SESSION 6 EXCURSION 3 Tour I Ingenious and Severe Tests

Tour I Ingenious and Severe Tests p. 119

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

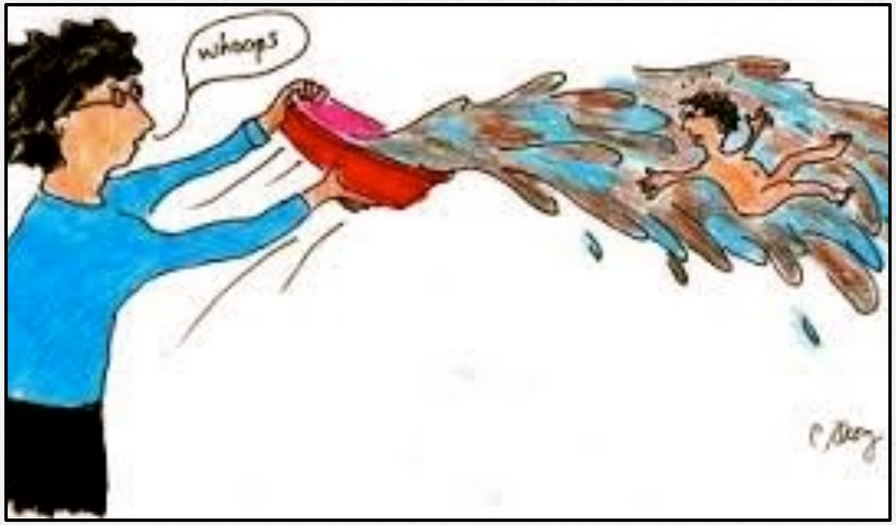
“**The statistical community has been deeply concerned about issues of reproducibility and replicability of scientific conclusions**. …. much confusion and even doubt about the validity of science is arising. **Such doubt can lead to radical choices such as…to ban P-values**…(ASA 2016)

Note 4, SIST 2015-16

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



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)

Simple significance tests (Fisher)

P-value. …to test the conformity or consistency of the data under analysis with H_0 in some respect:

…we find a function $d(\boldsymbol{X})$ of the data, the **test statistic**, such that (I add a third)

- $d(\boldsymbol{X})$ reduces the data as much as possible
- the larger the value of $d(\boldsymbol{X})$ the more inconsistent are the data with H_0;
- Must be able to compute $\Pr(d(\boldsymbol{X}) \underline{>} \mathrm{d}(\mathbf{x}); H_0)$

Problems: *fallacies of rejection p. 94*

- The reported (nominal) statistical significance result is ***spurious*** (it’s not even an actual P-value).

(This can happen in two ways: biasing selection effects, or violated assumptions of the model.)

Problems: 3 *fallacies of rejection p. 94*

- The reported statistically significant result is genuine, but it’s an **isolated** effect not yet indicative of a genuine experimental phenomenon. (Isolated low P-value ≠> *H*: statistical effect)

- Stat-sub problem: either (i) the magnitude of the effect is less than purported, call this a *magnitude error*, or (ii) the substantive interpretation is unwarranted. (*H* ≠> *H**)

