Class 14
Inference errors and statistical power

Alison Presmanes Hill

POINTS OF SIGNIFICANCE

Power and sample size

The ability to detect experimental effects is undermined in studies that lack power.

Statistical testing provides a paradigm for deciding whether the data are or are not typical of the values expected when the hypothesis is true. Because our objective is usually to detect a departure from the null hypothesis, it is useful to define an alternative hypothesis that expresses the distribution of observations when the null is false. The difference between the distributions captures the experimental effect, and the probability of detecting the effect is the statistical power.

Statistical power is critically relevant but often overlooked. When power is low, important effects may not be detected, and in experiments with many conditions and outcomes, such as 'omics' studies, a large percentage of the significant results may be wrong. **Figure 1** illustrates this by showing the proportion of inference outcomes in two sets of experiments. In the first set, we optimistically assume that hypotheses have been screened, and 50% have a chance for an effect (**Fig. 1a**). If they are tested at a power of 0.2, identified as the median in a recent review of neuroscience literature[1], then 80% of true positive results will be missed, and 20% of positive results will be wrong (positive predictive value, PPV = 0.80), assuming testing was done at the 5% level (**Fig. 1b**).

In experiments with multiple outcomes (e.g., gene expression studies), it is not unusual for fewer than 10% of the outcomes to have an a priori chance of an effect. If 90% of hypotheses are null (**Fig. 1a**), the situation at a 0.2 power level is bleak—over two-thirds of the positive results are wrong (PPV = 0.31; **Fig. 1b**). Even at the conventionally acceptable minimum power of 0.8, more than one-third of positive results are wrong (PPV = 0.64) because although we detect a greater fraction of the true effects (8 out of 10), we declare a larger absolute number of false positives (4.5 out of 90 nulls).

Fiscal constraints on experimental design, together with a commonplace lack of statistical rigor, contribute to many underpowered studies with spurious reports of both false positive and false negative effects. The consequences of low power are particularly dire in the search for high-impact

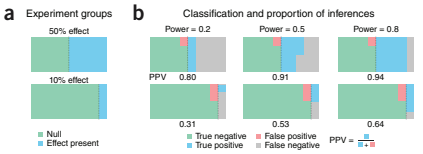


Figure 1 | When unlikely hypotheses are tested, most positive results of underpowered studies can be wrong. (**a**) Two sets of experiments in which 50% and 10% of hypotheses correspond to a real effect (blue), with the rest being null (green). (**b**) Proportion of each inference type within the null and effect groups encoded by areas of colored regions, assuming 5% of nulls are rejected as false positives. The fraction of positive results that are correct is the positive predictive value, PPV, which decreases with a lower effect chance.

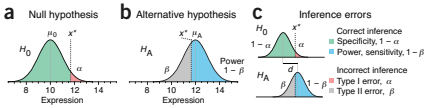


Figure 2 | Inference errors and statistical power. (**a**) Observations are assumed to be from the null distribution (H_0) with mean μ_0. We reject H_0 for values larger than x^* with an error rate α (red area). (**b**) The alternative hypothesis (H_A) is the competing scenario with a different mean μ_A. Values sampled from H_A smaller than x^* do not trigger rejection of H_0 and occur at a rate β. Power (sensitivity) is $1 - \beta$ (blue area). (**c**) Relationship of inference errors to x^*. The color key is same as in **Figure 1**.

results, when the researcher may be willing to pursue low-likelihood hypotheses for a groundbreaking discovery (**Fig. 1**). One analysis of the medical research literature found that only 36% of the experiments examined that had negative results could detect a 50% relative difference at least 80% of the time[2]. More recent reviews of the literature[1,3] also report that most studies are underpowered. Reduced power and an increased number of false negatives is particularly common in omics studies, which test at very small significance levels to reduce the large number of false positives.

Studies with inadequate power are a waste of research resources and arguably unethical when subjects are exposed to potentially harmful or inferior experimental conditions. Addressing this shortcoming is a priority—the Nature Publishing Group checklist for statistics and methods (http://www.nature.com/authors/policies/checklist.pdf) includes as the first question: "How was the sample size chosen to ensure adequate power to detect a pre-specified effect size?" Here we discuss inference errors and power to help you answer this question. We'll focus on how the sensitivity and specificity of an experiment can be balanced (and kept high) and how increasing sample size can help achieve sufficient power.

Let's use the example from last month of measuring a protein's expression level x against an assumed reference level μ_0. We developed the idea of a null distribution, H_0, and said that x was statistically significantly larger than the reference if it exceeded some critical value x^* (**Fig. 2a**). If such a value is observed, we reject H_0 as the candidate model.

Because H_0 extends beyond x^*, it is possible to falsely reject H_0, with a probability of α (**Fig. 2a**). This is a type I error and corresponds to a false positive—that is, inferring an effect when there is actually none. In good experimental design, α is controlled and set low, traditionally at $\alpha = 0.05$, to maintain a high specificity $(1 - \alpha)$, which is the chance of a true negative—that is, correctly inferring that no effect exists.

Let's suppose that $x > x^*$, leading us to reject H_0. We may have found something interesting. If x is not drawn from H_0, what distribution does it come from? We can postulate an alternative hypothesis that characterizes an alternative distribution, H_A, for the observation. For example, if we expect expression values to be larger by 20%, H_A would have the same shape as H_0 but a mean of $\mu_A = 12$ instead of $\mu_0 = 10$ (**Fig. 2b**). Intuitively, if both of these distributions have similar means, we anticipate that it will be more difficult to reliably distinguish between them. This difference between the distributions is typically expressed by the difference in their means, in units of their s.d., σ. This measure, given by

Graphical Inference for Infovis

Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja

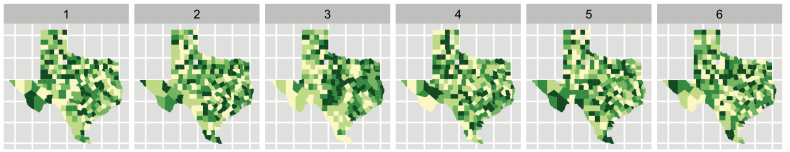


Fig. 1. One of these plots doesn't belong. These six plots show choropleth maps of cancer deaths in Texas, where darker colors = more deaths. Can you spot which of the six plots is made from a real dataset and not simulated under the null hypothesis of spatial independence? If so, you've provided formal statistical evidence that deaths from cancer have spatial dependence. See Section 8 for the answer.

Abstract— How do we know if what we see is really there? When visualizing data, how do we avoid falling into the trap of apophenia where we see patterns in random noise? Traditionally, infovis has been concerned with discovering new relationships, and statistics with preventing spurious relationships from being reported. We pull these opposing poles closer with two new techniques for rigorous statistical inference of visual discoveries. The "Rorschach" helps the analyst calibrate their understanding of uncertainty and the "line-up" provides a protocol for assessing the significance of visual discoveries, protecting against the discovery of spurious structure.

Index Terms—Statistics, visual testing, permutation tests, null hypotheses, data plots.

1 Introduction

What is the role of statistics in infovis? In this paper we try and answer that question by framing the answer as a compromise between curiosity and skepticism. Infovis provides tools to uncover new relationships, tools of curiosity, and much research in infovis focuses on making the chance of finding relationships as high as possible. On the other hand, most statistical methods provide tools to check whether a relationship really exists: they are tools of skepticism. Most statistics research focuses on making sure to minimize the chance of finding a relationship that does not exist. Neither extreme is good: unfettered curiosity results in findings that disappear when others attempt to verify them, while rampant skepticism prevents anything new from being discovered.

Graphical inference bridges these two conflicting drives to provide a tool for skepticism that can be applied in a curiosity-driven context. It allows us to uncover new findings, while controlling for apophenia, the innate human ability to see pattern in noise. Graphical inference helps us answer the question "Is what we see really there?"

The supporting statistical concepts of graphical inference are developed in [1]. This paper motivates the use of these methods for infovis and shows how they can be used with common graphics to provide users with a toolkit to avoid false positives. Heuristic formulations of these methods have been in use for some time. An early precursor is [2], who evaluated new models for galaxy distribution by generating samples from those models and comparing them to the photographic plates of actual galaxies. This was a particularly impressive achievement for its time: models had to be simulated based on tables of random values and plots drawn by hand. As personal computers became available, such examples became more common.[3] compared computer generated Mondrian paintings with paintings by the true artist, [4] provides 40 pages of null plots, [5] cautions against over-interpreting random visual stimuli, and [6] recommends overlaying normal probability plots with lines generated from random samples of the data. The early visualization system Dataviewer [7] implemented some of these ideas.

The structure of our paper is as follows. Section 2 revises the basics of statistical inference and shows how they can be adapted to work visually. Section 3 describes the two protocols of graphical inference, the Rorschach and the line-up, that we have developed so far. Section 4 discusses selected visualizations in terms of their purpose and associated null distributions. The selection includes some traditional statistical graphics and popular information visualization methods. Section 5 briefly discusses the power of these graphical tests. Section 8 tells you which panel is the real one for all the graphics, and gives you some hints to help you see why. Section 7 summarizes the paper, suggests directions for further research, and briefly discusses some of the ethical implications.

2 What is inference and why do we need it?

The goal of many statistical methods is to perform inference, to draw conclusions about the population that the data sample came from. This is why statistics is useful: we don't want our conclusions to apply only to a convenient sample of undergraduates, but to a large fraction of humanity. There are two components to statistical inference: testing (is there a difference?) and estimation (how big is the difference?). In this paper we focus on testing. For graphics, we want to address the question *"Is what we see really there?"* More precisely, is what we see in a plot of the sample an accurate reflection of the entire population? The rest of this section shows how to answer this question by providing a short refresher of statistical hypothesis testing, and describes how testing can be adapted to work visually instead of numerically.

Hypothesis testing is perhaps best understood with an analogy to

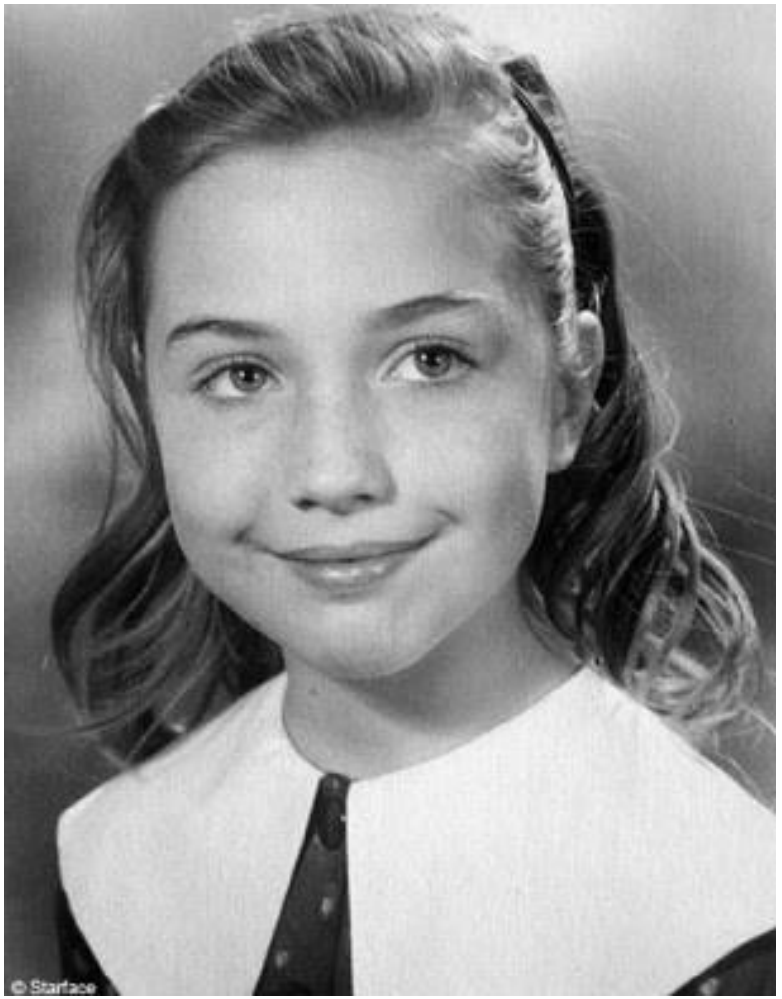
- *Hadley Wickham is an Assistant Professor of Statistics at Rice University. Email: hadley@rice.edu.*
- *Dianne Cook is a Full Professor of Statistics at Iowa State University.*
- *Heike Hofmann is an Associate Professor of Statistics at Iowa State University.*
- *Andreas Buja is the Liem Sioe Liong/First Pacific Company Professor of Statistics in The Wharton School at the University of Pennsylvania.*

Manuscript received 31 March 2010; accepted 1 August 2010; posted online 24 October 2010; mailed on 16 October 2010.
For information on obtaining reprints of this article, please send email to: tvcg@computer.org.

1077-2626/10/$26.00 © 2010 IEEE Published by the IEEE Computer Society

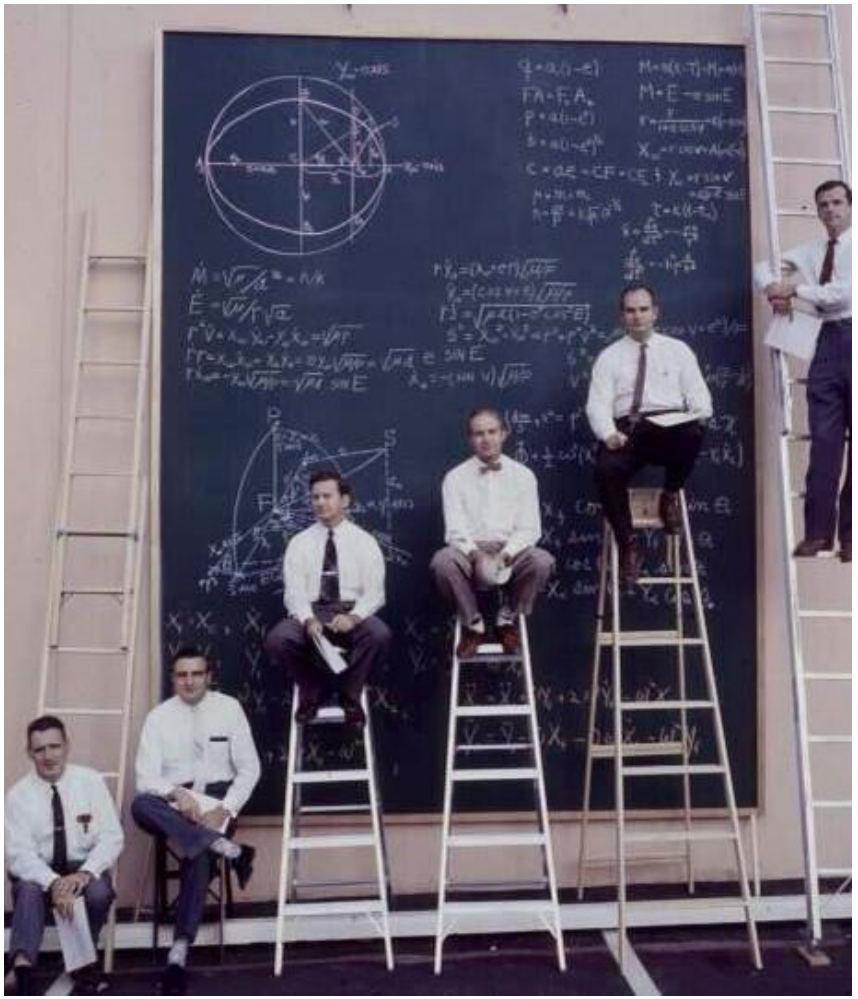
Aspiring astronauts

- Inspired by Alan Shepard, the first American to journey into space, a 14-year-old Hillary Rodham from suburban Chicago wrote a letter to NASA in 1961 asking what she needed to do to become an astronaut.



