

# Errors, Effect Size and Power

---

When the null hypothesis is true

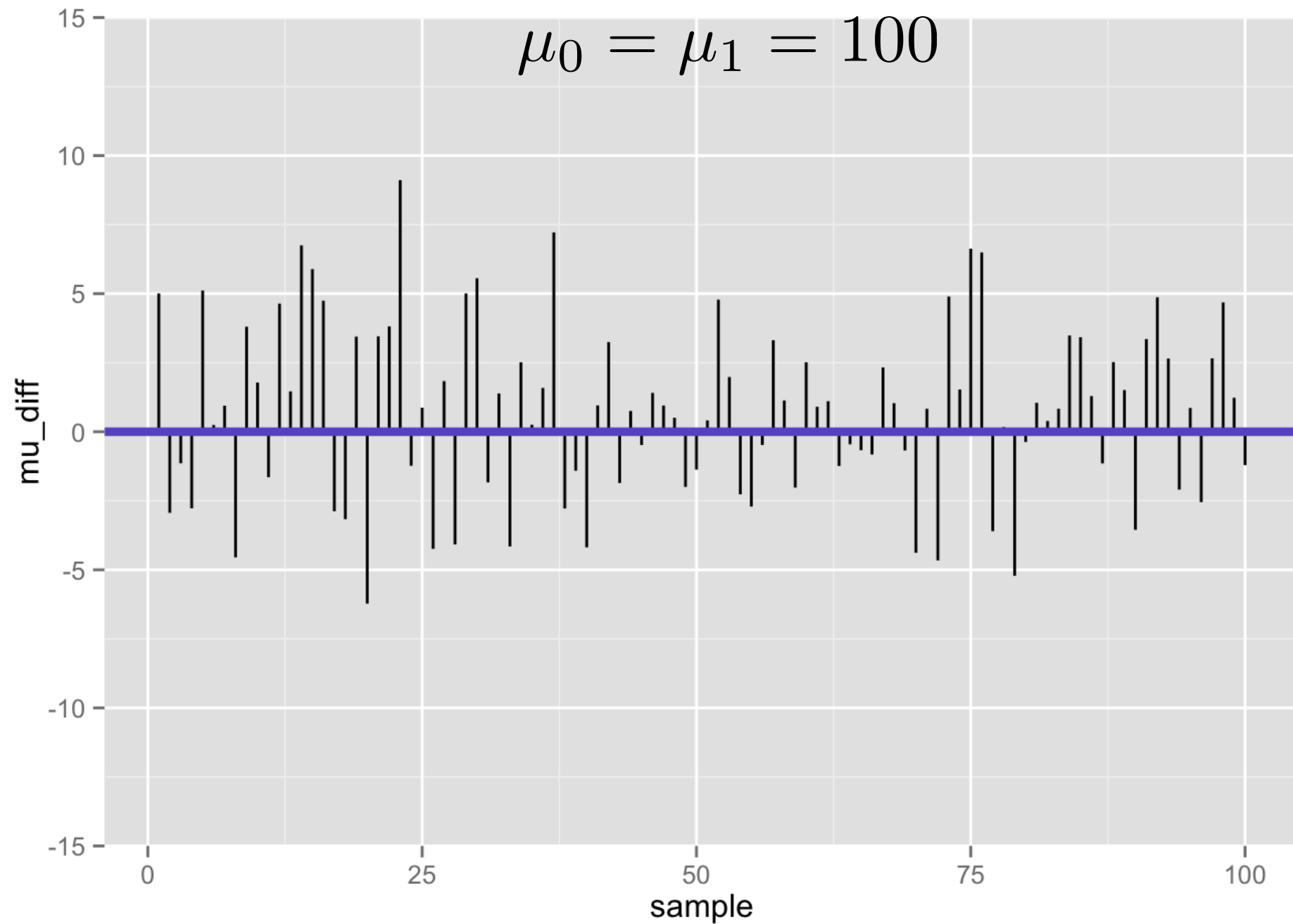
$$\mu_0 = \mu_1 = 100$$

---

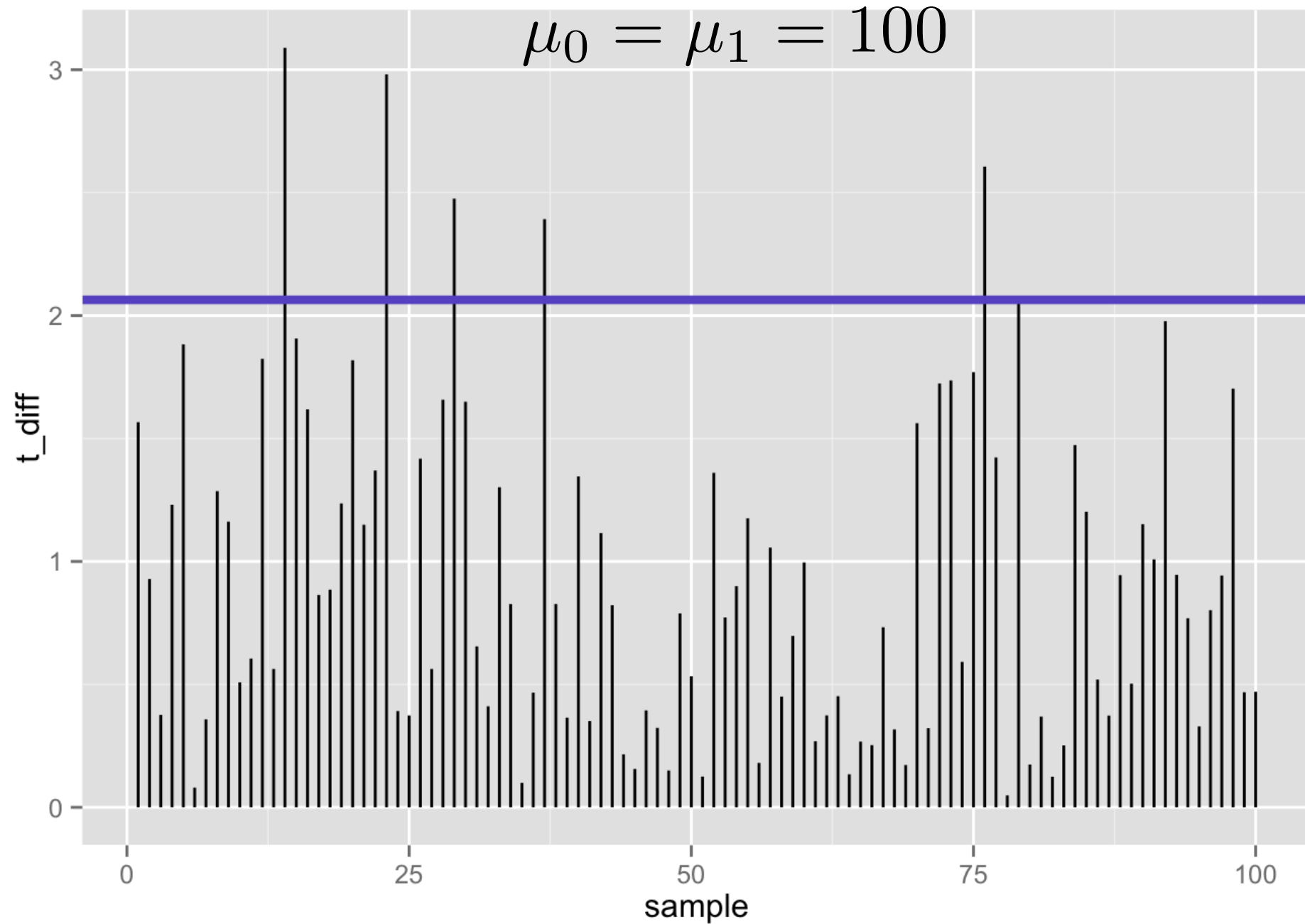
1-sample t-test

$$n = 25$$

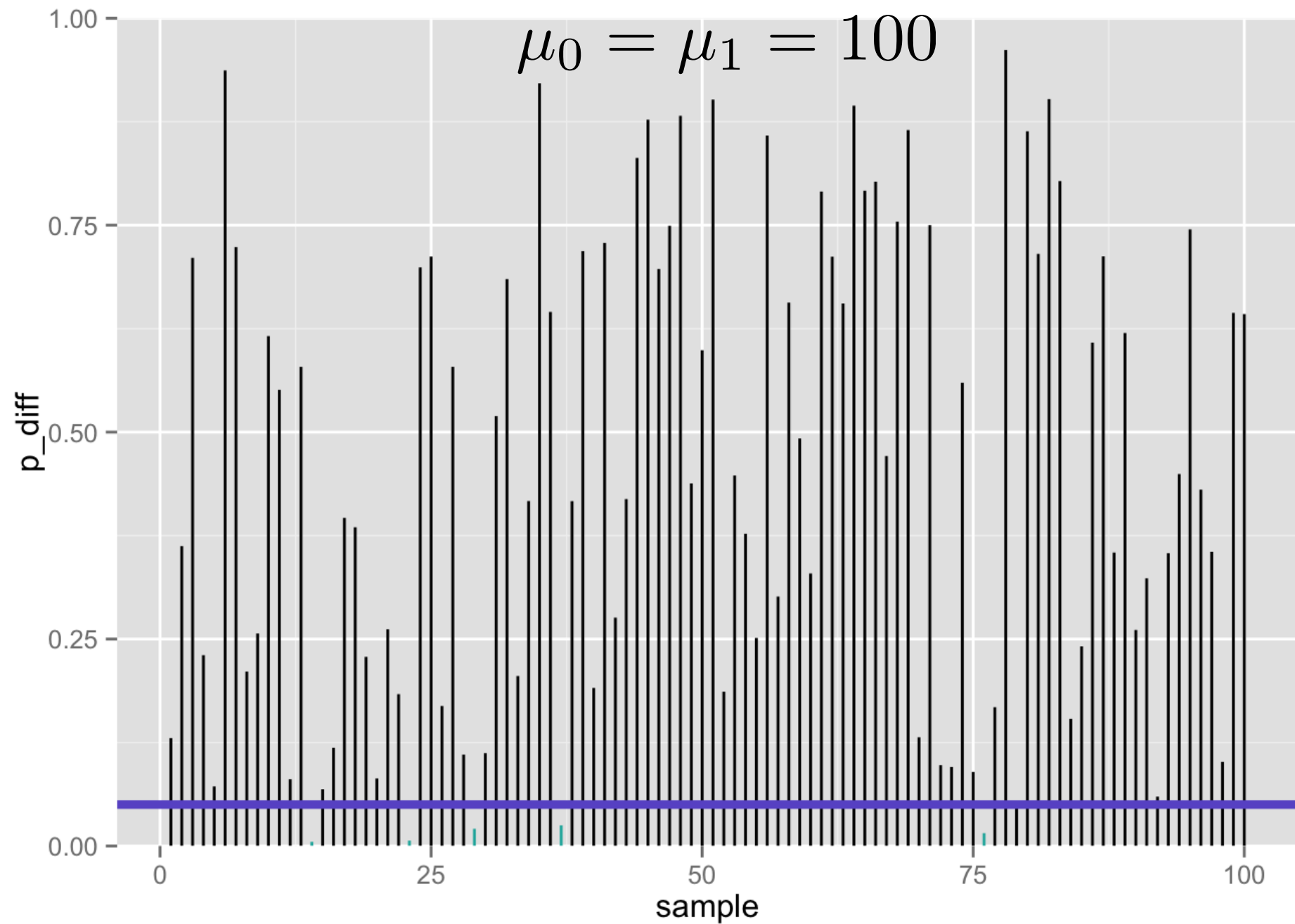
Differences between population mean and sample mean (100 samples) when null is **true**



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



100 p-values when null is **true**: 5% false positives using t-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 $\alpha$	# true $H_0$ 's
$H_1$	False negative Type II error $\beta$	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 $\alpha$	# true $H_0$ 's
$H_1$	False negative Type II error $\beta$	True positive $1 - \beta$	# 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 ( $\alpha$ )  
False positive



Type II error ( $\beta$ )  
False negative





# One way we can be wrong...

Type 1 error ( $\alpha$ )  
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:  $\alpha = ?$

# 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 ( $\beta$ )  
False negative



# In our aspiring astronauts example...

Type 1 error ( $\alpha$ )  
False positive



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

Type II error ( $\beta$ )  
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?

