# History and Philosophy of Statistical Testing: Fisher, Neyman and Pearson

**Deborah G. Mayo**

November 18, 2019

Rutgers, Dept. of Statistics & Biostatistics

Prof. Glenn Shafer: History of Probability and Statistics

Rm 211, SEC (T. Alexander Pond Science &
Engineering Resource Center)

# ******Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars***

(2018 CUP)

SIST* 119-121, 131-133, 137-140, 371-378, 382-391

- I am not a historian of statistics; but a philosopher of science

- Philosophers decided that philosophy of science without history of science was inadequate ~1970s

- The history (and the philosophy) of statistics, in the period that interests me, calls for disentangling philosophical preconceptions underlying received views

# How to Get Beyond the Stat Wars Requires Chutzpah

- I set sail with a very simple tool: If little if anything has been done to probe flaws in a claim, then there's poor evidence for it

- But today I'm just doing history

- Famous disputes between Neyman, Pearson and Fisher intertwine personality, professional, philosophical disagreements

# Where are members of our cast of characters in 1919? (p. 120)

***Fisher***

In 1919, Fisher accepts a job as a statistician at Rothamsted Experimental Station.

- A more secure offer by Karl Pearson (KP) required KP to approve everything Fisher taught or published
- A subsistence farmer

# Fisher & Family

*Plate 11. Mrs. Fisher 1938, with daughters, in order of age, Margaret (top right), Joan (bottom right), Phyllis (top left), Elizabeth (bottom left), Rose standing beside her chair, and June in her lap.*

*Plate 10. R. A. Fisher, 1938, with sons George (aged 18) and Harry (14).*

# *Neyman*

In 1919 Neyman is living a hardscrabble life in Poland, sent to jail for a short time for selling matches for food,

- Sent to KP in 1925 to have his work appraised.

## *Pearson*

Pearson (Egon) gets his B.A. in 1919, goes to study with Eddington at Cambridge the next year (on the theory of errors)

He describes the psychological crisis he’s going through when Neyman arrives in London:

“I was torn between conflicting emotions: a. finding it difficult to understand R.A.F., b. hating [Fisher] for his attacks on my paternal ‘god,’ c. realizing that in some things at least he was right”

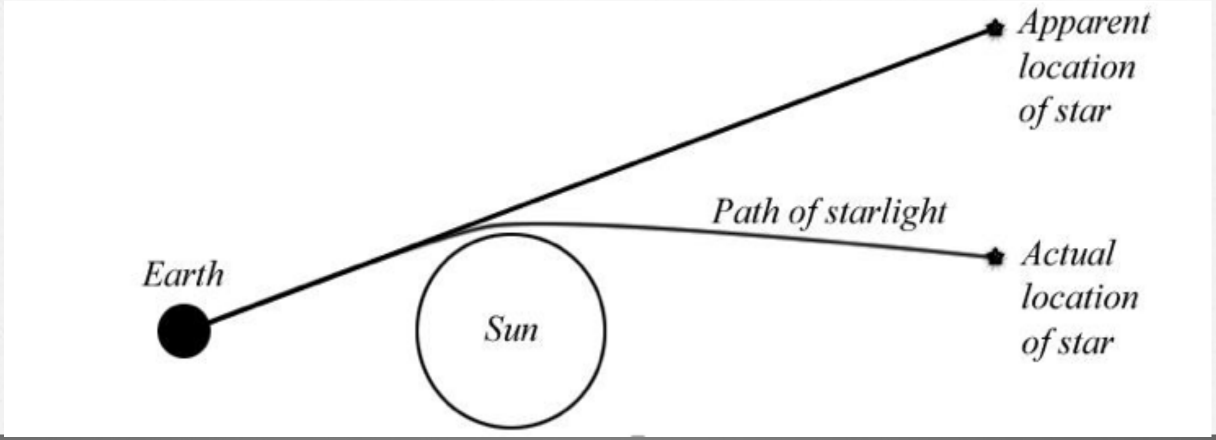
(Reid, C. 1998, p. 56).

# 100 Years Ago:

**6 November 2019:** the centenary of the joint meeting of the Royal Society and the Royal Astronomical Society when the eclipse results testing Einstein's General Theory of Relativity were announced

(the eclipse was in May 29, 2019)

On Einstein's theory, light passing near the sun is deflected by an angle λ, reaching 1.75", for light just grazing the sun.



# Two key stages of inquiry

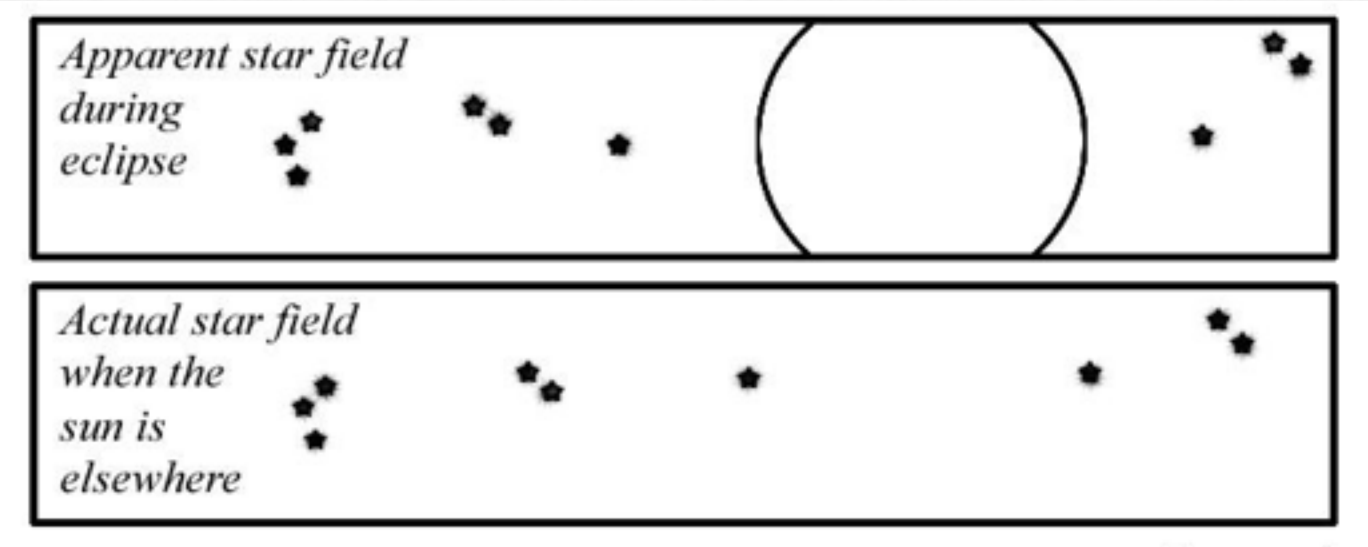
i. is there a deflection effect of the amount predicted by Einstein as against Newton (0.87")?

