Math 530/630: CM 4.5

Errors, Effect Size and Power

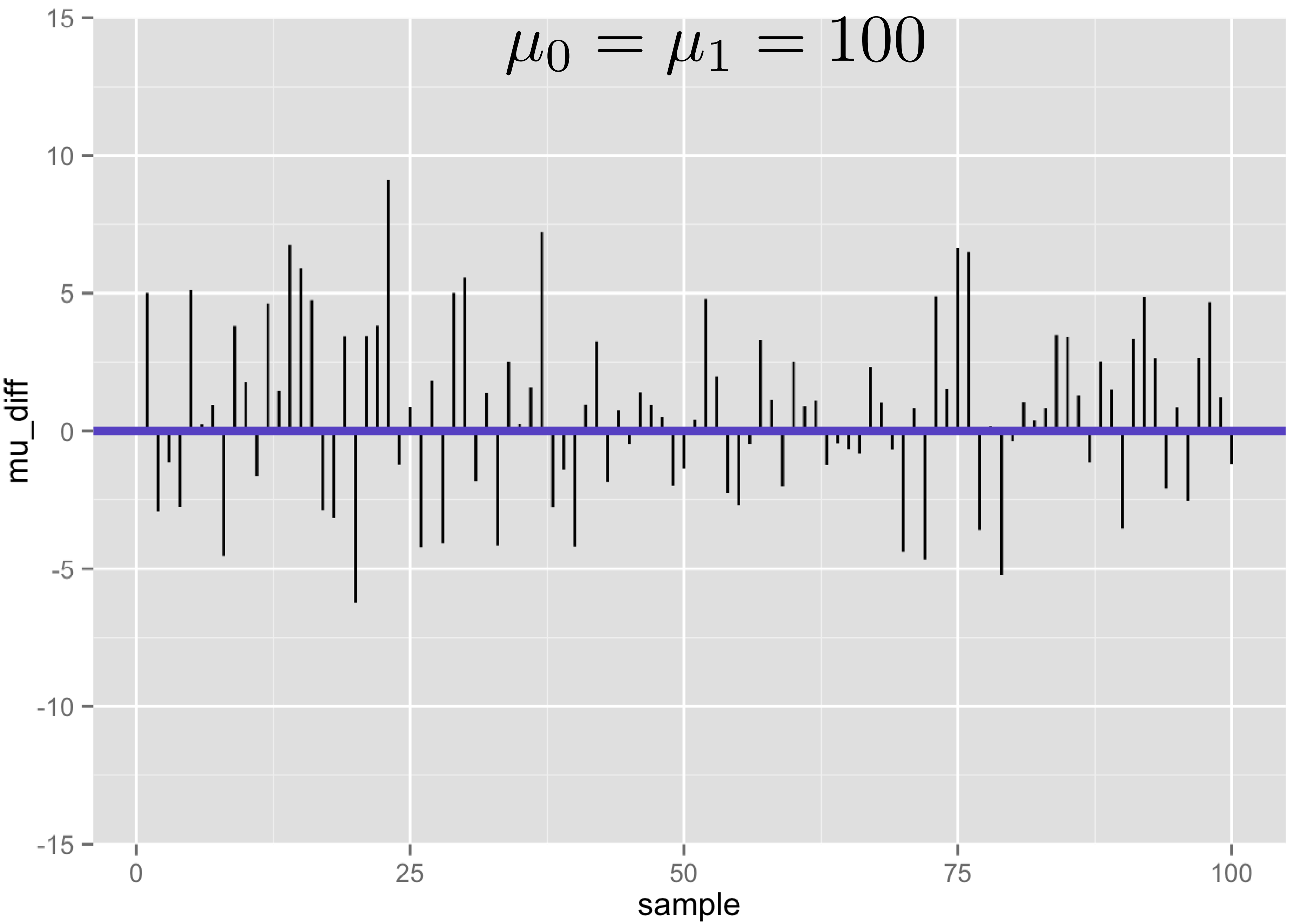
When the null hypothesis is true

$$\mu_0 = \mu_1 = 100$$

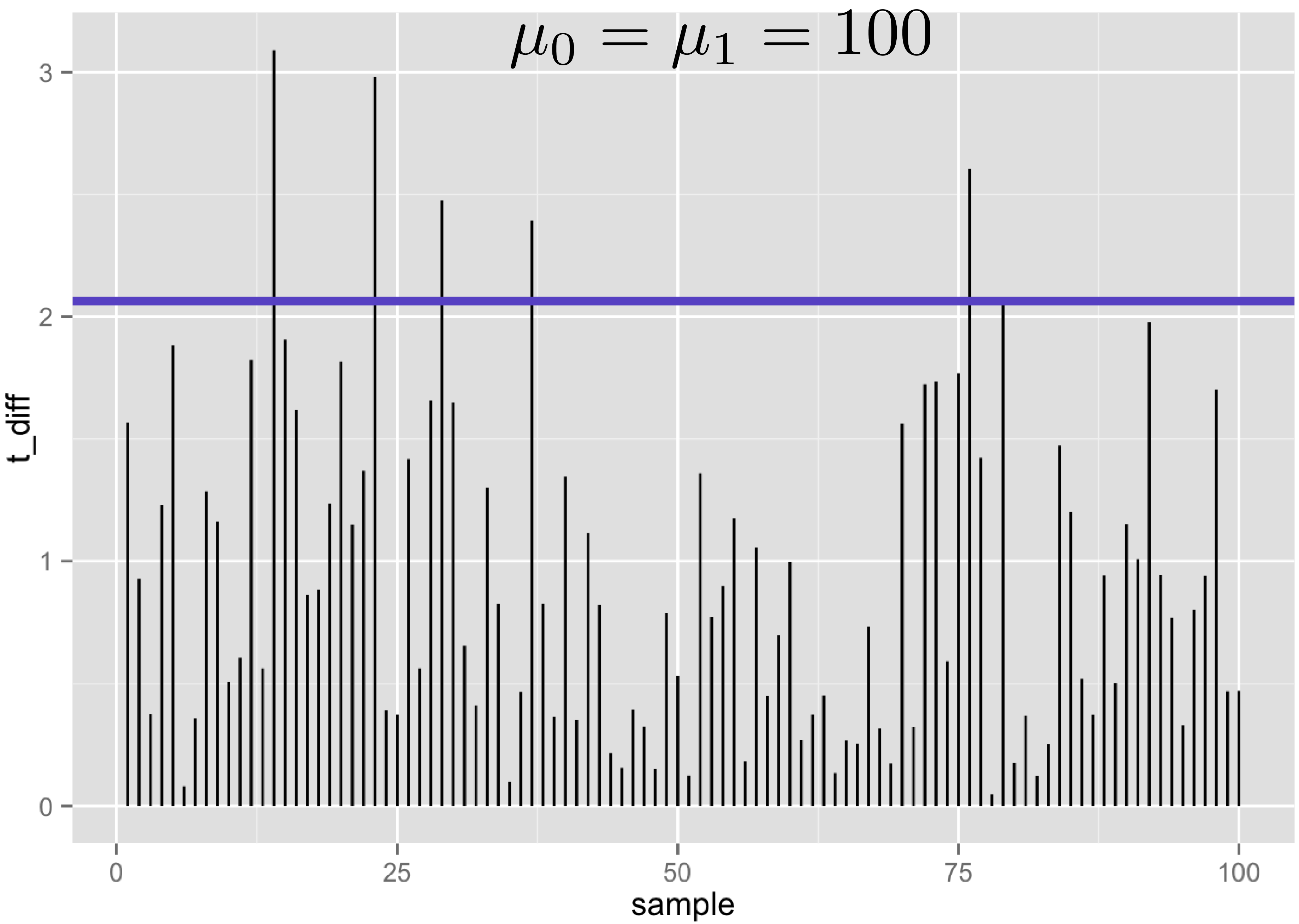
1-sample t-test

n = 25

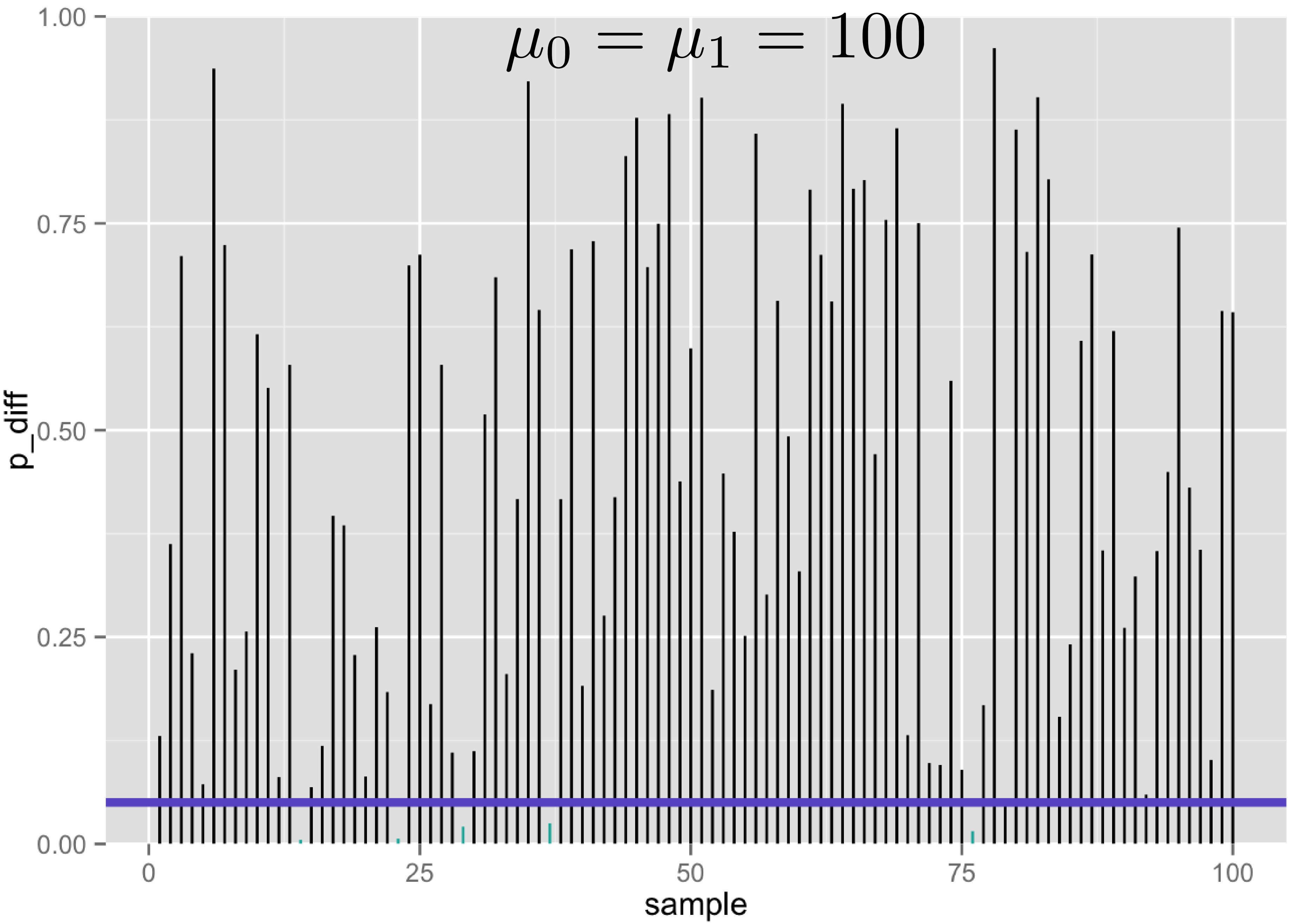
Differences between population mean and sample mean (100 samples) when null is **true**



100 t-statistics (absolute value) when null is **true**: 5% false positives using t-test (**α** = .05)



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



Confusion matrix

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

But… what if we are wrong??

Two ways we can be wrong…

