Excursion 3 Statistical Tests and Scientific Inference: Tour I Ingenious and Severe Tests (p. 119)

Will connect many of the themes of the seminar

- The role of probability (probabilism vs performance)

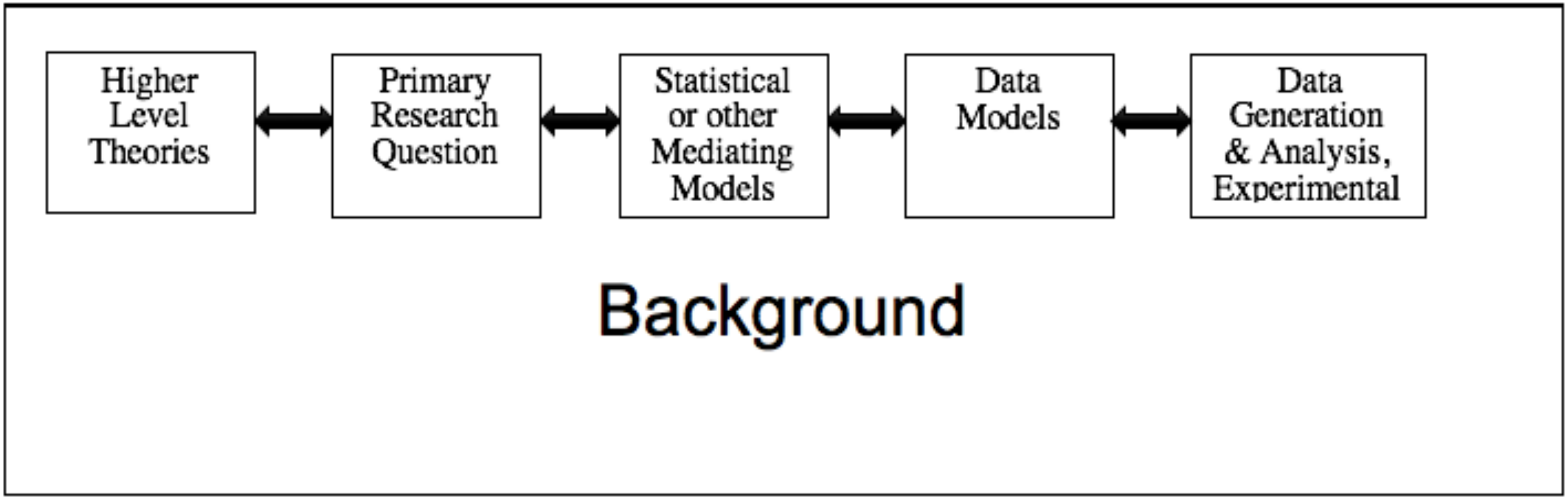
- Demarcating scientific inquiry

- Statistical inference in relation to scientific inference

By and large, Statistics is a prosperous and happy country, but it is not a completely peaceful one. Two contending philosophical parties, the Bayesians and the frequentists, have been vying for supremacy over the past two-and-a-half centuries. . . . Unlike most philosophical arguments, this one has important practical consequences. The two philosophies represent competing visions of how science progresses. (Efron 2013, pp. 130; emphasis added)

Statistical Inference and Sexy Science

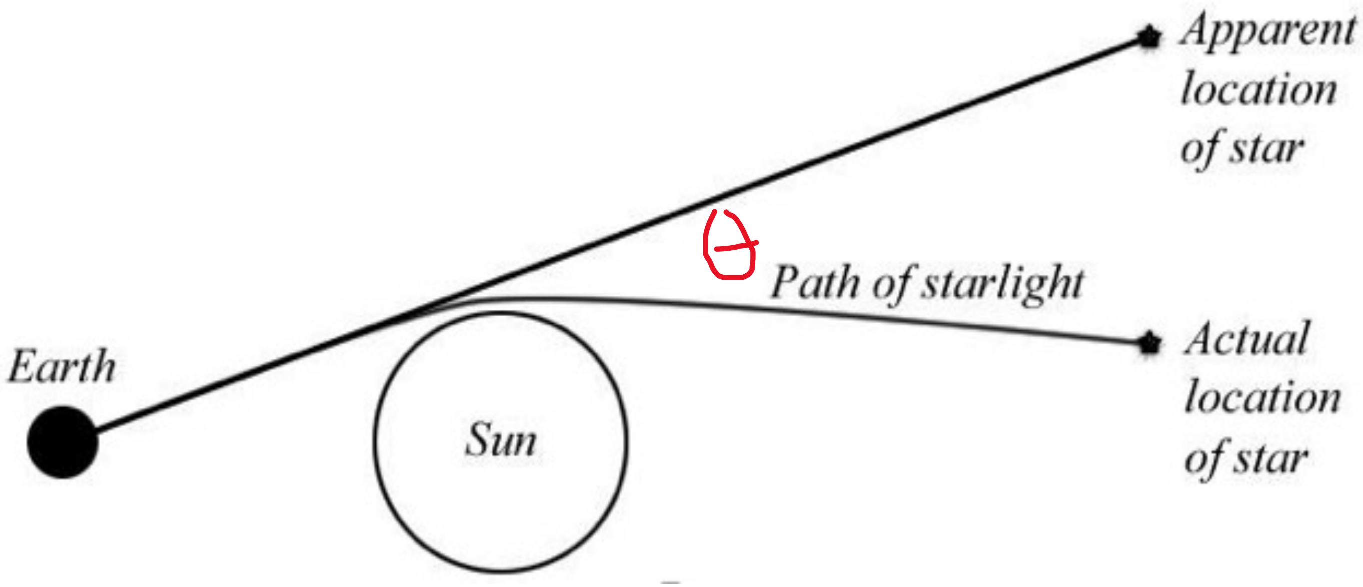
Testing (and developing) large scale theories connect with data only by intermediate hypotheses and models. (Souv E)


Higher Level Theories
Primary Research Question
Statistical or other Mediating Models
Data Models
Data Generation & Analysis, Experimental
Background

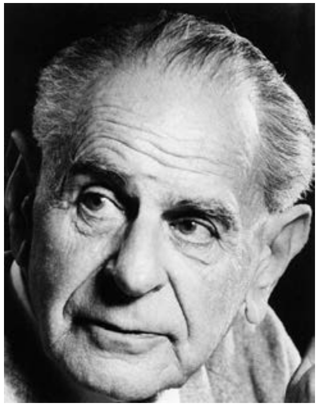
May 29, 1919: Testing GTR

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



Popper



[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. (Popper 1962, p. 36)

Two key stages of inquiry p. 122

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?

