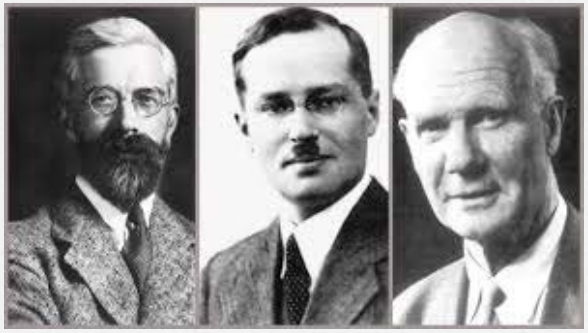
Meeting 2 (May 28) Tour I Ingenious and Severe Tests



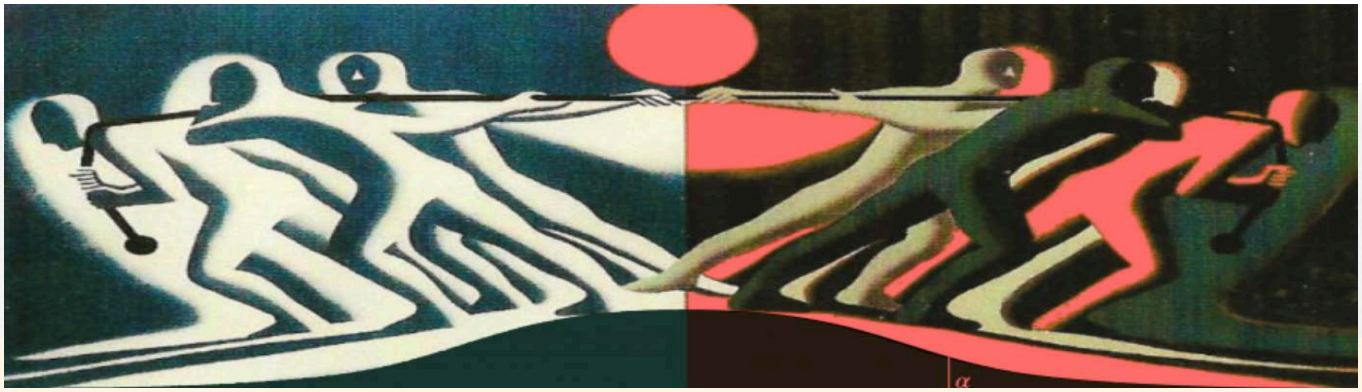
Tour I Ingenious and Severe Tests p. 119

But first, a bit on the goals of our seminar

This research seminar is intended as a lead-up to the workshop: *The Statistics Wars and Their Casualties*:

https://phil-stat-wars.com/2020/02/16/99/.

- to galvanize you to contribute to the ongoing evidence-policy controversies now taking place.



The Statistics Wars
and Their Casualties

~~19-20 June 2020~~ Delayed (a covid 19 casualty)

London School of Economics (CPNSS)

Alexander Bird (King's College London), **Mark Burgman** (Imperial College London), **Daniele Fanelli** (London School of Economics and Political Science), **Roman Frigg** (London School of Economics and Political Science), **David Hand** (Imperial College London), **Christian Hennig** (University of Bologna), **Katrin Hohl** (City University London), **Daniël Lakens** (Eindhoven University of Technology), **Deborah Mayo** (Virginia Tech), **Richard Morey** (Cardiff University), **Stephen Senn** (Edinburgh, Scotland), **Jon Williamson** (University of Kent)*

Some Questions

- What is the “replication crisis”? Are we in one? (2010-20)
- Should replication be required?
- Should we change “perverse incentives”? (open science, badges, preregistration)
- Do data dredging and P-hacking alter the import of data? Always? Sometimes?)
- Should P-values be redefined (lowered)? Banned/ replaced with (confidence intervals, Bayes factors)
- Should we stop saying “significant”?

