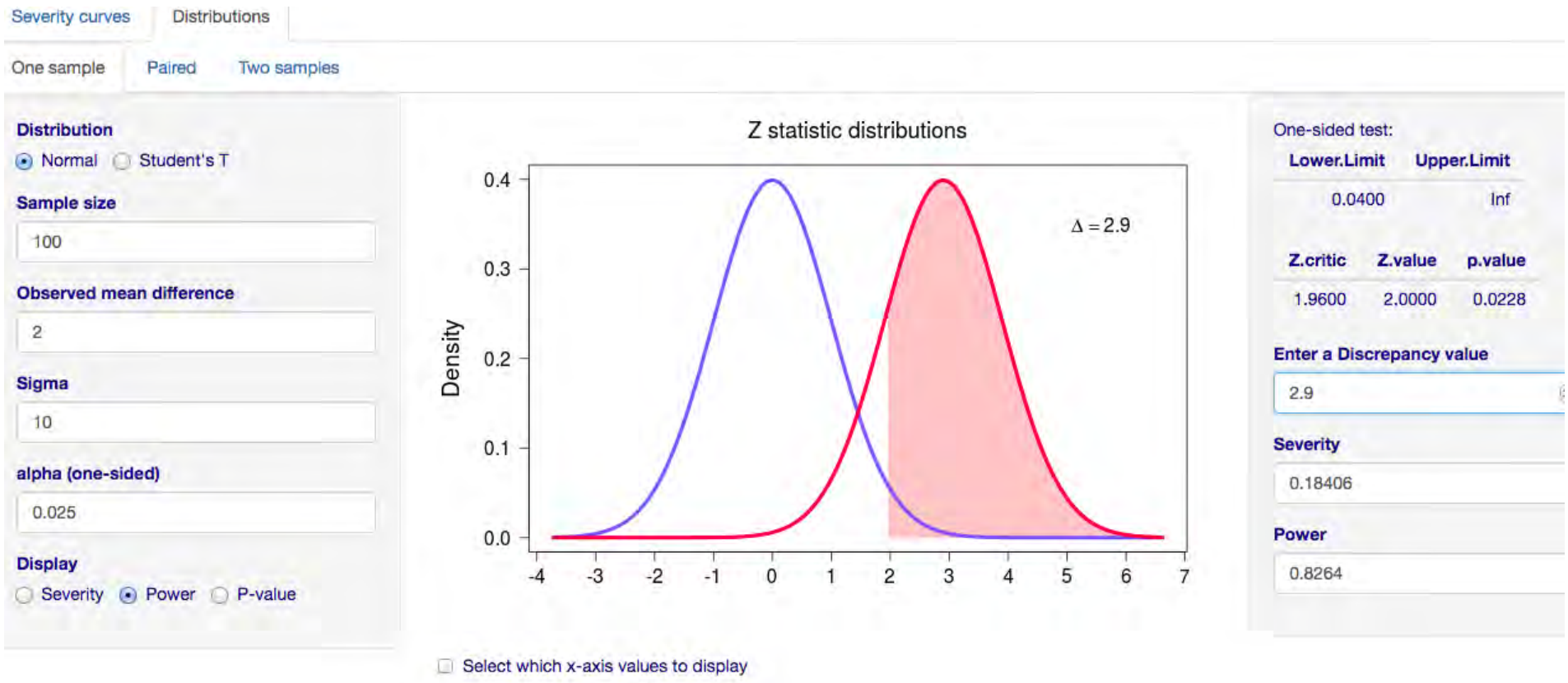
But now the denial of the alternative H_1 is not the same null hypothesis used to get *Type I error probability of .05.*

Instead it would be high, nearly as high as .84.



alternative is $\mu_{.84}$ (3, in our example)

e.g., let alternative be 2.9, *Type I error probability .82*



Likewise if the null $\mu \leq \mu_0$ is to have low α, its denial won't be one against which the test has high power (it will be close to α).

High power requires a μ exceeding the cut-off for rejecting *at level* α

We have to assume they have in mind a test between a point null H_0, or a small interval around it, and a *non-exhaustive* alternative hypothesis H_1: $\mu = \mu_{.84}$

Problem: To infer $\mu_{.84}$ based on $\alpha = .025$ (one-sided) is to be wrong 84% of the time.

We'd expect a more significant result 84% of the time were $\mu_{.84}$.

Same problem as with Johnson.

Back to the more general problem with the DS model

Is the PPV computation *relevant* to what working scientists want to assess: strength of the *evidence* for effects or its degree of corroboration?

Crud Factor. In many fields of social and biological science it's thought nearly everything is related to everything: "all nulls false".