Type 1 error (α)
False positive

Type II error (β)
False negative

One way we can be wrong…

Type 1 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

In our aspiring astronauts example…

Type 1 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

The boy who cried wolf caused both Type I and Type II errors, in that order.

First, everyone believed there was a wolf, when there actually was not a wolf.
Next, they believed there was no wolf, when there actually was a one.
Substitute "effect" for "wolf" - done!

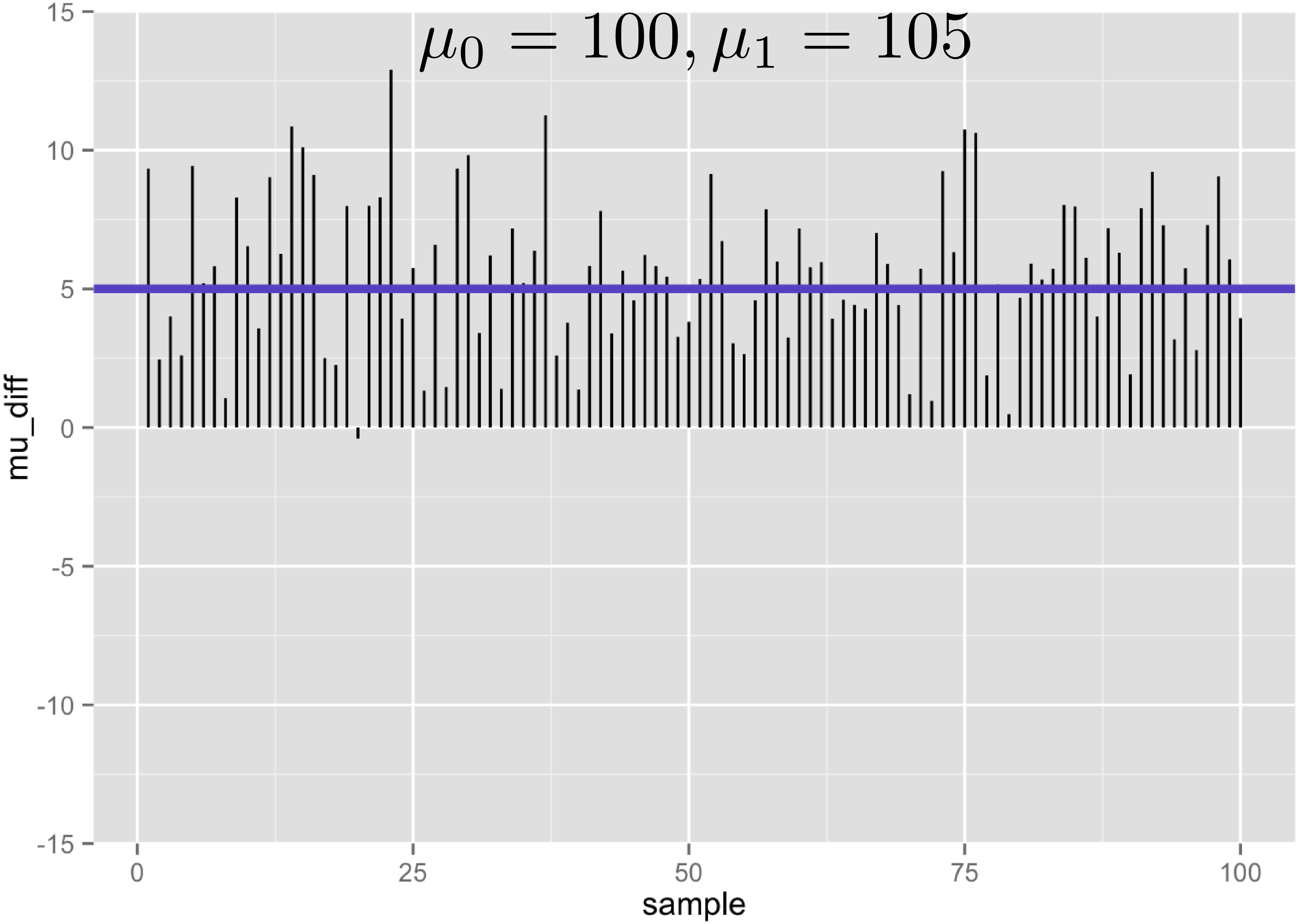
Thanks to @danolner and @JustinWolfers for this o

What if:
the null hypothesis is FALSE?