ii. is it "attributable to the sun's gravitational field" as described in Einstein's hypothesis?

Using known eclipse effect to explain it while saving Newton from falsification is unproblematic– if each conjecture is severely tested.

Eclipse photos of stars (eclipse plate) compared to their positions photographed at night when the effect of the sun is absent (the night plate)–a control.

Technique was known to astronomers from determining stellar parallax, "for which much greater accuracy is required" (Eddington 1920, pp. 115-16).


Apparent star field during eclipse
Actual star field when the sun is elsewhere

The problem in (i) is reduced to a statistical one: the observed mean deflections (from sets of photographs) are normally distributed around the predicted mean deflection μ.

$H_0$: $\mu \leq 0.87$ (Newton) vs $H_1$: $\mu > 0.87$

$H_1$: includes the Einsteinian value of 1.75.

2 expeditions, to Sobral, North Brazil and Principle, Gulf of Guinea (West Africa)

~5 months checking instrumental and other errors to announce “real effect”, GTR interpretation ~1921

**Sobral:** μ = 1.98" ± 0.18".
**Principe:** μ = 1.61" ± 0.45".

(in probable errors 0.12 and 0.30 respectively, 1 probable error is 0.68 standard errors SE.)

“It is usual to allow a margin of safety of about twice the probable error on either side of the mean.” [~1.4 SE]. The Principle plates are just sufficient to rule out the the ‘half-deflection’, the Sobral plates exclude it (Eddington 1920, p. 118).

# Simple significance tests (Fisher)

“**p-value**. …to test the conformity of the particular data under analysis with $H_0$ in some respect:

…we find a function $T = t(\mathbf{y})$ of the data, the **test statistic**, such that

- the larger the value of $T$ the more inconsistent are the data with $H_0$;
- $T = t(\mathbf{Y})$ has a known probability distribution when $H_0$ is true.

…the p-value corresponding to any $t_{0bs}$ as

$$p = p(t) = Pr(T \geq t_{0bs}; H_0)\text{”}$$

(Mayo and Cox 2006, 81)

# Testing reasoning

- If even larger differences than $t_{0bs}$ occur fairly frequently under $H_0$ (i.e., P-value is not small), there's scarcely evidence of incompatibility with $H_0$
- Small P-value indicates *some* underlying discrepancy from $H_0$ because **very probably you would have seen a less impressive** difference than $t_{0bs}$ were $H_0$ true.
- This still isn't evidence of a genuine statistical effect $H_1$, let alone a scientific conclusion $H^*$

Stat-Sub fallacy **$H \Rightarrow H^*$**

*Major "crisis" these days is said to be inferring* from a single (arbitrary) P-value that pertains to a *statistical* hypothesis $H_0$ *to a research claim H**

"[W]e need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result." (Fisher 1947, p. 14)

# More Test Reasoning

- Were $H_0$ a reasonable description of the process, then with very high probability you would not be able to regularly produce statistically significant results.

- So if you do, it's evidence $H_0$ is false in the particular manner probed.

- This is the basis for falsification in science.

# Neyman-Pearson (N-P) tests:

A null and alternative hypotheses $H_0$, $H_1$ that are exhaustive*

$H_0$: $\mu \leq 0$ vs. $H_1$: $\mu > 0$

"no effect" vs. "some positive effect"

A N-P test (naked math form): a rule that tells you when to "accept"/"reject" hypotheses so that the probability of erroneous rejections and non-rejections are controlled at low values. (SIST p. 140)

Introduces Type II error, and power

## ***So What's in a Test? (p. 129-130)*:**

We proceed by setting up a specific hypothesis to test, $H_0$ in Neyman's and my terminology, the null hypothesis in R. A. Fishers…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….:

**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. 143)

- After Neyman’s year at University College (1925/6), Pearson writes to him of his doubts (“suddenly smitten” with doubts due to Fisher)
- Neyman had just returned from his fellowship years to a hectic and difficult life in Warsaw, working multiple jobs in applied statistics. (SIST p. 131-2)

[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 .. . (C. Reid 1998, 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).

Criteria for the test statistic $\mathrm{d}(\boldsymbol{X})$ are:

**i.** it reduces the data as much as possible<br>
**ii.** the larger $\mathrm{d}(\boldsymbol{x}_0)$ the further the outcome from what’s expected under $H_0$, with respect to the particular question;<br>
**iii.** can compute P-value $p(\boldsymbol{x}_0)=\mathrm{Pr}(\mathrm{d}(\mathbf{X}) > \mathrm{d}(\boldsymbol{x}_0); H_0)$.