> These relationships are not, I repeat, Type I errors. They are facts about the world, and with N – 57,000 they are pretty stable. Some are theoretically easy to explain, others more difficult, others completely baffling. The 'easy' ones have multiple explanations, sometimes competing, usually not. (Meehl, 1990, p. 206).

He estimates the crud factor at around .3 or .4.

High prior prev gives high posterior prev

Will we be better able to replicate results in a field with a high crud factor?

By contrast: Even in a low prevalence situation, if I've done my homework, went beyond the one P-value, developed theories, I may have a good warrant for taking the effect as real.

Avoiding biasing selection effects and premature publication is what's doing the work, not prevalence.

The PPV doesn't tell us how valuable the statistically significant result is for predicting the truth or reproducibility of *that effect*.

We want to look at how well tested the particular hypothesis of interest is.

Suppose we find it severely tested.

Granted, we might assess the probability with which hypotheses pass so stringent a test, if false.

We have come full circle to evaluating the severity of tests passed. *Prevalence has nothing to do with it.*

SKIP The Story of Isaac

Isaac is a high school student who has passed (+) a battery of tests for D: "college-readiness": P(+|~D) is .05, P(+|D) ~1.

Isaac from Fewready town, Pr(D) = .001, so Pr(D|+) is still low.

Had Isaac been randomly selected from Manyready suburbs, P(D|+) is high.

Isaac would have to score quite a bit higher than if he had come from Manyready town for the same PPV.

There is a real policy question here. Should we demand higher test scores from students in Fewready town? Or is it reverse affirmative action?

The Dangers of the Diagnostic Screening Model for Science

> Large-scale evidence should be targeted for research questions where the pre-study probability is already considerably high, so that a significant research finding will lead to a post-test probability that would be considered quite definitive (Ioannidis, 2005, p. 0700).

The DS model has mixed up the probability of a Type I error (often called the "false positive rate") with the posterior probability: False Finding Rate FFR: Pr(H_0|H_0 is rejected).

In frequentist tests, reducing the Type II error probability results in *increasing* the Type I error probability: there is a trade-off.

In the DS model, the trade-off disappears: reducing the Type II error rate also reduces the FFR.

5.7 Statistical Theatre: "Les Miserables Citations" (SIST 371)

We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability* can by itself provide any valuable evidence of the truth or falsehood of that hypothesis.

But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior 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 1966, pp. 141-2/1933, pp. 290-1).

They are invariably put forward as proof that N-P tests are relevant only for a crude long-run performance goal.

I will deconstruct them

In a nutshell: I now see it as Neyman's attempt to avoid the skeptism over the possibility of inductively learning (that Fisher sought) but avoiding Fisher's problem regarding:

(a) the choice of a (possibly data dependent alternative) to sustain good error probability control

(b) fallacy of probabilistic instantiation in his fiducial inference

The paper opens with a discussion of two French probabilists– Joseph Bertrand and Émile Borel, author of *Le Hasard* (1914,1948)!

"Les Miserables Citations ". (Lehmann's translation from the French is used where needed.)

The curtain opens with a young Neyman and Pearson (from 1933) standing mid-stage, lit by a spotlight. (All speaking parts are exact quotes; Neyman does the talking).

Borel: "The particular form that problems of causes often take**...**is the following: **Is such and such a result due to chance or does it have a cause?** It has often been observed how much this statement lacks in precision**.** Bertrand has strongly emphasized this point**. But ....to refuse to answer under the pretext that the answer cannot be absolutely precise, is to... misunderstand the essential nature of the application of mathematics."** ..."**If one has observed a [precise angle between the stars]...in tenths of seconds...one would not think of asking to know the probability [of observing exactly this observed angle under chance] because one would never have asked that precise question before having measured the angle'...**

The question is whether one has the same reservations in the case in which one states that one of the angles of the triangle formed by three stars has "*une valeur remarquable*" [a striking or noteworthy value], and is for example equal to the angle of the equilateral triangle…. (Lehmann 1993/2012, p. 964.)

Here is what one can say on this subject: **One should carefully guard against the tendency to consider as striking an event that one has not specified *beforehand*, because the number of such events that may appear striking, from different points of view, is very substantia**l (ibid., p. 968).

The stage fades to black, then a spotlight beams on Neyman and Pearson mid-stage.

N-P: [W]e may consider some specified hypothesis, as that concerning the group of stars, and **look for a method which we should hope to tell us, *with regard to a particular group of stars*, whether they form a system, or are grouped 'by chance,'…their relative movements unrelated."** (1933, p. 140/290)

"If this were what is required of 'an efficient test', we should agree with Bertrand in his pessimistic view. …Indeed, if ***x*** is a continuous variable—as for example is the angular distance between two stars—then any value of ***x*** is a singularity of relative probability equal to zero.

We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis. But we may look at the purpose of tests from another view-point."