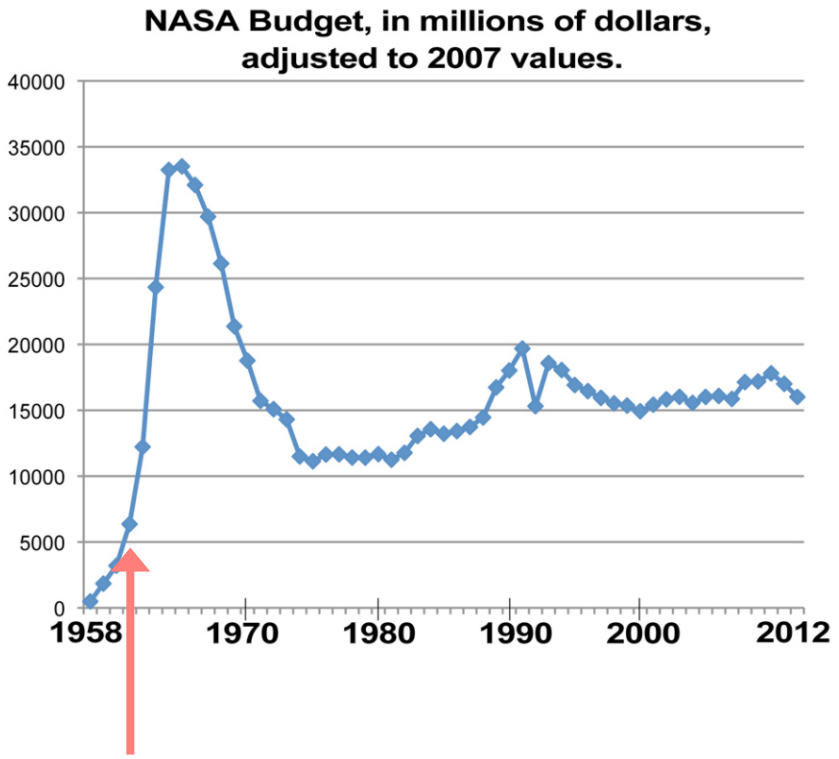
Let's pretend

- Upon receipt of the letter in 1961, NASA decided to conduct a study to test whether girls who are aspiring astronauts in high school have "above average" IQ



The (fictitious) study

- Unfortunately, the NASA budget in 1961 was pretty low
- So they studied only 25 high school girls, all of whom were aspiring astronauts (AA)
- Population (normally distributed):
  μ = 100; σ = 15



A (non-directional) alternative hypothesis

- H_0:

IQ scores for AA will not differ from population;

$$\mu_{aa} = \mu_0$$

$$\mu_{aa} = 100$$

$$\mu_{aa} - 100 = 0$$

- H_1:

IQ scores for AA will be different from population;

Two-tailed test

$$\mu_{aa} \neq \mu_0$$

$$\mu_{aa} \neq 100$$

$$\mu_{aa} - 100 \neq 0$$

One-sample z-test

- For known: $\mu, \sigma, \overline{x}$

$$z_x = \frac{\overline{x} - \mu_X}{\sigma / \sqrt{n}}$$

$$= \frac{\square - \square}{\square / \sqrt{\square}}$$

$$= \square$$

One-sample z-test

- For known: $\mu, \sigma, \overline{x}$

$$
\begin{aligned}
z_x &= \frac{\overline{x} - \mu_X}{\sigma/\sqrt{n}} \\
&= \frac{105 - 100}{15/\sqrt{25}} \\
&= \frac{5}{3} = 1.667
\end{aligned}
$$

Obtaining the p-value

- One-tailed

- Two-tailed

```
c(pz_up, pz_2)
```

Obtaining the p-value

- One-tailed

```
pz_up <- 1 - pnorm(z)
```

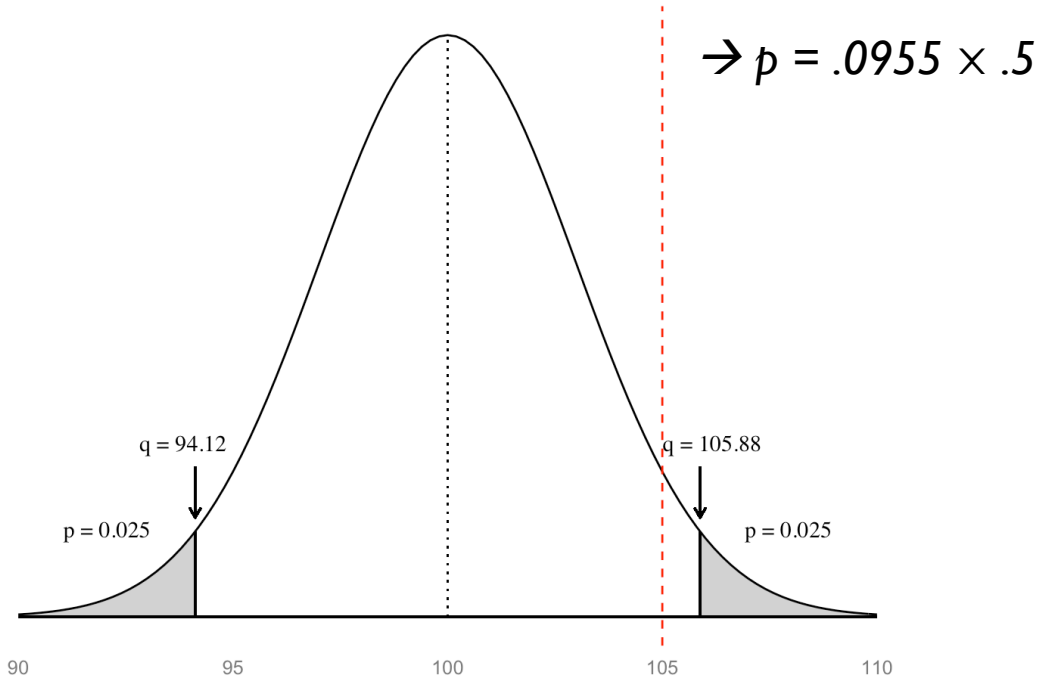
- Two-tailed

```
pz_2 <- 2*pz_up
```

```
c(pz_up, pz_2)
[1] 0.04779035 0.09558070
```

Two-tailed p-values more generally…

```
pz_1tail <- min(pnorm(z), 1 - pnorm(z))
pz_2tail <- 2 * pz_1tail
pz_2tail
[1] 0.0955807
```



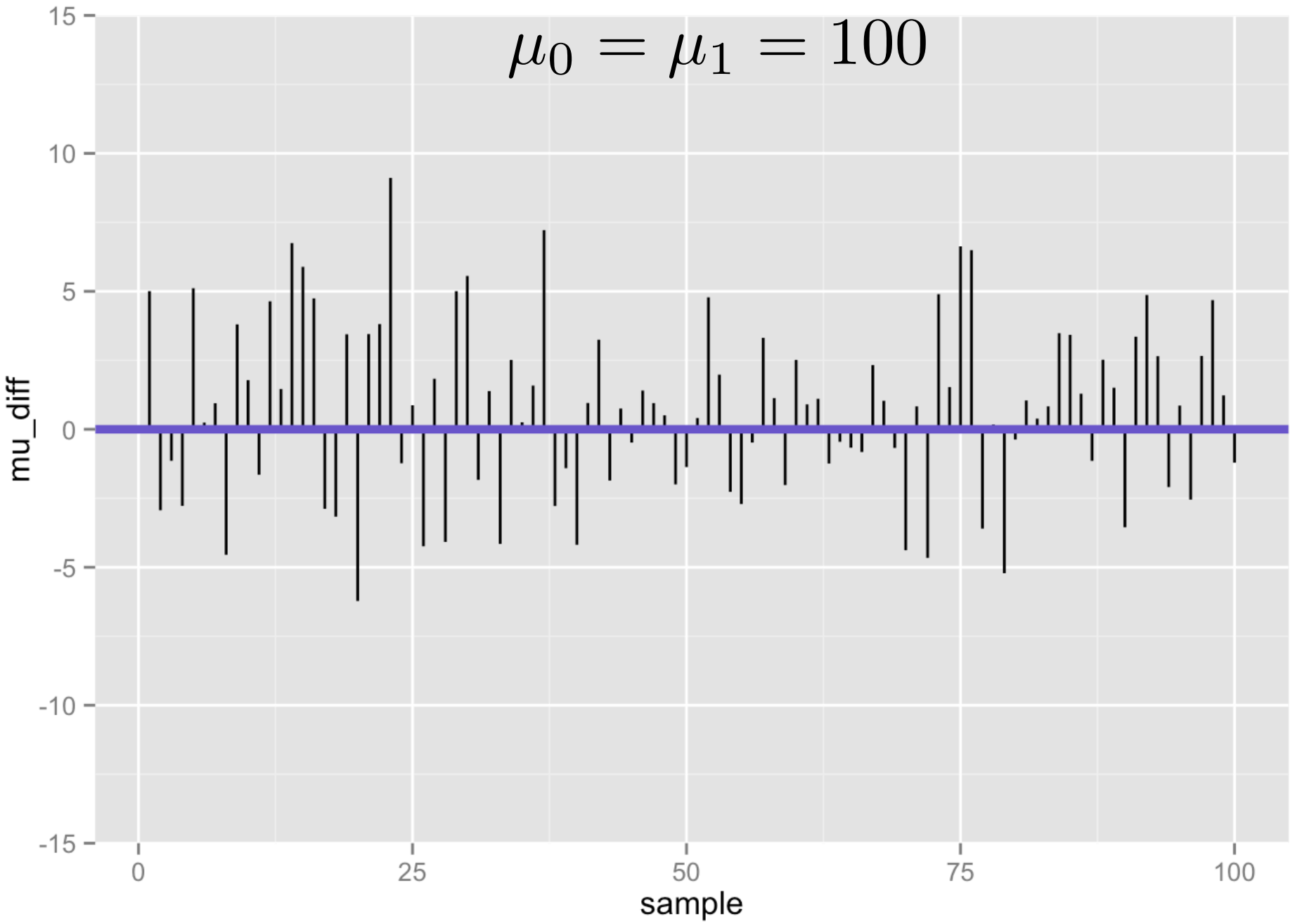
WHEN THE NULL HYPOTHESIS IS TRUE

$$\mu_0 = \mu_1 = 100$$

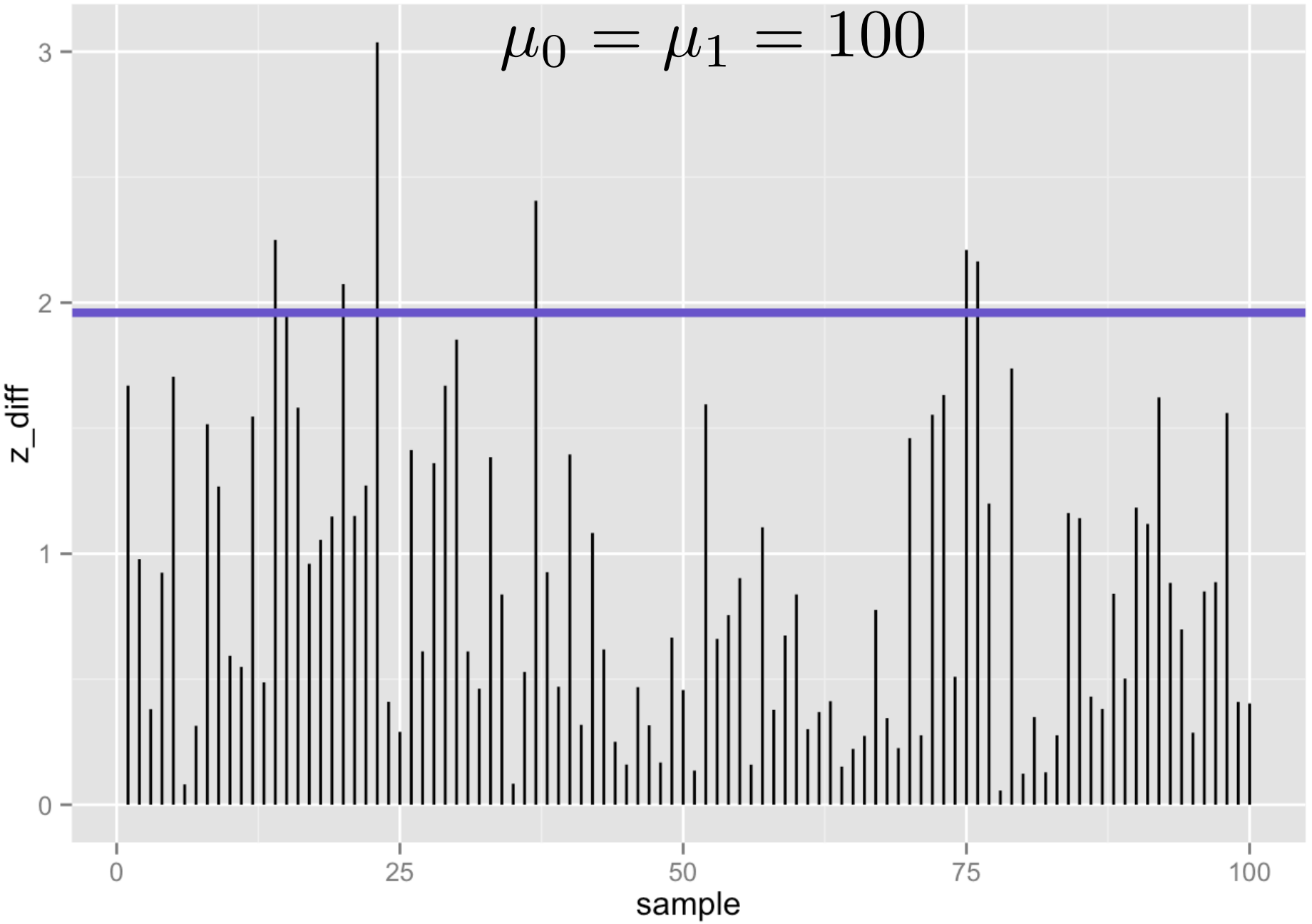
1-sample z-test

n = 25

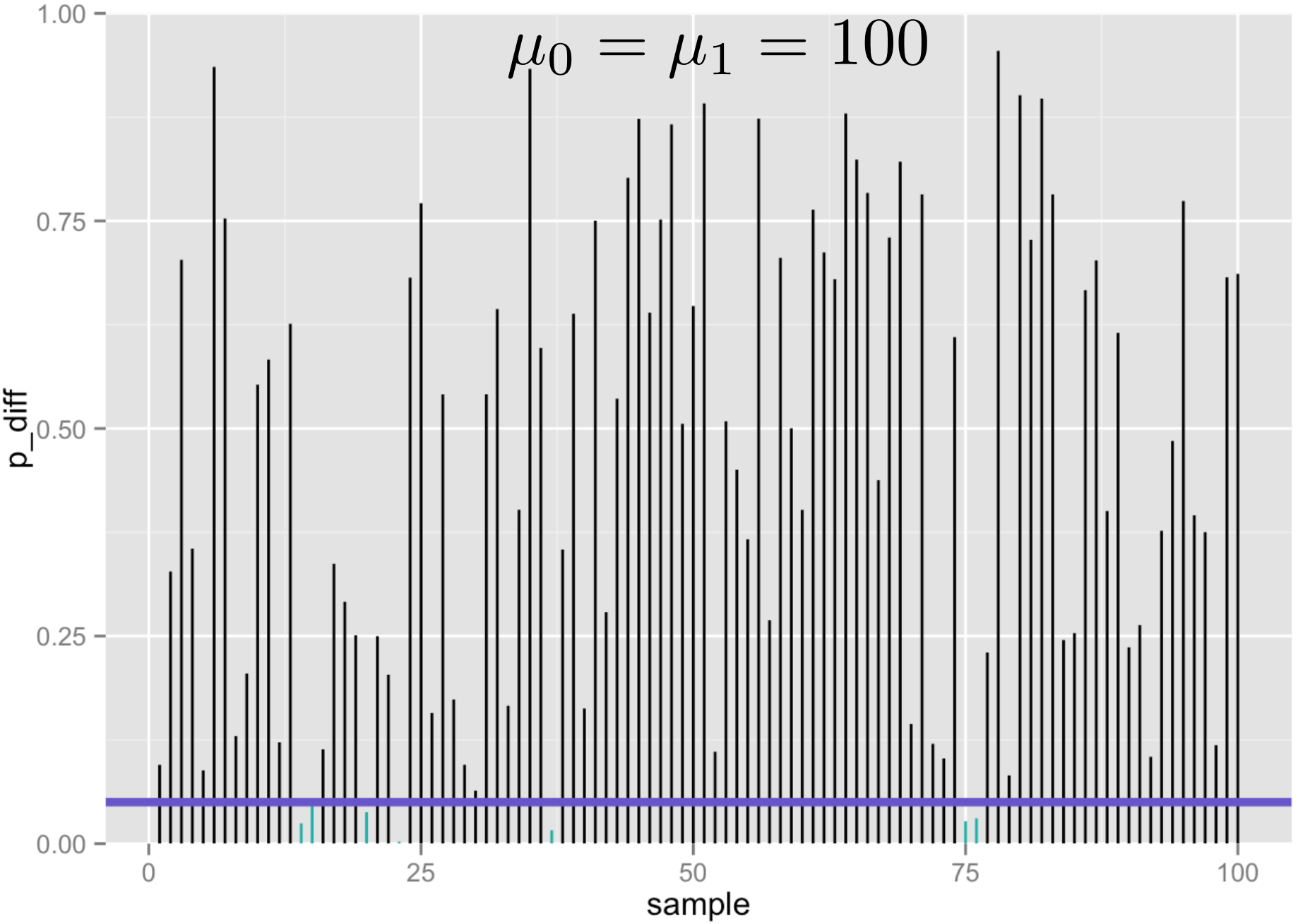
Differences between population mean and sample mean (100 samples) when null is **true**



100 z-statistics (absolute value) when null is **true**: 7% false positives using z-test ($\alpha = .05/2$)



100 p-values when null is **true**: 7% false positives using z-test ($\alpha = .05/2$)



Confusion matrix

	Call based on observed data		
True state of the world	Fail to reject H_0	Reject H_0	
H_0	True negative $1 - \alpha$	False positive Type I error α	# true H_0's
H_1	False negative Type II error β	True positive $1 - \beta$	# true H_1's
		# rejected H_0's	# total tests

If the null is true...

