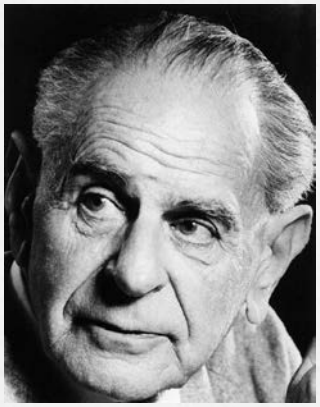
Day #5: Aug 1, 2019
Excursion 3 Statistical Tests and Scientific Inference: Tour I Ingenious and Severe Tests

Tour I Ingenious and Severe Tests p. 119

Popper, GTR and Severity



[T]he impressive thing about [the 1919 tests of Einstein's theory of gravity] is the *risk* involved in a prediction of this kind. … The theory is *incompatible* with certain possible results of observation–in fact with results which everybody before Einstein would have expected. This is quite different from [Freud and Adlerian psychology] (Popper 1962, p. 36)

The problem with Freudian and Adlerian psychology

- Any observed behavior – jumping in the water to save a child, or failing to save her–can be a accounted for by Adlerian inferiority complexes, or Freudian theories of sublimation or Oedipal complexes (Popper 1962, p. 35).

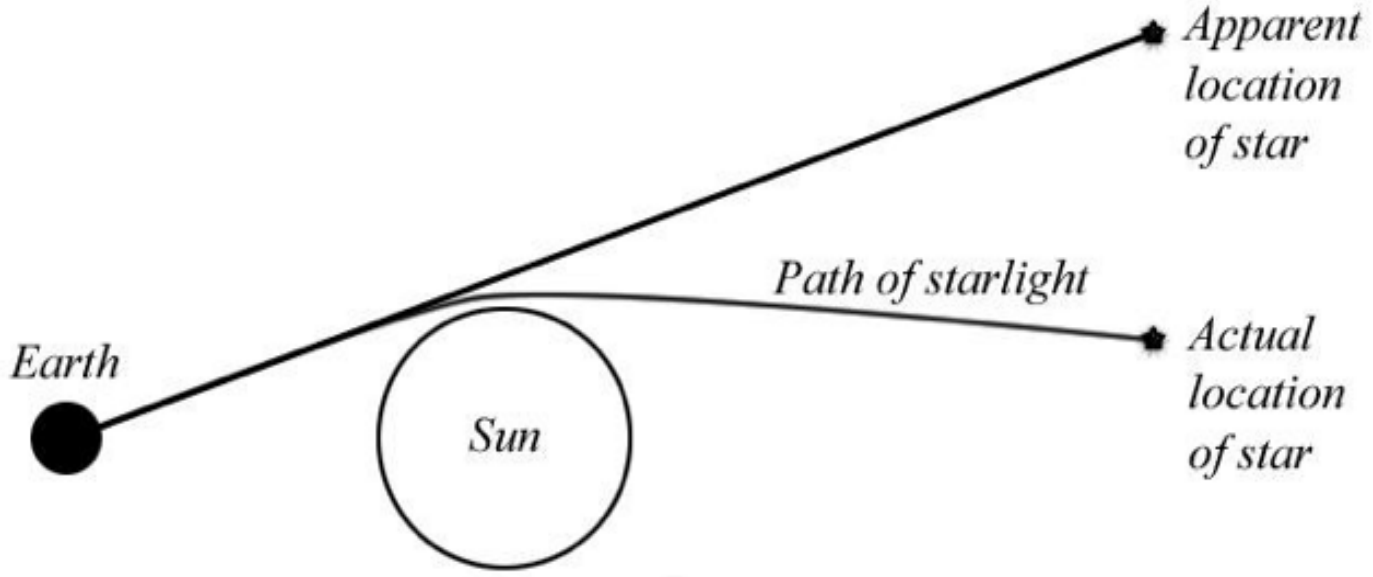
- We've already improved on Popper: it needn't be the flexibility of the theory but of the overall inquiry: research question, auxiliaries, and interpretive rules.