Both involve statistical inference p. 122

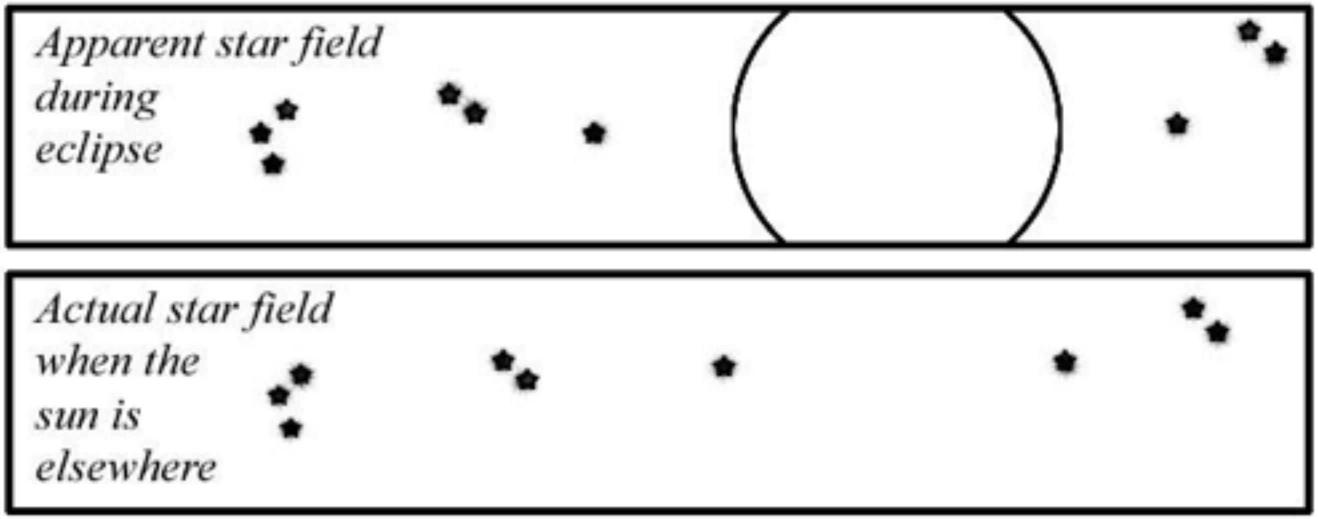
to *infer* what the deflection effect would have been

- for stars near the sun
- without the sun

"The bugbear of possible systematic error affects all investigations of this kind. How do you know there is not something in your apparatus responsible…" (Eddington)

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


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: μ ≤ 0.87 and the H_1: μ > 0.87

The Newtonian value is *.87;* H_1: includes the Einsteinian value of 1.75.

2 expeditions, to Sobral, North Brazil and Principe, 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).

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

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

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 statistically significant result is genuine, but it’s an **isolated** effect not yet indicative of a genuine experimental phenomenon. (eclipse tests run for many years)

- There’s evidence of a genuine statistical phenomenon but the substantive interpretation is unwarranted. (*H* ≠> *H**)

An *audit* of a P-value: a check of any of these concerns, generally in order, depending on the inference.

So I place the background information for auditing throughout our ‘series of models’ representation (figure 2.3, p. 87).

It was auditing an piecemeal checking that made the overall inquiry scientific

Some 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 the deflection effect had not been found in 1919, it would have been blamed on the sheer difficulty in discerning so small an effect.

Popperian Meehl is wrong (SIST p. 125)

Controversy about the Sobral results at stage (i): Did Eddington selectively report? (SIST 156)

Eddington: these results give, "all too good agreement to the 'half-deflection,' that is to say, the Newtonian value" (p. 117).

Earman and Glymour (1980) alleged that Dyson and Eddington threw out the results unwelcome for GTR for political purposes ("... one of the chief benefits to be derived from the eclipse results was a rapprochement between German and British scientists" (p. 83)).

Scotching a famous controversy

May 30,3 a.m., four of the astrographic plates were developed ... It was found that there had been a serious change of focus ... This change of focus can only be attributed to the unequal expansion of the mirror through the sun's heat ... It seems doubtful whether much can be got from these plates. (Dyson et al. 1920, p. 309)

Note: same data used to find and test a hypothesis, but severely:

The data x_0 (from Sobral astrographic plates)
were due to systematic distortion by the sun's
heat, not to the deflection of light,

passes with severity.

Scotching a famous controversy

An even weaker claim is all that's needed: we can't compute a valid estimate of error.

(We falsify a claim that the statistical analysis is valid, p. 157)

There was a reanalysis in 1979 by the Royal Greenwich observatory (Kennefick)

.

Stage (ii) Is the effect "attributable to the sun's gravitational field”?

Using the eclipse effect (from (i)) 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 (quasi-formal):

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.

Demarcating scientific inquiry

A severe tester says: A scientific inquiry or test must be able:

(a) to block inferences that fail the minimal requirement for evidence

You don’t have evidence for a claim if little if anything has been done to probe and rule out how it may be flawed.

Weak severity requirement

Demarcating scientific inquiry

A scientific inquiry or test must be able:

(b) to embark on a reliable probe to pinpoint blame for anomalies

The deflection effect was anomalous for Newton

Demarcating scientific inquiry (4 requirements)

A severe tester says: A scientific inquiry or test must be able:

(a) to block inferences that fail the minimal requirement for severity

(b) to embark on a reliable probe to pinpoint blame for anomalies

(c) (from (a)) to directly pick up on altered error probing capacities due to biasing selection effects, optional stopping, cherry picking, data-dredging etc.

(d) (from (b)) to test and falsify claims.
So we get four requirements for an inquiry to be scientific

Severe Tests of the Einstein deflection effect didn’t come until 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.

- Even where the Einstein deflection effect “passed” the test, they couldn’t infer all of GTR *severely*—merely the deflection effect

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

Severe Tests: Data x provides a good indication or evidence for hypothesis H only if x results from a test procedure that probably would have uncovered the falsity of H, and yet no such error is detected.

“Probability” in defining severity may be qualitative, and always combines several pieces together

That **“H is severely tested**” will always abbreviate that H has *passed* the severe or stringent probe, not, for example merely that H was subjected to one.

Is severity too severe?

“When high-level theoretical hypotheses are at issue, we are rarely in a position to justify a judgment to the effect that [such a passing result is improbable under the assumption that *H* is false].

If we take *H* to be Einstein’s general theory of relativity and E to be the outcome of the eclipse test, then in 1918 and 1919 physicists were in no position to be confident that the vast and then unexplored space of possible gravitational theories…does not contain alternatives to GTE that yield the same prediction for the bending of light as GTR” (Earman 1993, p. 117).

- scientists in the 1960s *were* confident that rivals to GTR *would* yield the same prediction about light-bending--they deliberately developed rivals to GTR that *would!*

- "*H* is false"—as it enters a severity assessment-- is not the so-called (Bayesian) *catchall* hypothesis, but refers, to specific errors or discrepancies

Experimental relativists designed a framework that would not bias them toward accepting GTR prematurely -- by partitioning the space of alternative gravity theories

- **Parameterized Post Newtonian (PPN) framework**: a list of parameters that allowed a systematic articulation of violation of, or alternatives to, what GTR says about specific gravity effects

- Set up largely as straw men with which to set firmer constraints on these parameters, check which portions of GTR have and have not been severly-tested

Souvenir K: Probativism

“A fundamental tenet of the conception of inductive learning most at home with the frequentist philosophy is that inductive inference requires building up incisive arguments and inferences by putting together several different piece-meal results . . . the payoff is an account that approaches the kind of full-bodied arguments that scientists build up in order to obtain reliable knowledge and understanding of a field. (Mayo and Cox 2006, p. 82)

…Learning from evidence turns …on piecemeal tasks of data analysis: estimating backgrounds, modeling data, and discriminating signals from noise.