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)*:**

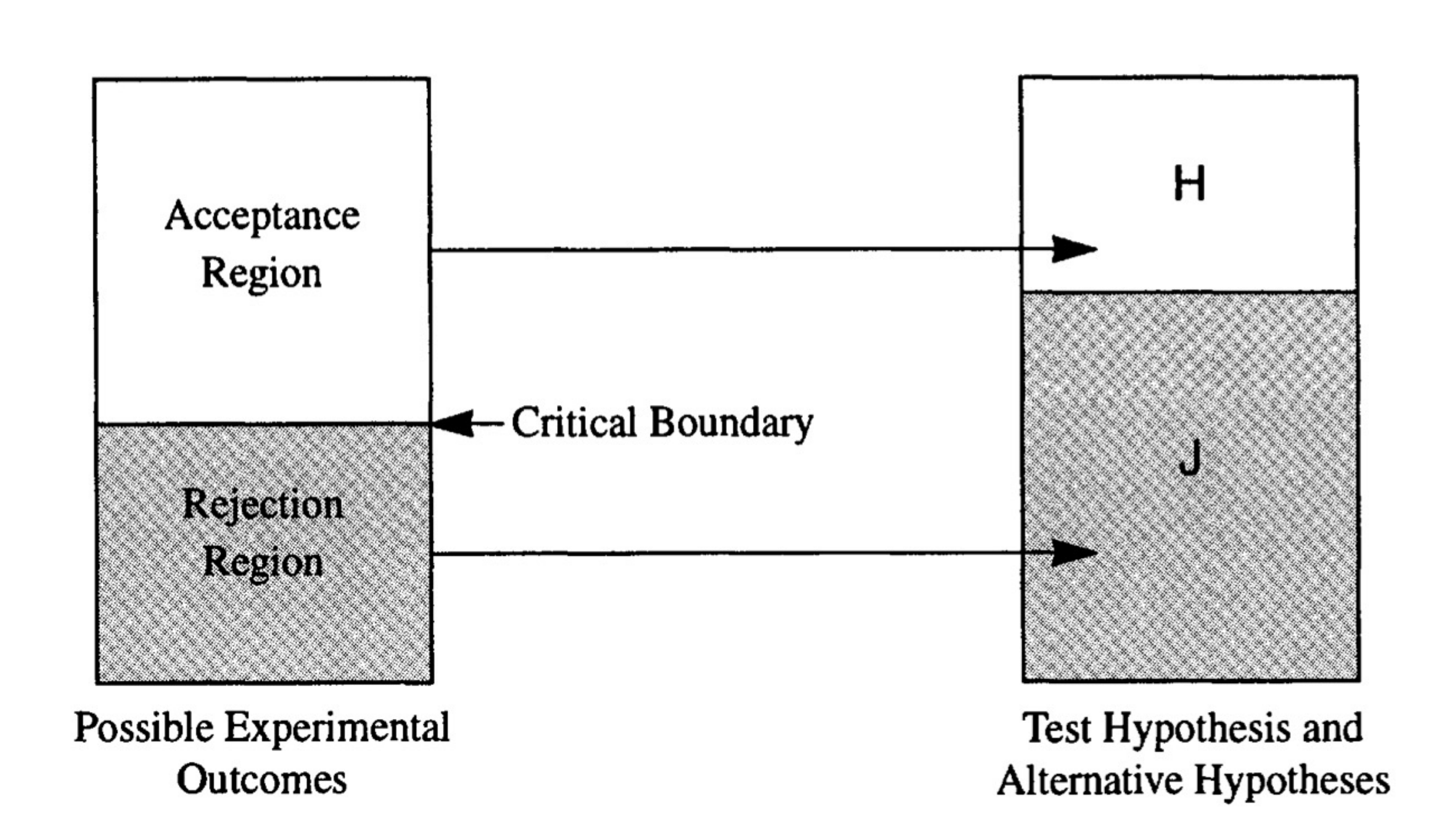
We proceed by setting up a specific hypothesis to test, H_0 in Neyman's and my terminology, the null hypothesis in R. A. Fishers…**in choosing the test, we take into account alternatives to** H_0 which we believe possible or at any rate consider it most important to be on the look out for….:

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] …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)**


Acceptance Region
Critical Boundary
Rejection Region
H
J
Possible Experimental Outcomes
Test Hypothesis and Alternative Hypotheses

Rejection Region, $\mathbf{RR}$: $= \{\boldsymbol{x}\colon \mathbf{d}(\boldsymbol{x}) \geq \boldsymbol{c_{\alpha}}\}$

the set of data points where $\mathrm{d}(\boldsymbol{x}) \geq c_{\alpha}$.