What if we follow Borel who insisted that: (a) the criterion to test a hypothesis (a 'statistical hypothesis') using some observations must be selected *not after the examination of the results of observation*, but before, and (b) this criterion should be a function of the observations 'en quelque sorte remarquable' [of a remarkable sort].

It is these remarks of Borel that served as an inspiration to Egon S. Pearson and myself in our effort to build a frequentist theory of testing hypotheses."(Neyman 1977, pp. 102-103.)

Inferential Rationales for Test Requirements

It's not hard to see that "*as far as a particular*" star grouping is concerned, we cannot expect a reliable inference to just any non-chance effect discovered in the data.

To cope with the fact that any sample is improbable in some respect, statistical methods do one of two things: appeal to prior probabilities or to error probabilities of a procedure.

…. The latter says, we need to consider the problem as of a *general* type. It's a general method, from a test statistic to some assertion about an alternative hypothesis, expressing the non-chance effect.

The Deconstruction So Far

If we accept the words, "an efficient test of the hypothesis *H*" to mean a statistical (methodological) falsification rule that controls the probabilities of erroneous interpretations of data, and ensures the rejection was *because* of the underlying cause (as modeled), then efficient tests are possible.

This requires (i) a prespecified test criterion to avoid verification biases while ensuring power (efficiency), and

(ii) consideration of alternative hypotheses to avoid fallacies of acceptance and rejection.

Fisher is to be credited, Pearson remarks, for his "emphasis on planning an experiment, which led naturally to the examination of the power function, (1962, p. 277). If you're planning, you're prespecifying.

Moreover, the test "criterion should be a function of the observations," and the alternatives, such that there is a known statistical relationship between the characteristic of the data and the underlying distribution (Neyman 1977, pp. 102-103).

An exemplary characteristic of this sort are the remarkable properties offered by pivotal test statistics such as Z or T, whose distributions are known.

$$Z = \sqrt{n}\,(\bar{X} - \mu)/\sigma$$

$$T = \sqrt{n}\,(\bar{X} - \mu)/\sigma$$

Z is the standard Normal distribution, and T the Student's T distribution, where σ is unknown and thus replaced by the estimator.

Consider the pivot Z. We know it's distribution is standard Normal. The probability Z > 1.96 is .025. But by pivoting, the Z > 1.96 is equivalent to

$$\mu < \bar{X} - 1.96\,\sigma/\sqrt{n},$$

so it too has probability .025.

Therefore, the procedure that asserts $\mu > \bar{X} - 1.96\sigma/\sqrt{n}$ asserts correctly 95% of the time!

We can make valid probabilistic claims about the method that hold post-data, *if interpreted correctly*.

This leads us to Fisher's Fiducial territory, and the initial development of the behavioral performance idea.

5.8 Neyman's Performance and Fisher's Fiducial Probability (SIST 382)

So what is fiducial inference? I begin with Cox's contemporary treatment:

> We take the simplest example,…the normal mean when the variance is known, but the considerations are fairly general. The lower limit
>
> $\bar{x}_0 - z_c\sigma/\sqrt{n}$
>
> derived from the probability statement
>
> $\Pr(\mu > \bar{X} - z_c\sigma/\sqrt{n}) = 1 - c$
>
> is a particular instance of a *hypothetical* long run of statements a proportion 1 – c of which will be true, assuming our model is sound.
>
> (Cox 2006, p. 66)

Once $\bar{x}_0$ is observed, $\bar{x}_0 - z_c\sigma/\sqrt{n}$ is what Fisher calls the *fiducial c per cent limit* for μ. The collection of such statements for different c's yields a fiducial distribution.

Here's Fisher in the earliest paper on fiducial inference in 1930. He sets 1 – c as .95 per cent.

> [W]e have a relationship between the statistic [$\bar{X}$] and the parameter μ, such that $\bar{x}_{.95}$ is **the 95 per cent. value corresponding to a given μ, a**nd this relationship implies the perfectly objective fact that in 5 per cent. of samples $\bar{X} > \bar{x}_{.95}$. (That is, $\Pr(\bar{X} < \mu + 1.65\sigma/\sqrt{n}) = .95$.)] (Fisher 1930, p. 533)

The 95 per cent. value $\bar{x}_{.95}$.

In the normal testing example, $\bar{x}_{.95} = \mu + 1.65\sigma/\sqrt{n}$.

In 95% of samples $\bar{X} < \bar{x}_{.95}$.

$\bar{x}_{.95}$ is the cut-off for a .05 one-sided test T+ (of $\mu \leq \mu_0$ vs. $\mu > \mu_0$).

$\bar{X} \geq \bar{x}_{.95}$ occurs whenever $\mu < \bar{X} - 1.65\sigma/\sqrt{n}$.