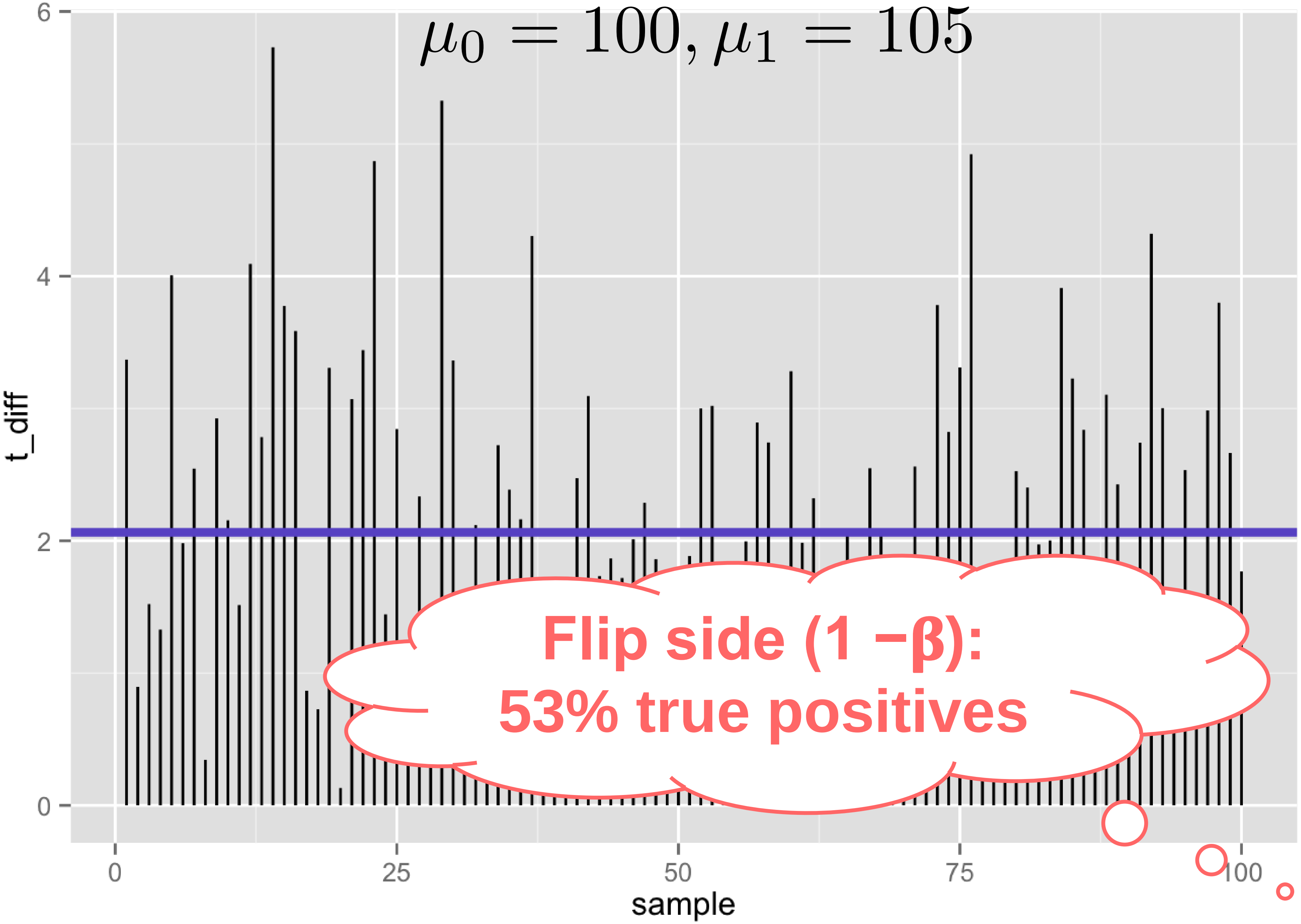
$$\mu_0 = 100, \mu_1 = 105$$

1-sample t-test
n = 25

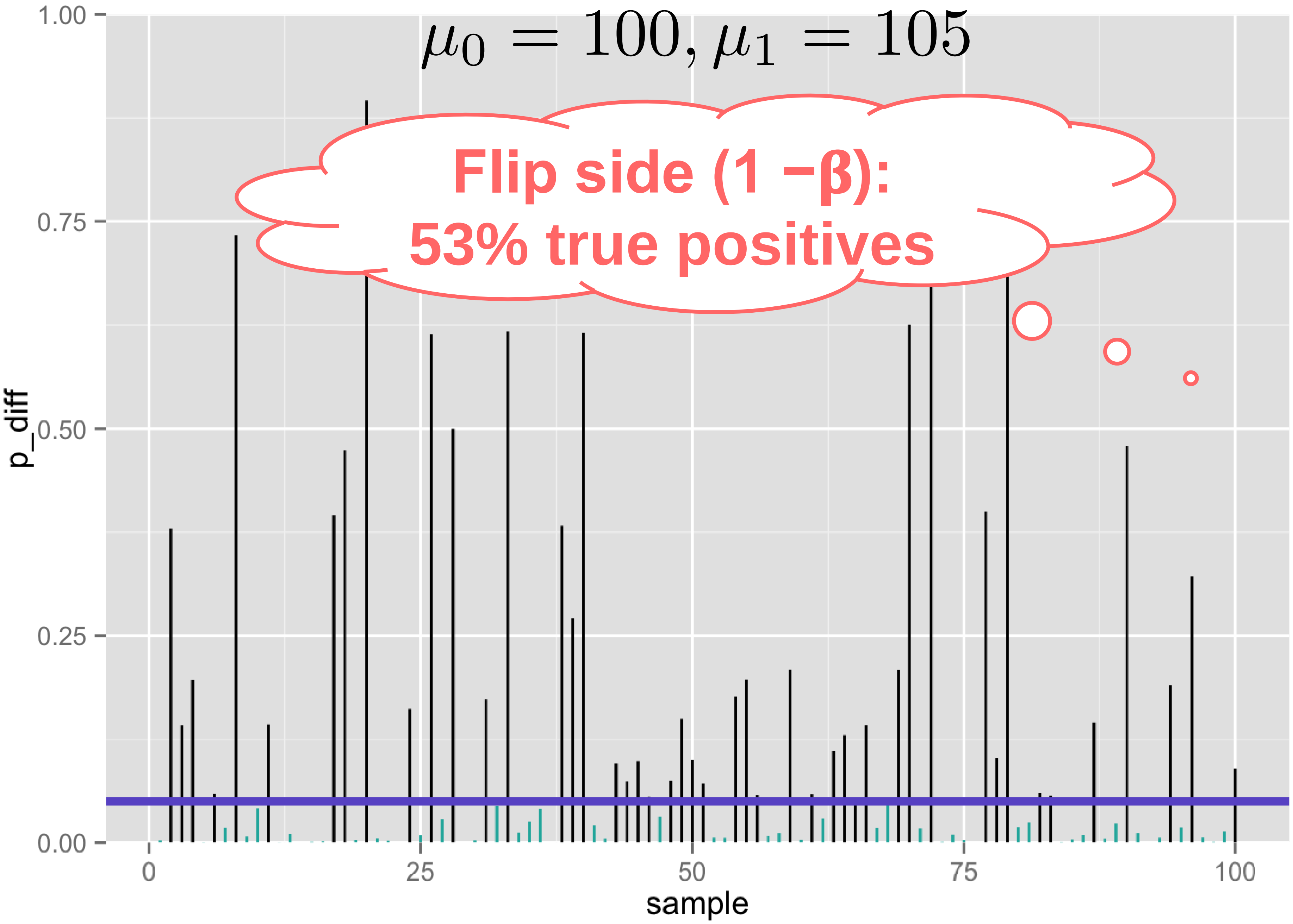
Differences between population mean and sample mean (100 samples) when null is **false**



100 t-statistics (absolute value) when null is false: 47% false negatives using t-test (α = .05)



100 p-values when null is false: 47% false negatives using t-test (α = .05)



Type II errors

- The one(s) that got away…

≈ 50% true positives seems low…

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

Power

- Power is the probability of correctly rejecting a false null hypothesis (i.e., true positive)

$$1 - b = P(\mathit{reject}\ H_0 \,|\, H_1\ \mathit{true})$$

- Suppose we wish to test (α = .025, 1-tailed):
  - H_0: $\mu \leq 100$
  - H_1: $\mu > 100$
- Let’s use the same sample of n=25 aspiring astronauts (we know σ = 15)
- First, what is β?