All other data points belong to the "nonrejection" or "acceptance" region

At first they had an "undecided" region, but tests are usually given as dichotomous (following Fisher)

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

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

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

The N-P test in terms of the P-value:

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

The “significance level” or “size” of the test (used ambiguously) (predesignated cut-off vs “attained” level)

Let hypotheses be $H_0: \theta = \theta_0$ vs. $H_1: \theta > \theta_0$.

Same N-P test as having $\underline{H_0: \theta < \theta_0}$ **< in** $\underline{H_0}$

Type II error probability (at $\theta_1) = \Pr(\mathrm{d}(\boldsymbol{X}) < c_\alpha; \theta_1) = \Re(\theta_1)$, for $\theta_1 \text{in J}$

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 $\theta_1) = \Pr\big(\mathrm{d}(\boldsymbol{X}) \geq c_\alpha; \theta_1$).

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

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

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

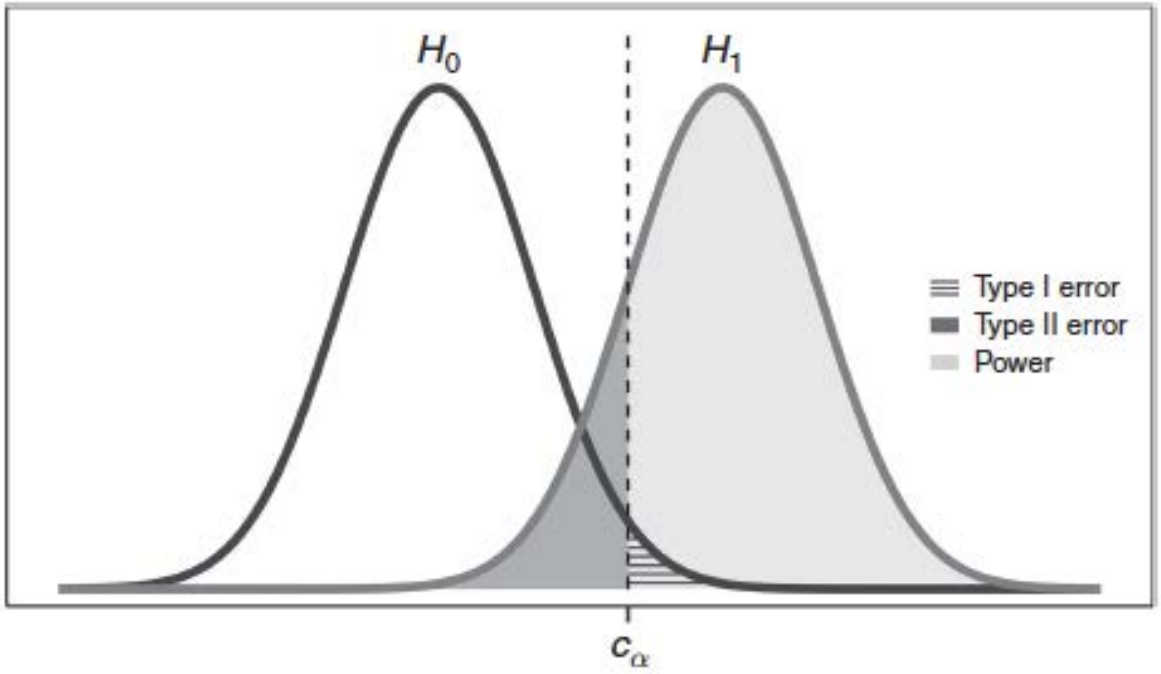


Figure 3.2 Type II error and power.