In evaluating formal statistical methods, you need to take account of the overall conception of scientific inference and learning---rarely discussed

Let’s go back to that….

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 (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).

Egon calls on Neyman to help stem the problems of Fisherian tests Neyman and Pearson (1933):

Introduces alternative hypotheses H_0, H_1

- Trade-off between Type I error (erroneous rejection of H_0,) and Type II errors (erroneously failing to reject H_0)

Tests of Statistical Hypotheses, statistical decision-making

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).
- She compares it to the effect of GTR in physics
- Overshadows Fisher's more informal tests

Fisher-Neyman (pathological) battles Fisher (1955), Pearson (1955) Neyman (1956)

- "being in the same building at University College London brought them too close to one another"! (Cox 2006, 195)

Problems of N-P tests

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*)

- I recommend tests be 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

- Initially developed in a N-P framework (Mayo 1991); it became more Fisherian with Cox

How to do all N-P tests do (and more) while a member of the Fisherian tribe

Reformulation of tests in terms of discrepancies (effect sizes) that are and are not severely-tested SEV(Test T, data *x*, claim *C*)

- In a nutshell: one tests several discrepancies from a test hypothesis and infers those well or poorly warranted

(Mayo1991-2018; Mayo and Spanos (2006) Mayo and Cox (2006); Mayo and Hand (2022)

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

FEV/SEV A small (i.e., statistically significant) *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.

Toy Example: Water Plant (SIST p. 142)

1-sided normal testing

H_0: $\mu \leq 150$ vs. H_1: $\mu > 150$ (Let $\sigma = 10$, $n = 100$)

let significance level $\alpha = .025$

Reject H_0 whenever $M \geq 150 + 2\sigma/\sqrt{n}$: $M \geq 152$

M is the sample mean, its value is M_0.
1SE = $\sigma/\sqrt{n} = 1$

Fisher would compute the P-value attained by the evidence d(*x*)

Let M = 152

Z = (152 – 150)/1 = 2

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

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

There's an indication of *some*
positive discrepancy from 150 because

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

Not very informative

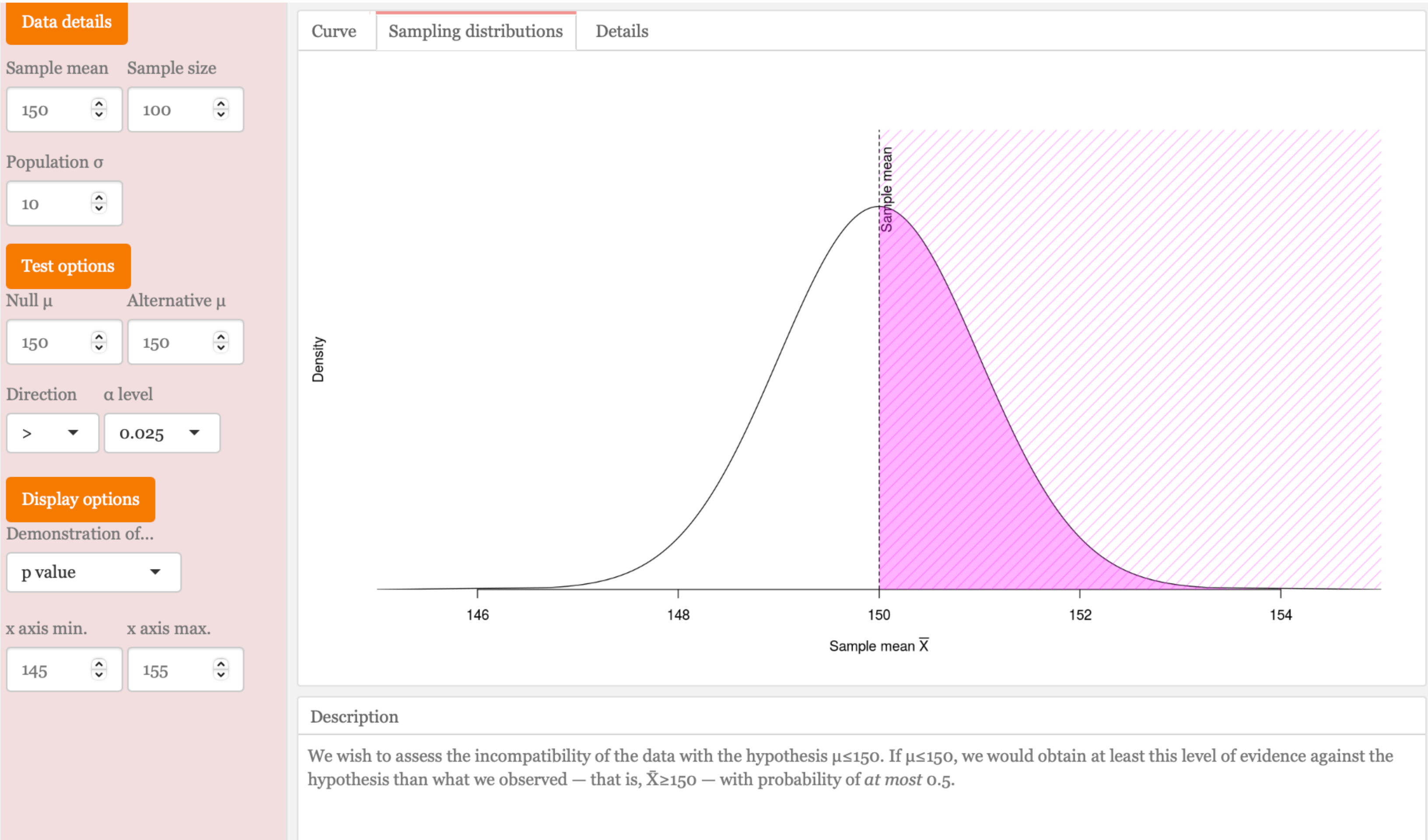
Are we warranted in inferring μ > 153 say?

https://errorstatistics.files.wordpress.com/2023/02/reading-questions-session-6-p-values_morey-appv2.pdf

P-value and severity app: Richard Morey (Professor of psychology, Cardiff)

First, P-values:

Observed $M_0 = 150$, P-value = .5 (testing $H_0: \mu \leq 150$)



Sample mean: 150
Sample size: 100
Population σ: 10
Null μ: 150
Alternative μ: 150
Direction: >
α level: 0.025
Demonstration of...: p value
x axis min.: 145
x axis max.: 155