$$b = P(\mathit{fail\ to\ reject}\ H_0 \,|\, H_1\ \mathit{true})$$

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
```

One-tailed test

If the null is true…

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

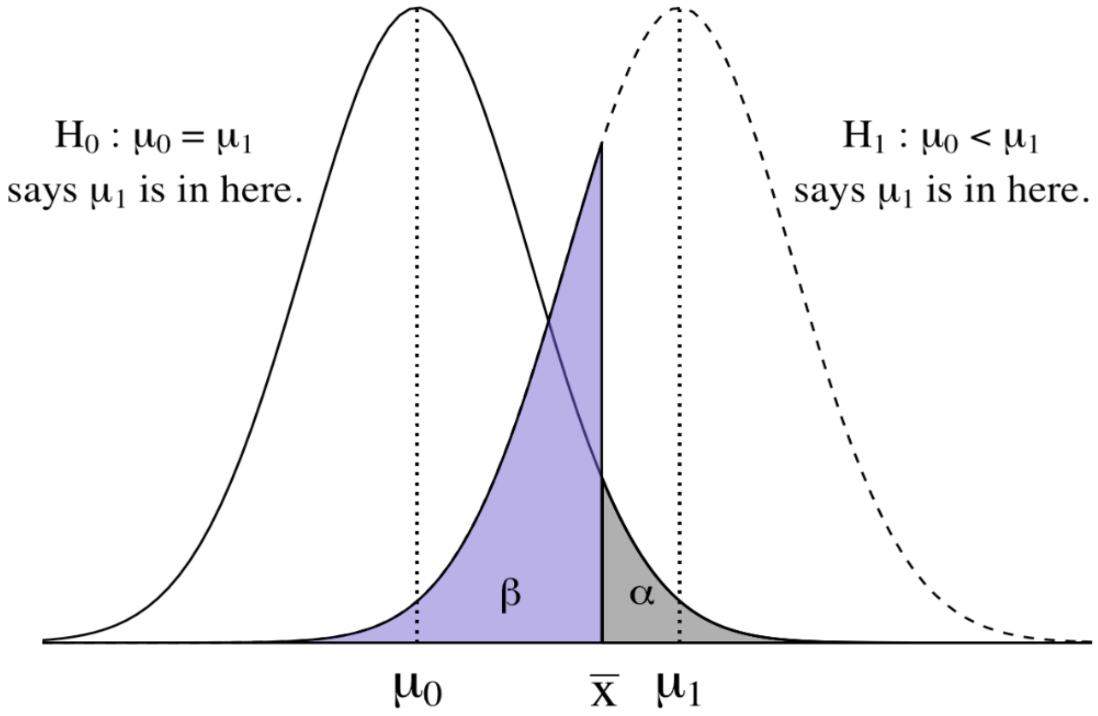
One-tailed test

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

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

Dueling distributions

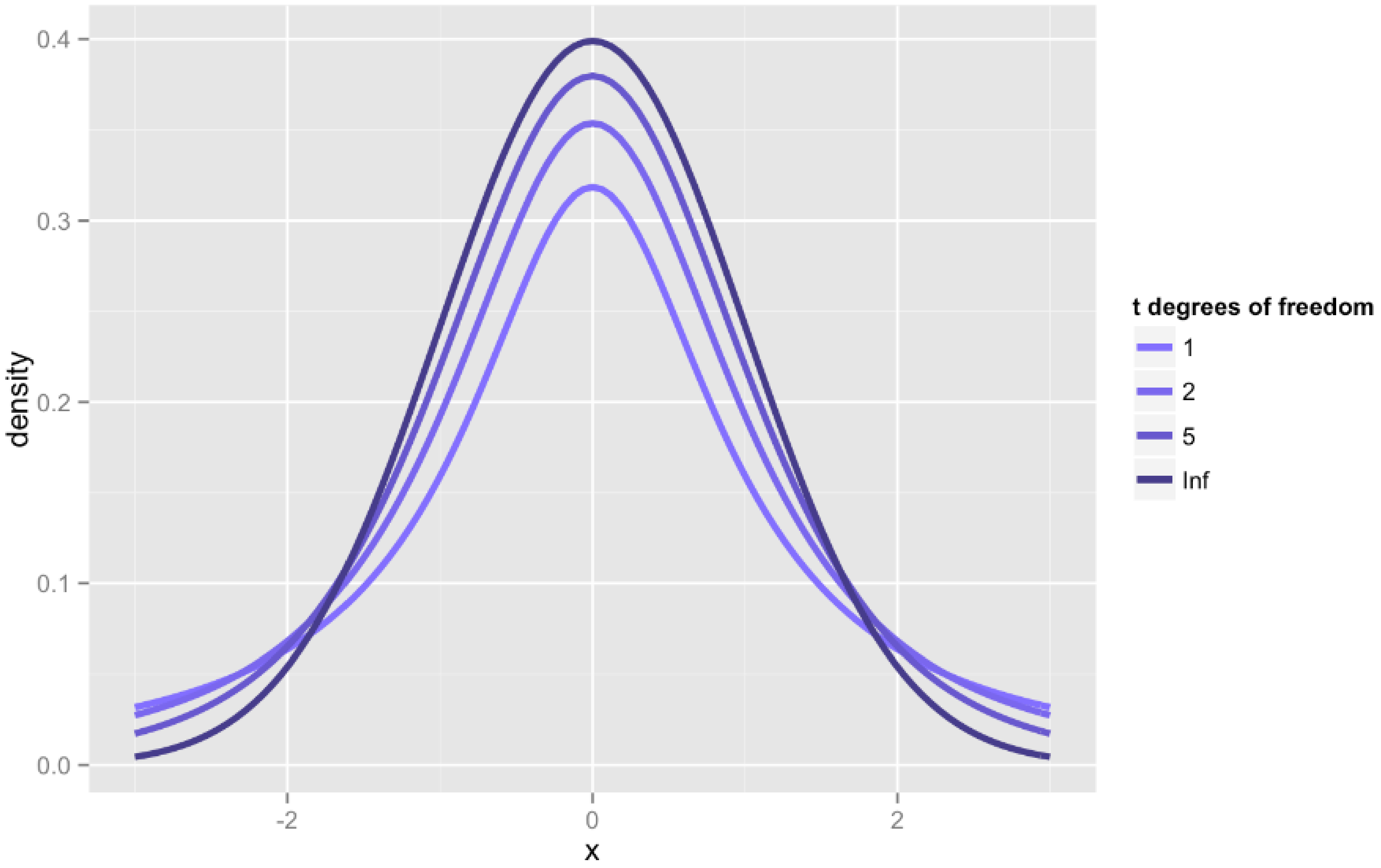
- In a one-tailed test, we have two dueling hypothetical distributions:
  - The null distribution, centered at μ_0
  - The alternative distribution, centered around some specific or unspecified other mean (higher or lower?) μ_1



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