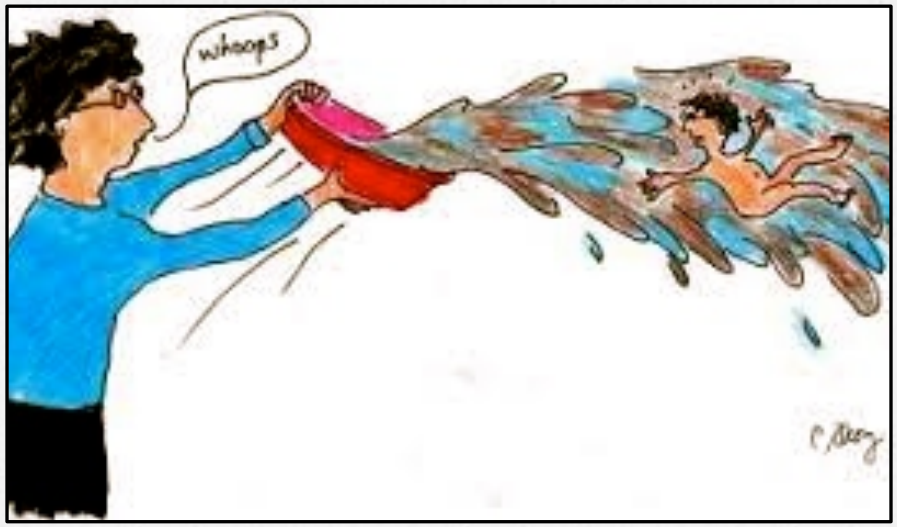
American Statistical Society (ASA): Statement on P-values

"**The statistical community has been deeply concerned about issues of reproducibility and replicability of scientific conclusions**. …. much confusion and even doubt about the validity of science is arising. **Such doubt can lead to radical choices such as…to ban P-values**…(ASA 2016)

Is philosophy of science relevant?

I was a philosophical observer at the ASA P-value “pow wow”

“Don’t throw out the error control baby with the bad statistics bathwater”
The American Statistician



The [ASA] is to be credited with opening up a discussion into p-values; …We should oust recipe-like uses of P-values that have long been lampooned, but without understanding their valuable (if limited) roles, there's a danger of blithely substituting "alternative measures of evidence" that throw out the error control baby with the bad statistics bathwater".

They only give an informal treatment based on a popular variant on simple (Fisherian) tests (NHST)

A bit of history: Where are members of our cast of characters in 1919?

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.

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. 1997, p. 56).

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

After Neyman's year at University College (1925/6), Pearson is suddenly "smitten" with doubts due to Fisher

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).

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)**

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) = $\Pr(d(X) \geq c_\alpha; \theta_1)$.

This is the area to the right of the vertical dotted line, under the H_1 curve, in Figure 3.2. Note $d(x_0)$ and c_α are always approximations expressed as decimals. For continuous cases, Pr is the probability density.