Description

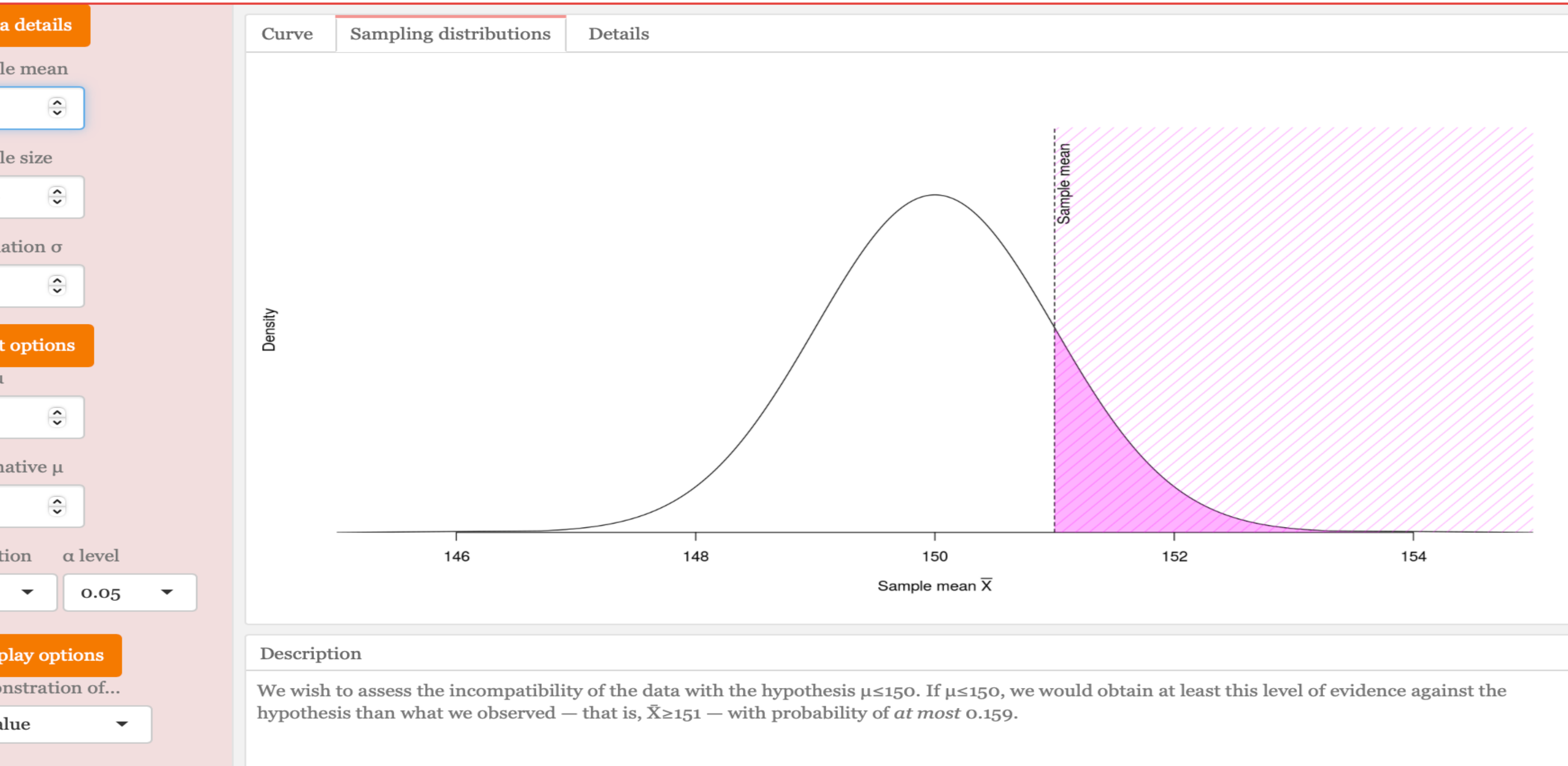
We wish to assess the incompatibility of the data with the hypothesis μ≤150. If μ≤150, we would obtain at least this level of evidence against the hypothesis than what we observed — that is, $\bar{X} \geq 150$ — with probability of *at most* 0.5.

Let $M = 150$ (i.e., observed $M_0 = 150$)

$Z = (150 – 150)/1 = 0$

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

Observed M_0 = 151, P-value = .16 (testing H_0: μ ≤ 150)



tion

α level 0.05

onstration of...

lue

Description

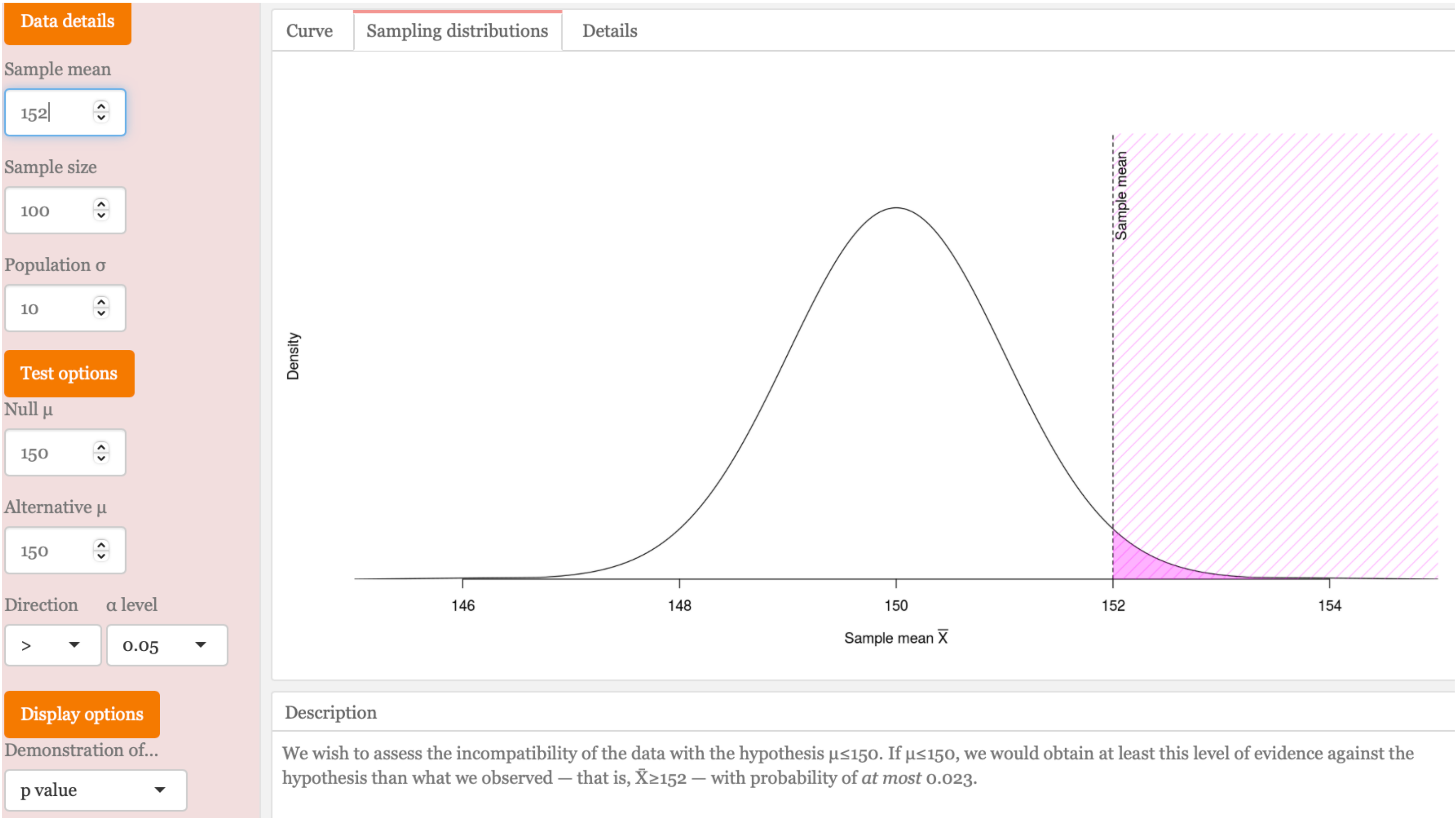
We wish to assess the incompatibility of the data with the hypothesis μ≤150. If μ≤150, we would obtain at least this level of evidence against the hypothesis than what we observed — that is, $\bar{X}$≥151 — with probability of *at most* 0.159.