# SIST p. 176 (not in reading)

It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result.. . It is usual and convenient for the experimenters to take 5 per cent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results.

(Fisher 1935a, pp. 13–14)

# N-P emulate Fisher

SIST p. 176 (not in reading)

“Our examination of the possible results of the experiment has therefore led us to a statistical test of significance, by which these results are divided into two classes with opposed interpretations . . . those which show a significant discrepancy from a certain hypothesis; . . . and on the other hand, results which show no significant discrepancy from this hypothesis.” (pp. 15–16)

(Fisher 1935a, DOE pp. 15–16)

# SIST p. 177 (not in reading)

“Lehmann is flummoxed by the association of fixed levels of significance with N-P since ‘[U]nlike Fisher, Neyman and Pearson (1933, p. 296) did not recommend a standard level but suggested that ‘ how the balance [between the two kinds of error] should be struck must be left to the investigator’” (Lehmann 1993b, p. 1244).

Neyman and Pearson stressed that the tests were to be “used with discretion and understanding” depending on the context (Neyman and Pearson 1928, p. 58).

# Water Plant (SIST p. 142)

1-sided testing the mean of a Normal distribution

$H_0$: μ ≤ 150  vs. $H_1$: μ > 150  (Let σ = 10, $n$ = 100)

let significance level α = .025 (round to 2 SE)

Reject $H_0$ whenever $\bar{X} \geq 150 + 2\sigma/\sqrt{n}$

↘ $\bar{x}_{.05}$

$\bar{X}$ is the sample mean, its value is $\bar{x}$ .

1SE = $\sigma/\sqrt{n} = 1$

# Rejection rules:

Reject iff $\bar{X}$ > 150 + 2SE (N-P)

(infer evidence against, take action)

**In terms of the P-value:**

Reject (or declare stat significance) iff
P-value ≤ .025 (Fisher)

(P-value a distance measure, but inverted)

Let $\bar{X}$ = 152, so I reject $H_0$.

# Some P-values

Let $\bar{X}$ = 152

Z = (152 – 150)/1 = 2

Z = (obs – $H_0$)/1 = 2

The P-value is Pr(Z > 2) = .025

# Some P-values

Let $\bar{X}$ = 151

Z = (151 – 150)/1 = 1

The P-value is Pr(Z > 1) = .16

# Some P-values

Let $\bar{X}$ = 150.5

Z = (150.5 – 150)/1 = .5

The P-value is Pr(Z > .5) = .3

# *Level of significance (SIST p. 174)*

For given observations y we calculate t = t(y),
…and the level of significance $p_{obs}$
by $p_{obs} = \Pr(T \geq t_{obs}; H_0)$.
. . . Hence pobs is the probability that we would mistakenly declare there to be evidence against $H_0$, were we to regard the data under analysis as just decisive against $H_0$. (Cox and Hinkley (1974):p. 66)

Significance level (under NP) is generally the fixed Type I error probability of a test