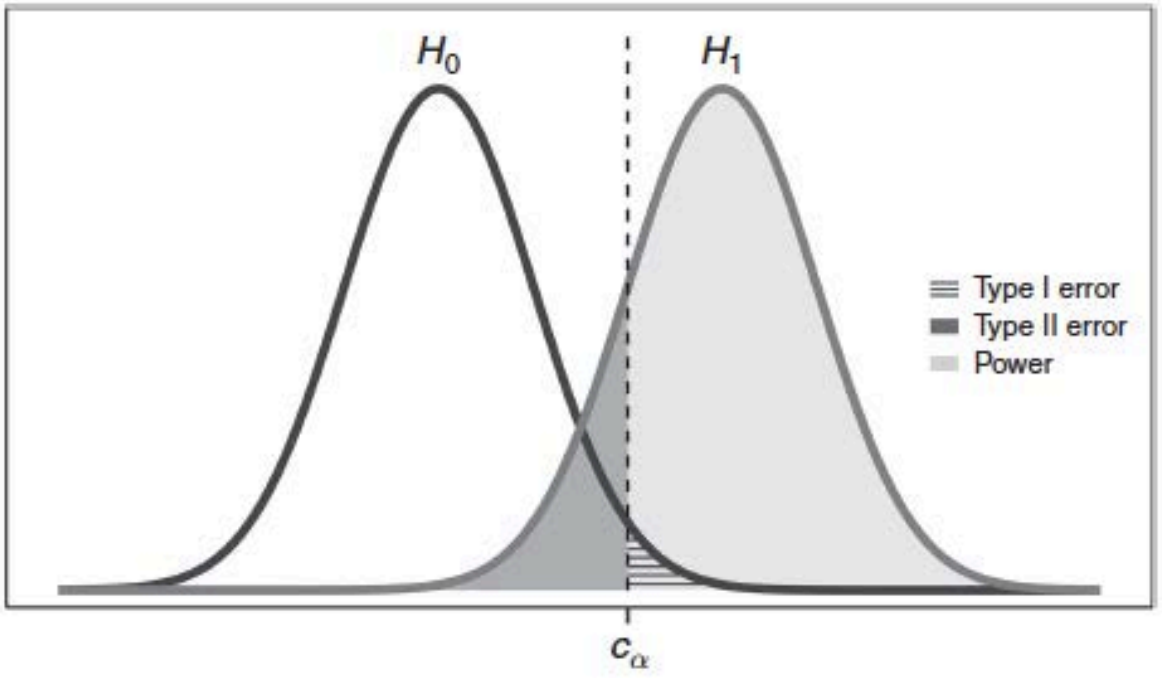


Figure 3.2 Type II error and power.

They were not intended to be used as Accept-Reject Routines (behavioristic performance)

‘[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 normal testing

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

let significance level α = .025

Reject H_0 whenever M ≥ 150 + 2σ/√n: M ≥ 152

M is the sample mean, $\bar{X}$, its value is M_0.
1SE = σ/√n = *1*

Rejection rules:

Reject iff M > 150 + 2SE (N-P)

In terms of the P-value:

Reject iff P-value ≤ .025 (Fisher)

(P-value a distance measure, but inverted)

Let M = 152, so I reject H_0.

SOME *P* –VALUES

Let M = 152

$Z = (152 – 150)/1 = 2$

$Z = (\text{mean obs} – H_0)/1 = 2$

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

SOME *P* –VALUES

Let M = 151

Z = (151 – 150)/1 = 1

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

SOME *P* –VALUES

Let M = 150.5

$Z = (150.5 – 150)/1 = .5$

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

SOME *P* –VALUES

Let M = 150

Z = (150 – 150)/1 = 0

The P-value is Pr(Z > 0) = .5

(important benchmark)

Major Criticism: P-values Don't Measure an Effect Size!

"A p-value, or statistical significance, does not measure the size of an effect or the importance of a result". (ASA Statement)

True, a P-value is a probability.

Reformulation

Severity function: SEV(Test T, data ***x***, claim *C*)

- Tests are reformulated in terms of a discrepancy γ from H_0
- Instead of a binary cut-off (significant or not) the particular outcome is used to infer discrepancies that are or are not warranted

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

The usual test infers there's an indication of *some* positive discrepancy from 150 because

$$Pr(M < 152: H_0) = .97$$

Not very informative

Are we warranted in inferring μ > 153 say?

- Note: Our inferences are not to point values, but we agree to the need to block inferences to discrepancies beyond those warranted with severity.

Consider: How severely has μ > 153 passed the test?

SEV(μ > 153) (p. 143)

M = 152, as before, claim *C*: μ > 153

The data "accord with C" but there needs to be a reasonable probability of a worse fit with C, if C is false

Pr(*"a worse fit"*; *C* is false)

Pr(M ≤ 152; μ ≤ 153)

Evaluate at μ = 153, as the prob is greater for μ < 153.

Consider: How severely has $\mu > 153$ passed the test?