Let $M = 151$

$Z = (151 - 150)/1 = 1$

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

Observed M_0 = 152, P-value = .025 (testing H_0: μ ≤ 150)



Data details

Sample mean: 152

Sample size: 100

Population σ: 10

Test options

Null μ: 150

Alternative μ: 150

Direction: >

α level: 0.05

Display options

Demonstration of...: p value

Description

We wish to assess the incompatibility of the data with the hypothesis μ≤150. If μ≤150, we would obtain at least this level of evidence against the hypothesis than what we observed — that is, $\bar{X}$≥152 — with probability of *at most* 0.023.

Let $M = 152$

$Z = (152 – 150)/1 = 2$

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

P-values Don’t Give an Effect Size: consider several inferences with the *same* observed mean

Recall:

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

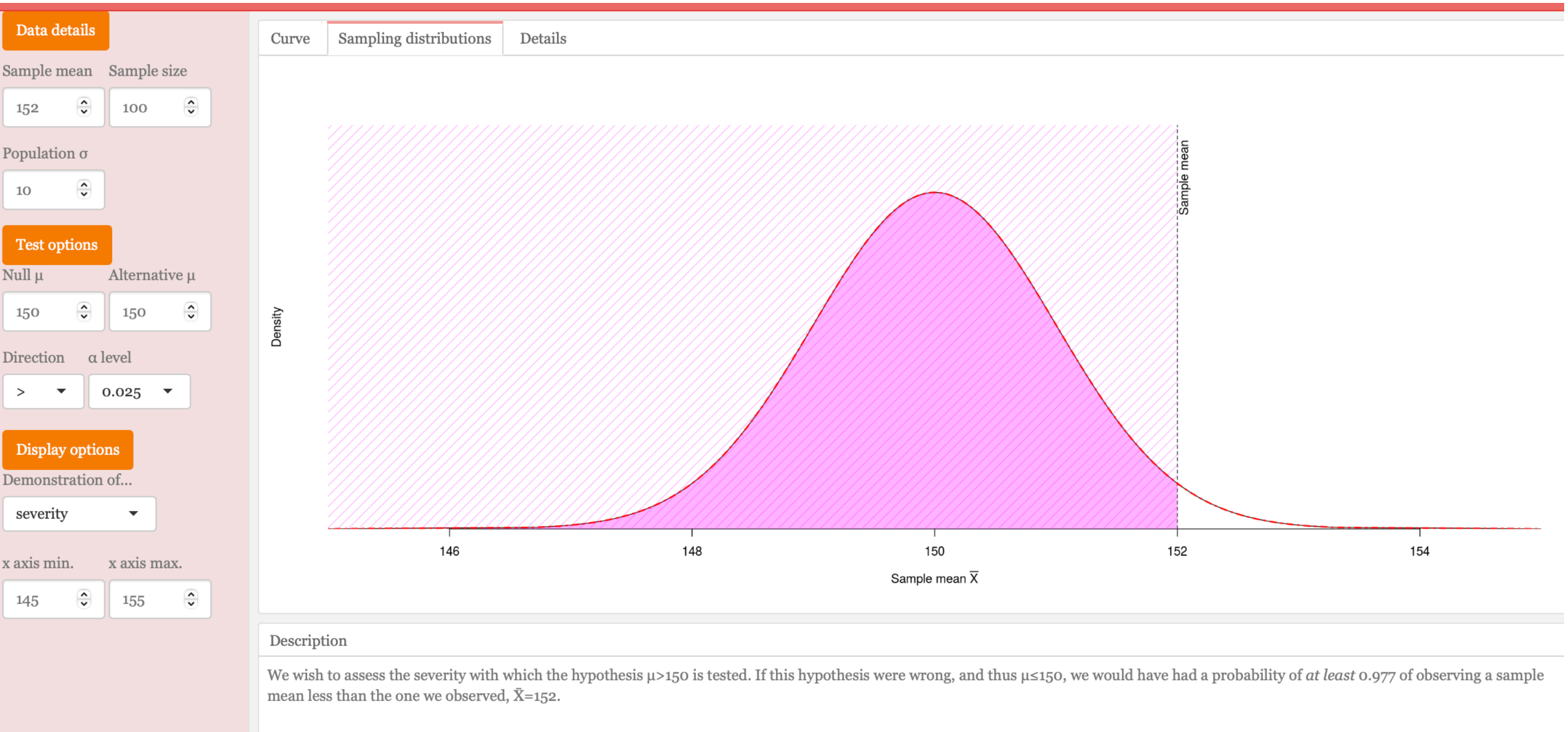
The usual test infers there's an indication of *some* positive discrepancy from 150 because

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

Not very informative

Are we warranted in inferring μ > 150, μ > 151, μ > 152, μ > 153?