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

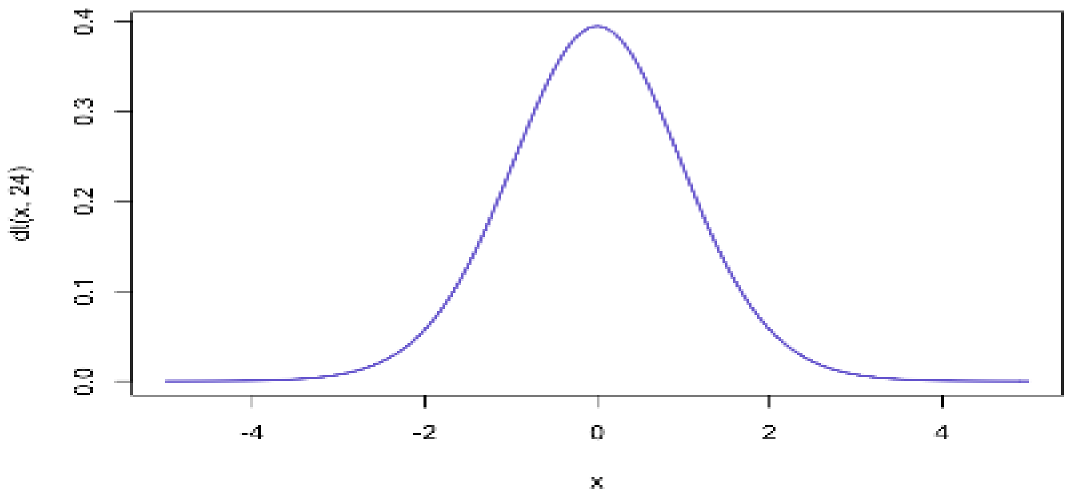
```
dt(x, df, ncp, log = FALSE)
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)
rt(n, df, ncp)
```

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 (v= degrees of freedom)

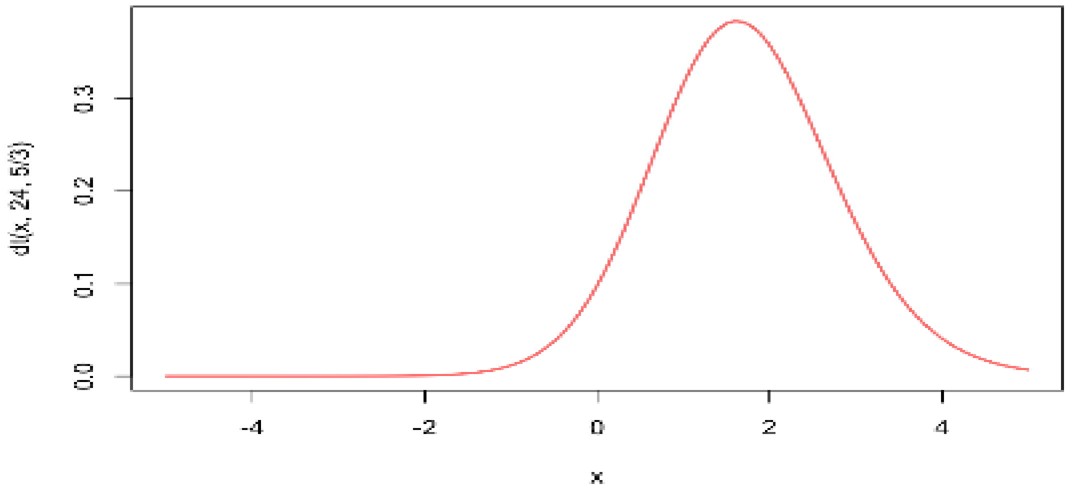
- $\boldsymbol{\delta} = 0$
- Mean = 0
- Variance slightly > than N(0, 1)
- Kurtosis (biased) is > 3
- Symmetric

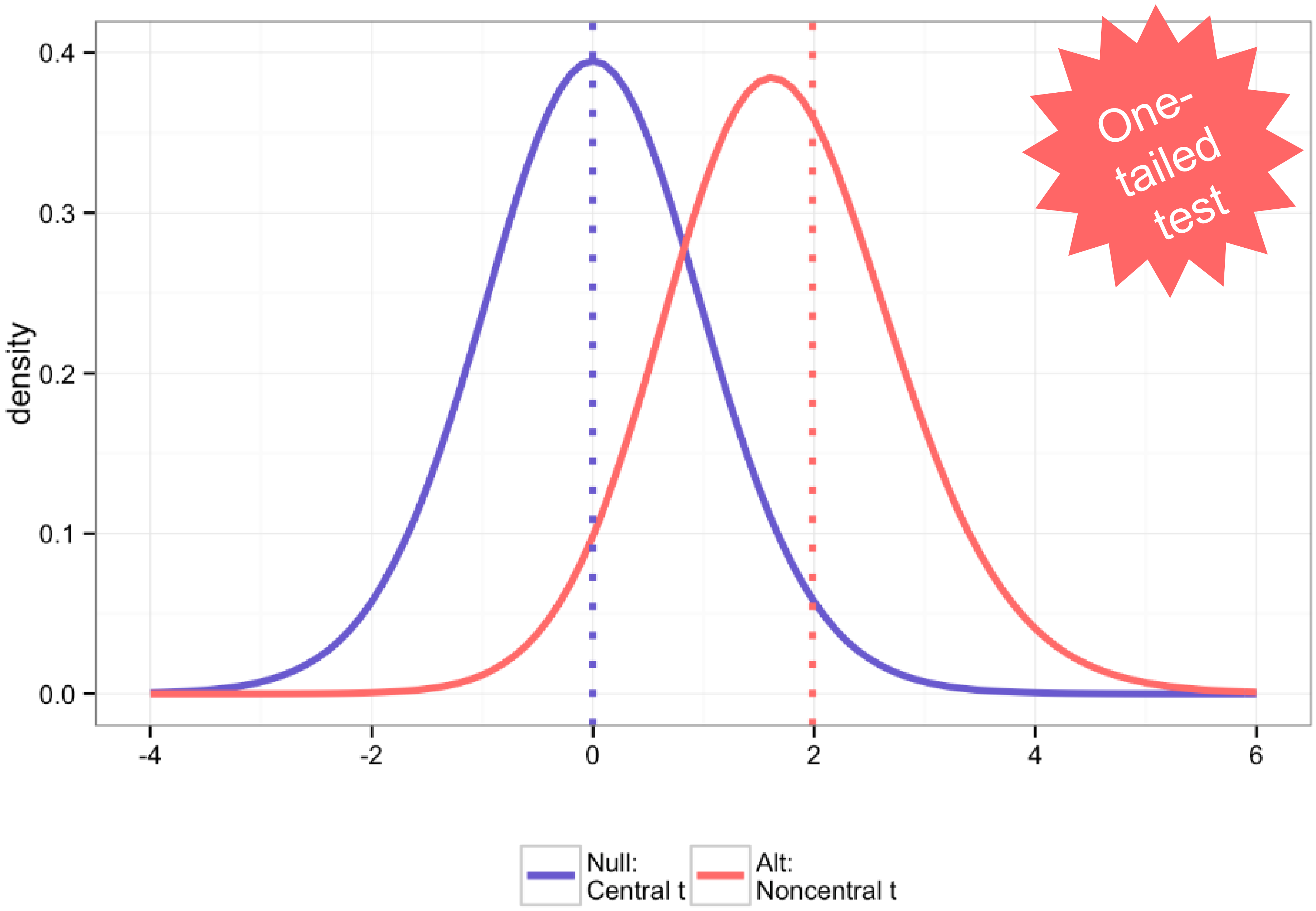


Noncentral t distribution (v= degrees of freedom)

- **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:

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

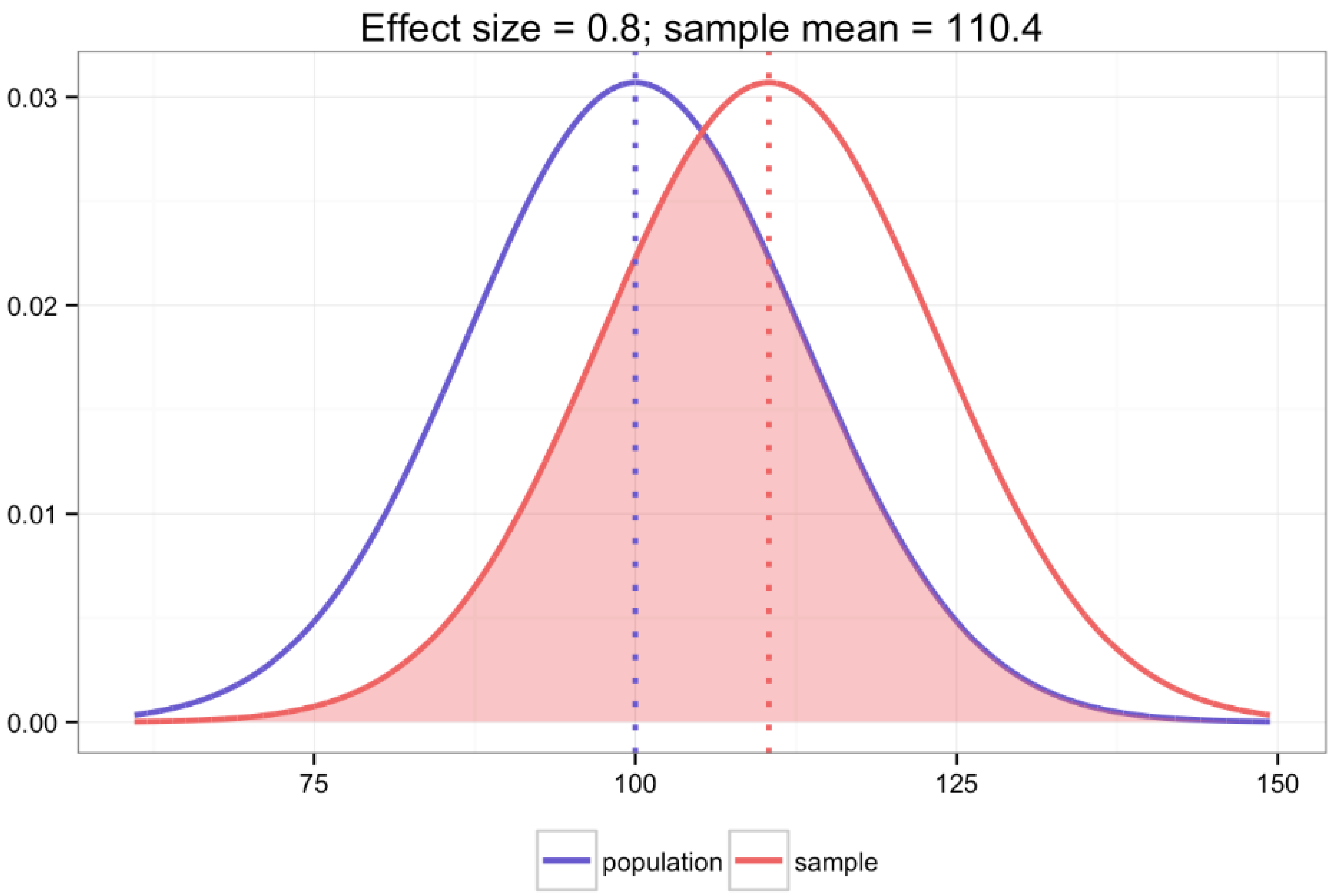