To get $Pr(M \leq 152: \mu = 153)$, standardize:
$Z = \sqrt{100}\,(152 - 153)/1 = -1$

$Pr(Z < -1) = .16$ Terrible evidence

Consider: How severely has $\mu > 150$ passed the test?

To get $Pr(M \leq 152: \mu = 150)$, standardize:
$Z = \sqrt{100}\,(152 - 150)/1 = 2$

$Pr(Z < 2) = .97$

Notice it's 1 – P-value

Now consider SEV(μ > 150.5) (still with M = *152)*

Pr (A worse fit with *C*; claim is false) = .97

Pr(M < 152; μ = 150.5)

Z = (152 – 150.5) /1 = 1.5

Pr (Z < 1.5)= .93 Fairly good indication μ > 150.5

Table 3.1 Reject in test T+: H_0: $\mu \leq 150$ vs. H_1: $\mu > 150$ with $\bar{x} = 152$

Claim $\mu > \mu_1$	Severity $\Pr(\bar{X} \leq 152; \mu = \mu_1)$
$\mu > 149$	0.999
$\mu > 150$	0.97
$\mu > 151$	0.84
$\mu > 152$	0.5
$\mu > 153$	0.16

μ > 150.5

.093

FOR PRACTICE:

Now consider SEV(μ > 151) (still with *M = 152)*

Pr (A worse fit with *C*; claim is false) = __

Pr(M < 152; μ = 151)

Z = (152 – 151) /1 = 1

Pr (*Z* < 1)= .84

MORE PRACTICE:

Now consider SEV(μ > 152) (still with M = 152)

Pr (A worse fit with *C*; claim is false) = __

Pr(M < 152; μ = 152)

Z = 0

Pr (Z < 0)= .5–important benchmark

Terrible evidence that μ > 152

Table 3.2 has exs with M = 153.

Example for you (change sample size)

1-sided normal testing

H_0: μ ≤ 150 vs. H_1: μ > 150 (Let σ = 10, n = 25)

let significance level α = .025

Reject H_0 whenever M ≥ 150 + 2σ/√n:

Assess SEV for the claims in Table 3.1 (p. 144)

Criticism: a P-value with a large sample size may indicate a trivial discrepancy or effect size

- Fixing the P-value, increasing sample size *n*, the cut-off gets smaller
- Get to a point where ***x*** is closer to the null than various alternatives
- Many would lower the P-value requirement as *n* increases-can always avoid inferring a discrepancy beyond what's warranted:

Severity tells us:

- an α-significant difference indicates *less* of a discrepancy from the null if it results from larger (n_1) rather than a smaller (n_2) sample size ($n_1 > n_2$)
- What's more indicative of a large effect (fire), a fire alarm that goes off with burnt toast or one that doesn't go off unless the house is fully ablaze?

- [The larger sample size is like the one that goes off with burnt toast]

Compare *n* = 100 with *n* = 10,000

H_0: μ ≤ 150 vs. H_1: μ > 150 (Let σ = 10, *n* = 10,000)

Reject H_0 whenever M ≥ 2SE: M ≥ 150.2

M is the sample mean (significance level = .025)

1SE = σ/√*n* = 10/√*10,000* = *.1*

Let M = 150.2, so I reject H_0.

(return to in Meeting 4)

Comparing n = 100 with n = 10,000

Reject H_0 whenever M ≥ 2SE: M ≥ 150.2

$\boldsymbol{SEV_{10,000}(\mu > 150.5) = 0.001}$

Z = (150.2 – 150.5) /.1 = -.3/.1 = -3
P(Z < -3) = .001

Corresponding 95% CI: [0, 150.4]

A .025 result is terrible indication μ > 150.5
When reached with n = 10,000

$\boldsymbol{While\ SEV_{100}(\mu > 150.5) = 0.93}$

Fallacies of Non-rejection.
Let M = 151, the test does not reject H_0.

We want to be alert to a fallacious interpretation of a "negative" result: inferring there's no positive discrepancy from μ = 150.