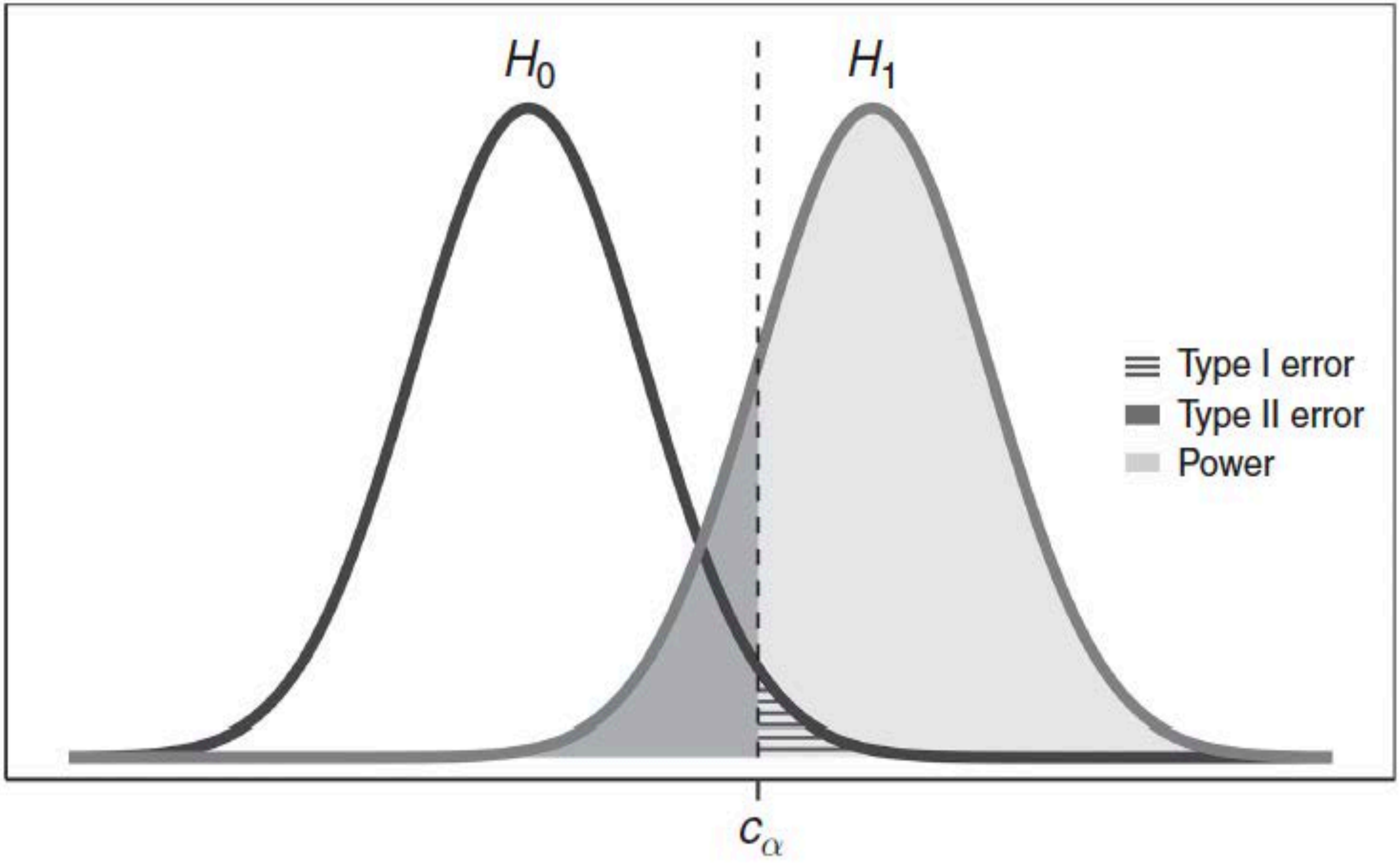


Figure 3.2 Type II error and power.

# *The power of test T+ against alternative μ = μ₁*

The prob the test would lead to reject $H_0$ when $\mu = \mu_1$.

Pr(Test T rejects $H_0$; $\mu = \mu_1$).

POW(T, $\mu_1$) = Pr(d($\boldsymbol{X}$) > $z_\alpha$; $\mu = \mu_1$): Power measures capability of a test to detect $\mu_1$
POW(T+, $\mu_1$ ) = Pr($\bar{X} > \bar{x}_\alpha$; $\mu_1$), $\bar{x}_\alpha = (\mu_0 + z_\alpha \sigma_{\bar{X}})$,

I put POW in terms of the $\bar{x}_\alpha$ value that just makes it to the cut-off, rather than using d

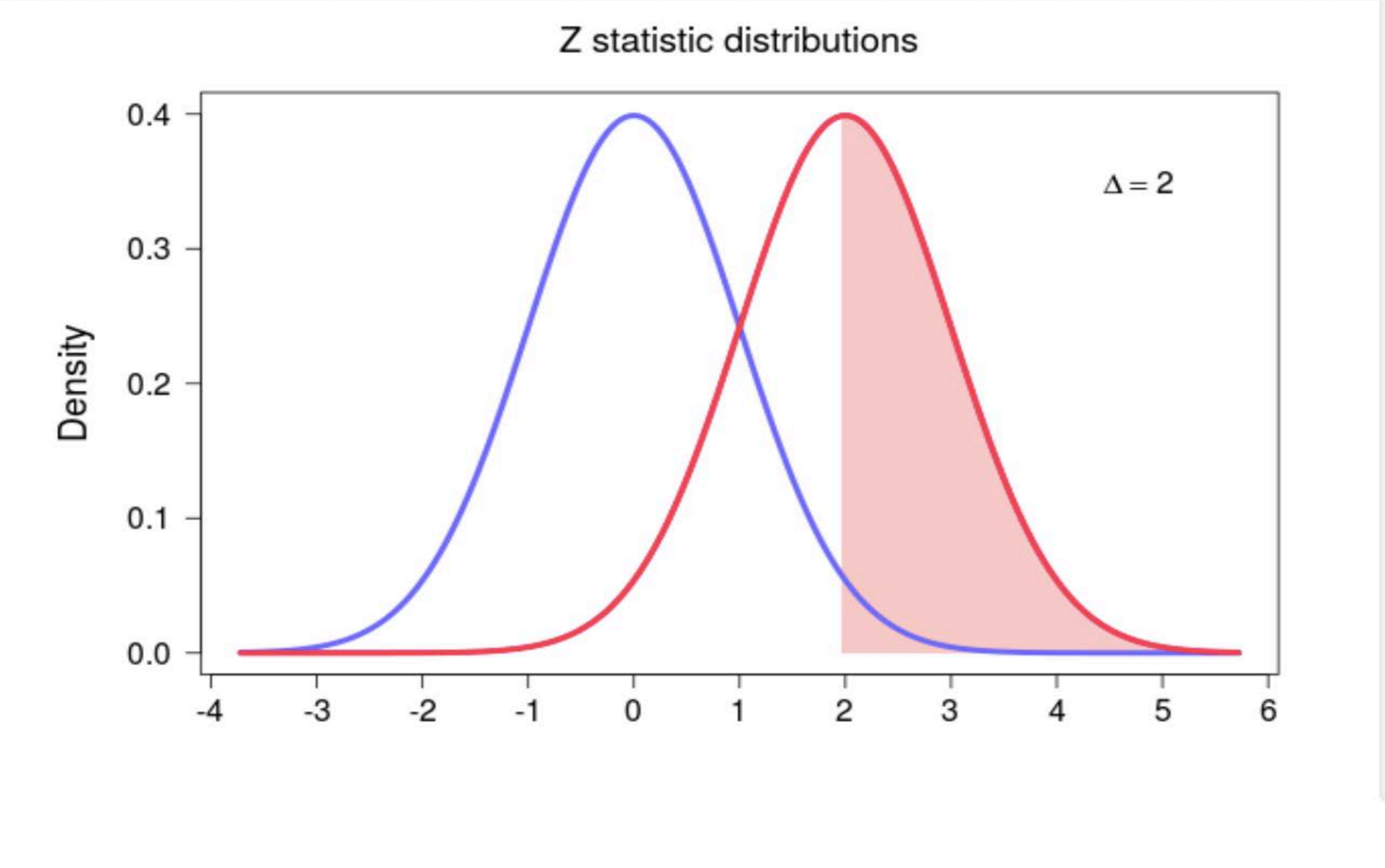
POW(T+, $\mu_0$ ) = Pr($\bar{X}$ > $\bar{x}_\alpha$; $\mu_1$), $\bar{x}_\alpha$ = ($\mu_0$ + $z_\alpha \sigma_{\bar{X}}$),