- Where E_s is the standardized measure of effect size…how do we calculate this?

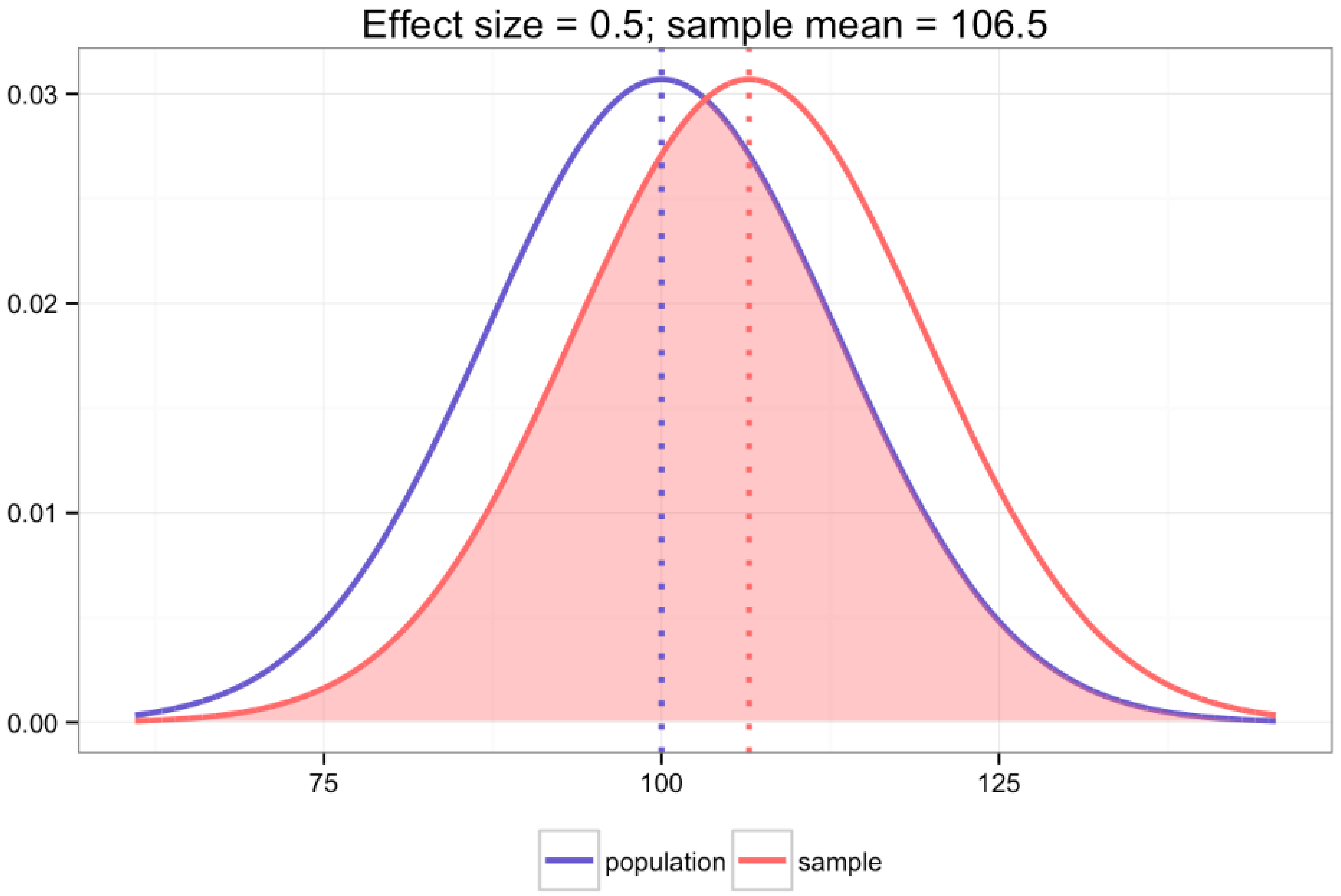
Effect size for *t-test*

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

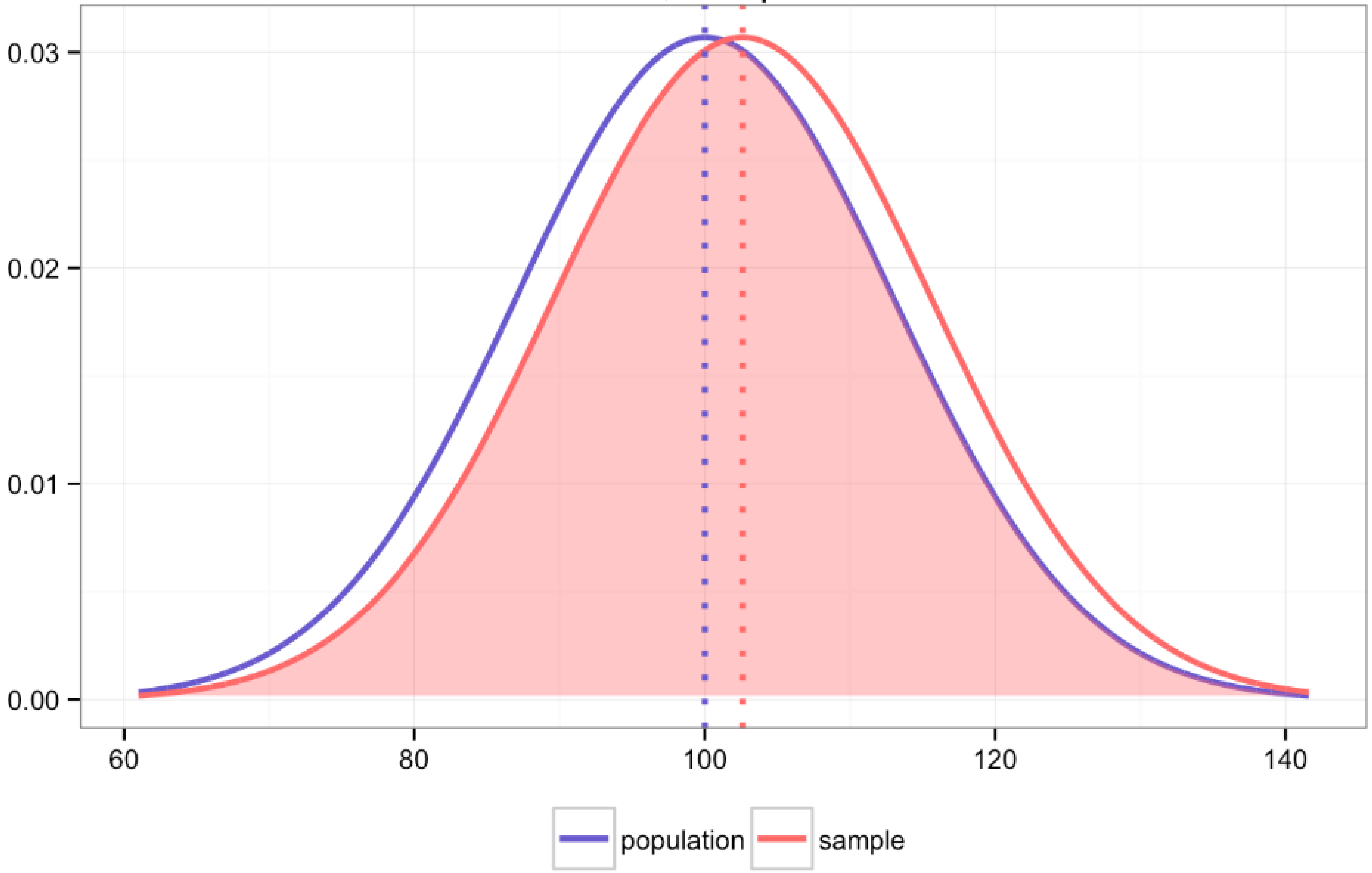
$$E_s = \frac{\mu_1 - \mu_0}{s}$$

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



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

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

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

Use alternative distribution
to solve for **β**

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

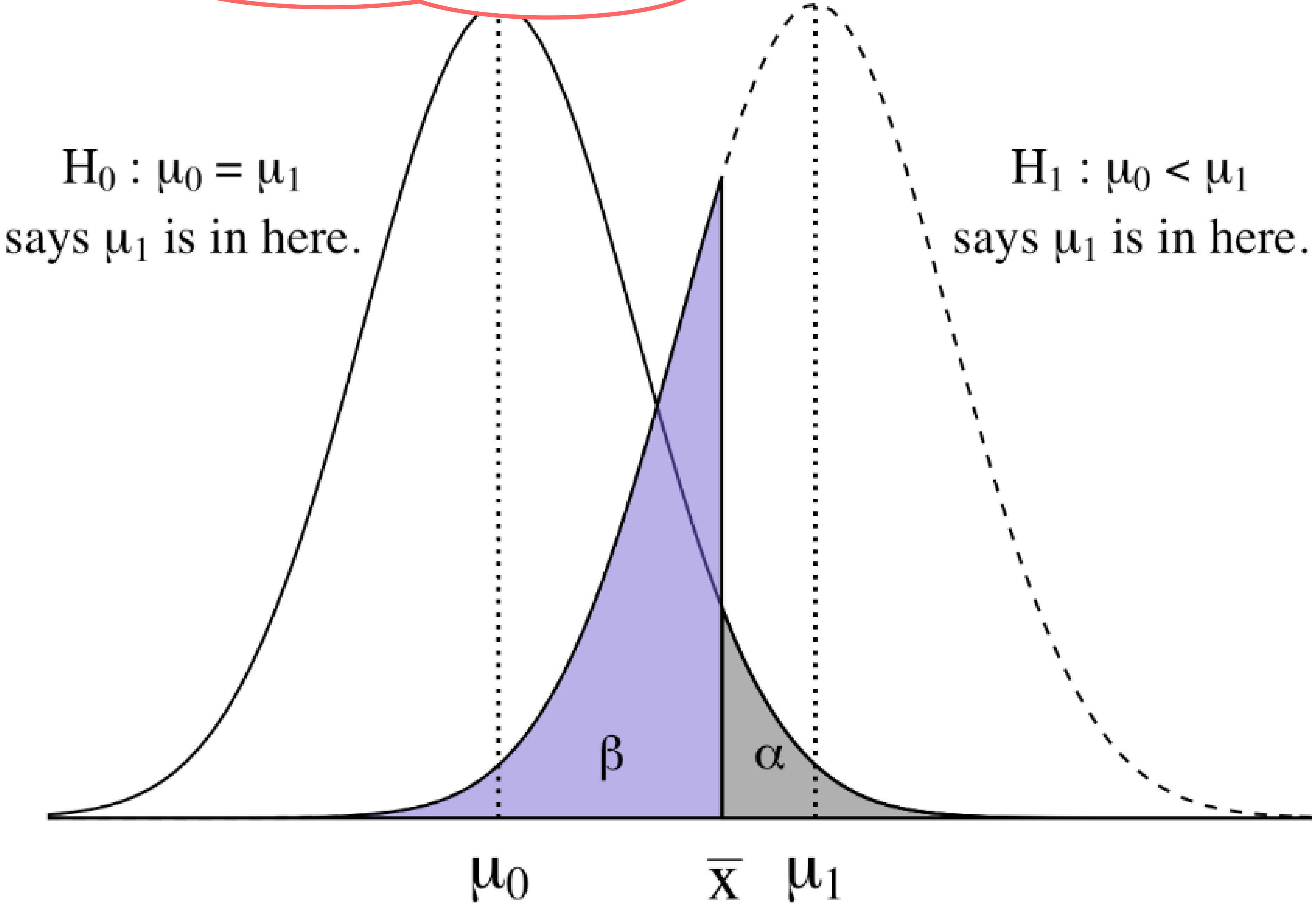
```
power.t.test(n = 25, delta = 5, sd = 13, type = "one.sample", alternative =
c("one.sided"))