Is there evidence of compliance? μ ≤ 150?
The data "accord with" H_0, but what if the test had little capacity to have alerted us to discrepancies from 150?

No evidence against H_0 is not evidence for it (Postcard).

We need to consider Pr(X > 151; 150), which is only .16.

Computation for SEV(T, M = 151, *C*: μ ≤ 150)
Z = (151 – 150)/1 = 1

Pr(Z > 1) = .16

SEV(*C*: μ ≤ 150) = low (.16).

- So there's poor indication of H_0

Can they say M = 151 is a good indication that μ ≤ 150.5?

No, SEV(T, M = 151, C: μ ≤ 150.5) = ~.3.

[Z = 151 – 150.5 = .5]

But M = 151 *is* a good indication that μ ≤ 152

[Z = 151 – 152 = -1; Pr (Z > -1) = .84]

SEV(μ ≤ 152) = .84

It's an even better indication μ ≤ 153 (Table 3.3, p. 145)

[Z = 151 – 153 = -2; Pr (Z > -2) = .97]

Retire Statistical Significance?

"researchers have been warned that a statistically non-significant result does not 'prove' the null hypothesis"

"…. we should never conclude there is 'no difference' or 'no association' just because a P value is larger than a threshold such as 0.05. (Amrhein et al., 2019)

Agreed

Neyman's fault?

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).

But he also set upper bounds (confidence intervals, power analysis)

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.

Frequentist Evidential Principle: FEV

FEV (i). ***x*** is evidence against H_0 (i.e., evidence of discrepancy from H_0), if and only if the P-value Pr(d > d_0; H_0) is very low (equivalently, Pr(d < d_0; H_0)= 1 - P is very high).

Contraposing FEV(i) we get our minimal principle

FEV (ia) ***x*** are poor evidence against H_0 (poor evidence of discrepancy from H_0), if there's a high probability the test would yield a more discordant result, if H_0 is correct.

Note the one-directional 'if' claim in FEV (1a)

(i) is not the only way **x** can be BENT.

P-value "moderate" (non-significance)

FEV(ii): A moderate *p* value is evidence of 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., smaller *P*- value) were a discrepancy γ to exist.

In the Neyman-Pearson theory of tests, the sensitivity of a test is assessed by the notion of *power*, defined as the probability of reaching a preset level of significance …for various alternative hypotheses. In the approach adopted here the assessment is via the distribution of the random variable *P*, again considered for various alternatives (Cox 2006, p. 25)

$\Pi(\gamma)$: "sensitivity function"

Computing $\Pi(\gamma)$ views the P-value as a statistic.
$\Pi(\gamma) = \Pr(P < p_{obs}; \mu_0 + \gamma)$.

The alternative $\mu_1 = \mu_0 + \gamma$.

Given that P-value inverts the distance, it is less confusing to write $\Pi(\gamma)$

$\Pi(\gamma) = \Pr(d > d_0; \mu_0 + \gamma)$.

Compare to the power of a test:

$POW(\gamma) = \Pr(d > c_\alpha; \mu_0 + \gamma)$ the N-P cut-off c_α.

The following slides are extra
FEV(ii) in terms of Π(γ)

P-value is modest *(not small):* Since the data accord with the null hypothesis, FEV directs us to examine the probability of observing a *result more discordant from* H_0 if $\mu = \mu_0 + \gamma$:

If $\Pi(\gamma) = \Pr(d > d_0; \mu_0 + \gamma)$ is very high, the data indicate that $\mu < \mu_0 + \gamma$.

Here $\Pi(\gamma)$ gives the severity with which the test has probed the discrepancy γ.

FEV (ia) in terms of $\Pi(\gamma)$

If $\Pi(\gamma) = \Pr(d > do; \mu_0 + \gamma)$ = moderately high (greater than .3, .4, .5), then there's poor grounds for inferring $\mu > \mu_0 + \gamma$.

This is equivalent to saying the SEV($\mu > \mu_0 + \gamma$) is poor.