Confusion matrix

	Call based on observed data		
True state of the world	Fail to reject H_0	Reject H_0	
H_0	True negative $1 - \alpha$	False positive Type I error α	# true H_0's
H_1	False negative Type II error β	True positive $1 - \beta$	# true H_1's
		# rejected H_0's	# total tests

But... what if we are wrong??

"Statistics does not tell us **WHETHER** we are right. It tell us the **CHANCES** of being wrong."

http://www.nature.com/nmeth/journal/v10/n9/full/nmeth.2613.html

Two ways we can be wrong…

Type I error (α)
False positive

Type II error (β)
False negative

One way we can be wrong…

Type I error (α)
False positive

Call: reject H_0

- If we had rejected the null, it is of course possible that we should not have!
- That is, the true state of the world may be H_0 (he's not pregnant), but our sample data leads us to reject H_0 and (incorrectly) conclude that he's pregnant
- This is really embarrassing, so we control this: α = ?

The other way we can be wrong…

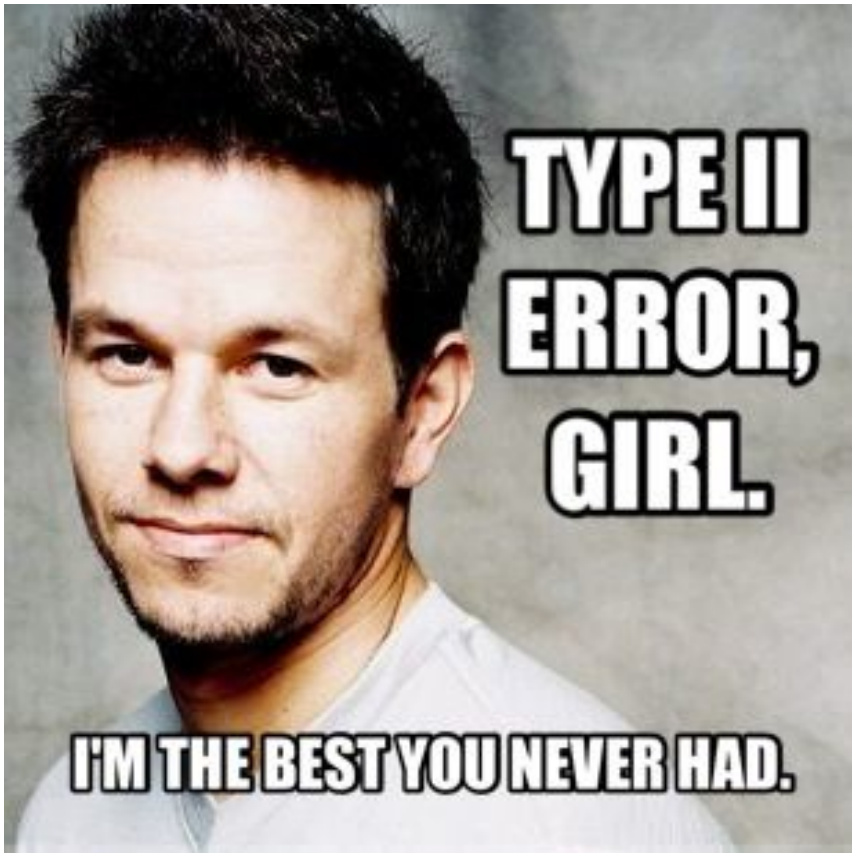
Call: fail to reject H_0

- If we conclude that we cannot reject the null, it is of course possible that we should have!
- That is, the true state of the world may be H_1 (she's pregnant), but our sample data says we don't have good enough evidence to reject H_0 (she's not pregnant)

Type II error (β)
False negative

Type II errors

- The one(s) that got away…



In our aspiring astronauts example…

Type I error (α)
False positive

Decide:
"You **are** smarter than average!"
Reality:
but you are actually **not**

Type II error (β)
False negative

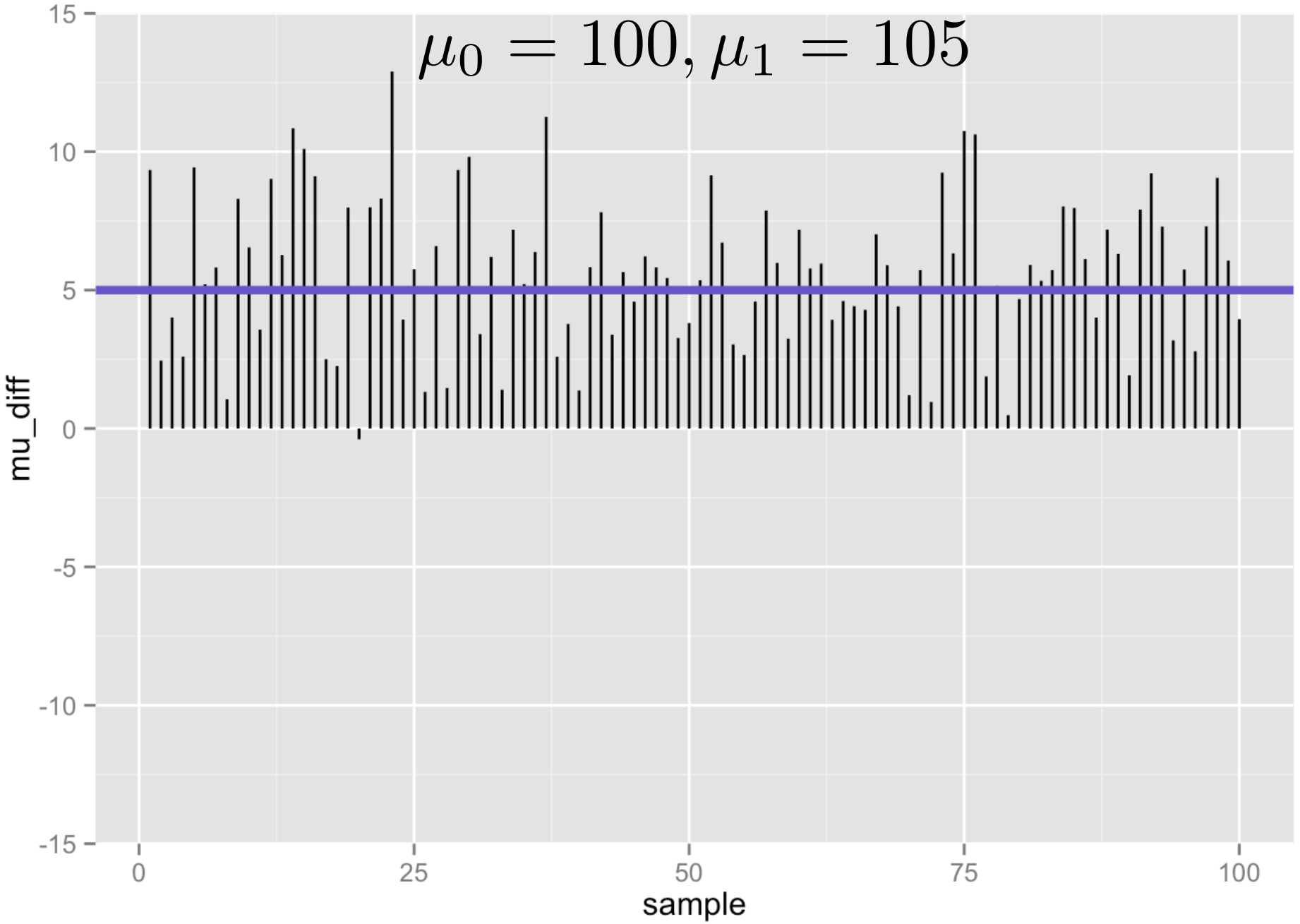
Decide:
"You're **not** smarter than average!"
Reality:
but you **are** actually

WHAT IF: THE NULL HYPOTHESIS IS FALSE?

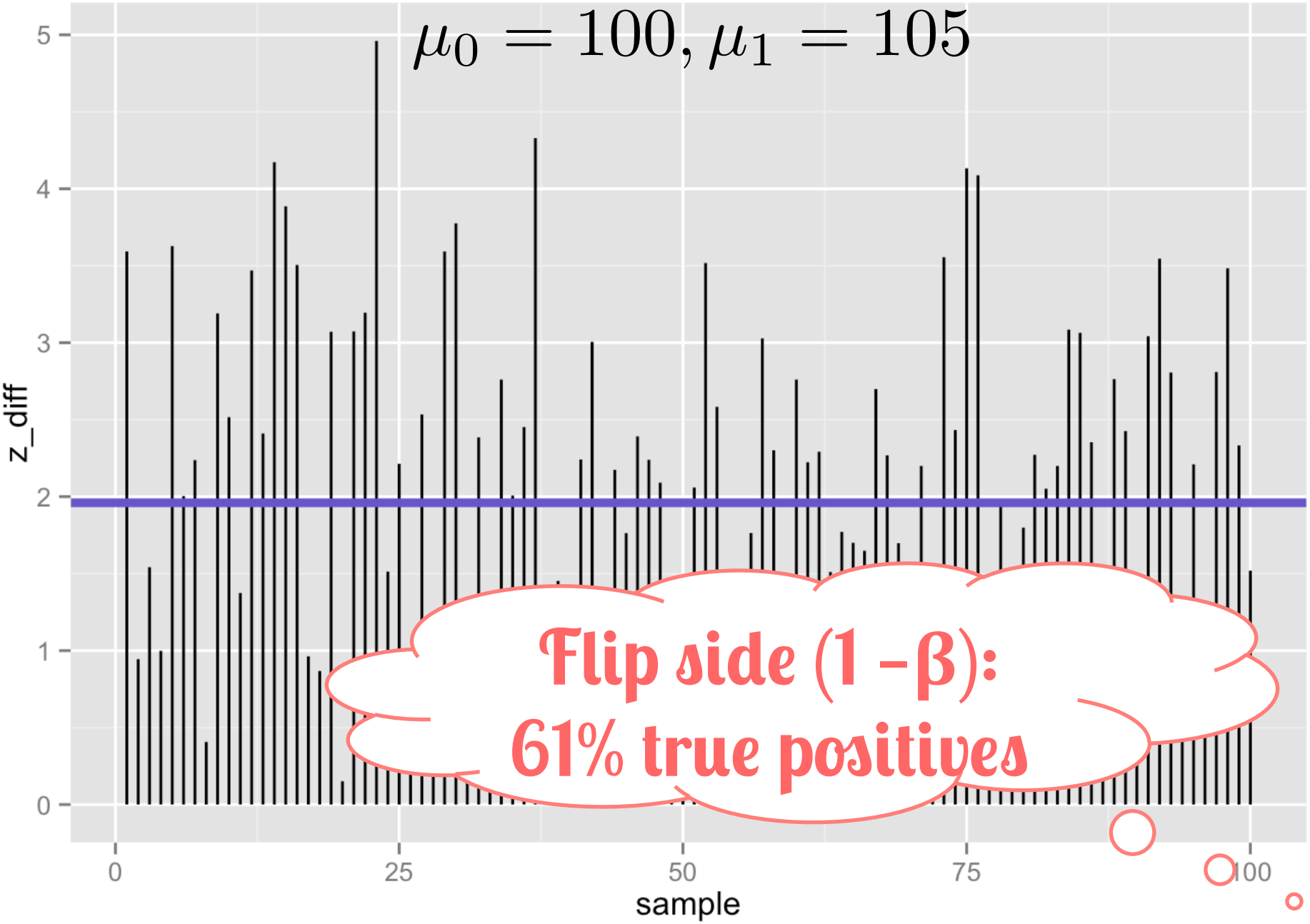
$$\mu_0 = 100, \mu_1 = 105$$

1-sample z-test

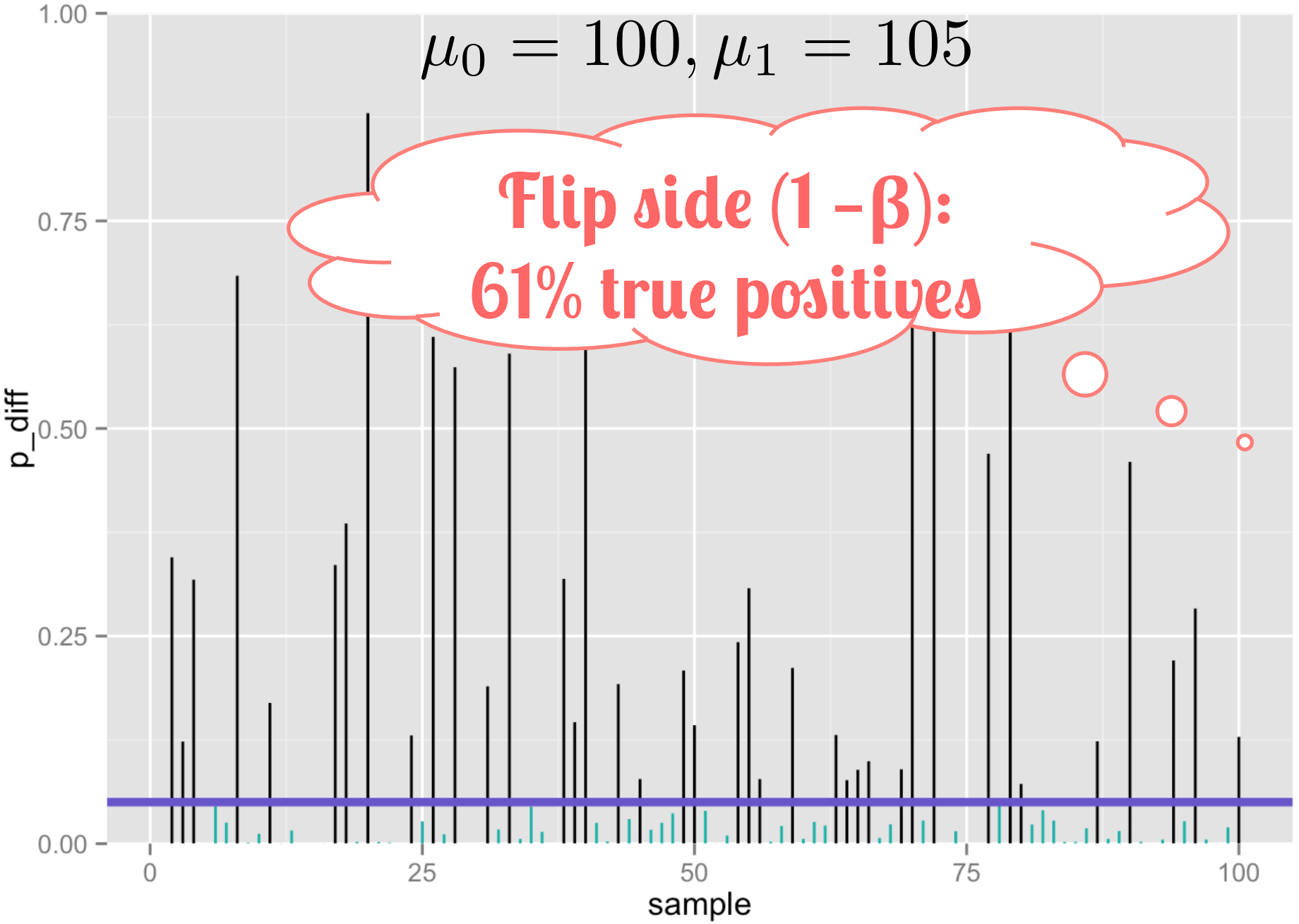
n = 25


$\mu_0 = 100, \mu_1 = 105$
15
10
5
0
-5
-10
-15
mu_diff
0
25
50
75
100
sample

100 z-statistics (absolute value) when null is **false**: 39% false negatives using z-test ($\alpha = .05/2$)



100 p-values when null is **false**: 39% false negatives using z-test ($\alpha = .05/2$)



Now let's vindicate our poor NASA interns: we have found the actual sample data, and now can calculate both the sample mean and standard deviation. We'll use the sample standard deviation (s = 13) to estimate the population s.d. (σ).

Obtaining a test statistic

- Remember our general formula for any test statistic about some parameter, θ :

$$\frac{\hat{\theta} - \theta_0}{SE_{\theta_0}}$$

$$\frac{\hat{\theta} - \theta_0}{\widehat{SE}_{\theta_0}}$$

One-sample t-test

$$t_{df=24} = \frac{\overline{x} - \mu_0}{s_x / \sqrt{n}}$$

$$= \frac{\quad - \quad}{\quad / \sqrt{\quad}}$$

$$= \frac{\quad}{\quad} =$$

One-sample t-test

$$t_{df=24} = \frac{\overline{x} - \mu_0}{s_x/\sqrt{n}}$$

$$= \frac{105 - 100}{13/\sqrt{25}}$$

$$= \frac{5}{2.6} = 1.923$$

What is the p-value for the *t* statistic?