$M_0 = 152$, SEV$(\mu > 150)$= .975



Data details

Sample mean: 152

Sample size: 100

Population σ: 10

Test options

Null μ: 150

Alternative μ: 150

Direction: >

α level: 0.025

Display options

Demonstration of...: severity

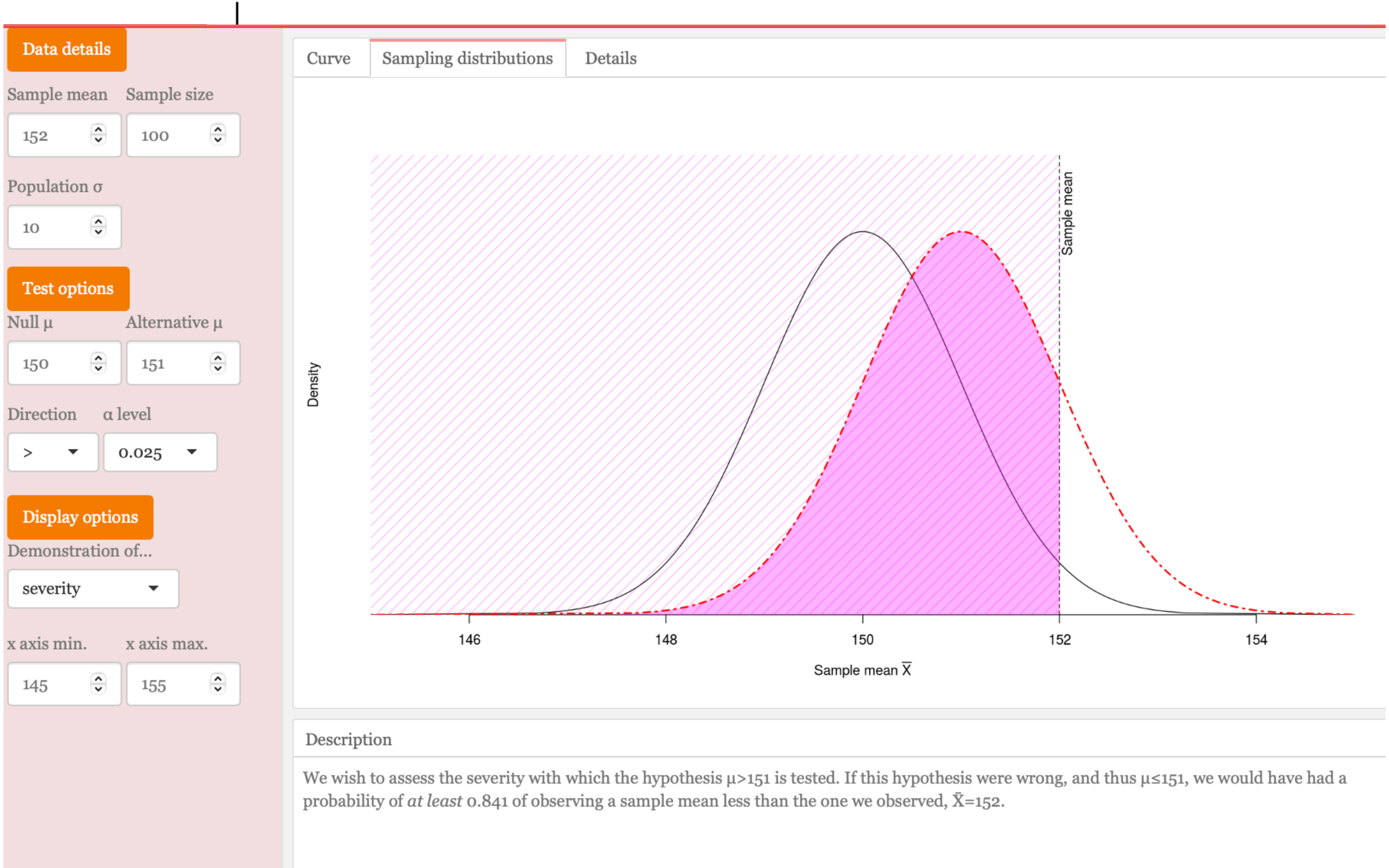
x axis min.: 145

x axis max.: 155

Description

We wish to assess the severity with which the hypothesis μ>150 is tested. If this hypothesis were wrong, and thus μ≤150, we would have had a probability of *at least* 0.977 of observing a sample mean less than the one we observed, $\bar{X}$=152.

$M_0 = 152$, SEV($\mu > 151$)= .84



Sample mean 152

Sample size 100

Population σ 10

Null μ 150

Alternative μ 151

Direction >

α level 0.025

Demonstration of... severity

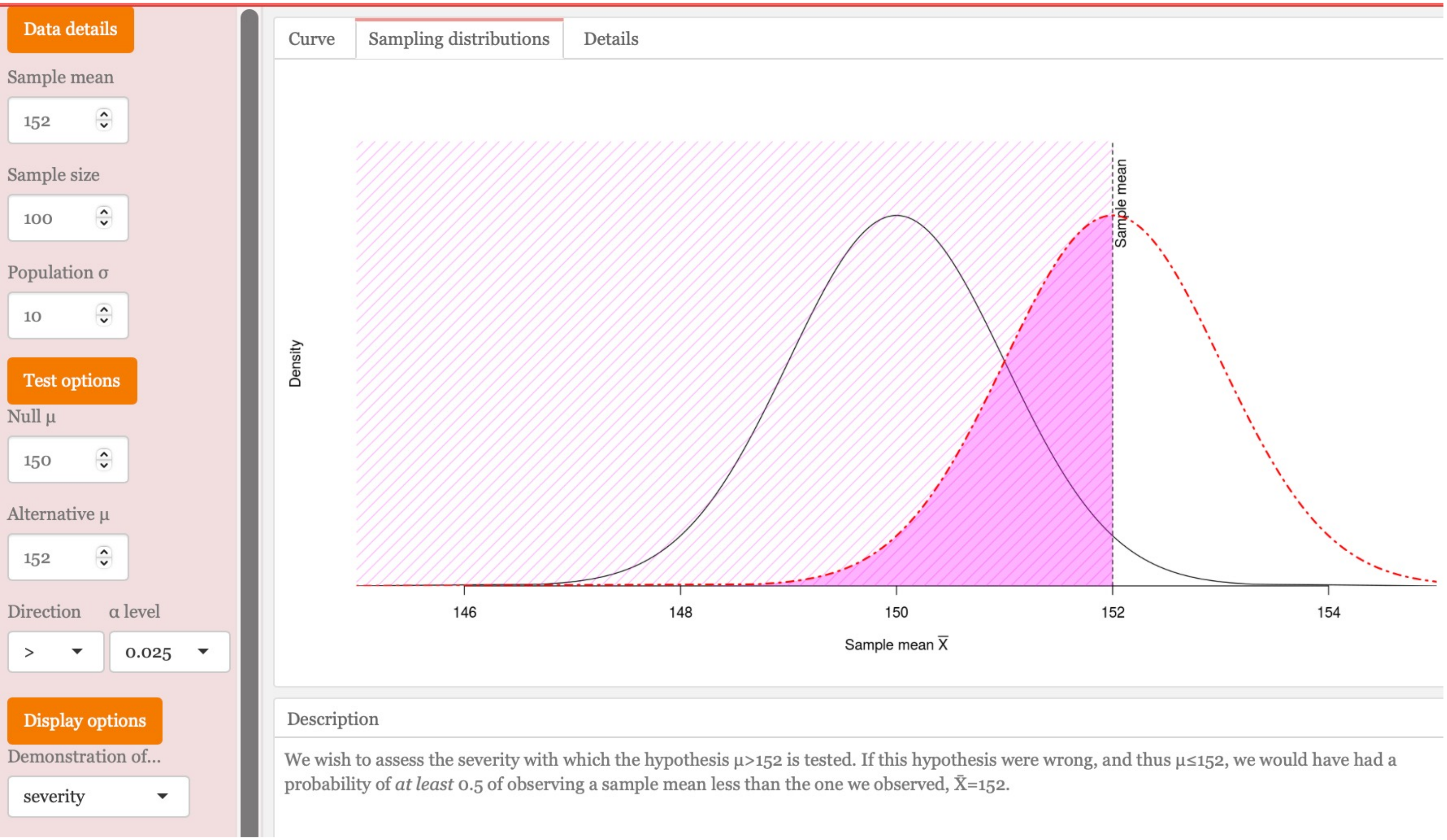
x axis min. 145

x axis max. 155

Description

We wish to assess the severity with which the hypothesis $\mu>151$ is tested. If this hypothesis were wrong, and thus $\mu\leq151$, we would have had a probability of *at least* 0.841 of observing a sample mean less than the one we observed, $\bar{X}=152$.