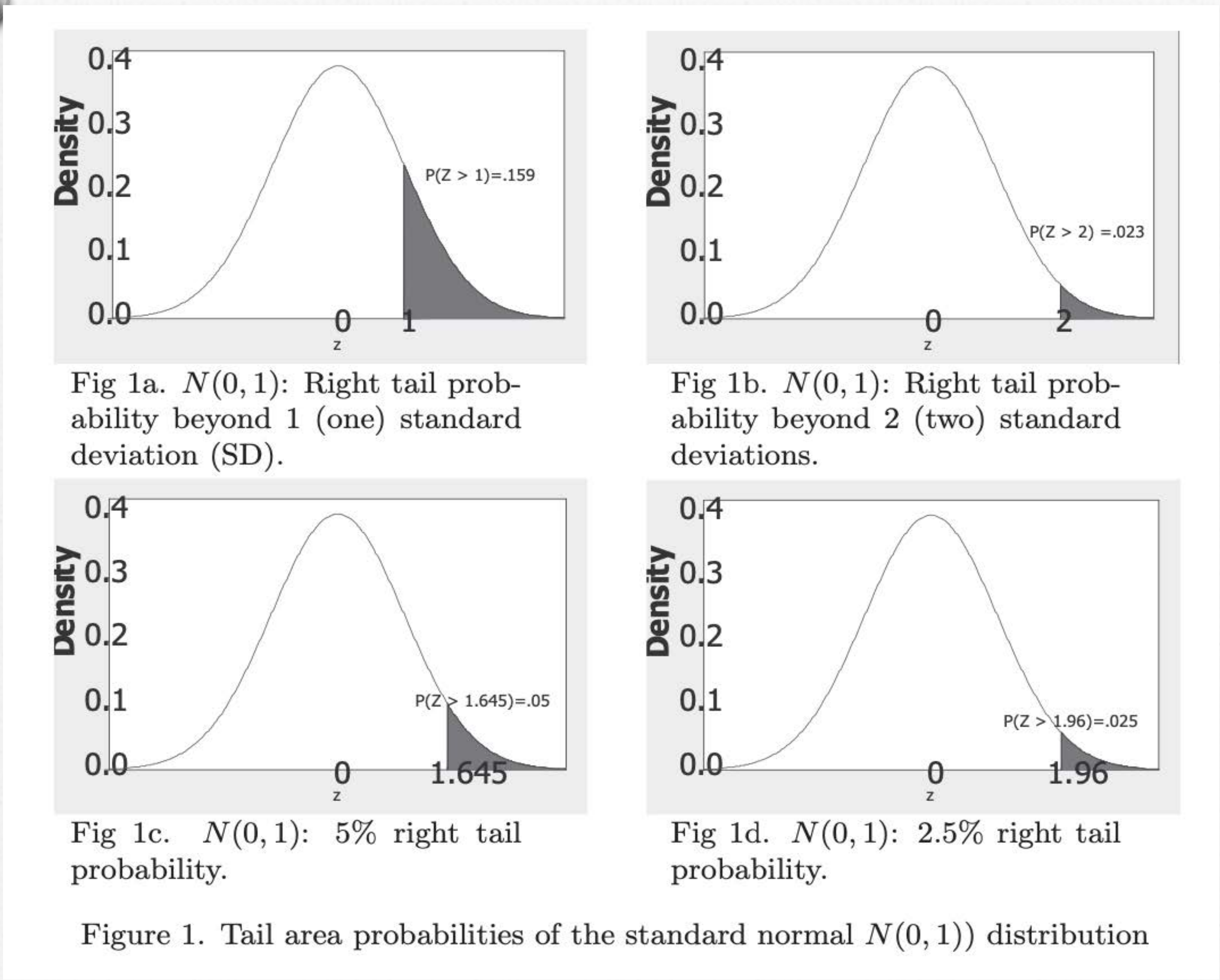


Fig 1a. $N(0,1)$: Right tail probability beyond 1 (one) standard deviation (SD).

Fig 1b. $N(0,1)$: Right tail probability beyond 2 (two) standard deviations.

Fig 1c. $N(0,1)$: 5% right tail probability.

Fig 1d. $N(0,1)$: 2.5% right tail probability.

Figure 1. Tail area probabilities of the standard normal $N(0,1))$ distribution

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 = (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)

Statistical methods begin to be assessed by optimality (e.g., uniformly most powerful tests)

- "The work [of N-P] quite literally transformed mathematical statistics" (C. Reid 1998, p. 104).
- Overshadows Fisher's more informal tests

Fisher (1955): He (Neyman) turned my tests into acceptance-sampling rules

E. S. Pearson: The “heresy” was mine!

What was he referring to?