- One-tailed, upper

- Two-tailed

What is the p-value for the *t* statistic?

- One-tailed, upper

```
pt_up <- 1 - pt(t, n - 1)
```

- Two-tailed

```
pt_2 <- 2 * pt_up
```

```
c(pt_up, pt_2)
[1] 0.03320682 0.06641363
```

Two-tailed p-values more generally…

```
pt_1tail <- min(pt(t, n - 1), 1 - pt(t, n - 1))
pt_2tail <- 2*pt_1tail
pt_2tail
[1] 0.06641363
```

Constructing a confidence interval for μ

$$\overline{x} \pm (q_{t,1-\alpha}) \left(\frac{s}{\sqrt{n}} \right)$$

Calculating 95% CI for μ in R, σ unknown

```
# sample statistics
xbar <- 105
s <- 13
n <- 25

# margin of error
me <- qt(.975, n - 1) * (s/sqrt(n)) # .975 --> .025 at EACH tail

# 95% confidence intervals
lowert <- xbar - me
uppert <- xbar + me
c(lowert, uppert)
```

Is μ = 100 in there??

Calculating 95% CI for μ in R, σ unknown

```
# sample statistics
xbar <- 105
s <- 13
n <- 25

# margin of error
me <- qt(.975, n - 1) * (s/sqrt(n)) # .975 --> .025 at EACH tail

# 95% confidence intervals
lowert <- xbar - me
uppert <- xbar + me
c(lowert, uppert)
[1]  99.63386 110.36614
```

One-sample t-test in R

- Have to have actual data- not just sample statistics
- So far, I have only been playing with sample statistics- I didn't actually have sample data! Let's make up some sample data with the sample mean/sd we need:

```
set.seed(1)
iq_aa <- seq(83.8, 126.2, length.out = 25)
mean(iq_aa) # perfect!
[1] 105
sd(iq_aa) # close enough!
[1] 13.00231
```

One-sample t-test in R

```
aat <- t.test(iq_aa, mu = 100) # Ho: mu = 100
aat

        One Sample t-test

data:  iq_aa
t = 1.9227, df = 24, p-value = 0.06646
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
  99.63291 110.36709
sample estimates:
mean of x
      105
# Recall our previous 95% confidence interval- pretty close!
c(lowert, uppert)
[1]  99.63386 110.36614
```

Mansplain it to me…

```
devtools::install_github(c("
hilaryparker/explainr",
"hilaryparker/mansplainr"))
```

```
mansplain(aat)
```

```
That's great that you were
able to do a hypothesis
test. You got a p-value of
0.066459. That means it's
not significant at alpha = .
05, but that's OK. The
important thing is that you
tried.
```

Complain about it…

```
devtools::install_github(c("
hilaryparker/explainr",
"hilaryparker/complainr"))
```

```
complain(aat)
```

```
This hypothesis test had a
p-value of 0.0664587.
```

```
That's if you can trust any
frequentist method. You
should really be doing a
Bayesian analysis. Did you
hear about that journal that
banned p-values?
```

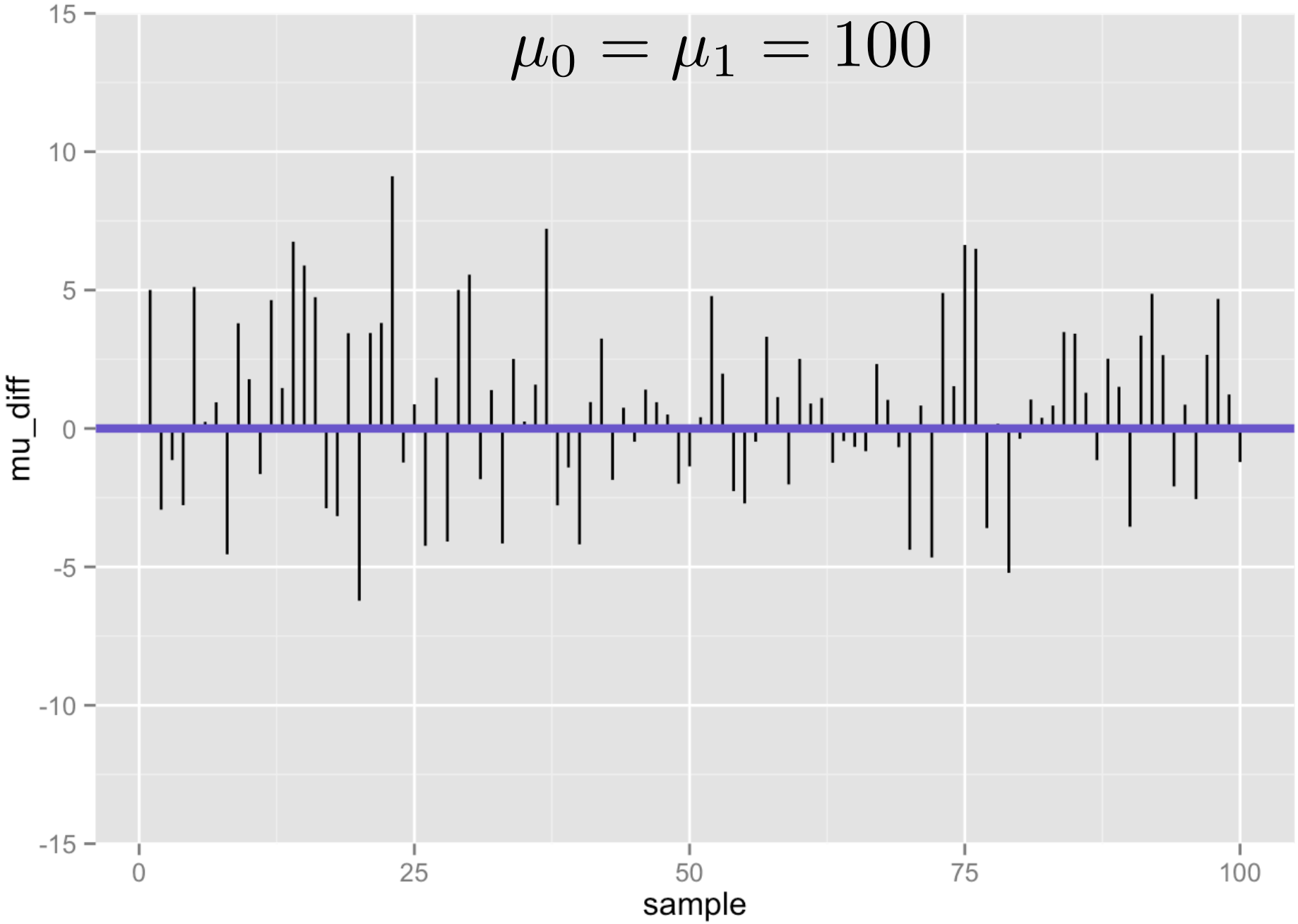
WHEN THE NULL HYPOTHESIS IS TRUE

$$\mu_0 = \mu_1 = 100$$

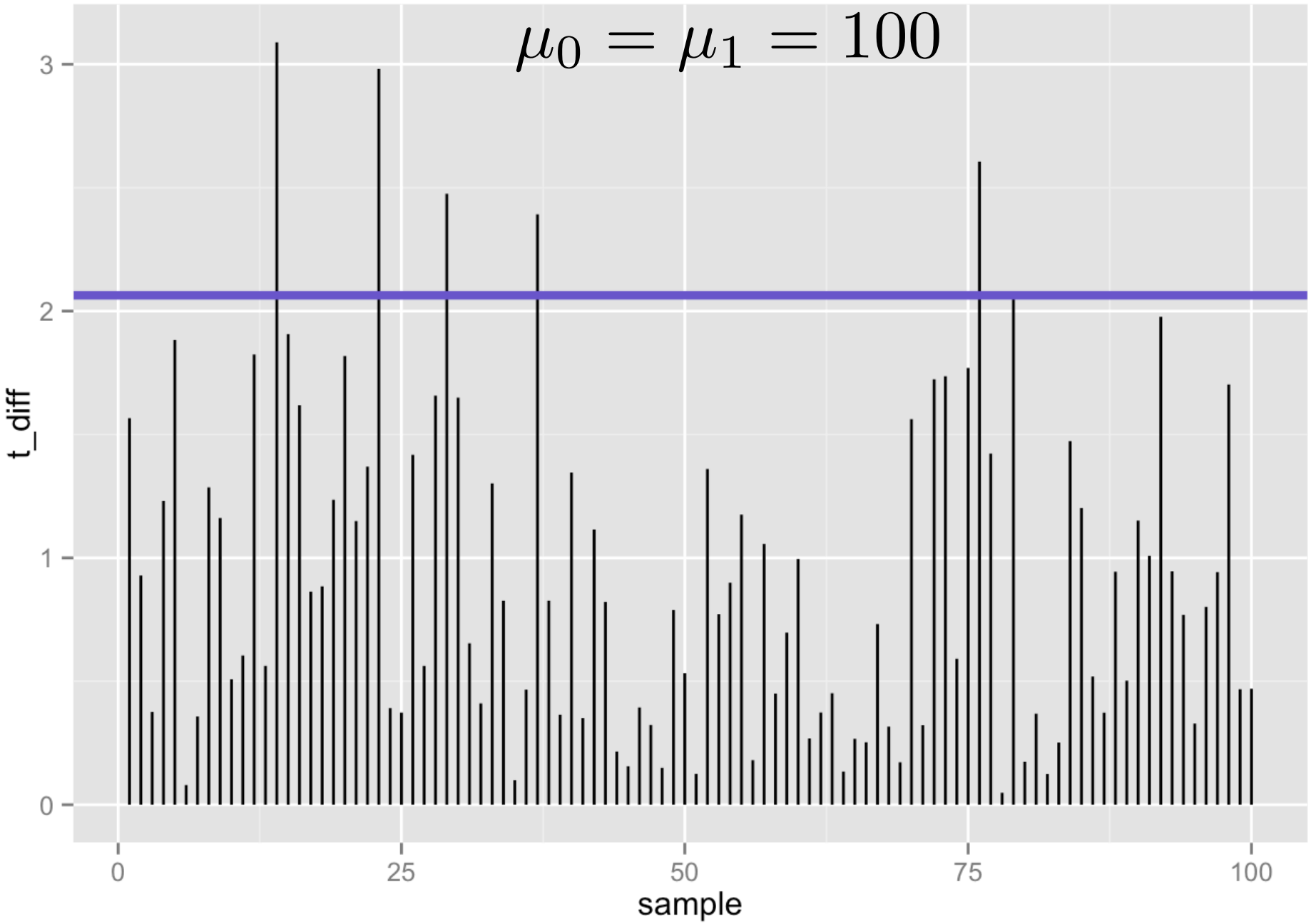
1-sample t-test

n = 25

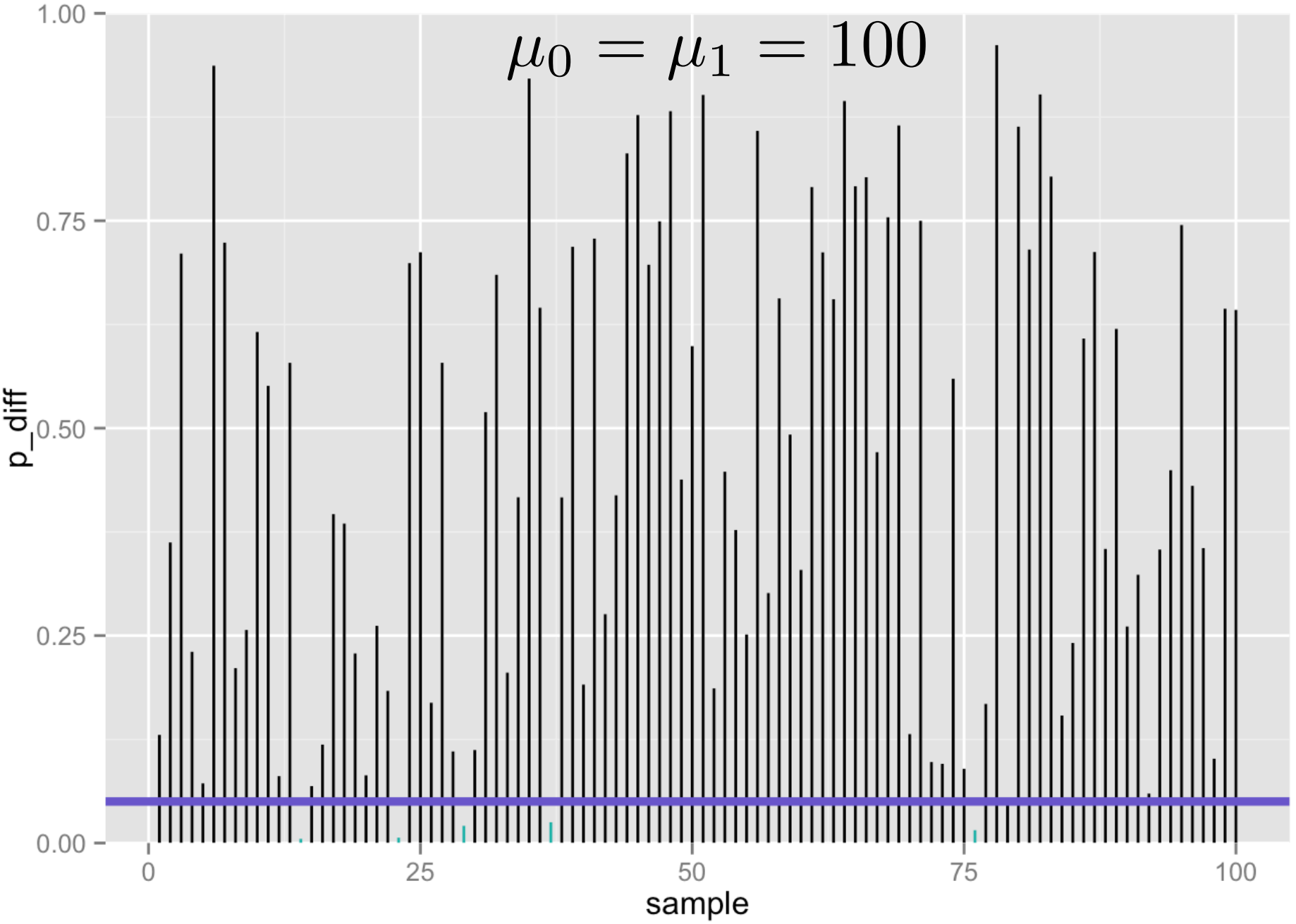
Differences between population mean and sample mean (100 samples) when null is **true**



100 t-statistics (absolute value) when null is **true**: 5% false positives using t-test ($\alpha = .05/2$)



100 p-values when null is **true**: 5% false positives using t-test (α = .05/2)



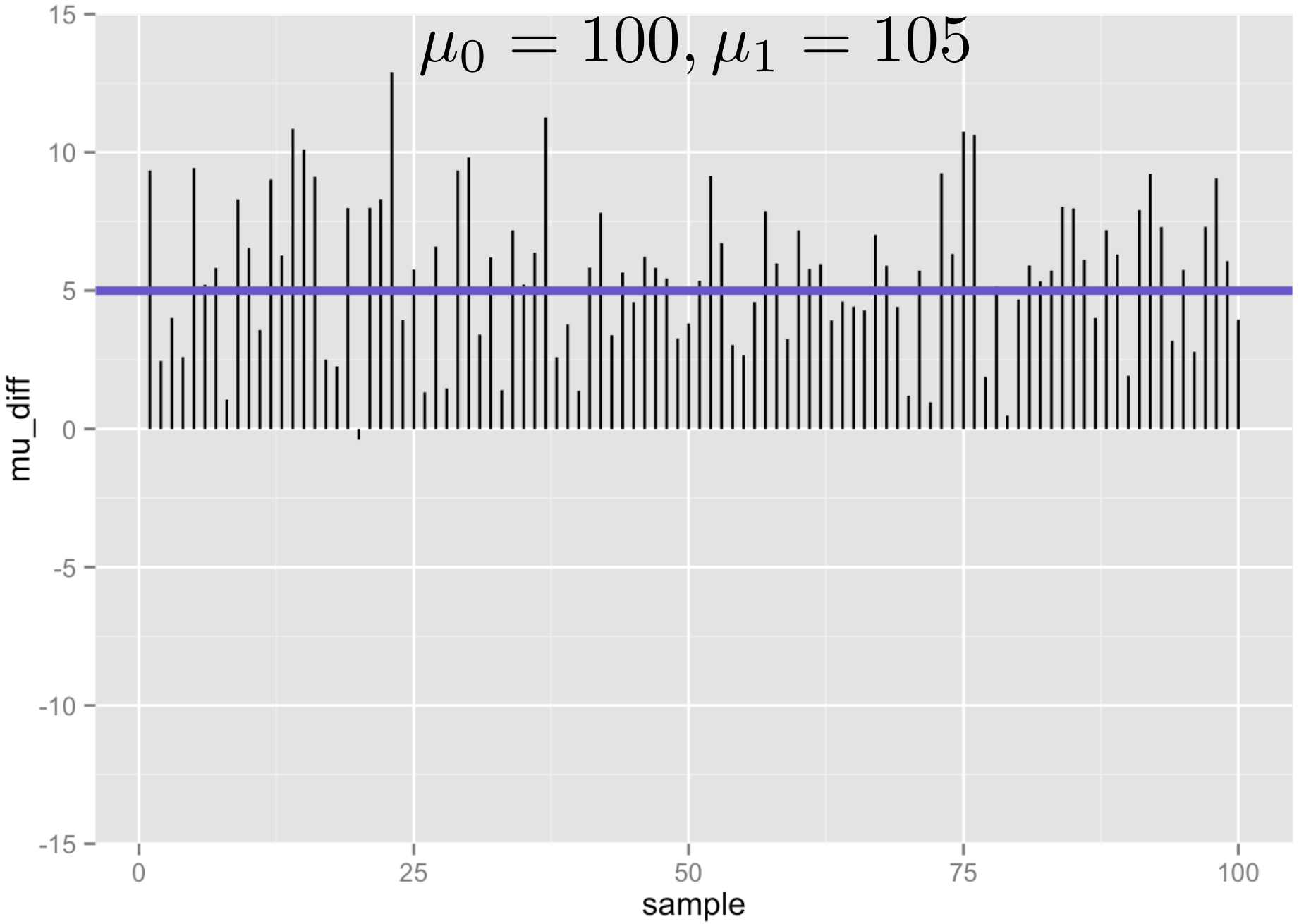
WHAT IF: THE NULL HYPOTHESIS IS FALSE?