Standardize to get Z= ( $\bar{x}_\alpha$ - $\mu_1$)/SE =
(152 - $\mu_1$)/1

The POW is the probability to the right of Z under the standard Normal curve

Our cut-off $\bar{x}_\alpha$; is 152. Consider $\mu_1$ = 152,

POW(T+, μ = 152) = Pr($\bar{X}$ > 152; μ = 152)



**a.** The power of T+ for $\mu_1 = \bar{x}_\alpha$ is 0.5.

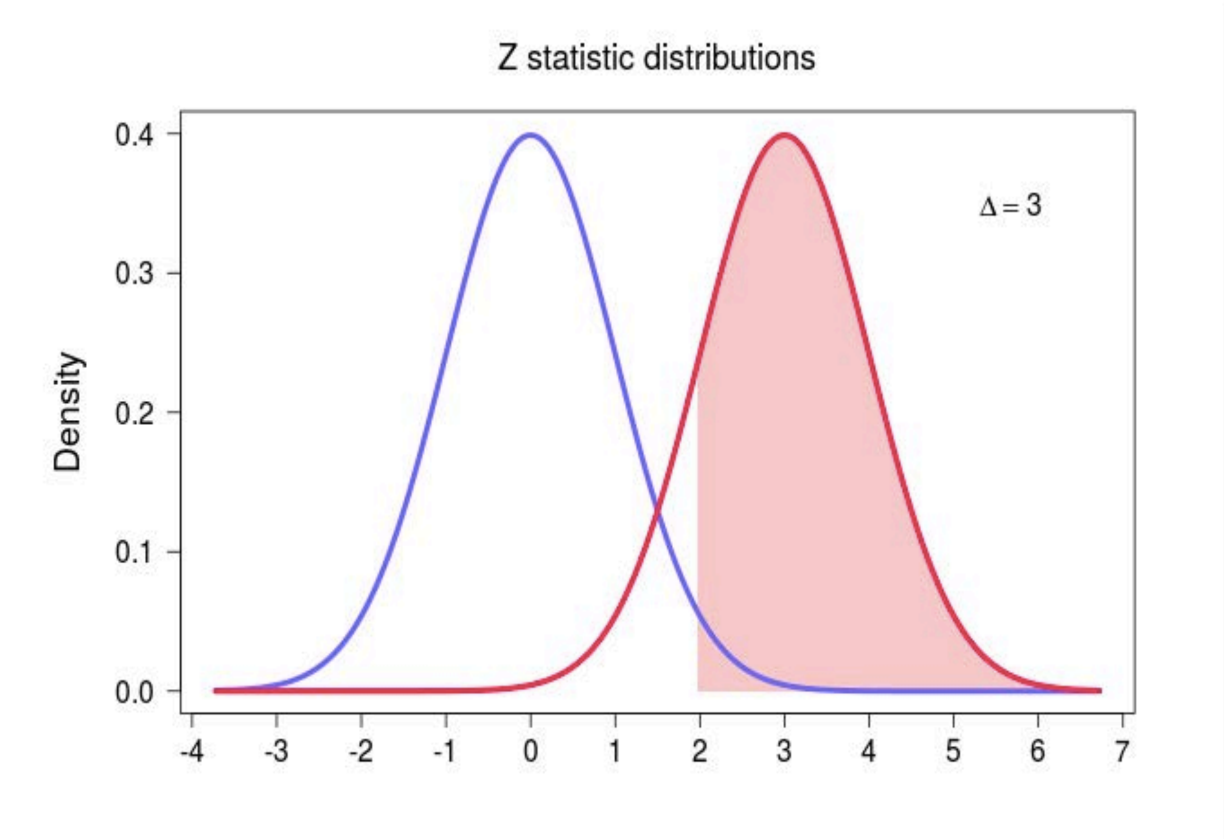
Z= ( 152 - 152)/SE = 0

Pr(Z > 0) = 0.5, so POW(T+, $\mu_1$ = 152) = 0.5.

***b.*** *The power of T+ for $\mu_1$*= 153.

Here, Z = (152 – 153) = -1

Pr(Z > -1) = .84, so in our example, POW(T+, $\mu_1$ = 153) = 0.84.



Neyman also developed confidence intervals ~1930

Back to the history SIST p. 146

Fisher writes to Neyman in summer of 1933 (cited in Lehmann 2011, p. 58):

> “You will be interested to hear that the Dept. of Statistics has now been …This arrangement will be much laughed at, but it will be rather a poor joke . . . I shall not lecture on statistics, but probably on ‘the logic of experimentation’.”

SIST p. 146:

When Karl Pearson retired in 1933, he refused to let his chair go to Fisher, so they split the department into two: Fisher Head of Eugenics, Egon Pearson Head of Applied Statistics.

They are one floor removed (Fisher on top)! The Common Room had to be “ carefully shared,”  as C. Reid puts it.

# Confidence Intervals CI’s

Just before Egon offers him a faculty position at University College starting 1934, Neyman gives a paper at the Royal Statistical Society (RSS) with a portion on confidence intervals, intending to generalize Fisher’s Fiducial intervals.

Arthur Bowley: “I am very glad Professor Fisher is present, as it is his work that Dr Neyman has accepted and incorporated.… I am not at all sure that the ‘confidence’ is not a confidence trick” (C. Reid 1998, p. 118).

Fisher was full of praise. “Dr Neyman…claimed to have generalized the argument of fiducial probability, and he had every reason to be proud of the line of argument he had developed for its perfect clarity (ibid).”