"From the start we shared Professor Fisher's view that in scientific enquiry, a statistical test is 'a means of learning'" (p. 206):

... it was not till after the main lines of this theory had taken shape …that the fact that there was a remarkable parallelism of ideas in the field of acceptance sampling became apparent. Abraham Wald's contributions to decision theory of ten to fifteen years later were perhaps strongly influenced by acceptance sampling problems, but that is another story. (ibid., pp. $204-5$)

Neyman: "acceptance” was merely shorthand: "The phrase 'do not reject H ' is longish and cumbersome ... My own preferred substitute for 'do not reject H' is 'no evidence against H is found" (Neyman 1976, p. 749).

That is the interpretation that should be used.

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).

Inductive behavior (akin to Popper?)

Neyman begins *First Course in Probability and Statistics*:

Claims are occasionally made that mathematical statistics and the theory of probability form the basis of some mental process described as "inductive reasoning." …the term "inductive reasoning" remains obscure and it is uncertain whether or not the term can be conveniently used to denote any clearly defined concept. On the other hand, …there seems to be room for the term "inductive behavior." This may be used to denote the adjustment of our behavior to limited amounts of observation.

Problems of N-P tests prevented via SEV

1. Fallacies of rejection: magnitude error, large n problem

2. Fallacy of acceptance (non-significance is not evidence for the null)

3. N-P tests are too coarse: need an interpretation that picks up on the observed difference

4. N-P tests are too behavioristic: SEV gives an inferential rationale for the error probability control

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?

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

SEV($\mu > 153$) (p. 143)

$M = 152$, as before, claim C: $\mu > 153$

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 $\mu = 153$, as the prob is greater for $\mu < 153$.

- Richard Morey app set for our example: https://richarddmorey.shinyapps.io/severity/?mu0=150&mu1=150&sigma=10&n=100&xbar=152&xmin=145&xmax=155&alpha=0.05&dir=%3E

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

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

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

Consider: How severely has μ > 150 passed the test?