$$\mu_0 = 100, \mu_1 = 105$$

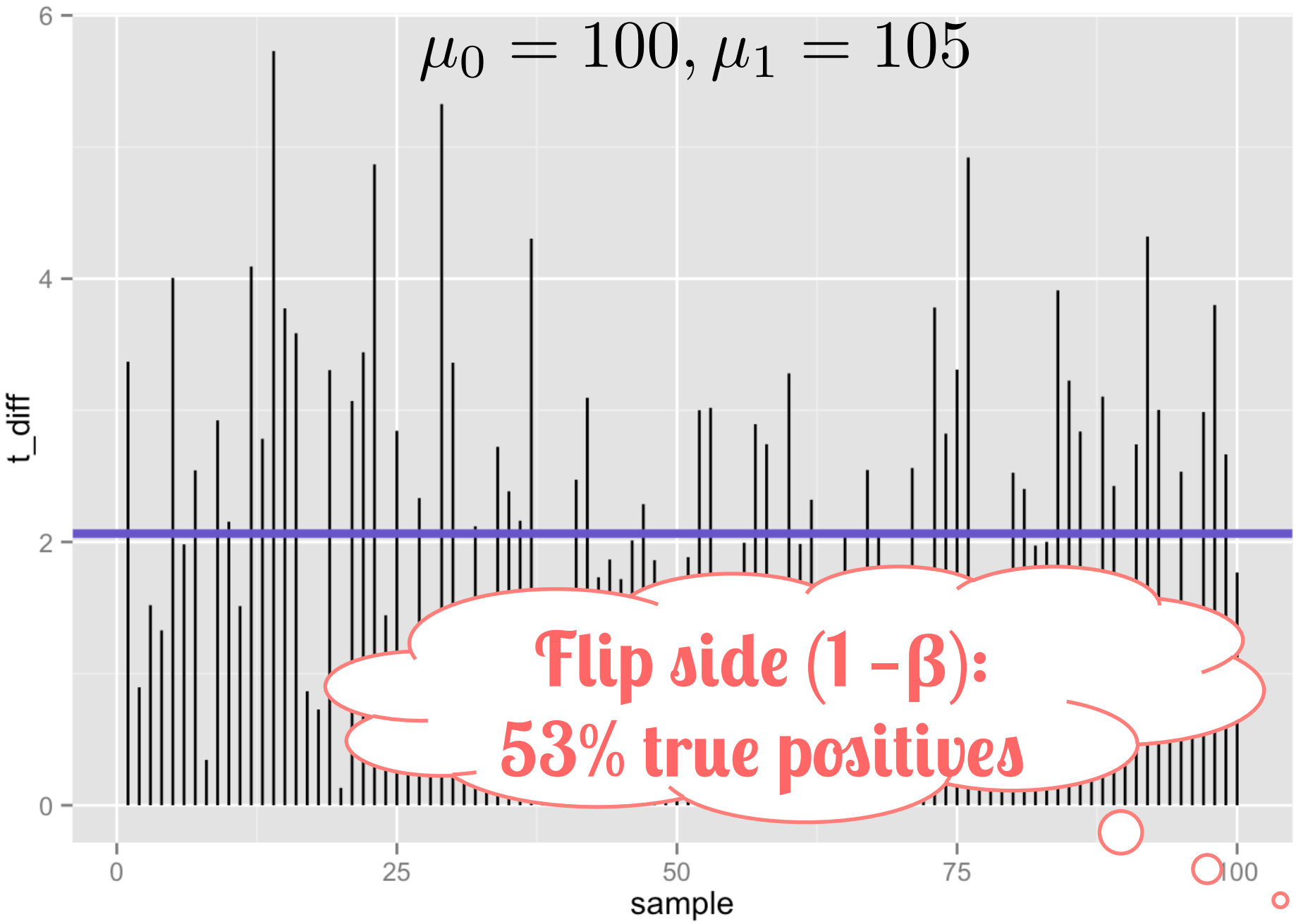
1-sample t-test

n = 25

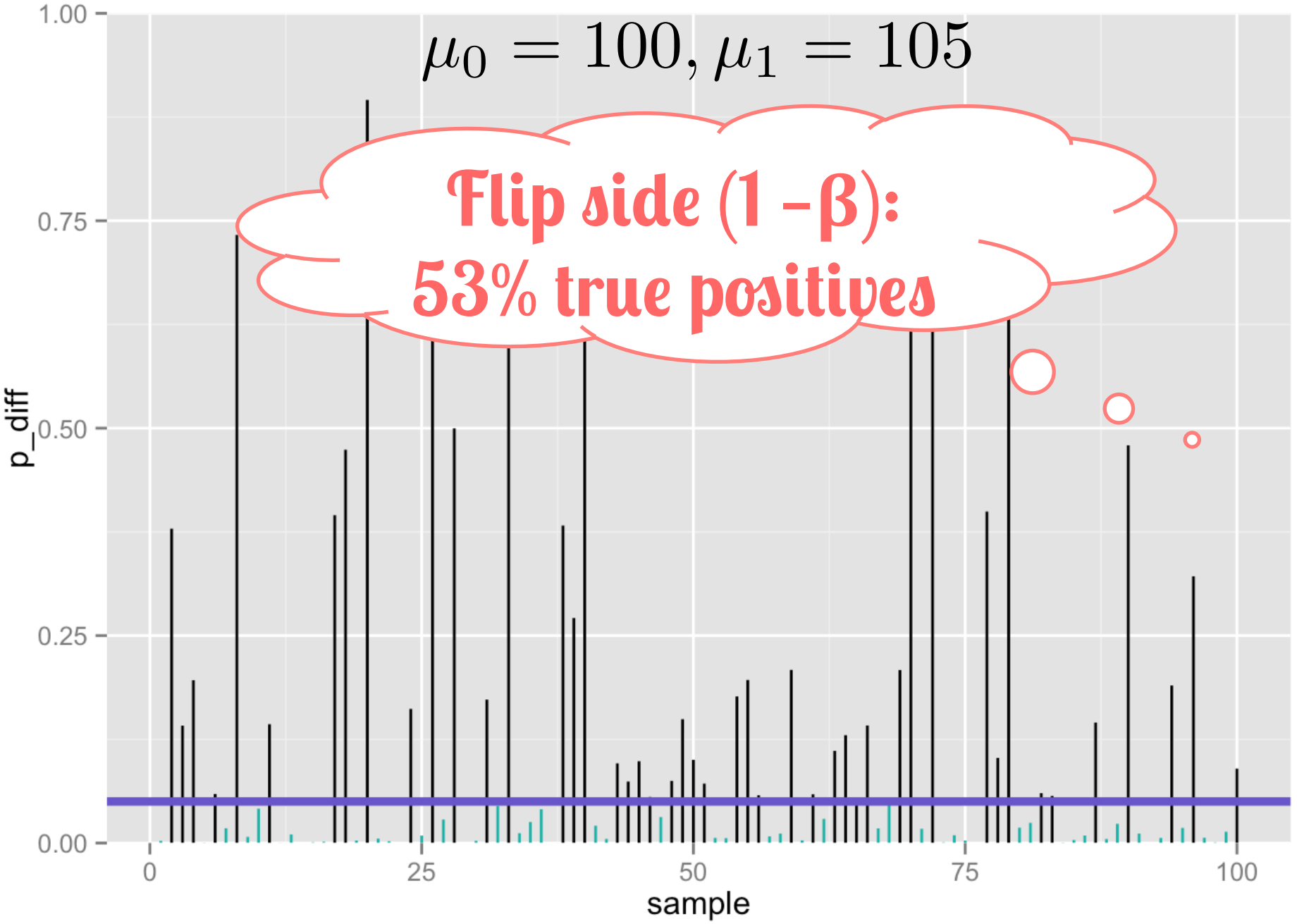
Differences between population mean and sample mean (100 samples) when null is **false**



100 t-statistics (absolute value) when null is **false**: 47% false negatives using t-test ($\alpha = .05/2$)



100 p-values when null is **false**: 47% false negatives using t-test ($\alpha = .05/2$)



≈ 50% true positives seems low…

- Only about half of our true effects would be detected in our study
- Why?
- Perhaps we lacked statistical power

Let's do a one-tailed t-test…

```
> aat_1 <- t.test(iq_aa, mu = 100, alternative =
c("greater"))
> aat_1

        One Sample t-test

data:  iq_aa
t = 1.9227, df = 24, p-value = 0.03323
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 100.5509      Inf
sample estimates:
mean of x
      105
```

If the null is true…

- The test statistic under the null will have a central t distribution with $\nu = n - 1 = 24$ degrees of freedom.
- The (one-tailed) critical value will be:

```
> qt(.95, 24) # t_critical, null dist
[1] 1.710882
```

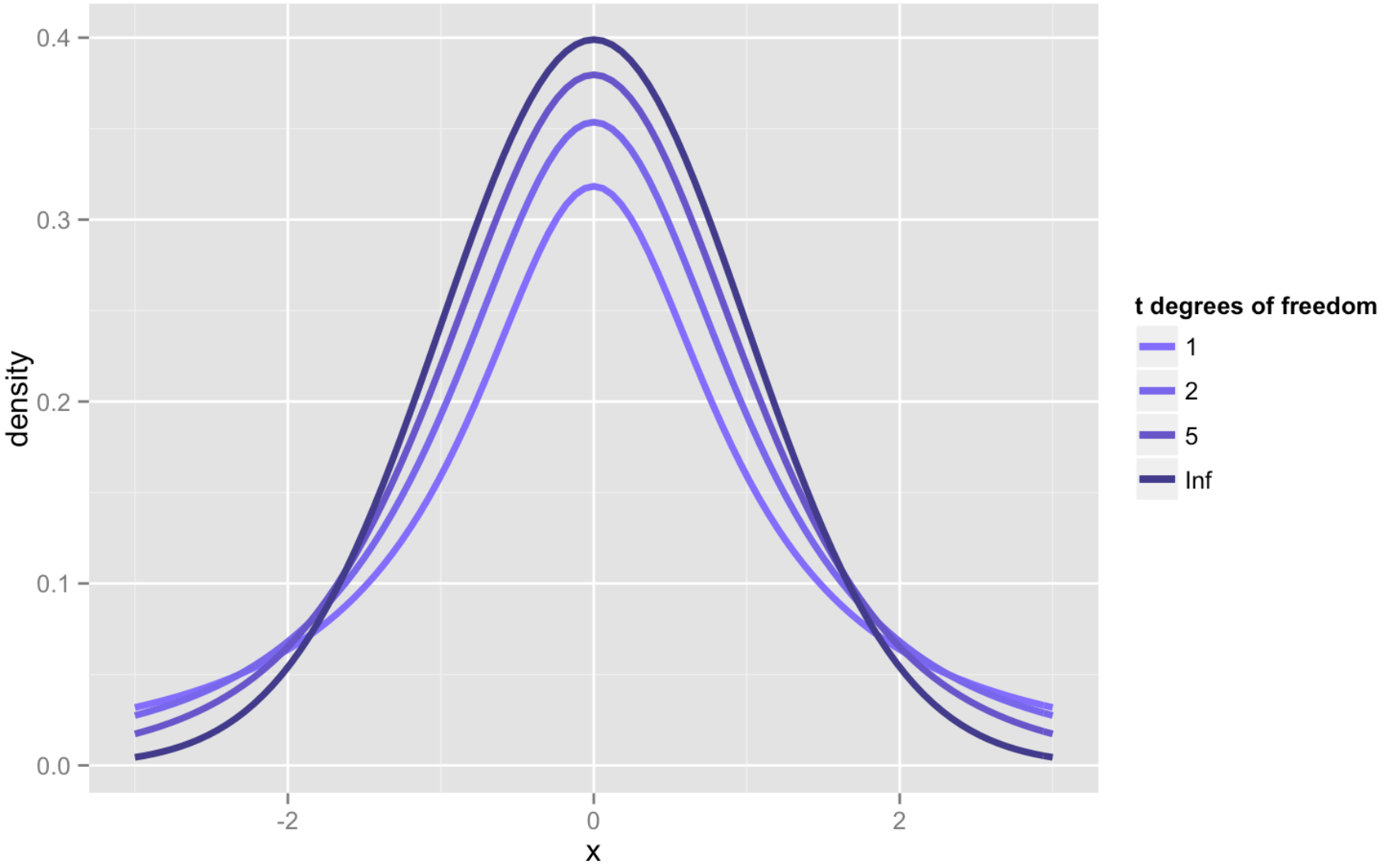
One-tailed test

- P(false positive) = α = Type I error rate = .05
- P(false negative) = β = Type II error = ?
- P(true positive) = $1 - \beta$ = power = ?

Finding β (and $1 - \beta$)

- Need to know exact **null** distribution (just as with NHST)
- Also need to know exact **alternative** distribution of the test statistic
  - Often requires some specialized statistical knowledge
- In general, it is much more likely that expressions for the null distribution of the test statistic will be available than expressions for the non-null distribution.

Recall student's *t*-distribution



TDist {stats} R Documentation

The Student t Distribution

Description

Density, distribution function, quantile function and random generation for the t distribution with `df` degrees of freedom (and optional non-centrality parameter `ncp`).

Usage

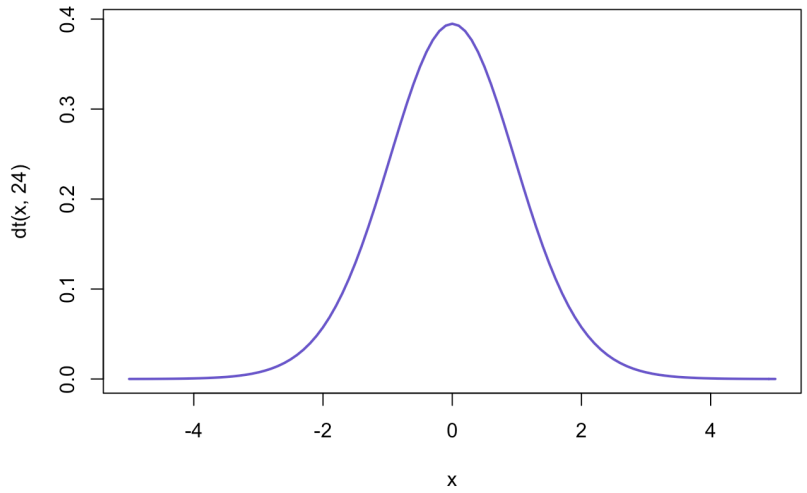
```
dt(x, df, ncp, log = FALSE)
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)
rt(n, df, ncp)
```

Arguments

`x, q`	vector of quantiles.
`p`	vector of probabilities.
`n`	number of observations. If `length(n) > 1`, the length is taken to be the number required.
`df`	degrees of freedom (> *0*, maybe non-integer). `df = Inf` is allowed.
`ncp`	non-centrality parameter *delta*; currently except for `rt()`, only for `abs(ncp) <= 37.62`. If omitted, use the central t distribution.