- Not picked up on in logics of induction

100 Years Ago: May 29, 1919: Testing GTR

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

Only detectable during a total eclipse, which "by strange good fortune” would occur on May 29, 1919



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

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

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

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

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

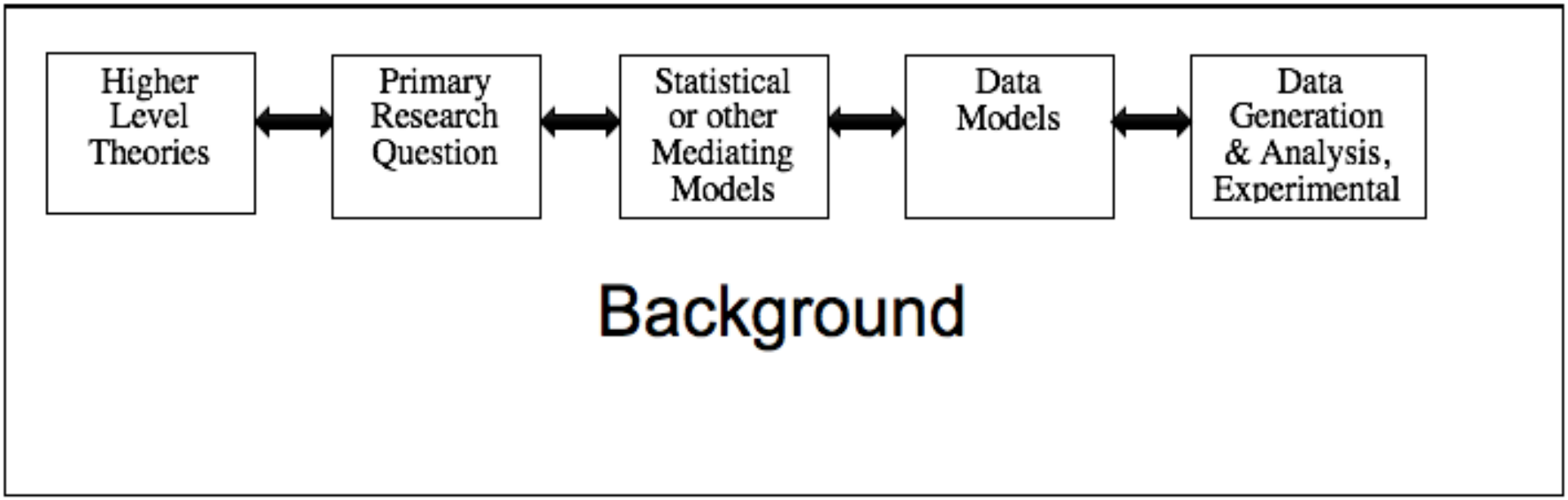
Pearson (Egon) gets his B.A. in 1919, goes to study with Eddington at Cambridge the next year (on the theory of errors)

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. 1997, p. 56).

…return to testing GTR

Statistical Inference and Sexy Science

Even large scale theories connect with data only by intermediate hypotheses and models. (Souv E)