To get Pr(M ≤ 152: μ = 150), standardize:
Z = (152- 150)/1 = 2

Pr(Z < 2) = .97

Notice it’s 1 – P-value

Now consider SEV*(*$\mu > 150.5$) (still with $M =$ *152)*

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

$\Pr(M < 152; \mu = 150.5)$

$Z = (152 – 150.5) /1 = 1.5$

$\Pr (Z < 1.5)= .93$ Fairly good indication $\mu > 150.5$

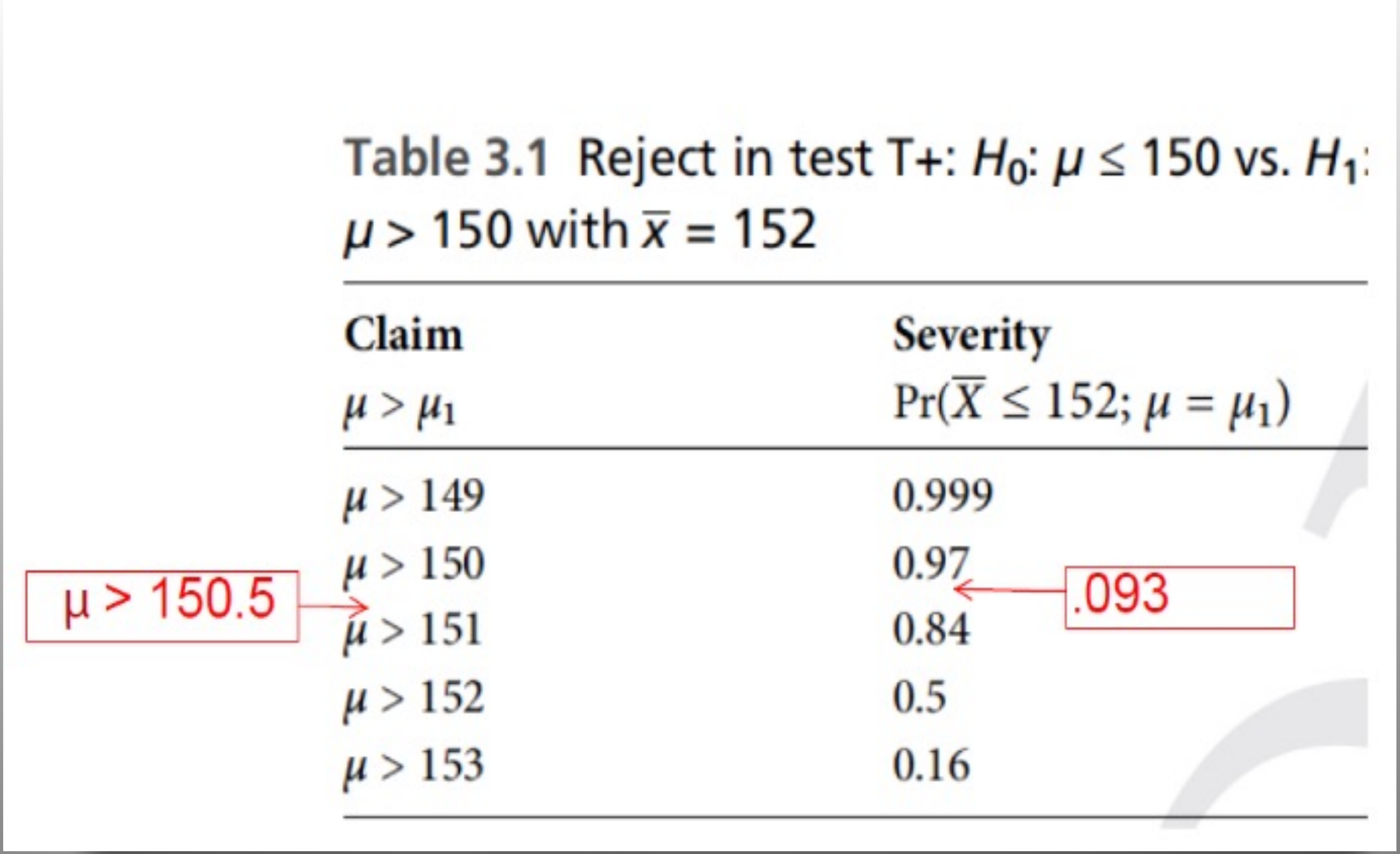


Table 3.1 Reject in test T+: H_0: $\mu \leq 150$ vs. H_1: $\mu > 150$ with $\bar{x} = 152$

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

μ > 150.5

.093

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.

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**

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?

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*: $\mu \leq 150$)
$Z = (151 – 150)/1 = 1$

$\Pr(Z > 1) = .16$

SEV(*C*: $\mu \leq 150$) = 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, $\text{SEV}(T, M = 151, C: \mu \leq 150.5) = \sim.3$.

$[Z = 151 – 150.5 = .5]$

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

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

$\text{SEV}(\mu \leq 152) = .84$

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

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

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.

https://richarddmorey.shinyapps.io/severity/?mu0=150&mu1=150&sigma=10&n=100&xbar=152&xmin=145&xmax=155&alpha=0.05&dir=%3E

Areas under the standard Normal distribution (to the right of z)

z	0	.5	1	1.5	1.65	1.96	2	2.5	3	4
Pr(Z $\geq$ z)	.5	.3	.16	.07	.05	.025	.023	.005	.001	~1

$$H_0: \mu \leq 150 \text{ vs. } H_1: \mu > 150.$$

To get the (one-sided) P-value associated with $\mu \leq 150$ for a given value of $\bar{X}$

1. Turn $\bar{X}$ into a standard Normal variable, i.e., a z score: subtract the hypothesized mean (150) from the observed sample mean $\bar{X}$ and divide by the standard deviation of $\bar{X}$, or the standard error SE. The SE is only $\sigma/\sqrt{n} = 10/\sqrt{100} = 1$

So $z = \frac{\bar{X}-150}{1}$

2. Find the area under the standard Normal curve to the right of z.

Example I: Find the P-value associated with $\mu \leq 150$ for different values of $\bar{X}$ (there’s no change to the SE). I did the first.