> Fisher had on the whole approved of what Neyman had said. If the impetuous Pole had not been able to make peace between the second and third floors of University College, he had managed at least to maintain a friendly foot on each! (E.S. Pearson, C. Reid 1998, p. 119)

In a CI estimation procedure, an observed statistic is used to set an upper or lower (1-sided) bound, or both upper and lower (2-sided) bounds for parameter μ.

Consider our test *T*+, $H_0$: $\mu \leq \mu_0$. against $H_1$: $\mu > \mu_0$.

The (1 – α) (uniformly most accurate) lower confidence bound for μ, which I write as $\hat{\mu}_{1-\alpha}(\bar{X})$, corresponding to test *T*+ is

$$\mu \geq \bar{X} - c_\alpha(\sigma/\sqrt{n})$$

$Pr(Z > c_\alpha) = \alpha$ where Z is the Standard Normal statistic.

## **Neyman develops CIs as inversions of tests**

(estimating μ in a Normal Distribution)

$\mu > \bar{x} - 1.96\sigma/\sqrt{n}$ CI-lower
$\mu < \bar{x} + 1.96\sigma/\sqrt{n}$ CI-upper

$\bar{x}$ : the observed sample mean

CI-lower: the value of μ that $\bar{x}$ is statistically significantly greater than at P= 0.025

CI-upper: the value of μ that $\bar{x}$ is statistically significantly lower than at P= 0.025

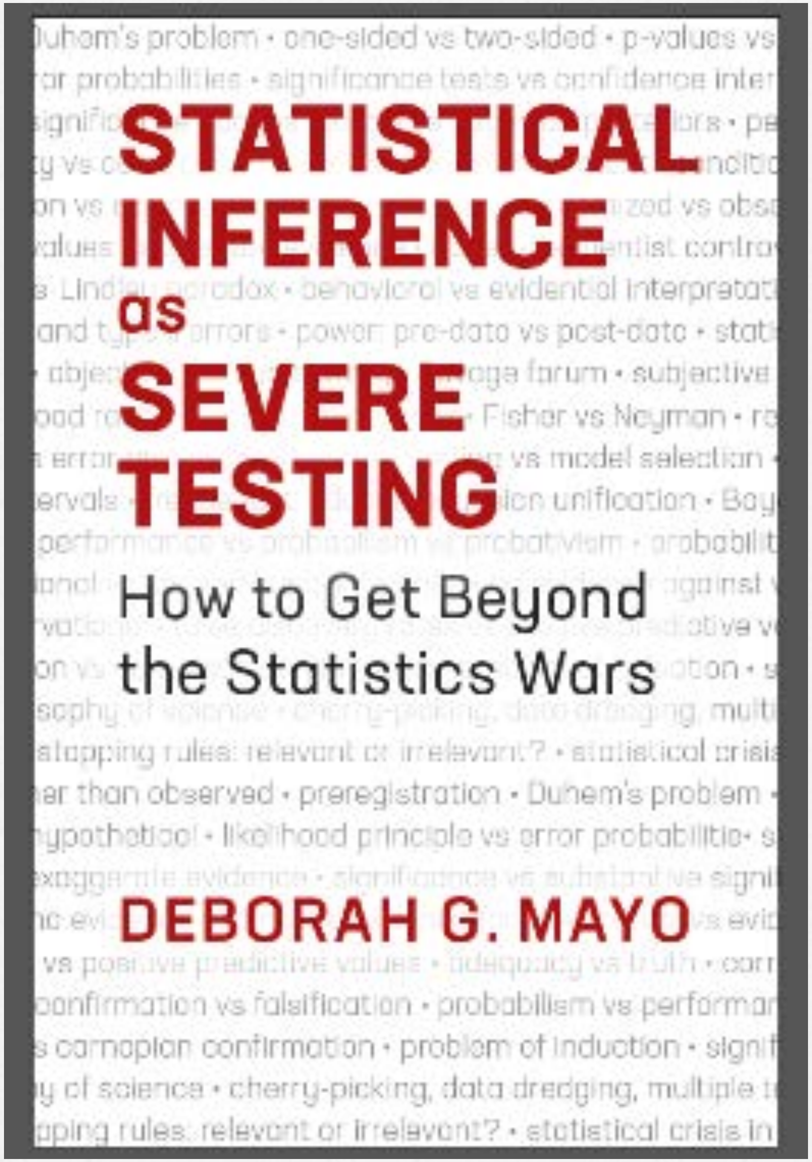
- You could get a CI by asking for these values, and learn indicated effect sizes with tests

# The N-P Revolution

“The work [of N-P] ..transformed mathematical statistics ” Some compared it  “to the effect of the theory of relativity upon physics ” (C. Reid 1998, p. 104).

Even when the optimal tests were absent, the optimal properties served as benchmarks to gauge the performance of methods could be gauged.

Paves the way for Abraham Wald ’s decision theory, and the work by Lehmann and others.



STATISTICAL INFERENCE as SEVERE TESTING

How to Get Beyond the Statistics Wars

DEBORAH G. MAYO

- Overshadows the more informal Fisherian tests.
- This irked Fisher–look what they’ve done to my baby.
- Famous feuds between Fisher and Neyman erupted as to whose paradigm would reign supreme starting 1935.

(SIST p. 139)

End Part I

# References

Eddington, A. ([1920]1987). *Space, Time and Gravitation: An Outline of the General Relativity Theory*, Cambridge Science Classics Series, CUP

Fisher, R. A. (1935a). *The Design of Experiments*, 1st edn., Edinburgh: Oliver and Boyd. Reprinted in Fisher 1990.

Fisher 1947 *The Design of Experiments*, Edinburgh: Oliver and Boyd

Lehmann, E. (2011). *Fisher, Neyman, and the Creation of Classical Statistics*, 1st edn. New York: Springer.