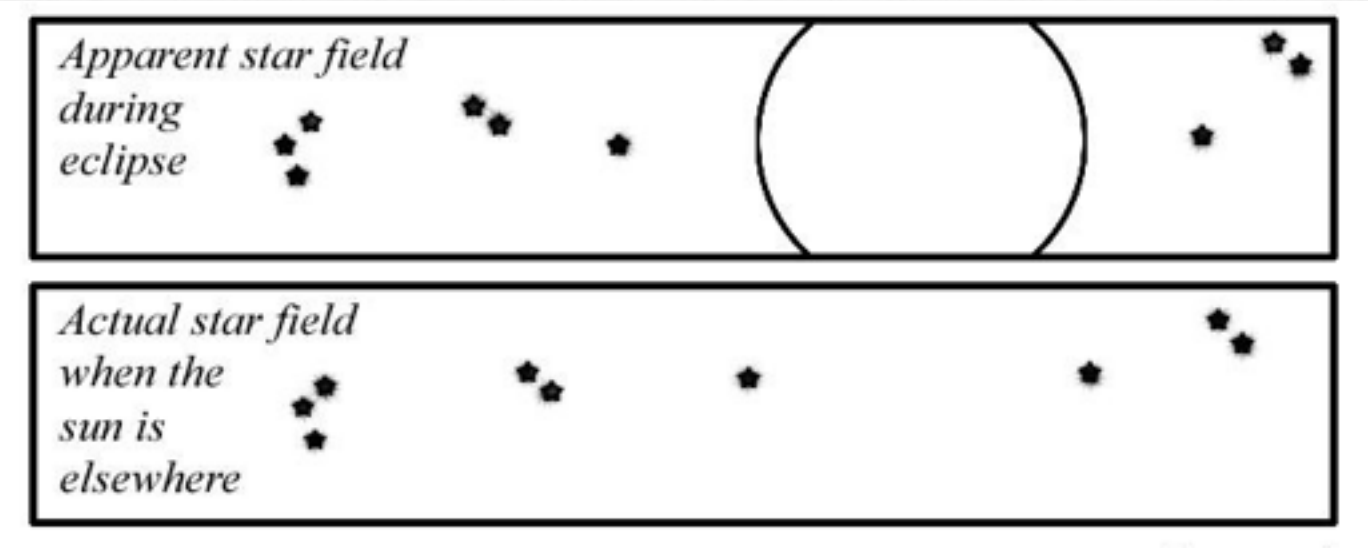
Two key stages of inquiry

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?

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" (ibid., pp. 115-16).



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: μ ≤ 0.87 and the H_1: μ > 0.87

H_1: includes the Einsteinian value of 1.75.

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

A year of checking instrumental and other errors…

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).

(ii) Is the effect "attributable to the sun's gravitational field"? (Can't assume H*)

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

Sir Oliver Lodge's "ether effect" was one of many (e.g., shadow, corona).

Were *any* other cause to exist that produced a considerable fraction of the deflection effect that alone would falsify the Einstein hypothesis (which asserts that *all* of the 1.75" are due to gravity) (Jeffreys 1919, p. 138).

Each Newton-saving hypothesis collapsed on the basis of a one-two punch:

1. the magnitude of effect that could have been due to the conjectured factor is far too small to account for the eclipse effect; and

2. if large enough to account for the eclipse effect, it would have false or contradictory implications elsewhere.

The Newton-saving factors might have been plausible but they were unable to pass severe tests.

Saving Newton this way would be bad science.

More Severe Tests of GTR in the 1970s

- Radio interferometry data from quasars (quasi-stellar radio sources) are more capable of uncovering errors, and discriminating values of the deflection than the crude eclipse tests.

- The Einstein deflection effect “passed” the test, but even then, they couldn’t infer all of GTR severely.

- The [Einstein] law is firmly based on experiment, even the complete abandonment of the theory would scarcely affect it. (Eddington 1920, p. 126)

Some Popperian confusions about falsification and severity

Popper lauds GTR as sticking its neck out, ready to admit its falsity were the deflection effect not found (1962, pp. 36-7).

Even if no deflection effect had been found in 1919, it would have been blamed on the sheer difficulty in discerning so small an effect.

Meehl is wrong (SIST p. 125)

If GTR, then the deflection effect is observed in the 1919 eclipse tests.

No deflection is observed in the 1919 eclipse tests.

Therefore ~GTR (or evidence against GTR).

The first premise of this valid argument is false, the argument is unsound.

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 $T = t(\boldsymbol{y})$ of the data, the **test statistic**, such that

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

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

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

(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* => *H****

Neyman-Pearson (N-P) tests:

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

H_0: μ ≤ 0 vs. H_1: μ > 0

"no effect" vs. "some positive effect"

introduces Type II error, and power

A test in its 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)

(performance)

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)

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 $p(\boldsymbol{x}_0)=\Pr(d(\mathbf{X}) > d(\boldsymbol{x}_0); H_0)$.

Meaning vs application distinction:

(SIST 147)Cox There's a distinction between N-P tests regarded as clarifying the meaning of statistical significance as opposed to an instruction on how to use the ideas…accepting and rejecting a hypothesis are strongly context dependent notions.