Central t-distribution (ν = degrees of freedom)

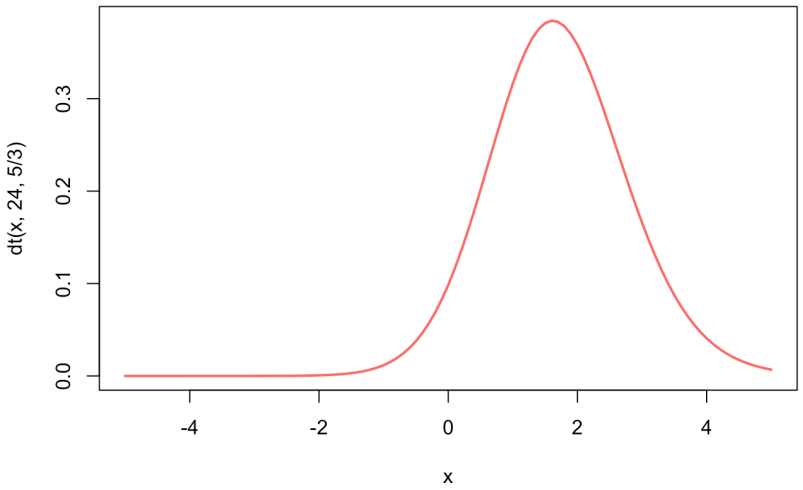
- $\boldsymbol{\delta}$ = 0
- Mean = 0
- Variance slightly > than N(0, 1)
- Kurtosis (biased) is > 3
- Symmetric

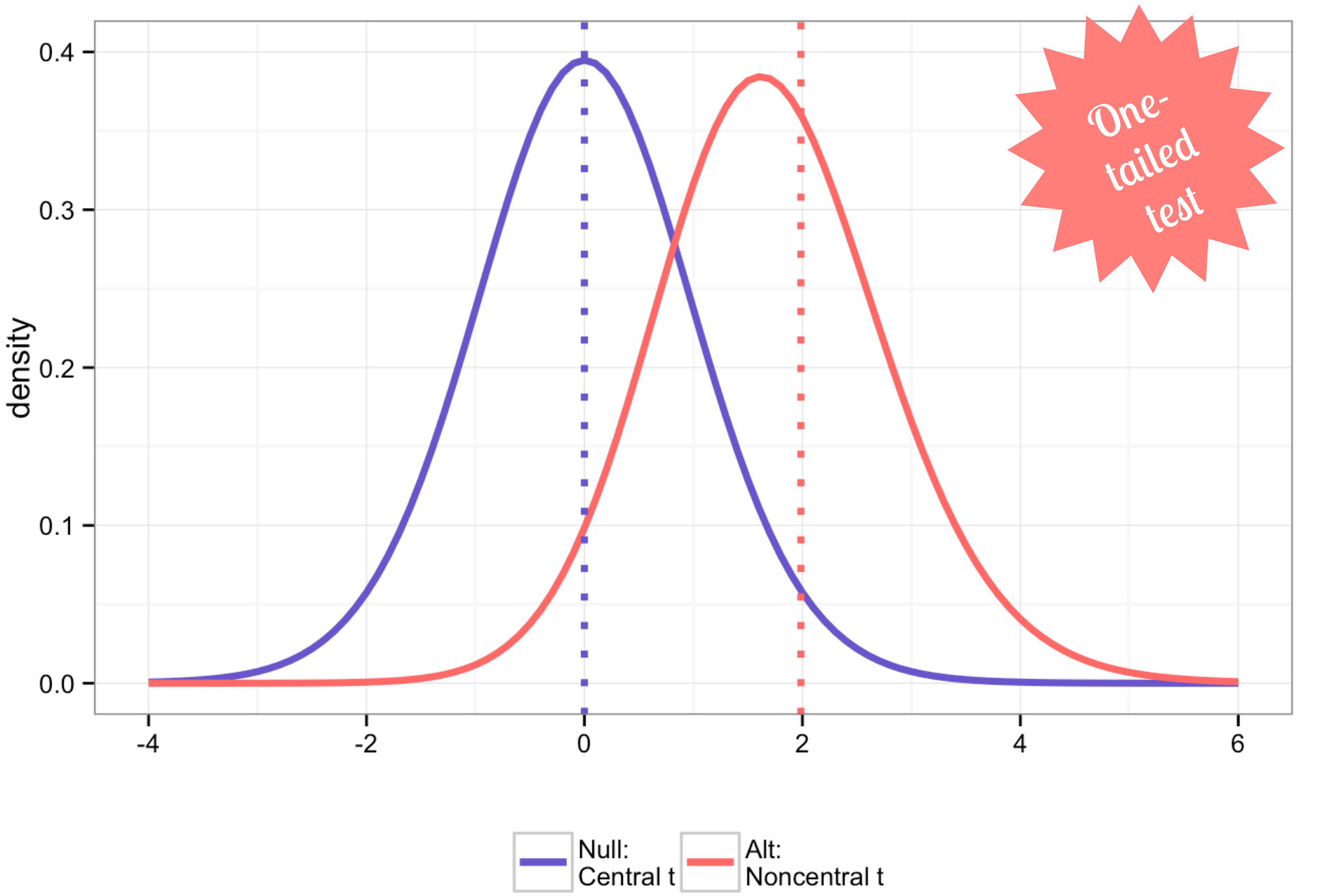


Noncentral t distribution (ν = degrees of freedom)

- **δ ≠ 0**
- Asymmetric: skewed in the direction of **δ**

$$E(T) = \begin{cases} \delta\sqrt{\frac{\nu}{2}}\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)} & \text{if } \nu > 1 \\ \text{Does not exist} & \text{if } \nu \leq 1 \end{cases}$$




One-tailed test
density
0.4
0.3
0.2
0.1
0.0
-4
-2
0
2
4
6
Null:
Central t
Alt:
Noncentral t

How do we calculate the ncp?

- The noncentrality parameter (ncp) is defined as:

$$\delta = \sqrt{n} E_s$$

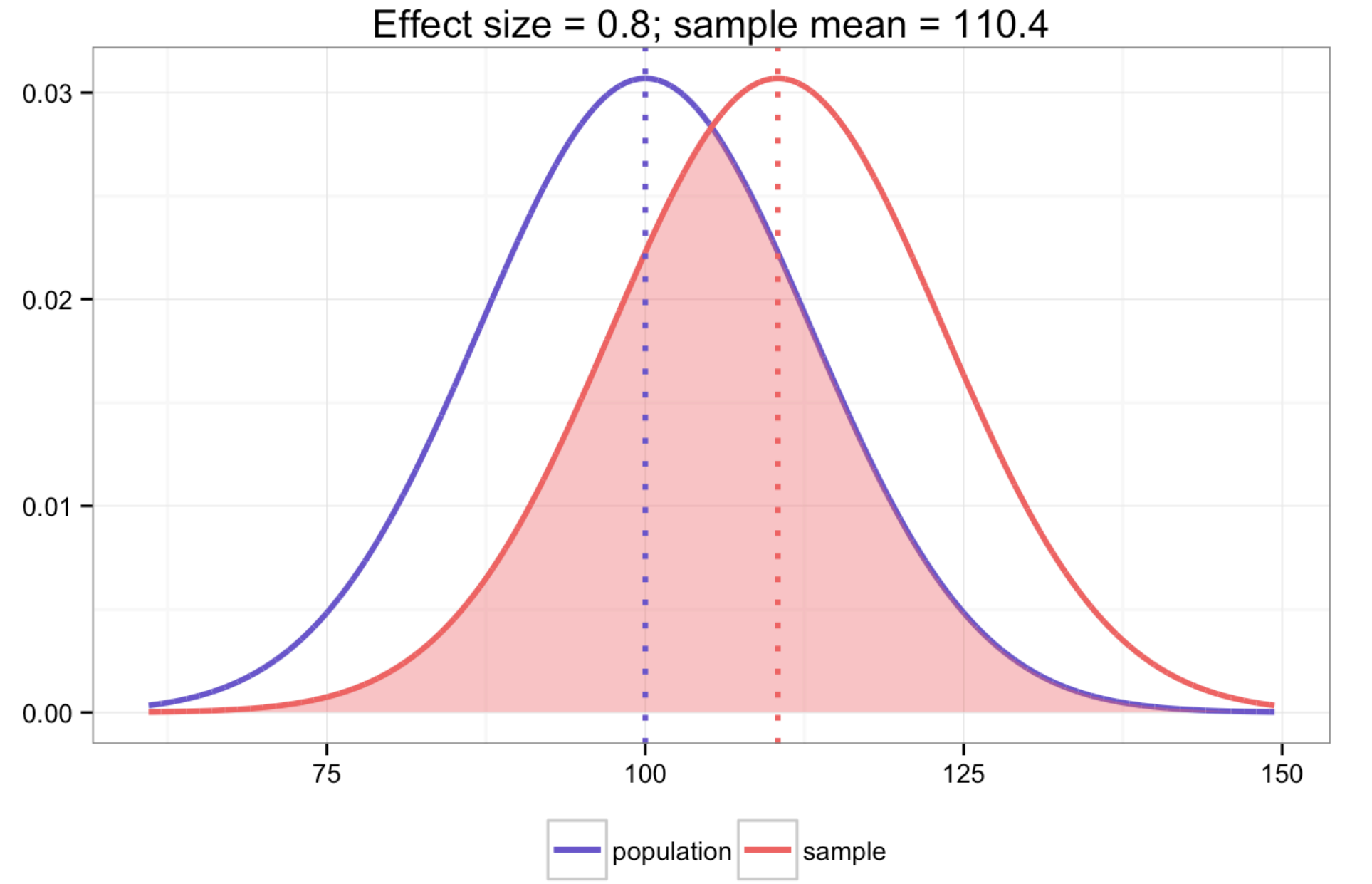
- Where E_s is the standardized measure of effect size…how do we calculate this?

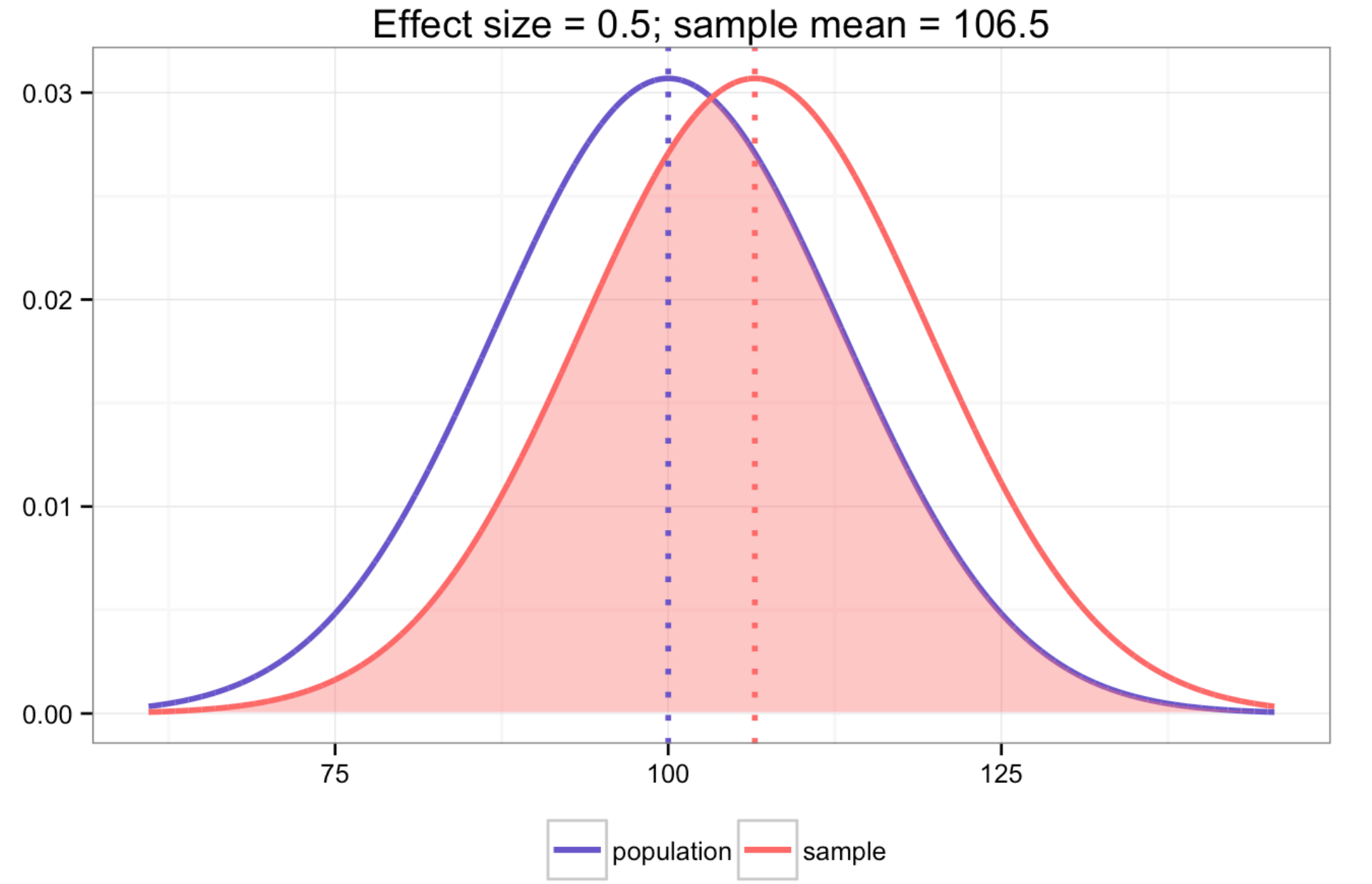
Effect size

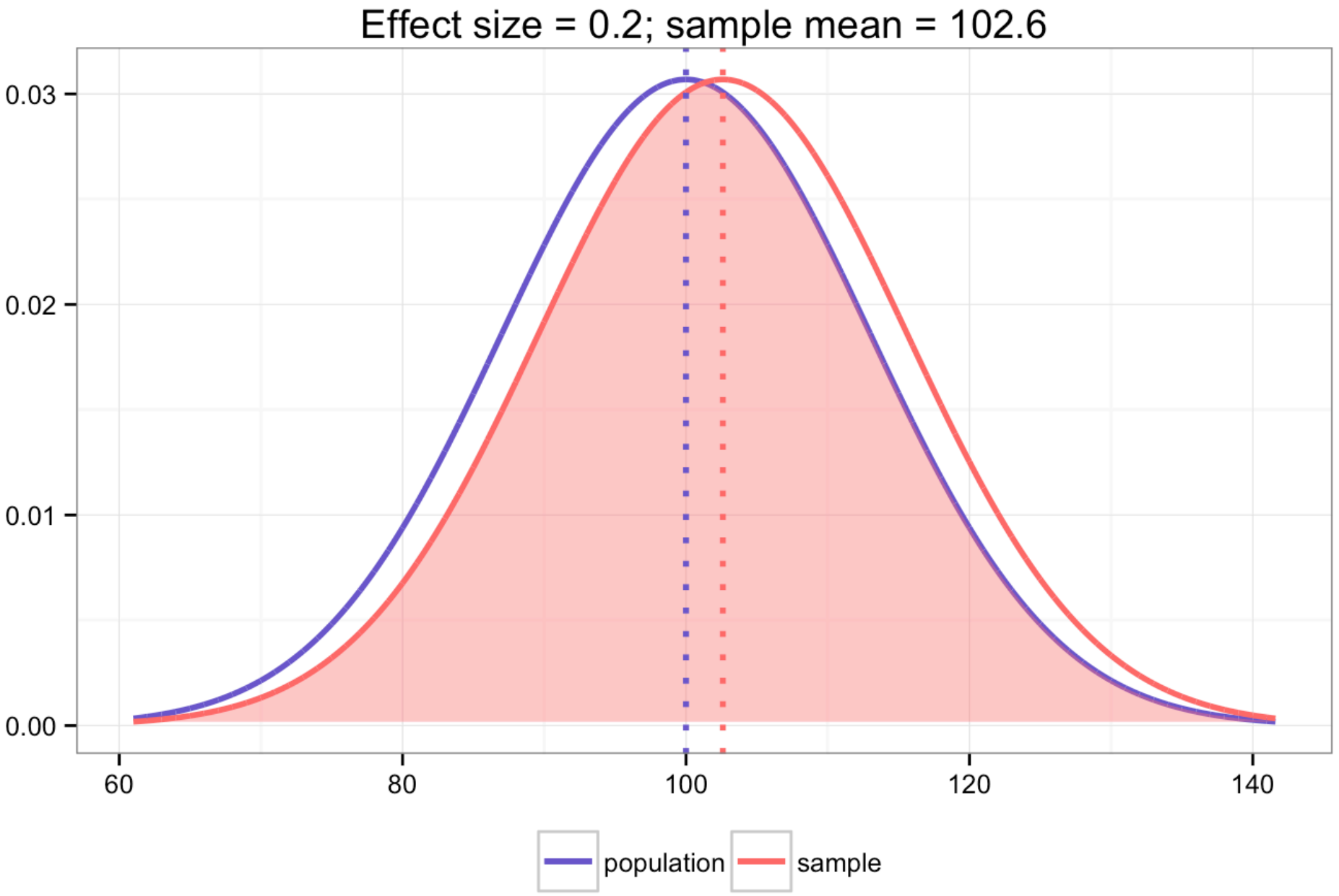
$$t_\nu = \frac{\overline{x} - \mu_0}{s_x / \sqrt{n}}$$

$$E_s = \frac{\mu_1 - \mu_0}{s}$$

What is not in this formula?