Reject the null at level .05 whenever $\mu <$ the lower bound of a .95 CI.

For a particular observed $\bar{x}_0$, $\bar{x}_0 - 1.65\sigma/\sqrt{n}$ is the 'fiducial 5 per cent. value of μ'.

> We may know as soon as $\bar{X}$ is calculated what is the fiducial 5 per cent. value of μ, *and that the true value of μ will be less than this value in just 5 per cent. of trials.* This then is a definite probability statement about the unknown parameter μ which is true irrespective of any assumption as to it's *a priori* distribution. (ibid., emphasis is mine).[i]

This seductively suggests $\mu < \mu_{.05}$ gets the probability .05–a fallacious probabilistic instantiation, for a frequentist.

However, a kosher probabilistic statement about Z is "a particular instance of a hypothetical long run of statements 95% of which will be true."

So, what is being assigned the fiducial probability?

SIST, 383 Fisher: "we may infer, without any use of probabilities a priori, a frequency distribution for μ which shall correspond with the with the aggregate of all such statements…to the effect that the probability μ is less than $\bar{x}$ – 2.145 s/$\sqrt{n}$ is exactly one in forty" (Fisher 1936, p. 253).

Suppose you're Neyman and Pearson working in the early 1930s aiming to clarify and justify Fisher's methods. 'I see what's going on':

The method outputs statements with a probability (some might say a propensity) of .975 of being correct.

"We may look at the purpose of tests from another viewpoint": probability ensures us of the performance of a method.

1955-6 Triad: Telling what's true about the Fisher-Neyman conflict SIST: 388

Fisher 1955, Pearson 1955, and Neyman 1956.

Neyman, thinking he was correcting and mproving my own early work on tests of significance as a means to the "improvement of natural knowledge", in fact reinterpreted them in terms of that technical and commercial apparatus which is known as an acceptance procedure. ... (pp. 69-70.)

Pearson's (1955) response: "To dispel the picture of the Russian technological bogey, [I was sitting on a gate at my cousins agricultural station, the one whose fiancé Pearson fell in love with] I was "smitten" by an absence of logical justification for some of Fisher's tests, and I turned to Neyman to help me solve the problem.

This takes us to where we began: the miserable passages, pinning down the type of character in the test statistic, the need for the alternative and power considerations.

The second part is Neyman fixing Fisher's assigning the probability to a particular interval in fiducial inference.

(**SIST** 389)

According to Fisher, Neyman violates "…the principles of deductive logic [by accepting a] general symbolical statement such as

[1] $\Pr\{(\bar{x} - ts) < \mu < (\bar{x} + ts)\} = \alpha$,

as rigorously demonstrated, and yet, when numerical values are available for the statistics $\bar{x}$ and s, so that on substitution of these and use of the 5 per cent. value of t, the statement would read

[2] $\Pr\{92.99 < \mu < 93.01\} = .95$ per cent.,

to deny to this *numerical* statement any validity. This evidently is to deny the syllogistic process" (Fisher 1955, p. 75).

But the move from (1) to (2) is fallacious!

I. J. Good describes how many felt, and still feel:

> It seems almost inconceivable that Fisher should have made the error which he did in fact make. [That is why] …so many people assumed for so long that the argument was correct. They lacked the *daring* the question it. (Good 1971, In reply to comments on his paper in Godambe and Sprott).

Neyman (1956):"It is doubtful whether the chaos and confusion now reigning in the field of fiducial argument were ever equaled in any other doctrine. The source of this confusion is the lack of realization that equation (1) does not imply (2)" (ibid. p. 293).

"Bartlett's revelation [1936, 1939] that the frequencies in repeated sampling … need not agree with Fisher's solution" in the case of a difference between two normal means with different variances.

It was the collapse of Fisher's rebuttals that led Fisher to castigate N-P for assuming error probabilities and fiducial probabilities *ought* to agree, declaring the idea "foreign to the development of tests of significance." (**SIST** 390)

Statistician Sandy Zabell (1992):"such a statement is curiously inconsistent with Fisher's own earlier work" ; because of Fisher's stubbornness "he engaged in a futile and unproductive battle with Neyman which had a largely destructive effect on the statistical profession" (Zabell 1992 p. 382).

The Fisher-Neyman dispute is pathological (SIST 390)
There's no disinterring the truth of the matter.

Fisher grew to renounce performance goals he himself had held when it was found fiducial solutions disagreed with them.

Inability to identify conditions wherein the error probabilities "rubbed off"–where there are no "recognizable subsets" with a different probability of success–led Fisher to apparently reject error probabilities.

Fisher may have started out seeing fiducial probability as both a frequency of correct claims in an aggregate and a rational degree of belief (1930,p. 532), but the difficulties in satisfying uniqueness led Fisher to give up the former.