Mayo, D. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.

Mayo, D. and Cox, D. (2006). 'Frequentist Statistics as a Theory of Inductive Inference', in Rojo, J. (ed.), *Optimality: The Second Erich L. Lehmann Symposium*, Lecture Notes-Monograph series, Institute of Mathematical Statistics (IMS), 49, pp. 77–97.

# References cont.

Neyman, J. and Pearson, E. (1928). ‘On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I’, Biometrika 20A(1/2), 175–240. Reprinted in *Joint Statistical Papers*, 1–66.

Pearson, E. (1947). ‘The Choice of Statistical Tests Illustrated on the Interpretation of Data Classed in a 2 Å~ 2 Table’, *Biometrika* 34 (1/2), 139–167. Reprinted 1966 in *The Selected Papers of E. S. Pearson*, pp. 169–200.

Pearson, E. (1966). *The Selected Papers of E. S. Pearson*. Berkeley, CA: University of California Press.

Reid, C. (1998).. *Neyman* New York: Springer Science & Business Media.

# History and Philosophy of Statistical Testing: Fisher, Neyman and Pearson (Part II)

**Deborah G. Mayo**

November 18, 2019

Rutgers, Dept. of Statistics & Biostatistics

Prof. Glenn Shafer: History of Probability and Statistics

Rm 211, SEC (T. Alexander Pond Science & Engineering Resource Center)

# Two goals: To explicate:

- The meaning of a N-P (1933) classic passage

- The famous Fisher-Neyman battle

# 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 1967, 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 deconstruct them

In a nutshell: I see them as Neyman’s attempt to avoid the skepticism over the possibility of inductively learning (that Fisher sought) but avoiding Fisher’s problem regarding:

(a) flexibility in the choice of an alternative (to preclude data dependent alternatives Fisher doesn't clearly rule out)

(b) fallacy of probabilistic instantiation (applying probability to the particular case) in his fiducial inference

# “Les Miserables Citations”

"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. 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 1933, pp. 141–2)

The 1933 paper opens with a discussion of two French probabilists–Joseph Bertrand (1889) and Émile Borel, author of *Le Hasard* (1914)!

4 players

*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).*

*Neyman and Pearson describe Bertrand's pessimism and Borel's response*

*Bertrand*: "How can we decide on the unusual results that chance is incapable of producing?" …

*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? ….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.

*.. 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*."

We may follow Borel: (a) the criterion to test a hypothesis (a 'statistical hypothesis') 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'.

"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.)

Skip: Relevant asides: SIST p. 376

Fisher acknowledges that "the same data may contradict the hypotheses in any of a number of different ways" (Fisher 1935a, p. 187) [T]he experimenter . . . is aware of what which interests him, and which he thinks may be statistically significant" (ibid., p. 190)

"By choosing the feature most unfavourable to $H_0$ out of a very large number of features examined it will usually be possible to find some reason for rejecting the hypothesis. ..[If we do that] We shall need to find an answer to the more difficult question. Is it exceptional that the most unfavourable criterion of the n, say, examined should have as unfavourable a value as this?" (Pearson and Chandra Sekar 1936, p. 127)

Skip

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).

**The passages can be read as calling for behavioral considerations but there are inferential reasons**

1. N-P rejects a unified rational measure of belief on hypotheses, but wanted to avoid pessimism of Bertrand*

*This is also what Fisher called for rejecting the principle of indifference needed for Bayesianism. (SIST pp. 386-7)
(*If statistical inference is Bayesian, Neyman will talk instead of inductive "behavior")

2. To avoid the pitfalls of Fisher's fiducial probability

skip
*Neyman (1962) says he's against "the inferential theory":

In the present paper … the term ‘inferential theory’ … will be used to describe the attempts to solve the Bayes’ problem with a reference to confidence, beliefs, etc., through … either a substitute a priori distribution [exemplified by the so called principle of insufficient reason] or a new measure of uncertainty [such as Fisher’s fiducial probability] (p. 16).

# 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$_c$σ/√*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.

This is the lower (1 - c) confidence limit.

> [W]e have a relationship between the statistic [$\bar{X}$] and the parameter μ, such that [$\bar{x}_{.05}$] 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 $\bar{X}$> $\bar{x}_{.05}$. (That is, Pr($\bar{X}$ < μ + 1.65σ/√*n*) = .95.) ] (Fisher 1930, p. 533)

**The 95 per cent. value** $\bar{x}_{.05}$.is the cut-off for rejection at the .05 value (one-sided)

In the normal testing example, $\bar{x}_{.05}$ = μ + 1.65σ/√*n*.

In 95% of samples $\bar{X}$< $\bar{x}_{.05}$.

$\bar{X}$ ≥ $\bar{x}_{.05}$ occurs whenever μ < $\bar{X}$ – 1.65σ/√*n*.

Reject the null at level .05 whenever μ < the lower bound of a .95 CI.

For a particular observed $\bar{x}_0$, $\bar{x}_0$ – 1.65σ/√*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 μ < $\bar{x}_0$ – 1.65σ/√*n* 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 aggregate of all such statements…to the effect that the probability μ is less than" $\bar{x}_0$ – 1.65σ/√*n* is equal to .05 (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)**

1. Big blow-up in 1935: Neyman wouldn't use Fisher's book in his class (SIST p. 387)
2. Stressing the performance view, to Fisher's ears, was to say he was wrong about fiducial probability

Fisher insisted (even in 1955) he was right

Fisher (1955), Neyman violates “…the principles of deductive logic" by accepting [1] and refusing [2].

> "[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.