Effect size = 0.8; sample mean = 110.4
0.03
0.02
0.01
0.00
75
100
125
150
population
sample


Effect size = 0.5; sample mean = 106.5
0.03
0.02
0.01
0.00
75
100
125
population
sample


Effect size = 0.2; sample mean = 102.6
0.03
0.02
0.01
0.00
60
80
100
120
140
population
sample

Calculating effect size and ncp

- In our example, the E_s is defined just by the sample mean, null mean, and the sample s.d., so the ncp is:

$$\begin{aligned}\delta &= \sqrt{n} E_s \\ &= \sqrt{25} \times \frac{105 - 100}{13} \\ &= 5 \times \frac{5}{13} = 1.923077\end{aligned}$$

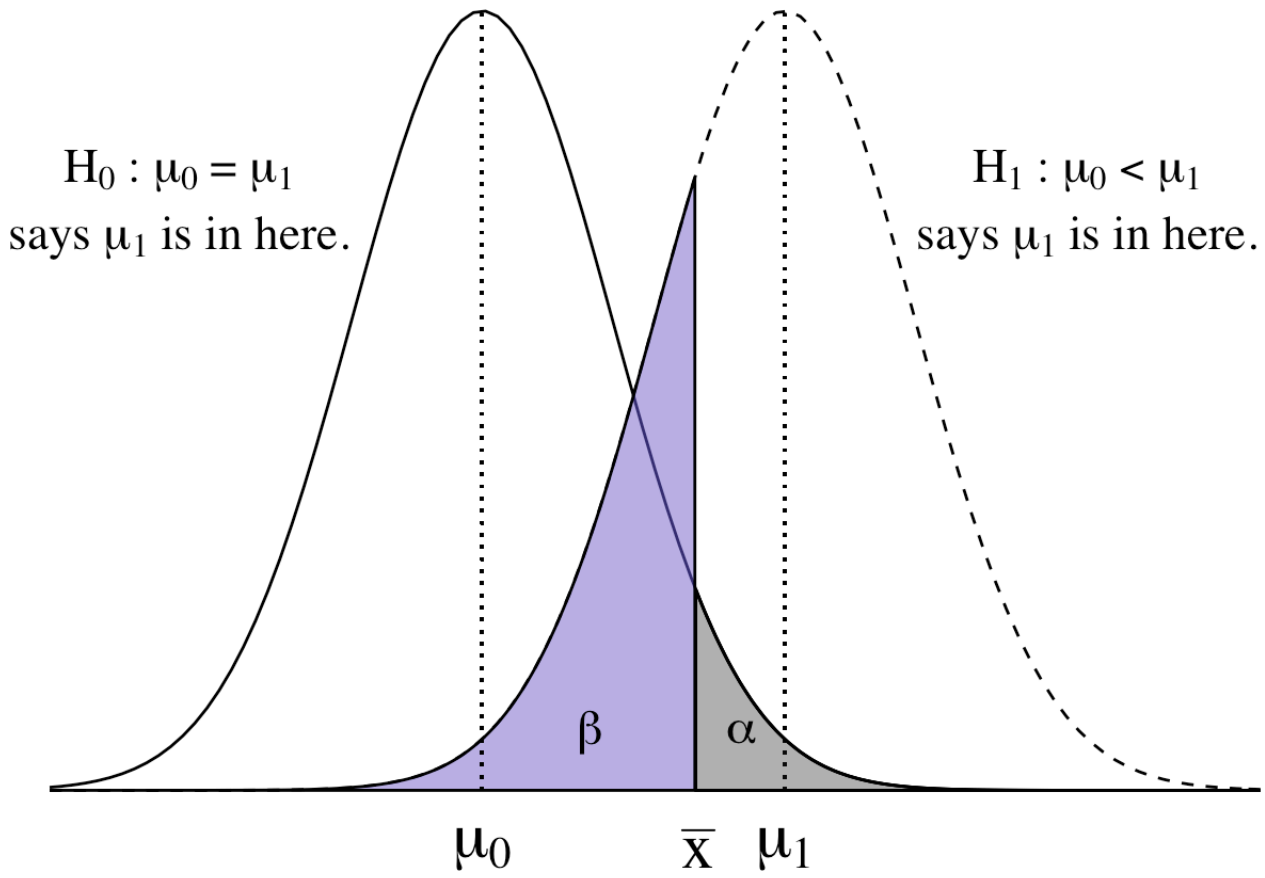
The null and alternative t-distributions

Distribution of t under H_0 is $t_{\nu=24,\delta=0}$

Distribution of t under H_1 is $t_{\nu=24,\delta=1.923}$

Use alternative distribution to solve for β

TYPE II ERROR: p(false negative) = β



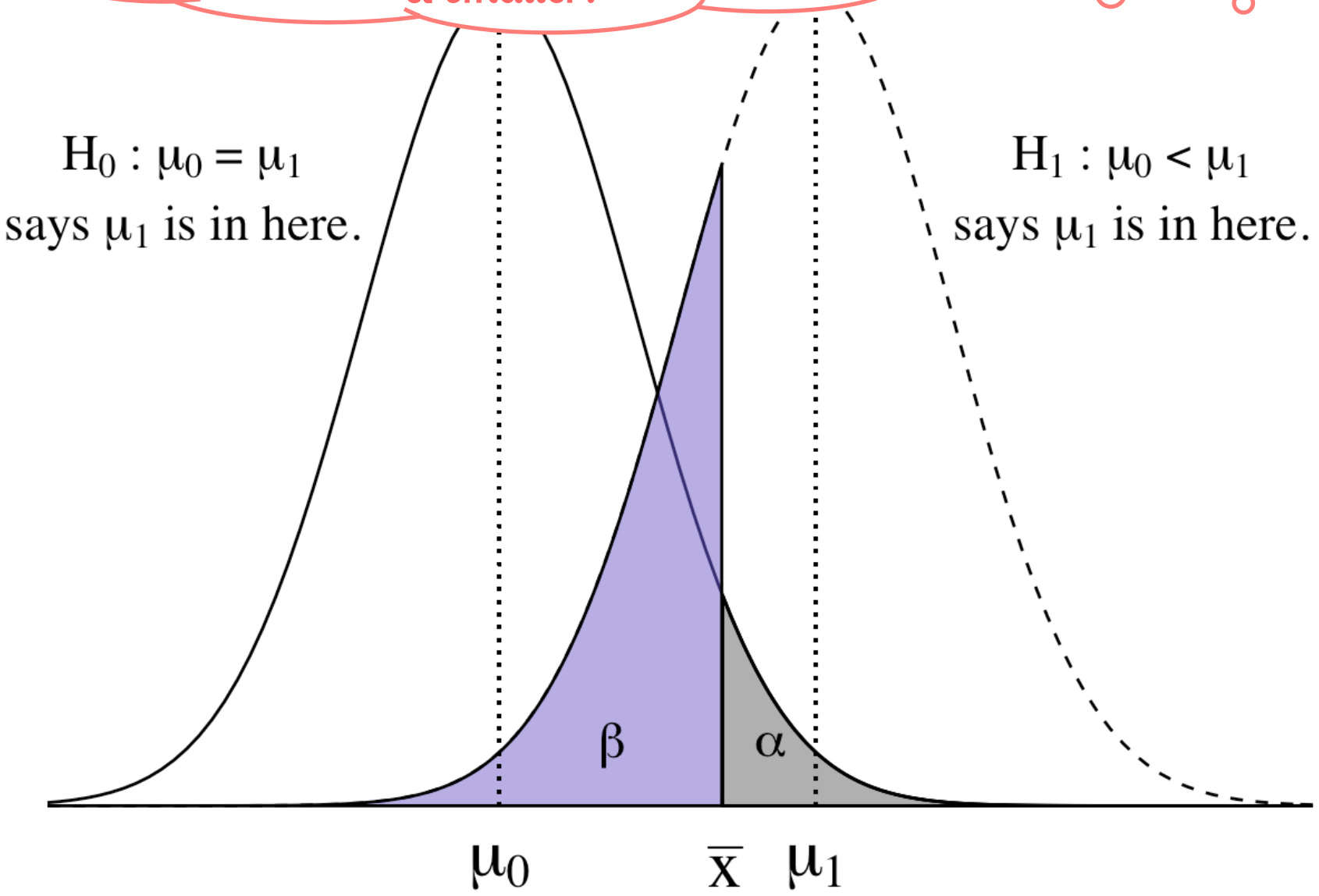
TYPE II ERROR: p(false negative) = β

For a given decision,

- P(false positive) = α = Type I error
- P(false negative) = β = Type II error

What will happen to β if I make α smaller?

What will happen to β if I make α smaller?



$$\alpha \uparrow \rightsquigarrow \beta \downarrow$$

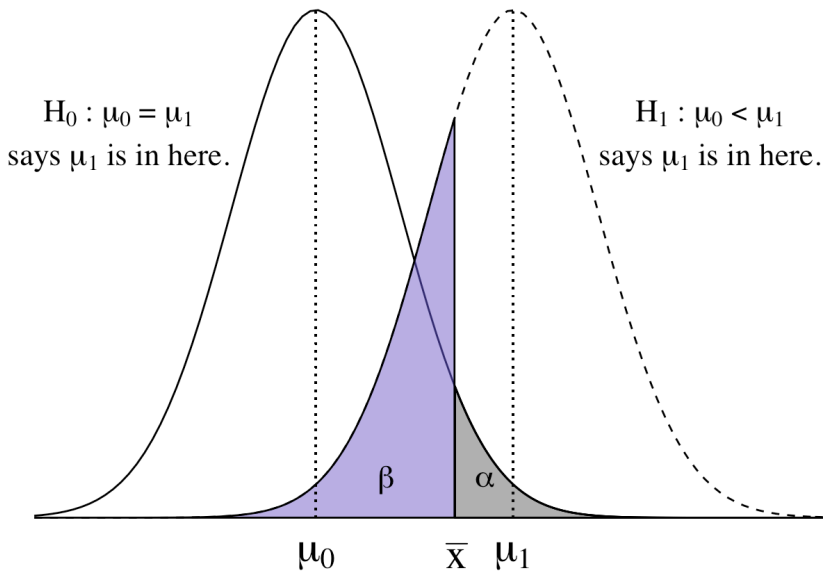
$$\alpha \downarrow \rightsquigarrow \beta \uparrow$$

TYPE II ERROR: p(false negative) = β

- Probability is defined as the chances of observing a t-statistic more extreme than the one we observed given the alternative t-distribution
- "How much more wrong could we have been?"

In our aspiring astronauts example…

```
qt(.95, 24)
[1] 1.710882
pt(qt(0.95, 24), 24, 25/13)
[1] 0.4115342 # beta
```



Type II error (β)
False negative

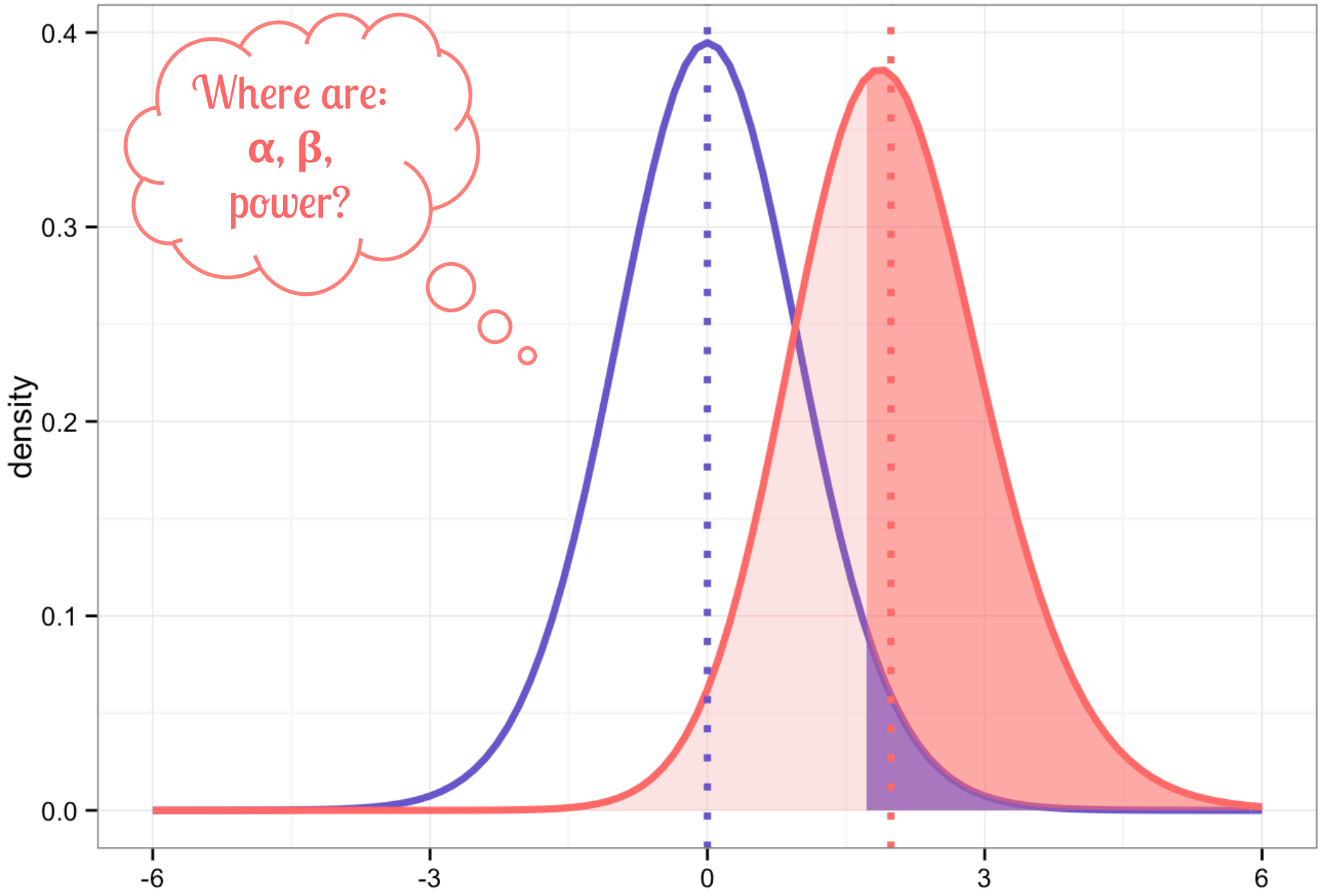
Decide:
"You're **not** smarter than average!"
Reality:
but you **are** actually

POWER: p(true positive) = 1 − β

- So power is the probability of exceeding the rejection point in this noncentral t distribution.