     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

What will happen to **β** if I make **α** smaller?

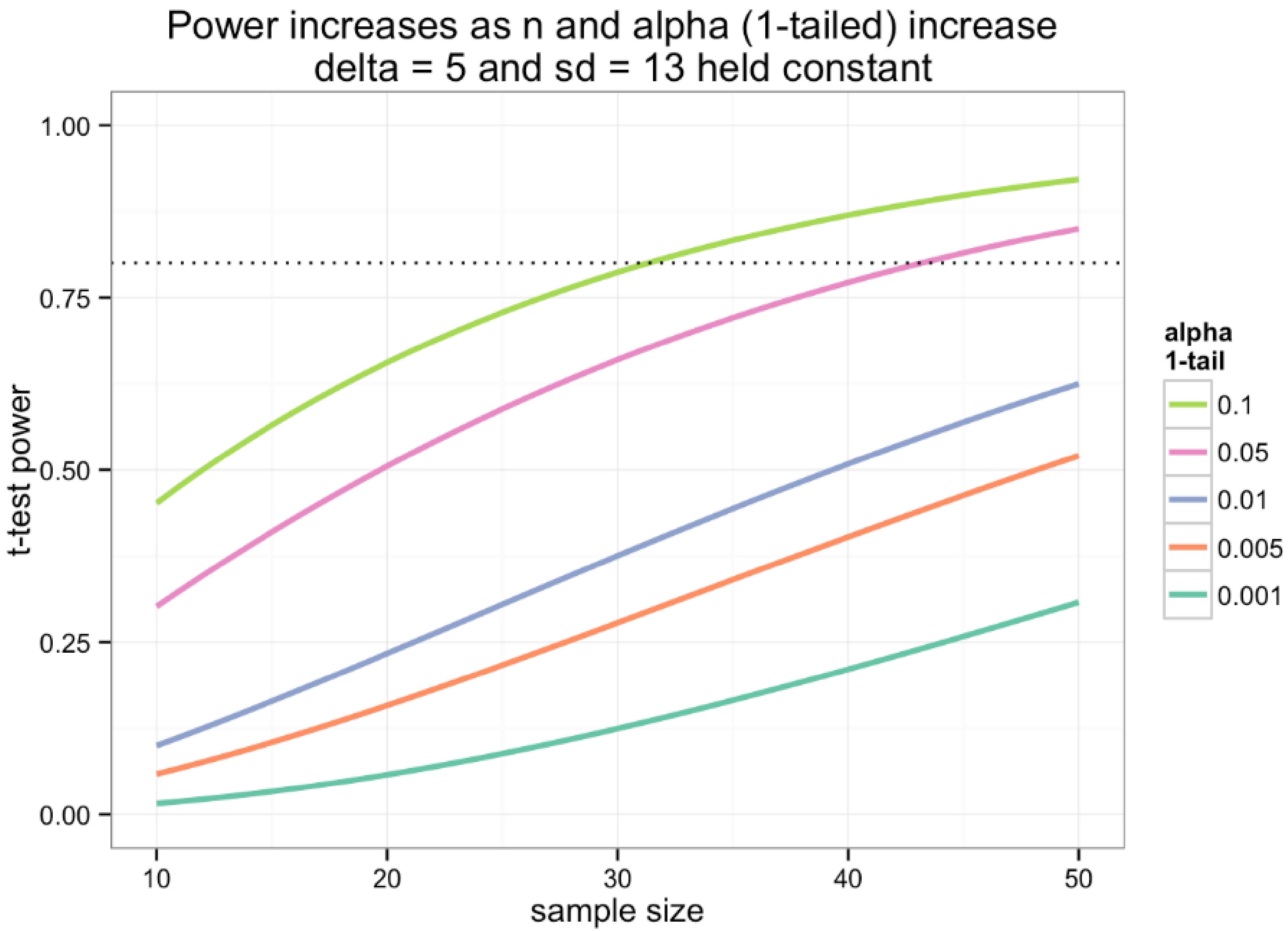


$$\alpha\uparrow \rightsquigarrow \beta\downarrow$$

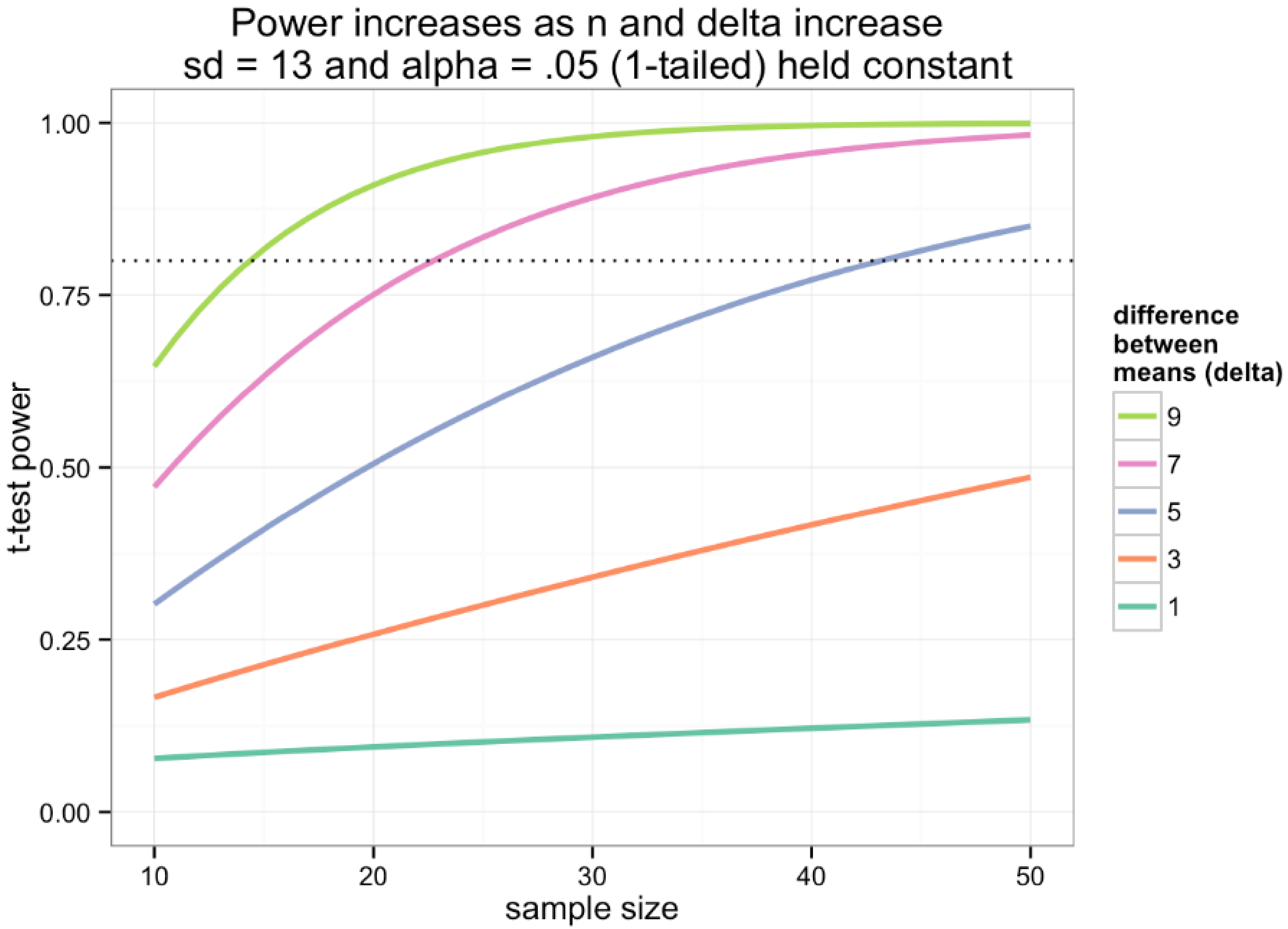
$$\alpha\downarrow \rightsquigarrow \beta\uparrow$$