Fisher's begins to castigate N-P for assuming error probabilities and fiducial probabilities *ought* to agree, declaring the former "foreign to the development of tests of significance." (**SIST** 390)

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" (p. 382).

# The Fisher-Neyman dispute is pathological (SIST 390)

There's no disinterring the truth of the matter.

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

Fisher may have started out seeing fiducial probability as both a frequency of correct claims in an aggregate (performance) and a degree of support or belief (probabilism) but the difficulties in satisfying uniqueness* led Fisher to give up the former.

*recognizable subsets with a different probability of success

# The assumption throughout about the role of probability: probabilism or performance

**Probabilism.** To assign a degree of probability, confirmation, support or belief in a hypothesis, given data $\boldsymbol{x}_0$ (absolute or comparative) (e.g., Bayesian, likelihoodist, Fisher (at times))

**Performance**. To assess and control long-run reliability of methods, coverage probabilities (frequentist, behavioristic Neyman-Pearson, Fisher (at times))

# New History?

A proper subset of methods that control performance are those that assess and control the capability a method would have found a specified flaw if present

This counterfactual requires a principle that says: if a method would have, with high probability, unearthed a given mistake or flaw in claim C, and yet it does not, then there's warrant for its absence.

A claim is warranted to the extent that it passes a test that, with high probability, it would have failed if false or flawed.

# FEV: Frequentist Principle of Evidence; Mayo and Cox (2006); SEV: Mayo 1991, Mayo and Spanos (2006)

**FEV/SEV** A small *P*-value indicates discrepancy γ from $H_0$, if and only if, there is a high probability the test would have resulted in a larger P-value were a discrepancy as large as γ absent.

**FEV/SEV** A moderate *P*-value indicates the absence of a discrepancy γ from $H_0$, only if there is a high probability the test would have given a worse fit with $H_0$ (i.e., a smaller P-value) were a discrepancy γ present.

# REFERENCES:

Bertrand, J. ([1889]/ 1907). *Calcul des Probabilités*. Paris: Gauthier-Villars.

Borel, E. ([1914]/ 1948). *Le Hasard*. Paris: Alcan.Cox 2006

Cox, D. (2006a). *Principles of Statistical Inference*. Cambridge: Cambridge University Press.

Fisher, R. A. (1930). 'Inverse Probability', *Mathematical Proceedings of the Cambridge Philosophical Society* 26(4), 528–35.

Fisher, R. A. (1936), 'Uncertain Inference', *Proceedings of the American Academy of Arts and Sciences* 71, 248–58.

Fisher, R. A. (1955). 'Statistical Methods and Scientific Induction', *Journal of the Royal Statistical Society*: Series B 17(1), 69–78.

Godambe, V. and Sprott, D. (eds.) (1971). *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston of Canada.

Good, I. J. (1971a). 'The Probabilistic Explication of Information, Evidence, Surprise, Causality, Explanation, and Utility' and 'Reply', in Godambe, V. and Sprott, D. (eds.), pp. 108–22, 131–41.

Lehmann, E. (1993a). ‘The Bertrand-Borel Debate and the Origins of the Neyman-Pearson Theory’, in Ghosh, J., Mitra, S., Parthasarathy, K. and Prak Ma Rao, L. (eds.), *Statistics and Probability: A Raghu Raj Bahadur Festschrift*, New Delhi: Wiley Eastern, 371–80. Reprinted in Lehmann 2012, pp. 965–74.

Lehmann, E. (2012). *Selected Works of E. L. Lehmann*, Rojo, J. (ed.). New York: Springer.

Mayo, D. (1991). ‘Novel Evidence and Severe Tests’, Philosophy of Science 58(4), 523–52.

Mayo, D. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. CUP.

Mayo, D. and Cox, D. (2006). ‘Frequentist Statistics as a Theory of Inductive Inference’, in Rojo, J. (ed.), *Optimality: The Second Erich L. Lehmann Symposium*, Lecture Notes-Monograph series, Institute of Mathematical Statistics (IMS), 49, pp. 77–97. (Reprinted 2010 in Mayo, D. and Spanos, A. (eds.), pp. 247–75.)

Mayo, D. and Spanos, A. (2006). ‘Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction’, *British Journal for the Philosophy of Science* 57(2), 323–57.

Neyman, J. (1956). ‘Note on an Article by Sir Ronald Fisher’, *Journal of the Royal Statistical Society*, Series B (Methodological) 18(2), 288–94.

Neyman, J. (1962). ‘Two Breakthroughs in the Theory of Statistical Decision Making’, *Revue De l’Institut International De Statistique / Review of the International Statistical Institute*, 30(1),11–27.

Neyman, J. (1977). ‘Frequentist Probability and Frequentist Statistics’, Synthese 36(1), 97–131.

Neyman, J. and Pearson, E. (1933). ‘On the Problem of the Most Efficient Tests of Statistical Hypotheses’, *Philosophical Transactions of the Royal Society of London* Series A 231, 289–337. Reprinted in Joint Statistical Papers, 140–85.

Neyman, J. and Pearson, E. (1967). *Joint Statistical Papers of J. Neyman and E. S. Pearson*. Berkeley, CA: University of California Press.

Pearson, E. (1955). ‘Statistical Concepts in Their Relation to Reality’, *Journal of the Royal Statistical Society* Series B 17(2), 204–7.

Pearson, E. (1962). ‘Some Thoughts on Statistical Inference’, *The Annals of Mathematical Statistics* 33(2), 394–403. Reprinted 1966 in *The Selected Papers of E. S. Pearson*, pp. 276–83.

Pearson, E. (1966). The Selected Papers of E. S. Pearson. Berkeley, CA: University of California Press.

Zabell, S. L. (1992). ‘R. A. Fisher and Fiducial Argument’, *Statistical Science* 7(3), 369–87.