Neyman develops CIs as inversions of tests
(estimating μ in a Normal Distribution)

$\mu > M_0 - 1.96\sigma/\sqrt{n}$ CI-lower
$\mu < M_0 + 1.96\sigma/\sqrt{n}$ CI-upper

M_0 : the observed sample mean

CI-lower: the value of μ that M_0 is statistically significantly greater than at P= 0.025

CI-upper: the value of μ that M_0 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

We get an inferential rationale absent from CIs

CI Estimator:

CI-lower < μ < CI-upper

Because it came from a procedure with good coverage probability

Severe Tester:

μ > CI-lower because with high probability (.975) we would have observed a smaller M_0 if μ ≤ CI-lower

μ < CI-upper because with high probability (.975) we would have observed a larger M_0 if μ ≥ CI-lower

We get an inferential rationale absent from CIs

CI Estimator:

CI-lower < μ < CI-upper

Because it came from a procedure with good coverage probability

Severe Tester:

μ > CI-lower because with high probability (.975) we would have observed a smaller M_0 if μ ≤ CI-lower

μ < CI-upper because with high probability (.975) we would have observed a larger M_0 if μ ≥ CI-lower

Days #5-7 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, 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.

PRACTICE WITH P-VALUES

Let M = 152

$Z = (152 - 150)/1 = 2$

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

PRACTICE WITH P-VALUES
Let M = 151

Z = (151 – 150)/1 = 1

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

SEV (μ > 150) = .84 = 1 – P-value

PRACTICE WITH P-VALUES

Let M = 150.5

$Z = (150.5 – 150)/1 = .5$

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

PRACTICE WITH P-VALUES

Let M = 150

$Z = (150 - 150)/1 = 0$

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

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 priniciple

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.

Reformulating Tests: P-values Don't Give an Effect Size

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?

- Recall the complaint of the Likelihoodist (p. 36)
- For them, inferring H_1: $\mu > 150$ means every value in the alternative is more likely than 150
- 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.

Using Severity to Avoid Fallacies: Fallacy of Rejection: Large *n* problem

- 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]

(looks ahead) 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.

Comparing n = 100 with n = 10,000

Reject H_0 whenever M ≥ 2SE: M ≥ 150.2

***SEV*$_{10,000}$(μ > 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

***While SEV*$_{100}$(μ > 150.5) = 0.93**

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

The standard formulation of N-P (as well as Fisherian) tests stops there.

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.

Condition (S-2) requires us to consider Pr(X > 151; 150), which is only .16.

P-value "moderate"

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.

For a Fisherian like Cox, a test's power only has relevance pre-data, they can measure "sensitivity".

> 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)

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 $\mu \leq 150.5$?

No, SEV(T, M = 151, C: $\mu \leq 150.5$) = ~.3.

[Z = 151 – 150.5 = .5]

But M = 151 is a good indication that $\mu \leq 152$

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

SEV($\mu \leq 152$) = .84

It's an even better indication $\mu \leq 153$ (Table 3.3, p. 145)

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

$\Pi(\gamma)$: "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.

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.

FEV/SEV (for Excur 3 Tour III)

Test T+: Normal testing: H_0: $\mu \leq \mu_0$ vs. H_1: $\mu > \mu_0$

σ known

(FEV/SEV): If d(x) is statistically significant (P- value very small), then test T+ passes $\mu > M_0 - k_\varepsilon \sigma/\sqrt{n}$ with severity $(1 - \varepsilon)$.

(FEV/SEV): If d(x) is *not* statistically significant (P- value moderate), then test T+ passes $\mu < M_0 + k_\varepsilon \sigma/\sqrt{n}$ with severity $(1 - \varepsilon)$,

where $P(d(X) > k_\varepsilon) = \varepsilon$.

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.

Note on the Likelihoodist computation compared to the significance tester

1. What do I believe, given x
2. What should I do, given x
3. How should I interpret this observation x as evidence? (comparing 2 hypotheses)

For #1–degrees of belief, Bayesian posteriors

For #2–frequentist performance

For #3–LL (p. 33)

Don’t confuse evidence and belief (in the case of the trick deck), p. 38

p. 33

Comparative logic of support

- **Ian Hacking (1965)** "Law of Likelihood":

  x support hypothesis H_0 less well than H_1 if,

  $\Pr(\boldsymbol{x};H_0) < \Pr(\boldsymbol{x};H_1)$

  (rejects in 1980)

- Any hypothesis that perfectly fits the data is maximally likely (even if data-dredged)

- The measure of comparative support for Royall is the Likelihood ratio.