One-tailed test

```
qt(.95, 24) # t_critical, null dist
[1] 1.710882
pt(qt(0.95, 24), 24, 25/13) # beta
[1] 0.4115342
1 - pt(qt(0.95, 24), 24, 25/13) # power
[1] 0.5884658
```


Where are:
α, β,
power?
density
0.4
0.3
0.2
0.1
0.0
-6
-3
0
3
6

Power

```
power.t.test(n = 25, delta = 5, sd = 13, type = "one.sample",
alternative = c("one.sided"))
```

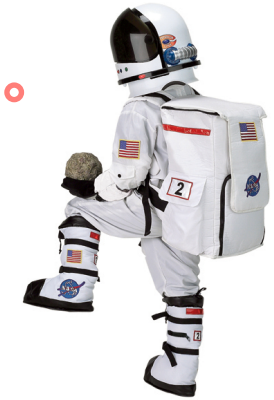
```
     One-sample t test power calculation

              n = 25
          delta = 5
             sd = 13
      sig.level = 0.05
          power = 0.5884658
    alternative = one.sided
```

One-tailed test

Good for:"post-mortem" power analysis

delta here is confusing: it is neither the ncp nor the effect size- it is the raw difference between means you wish to detect



Factors that affect Power (1 – β)

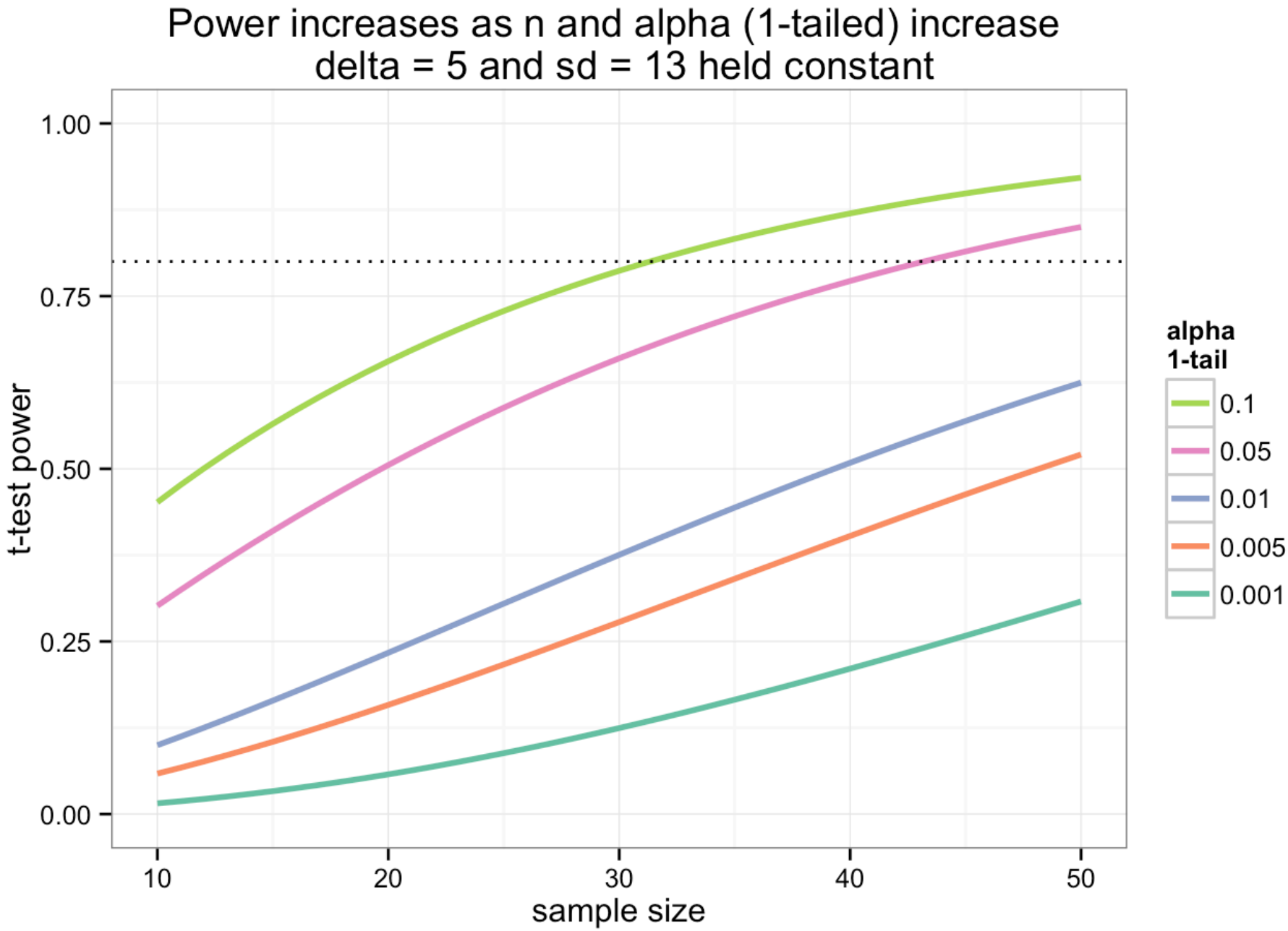
- Sample size
  - Increased n reduces SE_{mean}
- Level of significance
  - Power increases as **α** increases
- Reliability of your measure
  - Classical test theory:
    total variance = true score variance + error variance
- Effect size (sds between the true mean & the one hypothesized in H_0; $\mu - \mu_0$)
- Population variance
  - Decreased variance reduces SE_{mean}

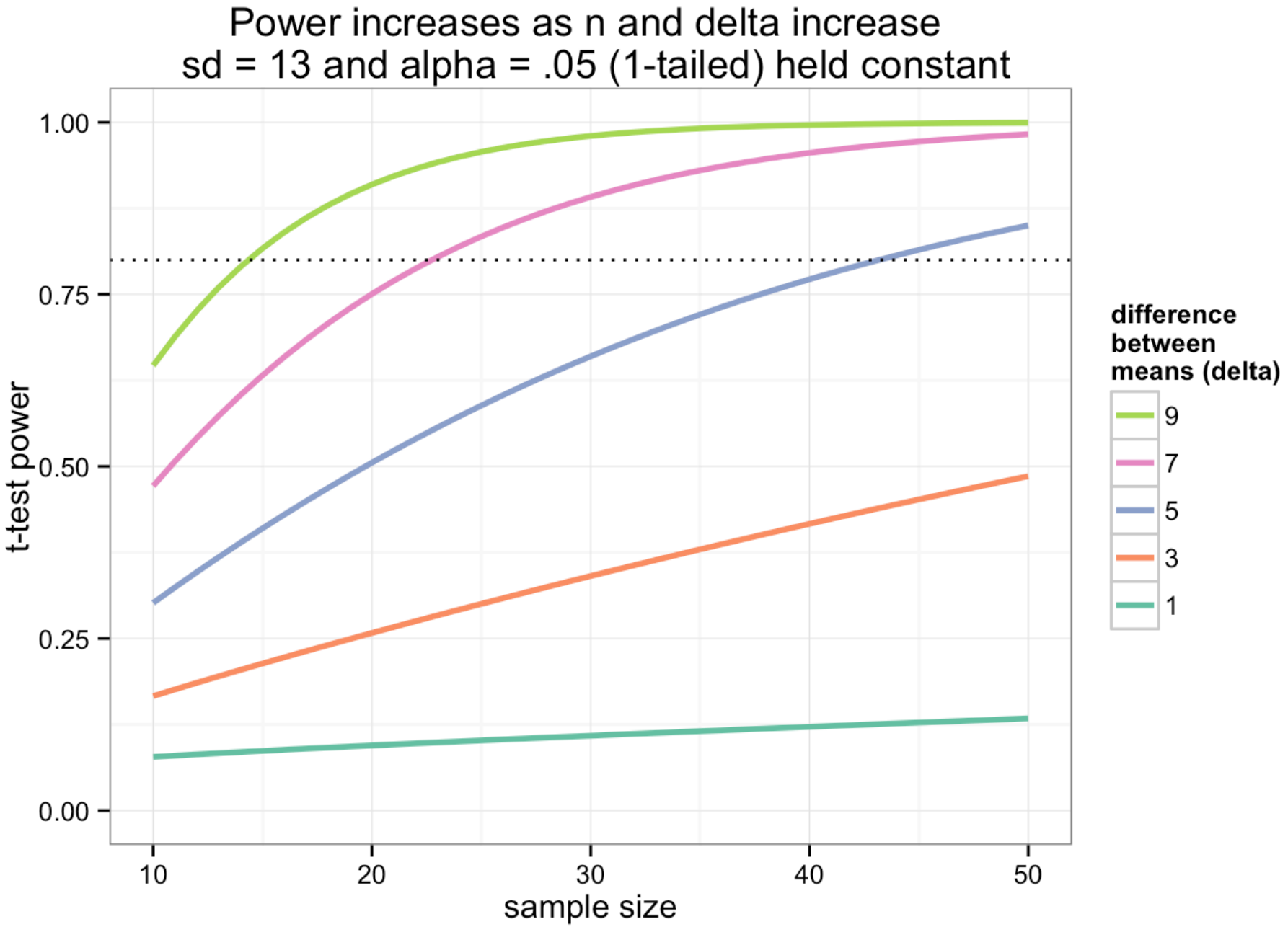
Power

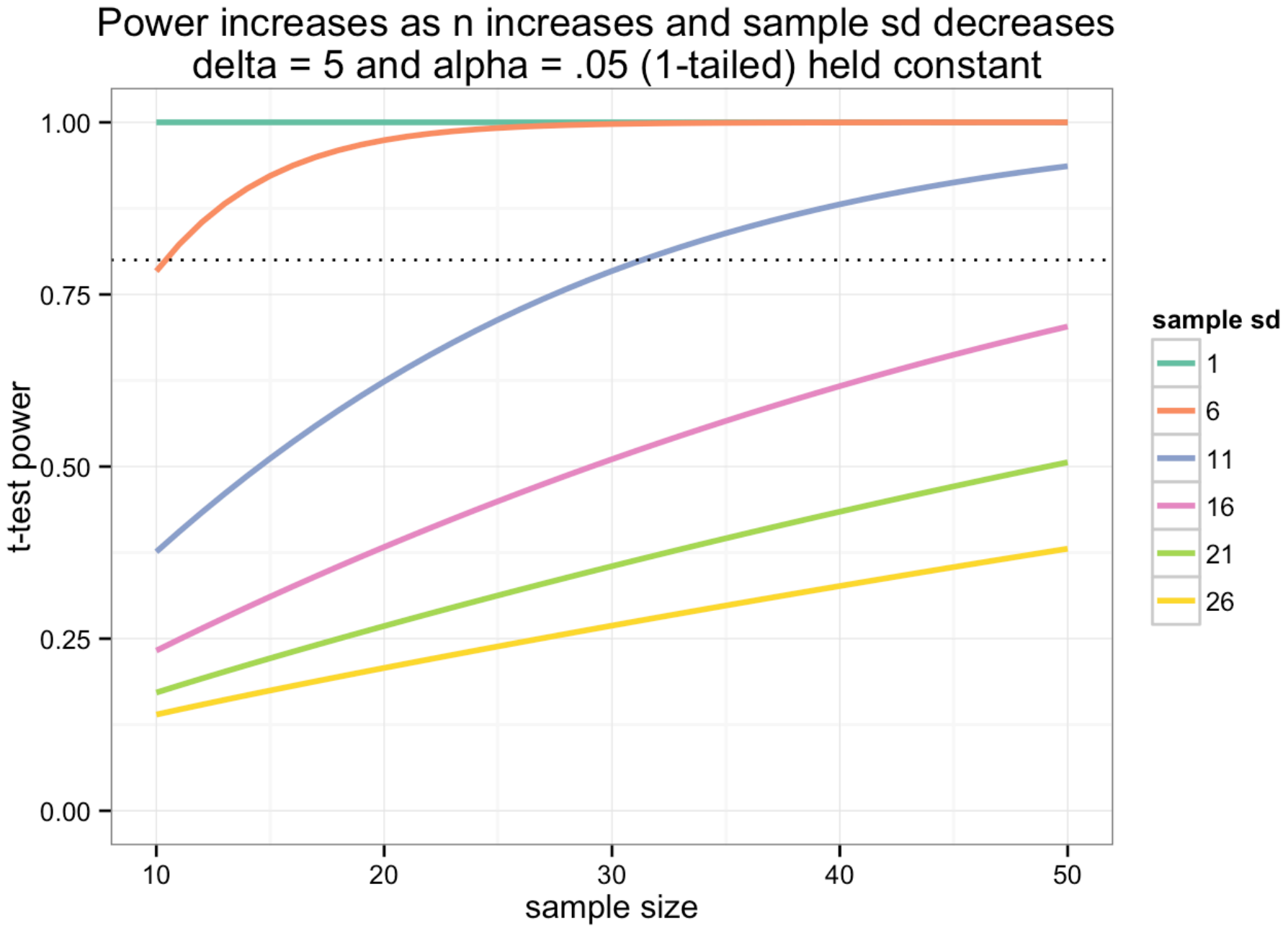
For a given decision,

- P(false positive) = α = Type I error
- P(true positive) = $1 - \beta$ = power

What will happen to $1 - \beta$ (power) if I make α smaller?


Power increases as n and alpha (1-tailed) increase
delta = 5 and sd = 13 held constant
1.00
0.75
0.50
0.25
0.00
t-test power
10
20
30
40
50
sample size
alpha
1-tail
0.1
0.05
0.01
0.005
0.001


Power increases as n and delta increase
sd = 13 and alpha = .05 (1-tailed) held constant
t-test power
1.00
0.75
0.50
0.25
0.00
10
20
30
40
50
sample size
difference between means (delta)
9
7
5
3
1


Power increases as n increases and sample sd decreases
delta = 5 and alpha = .05 (1-tailed) held constant
t-test power
1.00
0.75
0.50
0.25
0.00
10
20
30
40
50
sample size
sample sd
1
6
11
16
21
26

How large would our "n" have to be?

To detect:

- $\Delta = 5$
- $1 - \beta = .80$
- $\alpha = .05$

One-tailed test

With s.d. = 13

Good for: a priori sample size determination

Sample size determination

```
power.t.test(n =      , delta =   , sd =   , sig.level =    , power
=    , type =                  , alternative =                  )
```

Sample size determination

```
power.t.test(n = NULL, delta = 5, sd = 13, sig.level = .05, power
= .80, type = "one.sample", alternative = c("one.sided"))
```

```
     One-sample t test power calculation

              n =
          delta = 5
             sd = 13
      sig.level = 0.05
          power = 0.8
    alternative = one.sided
```

Sample size determination

```
power.t.test(n = NULL, delta = 5, sd = 13, sig.level = .05, power
= .80, type = "one.sample", alternative = c("one.sided"))
```

```
     One-sample t test power calculation

              n = 43.17957
          delta = 5
             sd = 15
      sig.level = 0.05
          power = 0.8
    alternative = one.sided
```

How small of an effect could we detect…

If we knew we could get:

- n = 100 high school girls who are aspiring astronauts

And we wanted:

- $1 - \beta = .80$
- $\alpha = .05$

One-tailed test

With s.d. = 13

Sample size determination

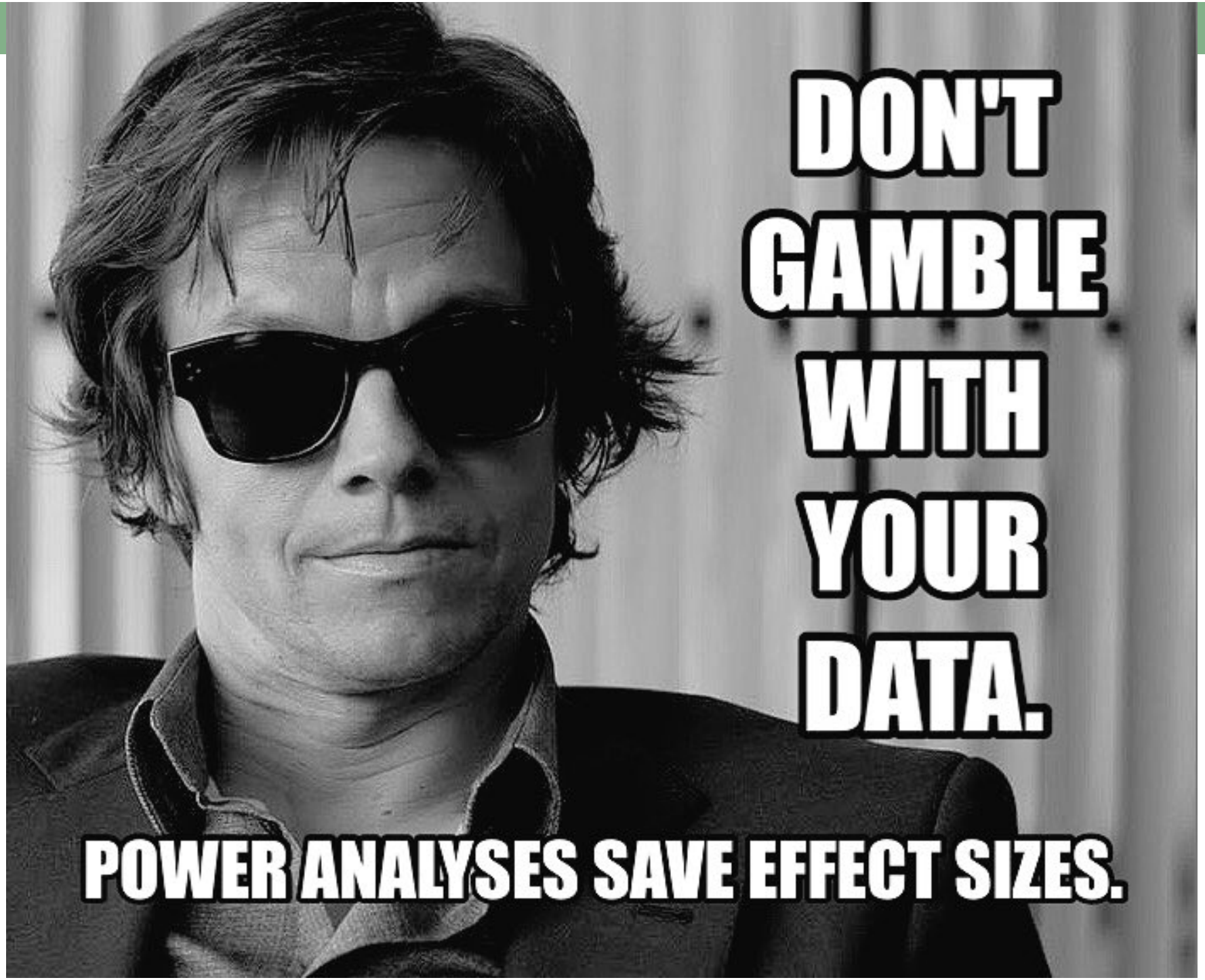
```
power.t.test(n =    , delta =    , sd =   , sig.level =   ,
power =   , type =                , alternative =                )
```

Effect size determination

```
power.t.test(n = 100, delta = NULL, sd = 13, sig.level = .05,
power = .80, type = "one.sample", alternative = c("one.sided"))
```

```
     One-sample t test power calculation

              n = 100
          delta = 3.254735
             sd = 13
      sig.level = 0.05
          power = 0.8
    alternative = one.sided
```


DON'T GAMBLE WITH YOUR DATA.
POWER ANALYSES SAVE EFFECT SIZES.