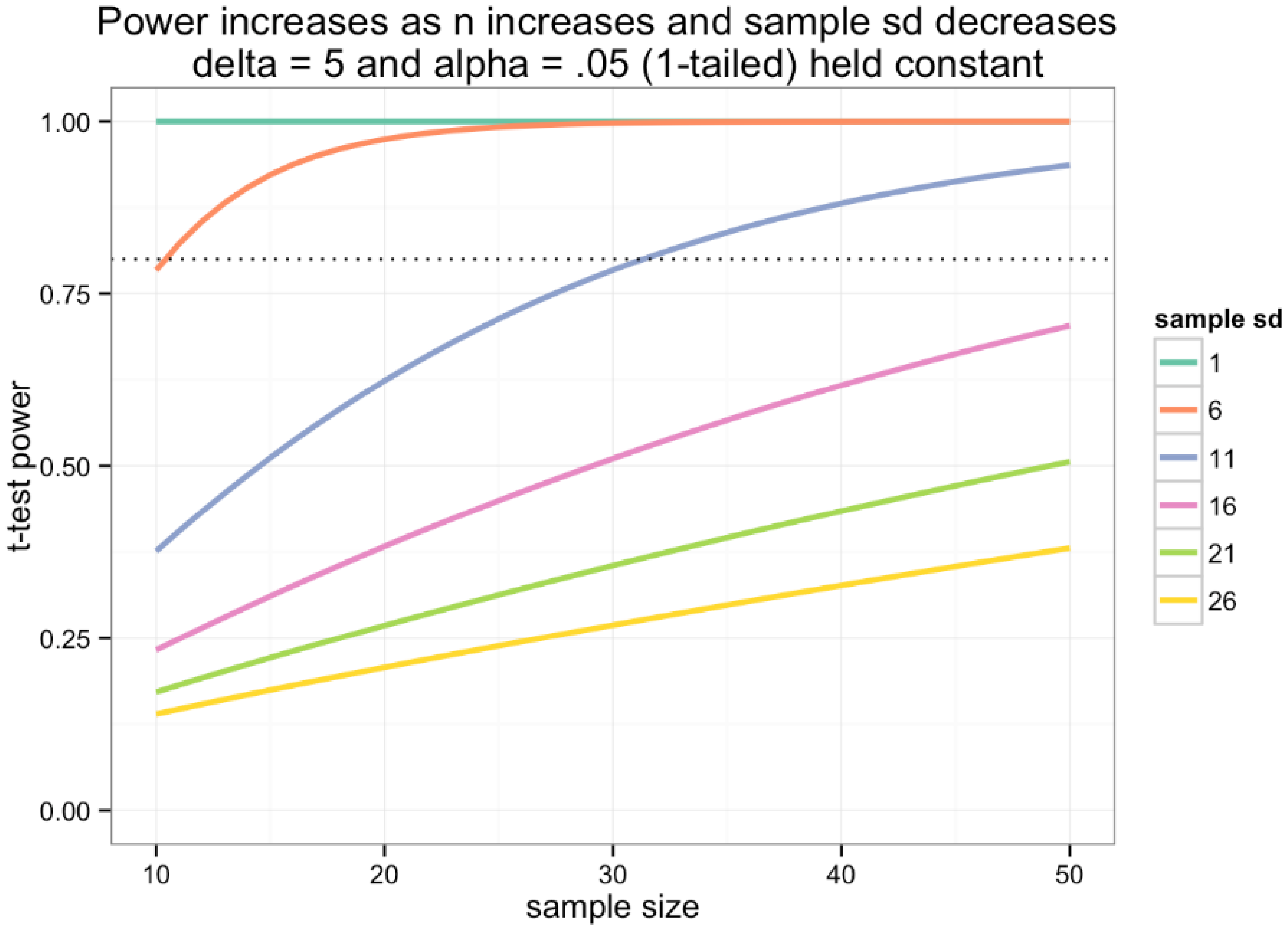
Factors that affect Power ($1 - \beta$)

- Sample size
  - Increased n reduces SE_{mean}
- Level of significance
  - Power increases as $\boldsymbol{\alpha}$ 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 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
1.00
0.75
0.50
0.25
0.00
t-test power
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:

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

One-tailed test

With s.d. = 13

Good for: a priori sample size determination

Sample size determination

```
power.t.test(n = [redacted], delta = [redacted] sd = [redacted], sig.level = [redacted], power = [redacted], type = 
[redacted], alternative = [redacted]
```

One-tailed test

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 = 13
      sig.level = 0.05
          power = 0.8
    alternative = one.sided
```

One-tailed test

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
```

One-tailed test

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:

- $1 - \boldsymbol{\beta} = .80$
- $\boldsymbol{\alpha} = .05$

One-tailed test

With s.d. = 13

Effect size determination

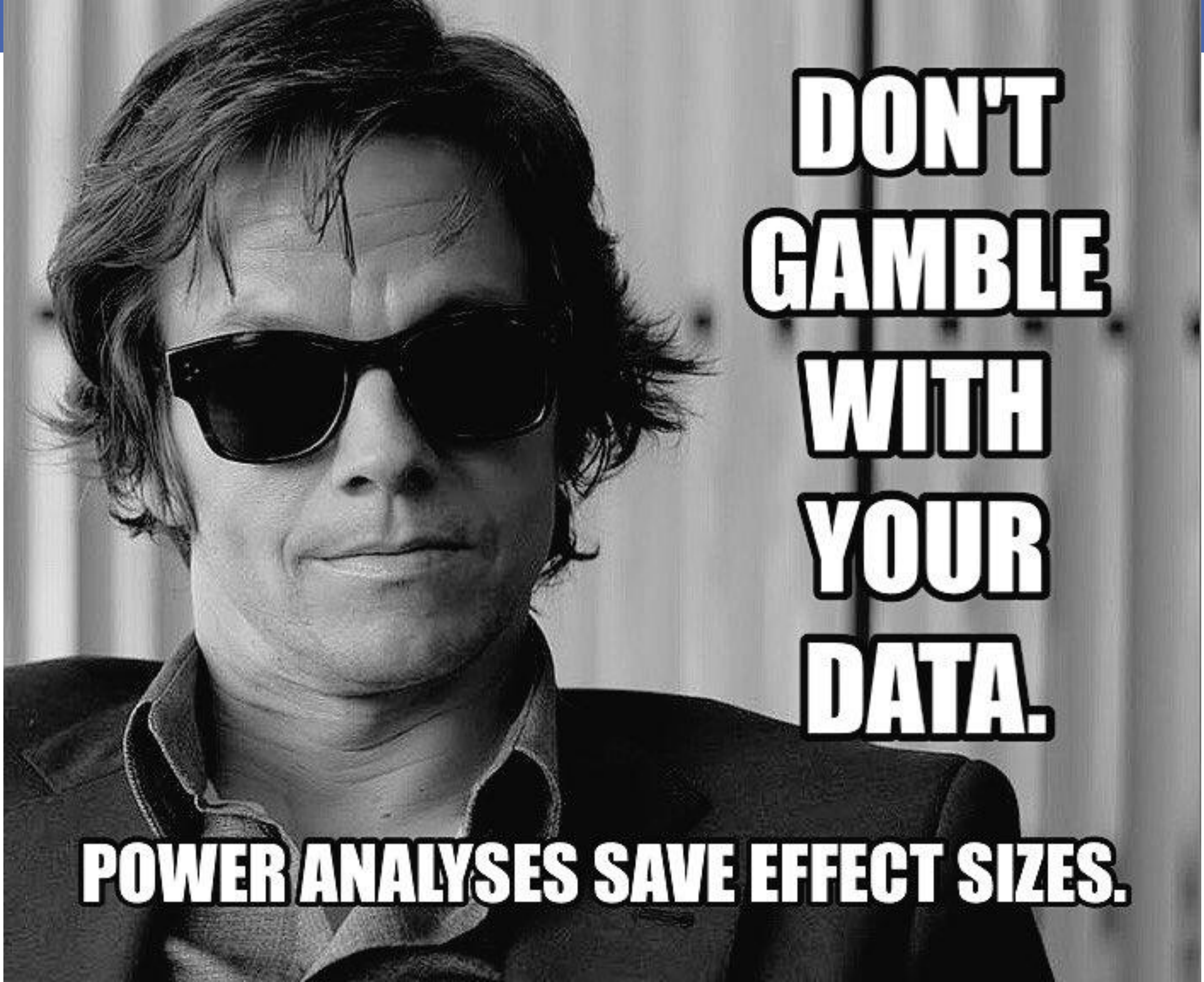
```
power.t.test(n = [redacted], delta = [redacted] sd = [redacted], sig.level = [redacted], power = [redacted], type = [redacted], alternative = [redacted]
```

One-tailed test

Effect size determination

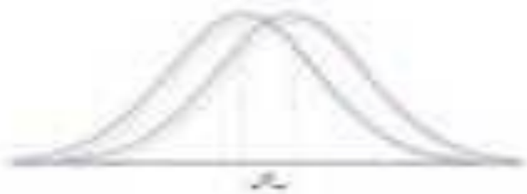
```
power.t.test(n = 100, delta = NULL, sd = 13, sig.level = .05, power = .80, type =
"one.sample", alternative = c("one.sided"))
```

```
     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.

Let’s play….

Effect size



- Effect size is a quantitative measure of the *strength of a phenomenon*.
- Effect size emphasizes the **size** of the difference or relationship
- Examples:
  - the correlation between two variables (specifically r^2)
    - $r=.1$ weak, $r=.5$ moderate, $r=.7$ strong, $r=.9$ very strong
  - the regression coefficient in a regression (B_0, B_1, B_2)
    - Relative to model and field
  - the mean differences in t tests (use Cohen's D)
    - $d = .2$ is small, $r = .5$ is medium; $r = .8$ is large
  - The mean differences in ANOVA (use eta)
    - .01 is small, .06 medium, .14 large