$\bar{X}$ = 152	Z = 2	P-value = .023
$\bar{X}$ = 151	Z = ?	P-value = ?
$\bar{X}$ = 150.5	Z = ?	P-value = ?
$\bar{X}$ = 150	Z = ?	P-value = ?

Negative z-values: What if $\bar{X}$ < 150 results in z being a minus number? Say $\bar{X}$ = 149, so z = -1.
Pr(Z $\geq$ -z) =1 – Pr(Z < -z), and because of symmetry of the Normal distribution, Pr(Z < -z)= Pr(Z > z). So the P-value is 1 – Pr(Z > z) = 1 -.16 = .84.

Don’t worry, you can use the SEV app by Richard Morey.

The Morey SEV app. Go to *sampling distribution* (although the *curve selection* is also very informative)

Change the sampling mean to be the observed $\bar{X}$. When asking for the P-value, ignore the *alternative* (it’s imagined to be a Fisherian test with just the null for this purpose), and ignore the *alpha level* box which is for power in a N-P test. Then, under *display options* ask for the *P-value*.

It’s useful also to go to the *curve selection* to see the P-value. (Keep the arrow choice to >, although you can also use it for < problems.)

Example II: Now fix $\bar{X}$ = 152, and find P-values associated with testing 3 different null hypotheses:

$\mu \leq 151,\ \mu \leq 152$,

For $\mu \leq 151$

$z = \frac{\bar{X} - 151}{1} \quad = \frac{152 - 150}{1}$ = 1

(a)If you were testing

$$H_0: \mu \leq 151 \text{ vs. } H_1: \mu > 151,$$

the P-value would be .16.

Now you do the other two:

(b)For $\mu \leq 152$,

$z = \frac{\bar{X}-152}{1}$ so the P-value is _______ if you were testing

$$H_0: \mu \leq 152 \text{ vs. } H_1: \mu > 152,$$

(c) For $\mu \leq 153$,

$z = \frac{\bar{X}-153}{1} =$ _____

so the P-value is _____ if you were testing

$$H_0: \mu \leq 153 \text{ vs. } H_1: \mu > 153,$$

Getting these P-values using the Morey app. The sample mean remains FIXED at $\bar{X}$ = 152, and the *alternative* and the *alpha score* boxes are irrelevant (it can be done in different ways, but let's just stick with one way). The ONLY thing you change is the value for the null μ . Then under display option click P-value (it's lower case in the app). You can do it by means of the *sampling distribution* display or the *curve selection*. The sampling distribution display also provides the reasoning at the bottom

Severity. The severity associated with $\mu > \mu'$. (see SIST p. 143)
Using the Morey app: Set the sample mean $\bar{X}$ and change ***the alternative value for*** $\boldsymbol{\mu\ to\ \mu'}$***.***
This alternative will be some discrepancy from the null value under test but, for simplicity, this computation app for severity does not pick up on changes you make to the *null box*—that is assumed fixed. Nor does it pick up on changes to the *alpha-level box,* used in N-P tests.
Then under *display option click severity* using either the *sampling distribution* display or the *curve*. The *sampling distribution* display also provides the reasoning at the bottom. The *curve* supplies SEV values for other discrepancies, so it's especially useful.

Compute the SEV values for the examples in Table 3.1, SIST p. 144. Here $\bar{X}$ = 152

Notice that in each case the SEV value for inferring $\mu > \mu'$ corresponds to 1 – the P-value associated with testing $\mu < \mu'$ with this observed sample mean $\bar{X}$.