$M_0 = 152$, SEV($\mu > 152$)= .5



Data details

Sample mean: 152

Sample size: 100

Population σ: 10

Test options

Null μ: 150

Alternative μ: 152

Direction: >

α level: 0.025

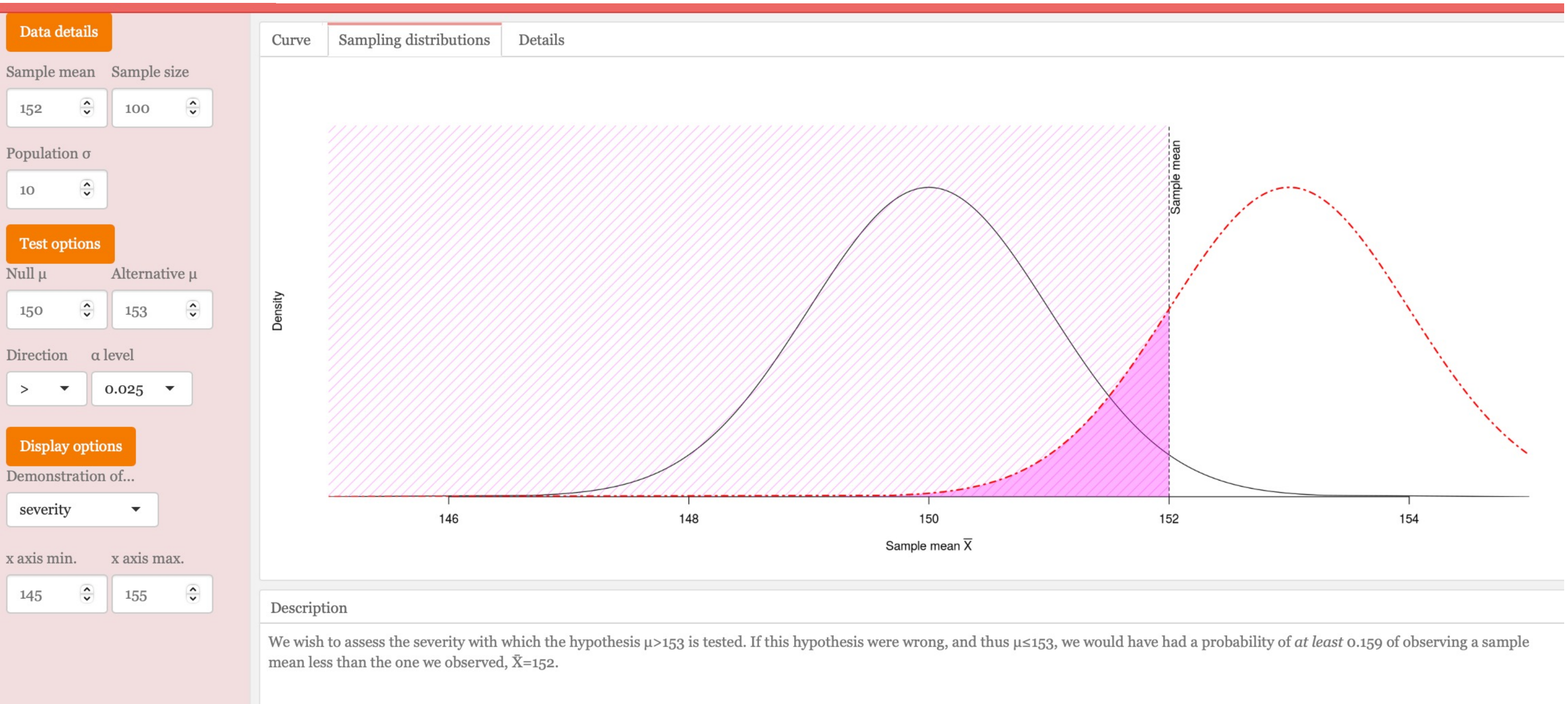
Display options

Demonstration of...: severity

Description

We wish to assess the severity with which the hypothesis $\mu > 152$ is tested. If this hypothesis were wrong, and thus $\mu \leq 152$, we would have had a probability of *at least* 0.5 of observing a sample mean less than the one we observed, $\bar{X} = 152$.

M_0 = 152, SEV(μ > 153)= .16



Sample mean: 152

Sample size: 100

Population σ: 10

Null μ: 150

Alternative μ: 153

Direction: >

α level: 0.025

Demonstration of...: severity

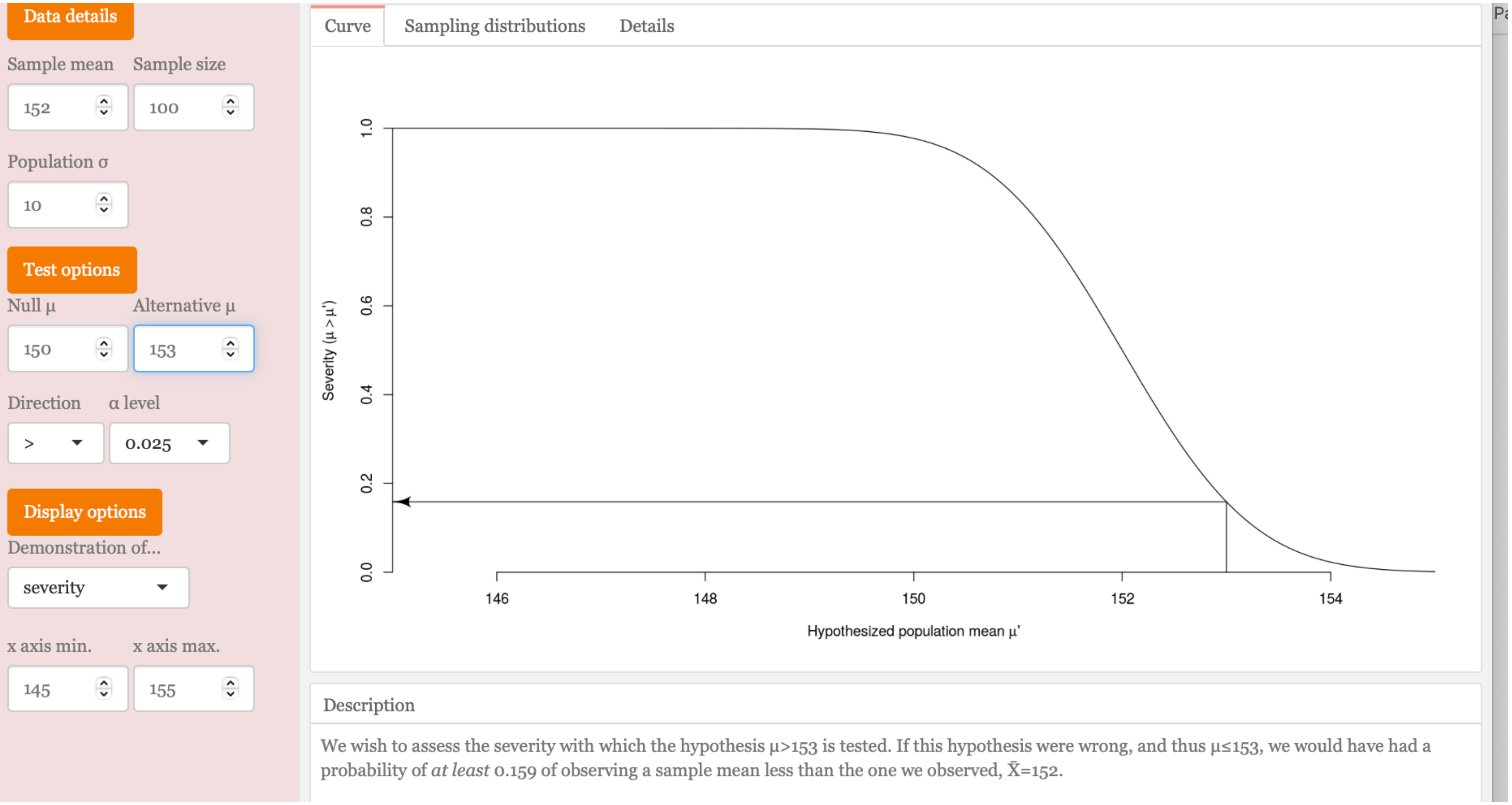
x axis min.: 145

x axis max.: 155

Description

We wish to assess the severity with which the hypothesis μ>153 is tested. If this hypothesis were wrong, and thus μ≤153, we would have had a probability of *at least* 0.159 of observing a sample mean less than the one we observed, X̄=152.

The SEV curve is informative



Sample mean: 152

Sample size: 100

Population σ: 10

Null μ: 150

Alternative μ: 153

Direction: >

α level: 0.025

Demonstration of...: severity

x axis min.: 145

x axis max.: 155

Description

We wish to assess the severity with which the hypothesis μ>153 is tested. If this hypothesis were wrong, and thus μ≤153, we would have had a probability of *at least* 0.159 of observing a sample mean less than the one we observed, $\bar{X}$=152.

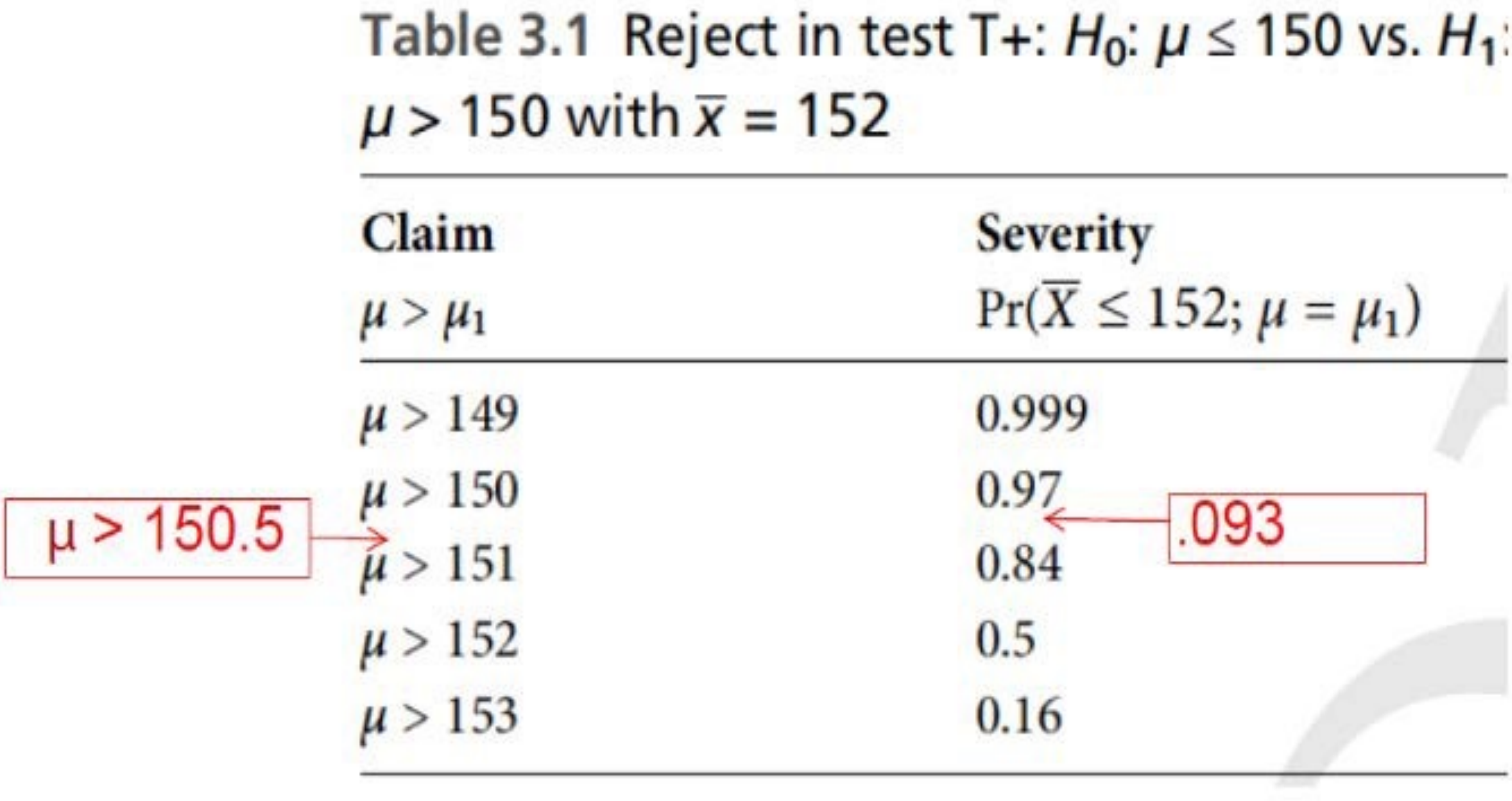


Table 3.1 Reject in test T+: H_0: $\mu \leq 150$ vs. H_1: $\mu > 150$ with $\bar{x} = 152$

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

$\mu > 150.5$

.093

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 ≥ z)	.5	.3	.16	.07	.05	.025	.023	.005	.001	~1

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

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 ≥ -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 $\mu \leq 151$

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

(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 $\boldsymbol{\mu}$ $\boldsymbol{to}$ $\boldsymbol{\mu'}$.***

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 $\underline{\mu < \mu'}$ with this observed sample mean $\bar{X}$.