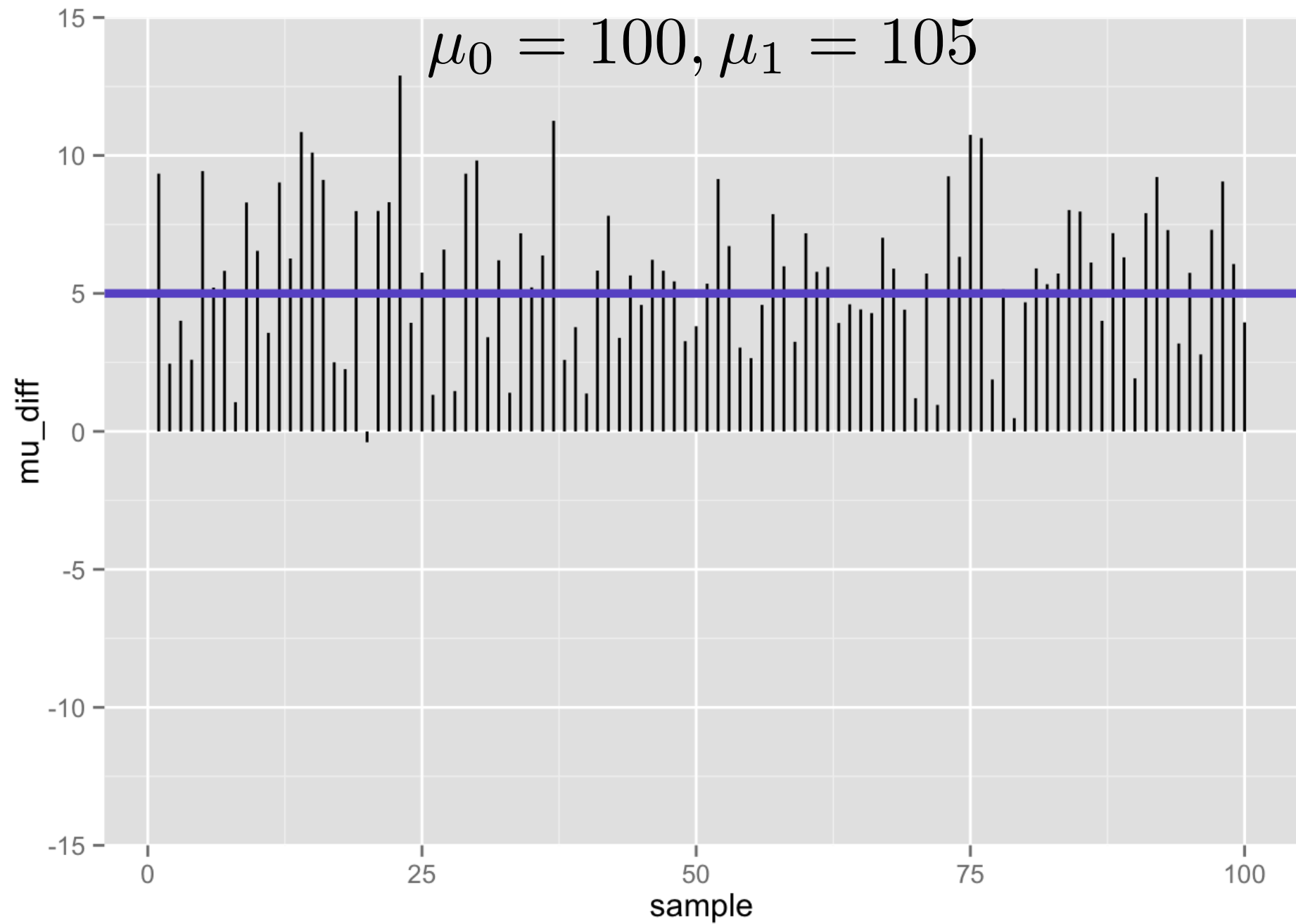
$$\mu_0 = 100, \mu_1 = 105$$

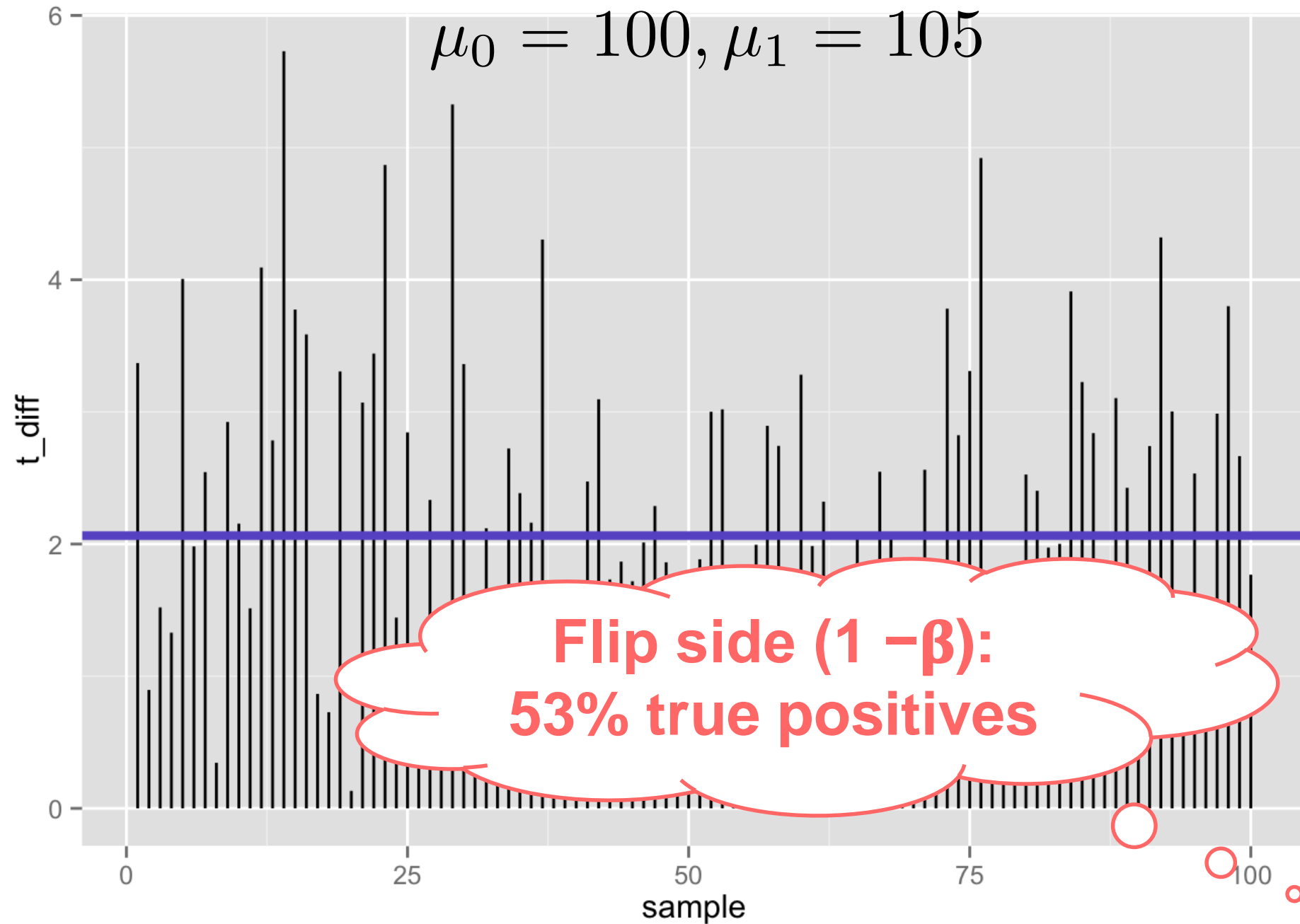
---

1-sample t-test

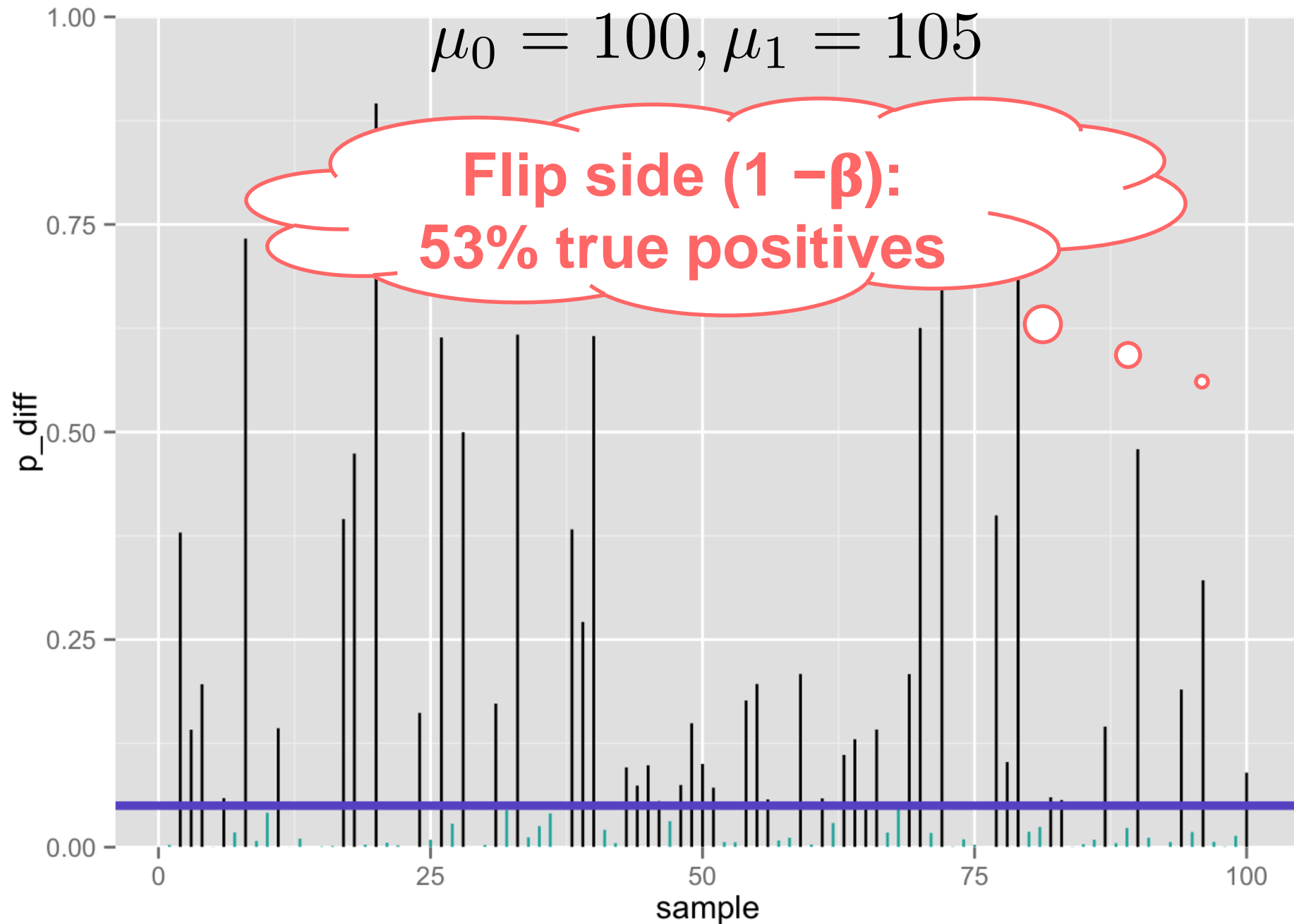
$$n = 25$$

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





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





# Type II errors

- The one(s) that got away...

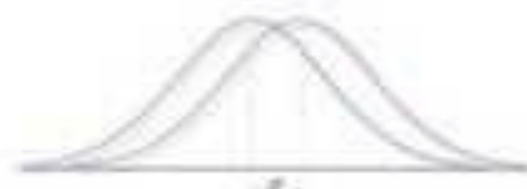


≈ 50% true positives seems low...

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



## 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,  $d = .5$  is medium,  $d = .8$  is large
  - The mean differences in ANOVA (use eta)
    - .01 is small, .06 medium, .14 large

# Power

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

$$1 - b = P(\text{reject } H_0 \mid H_1 \text{ true})$$

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

$$b = P(\text{fail to reject } H_0 \mid H_1 \text{ true})$$

# 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
```



# 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:

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

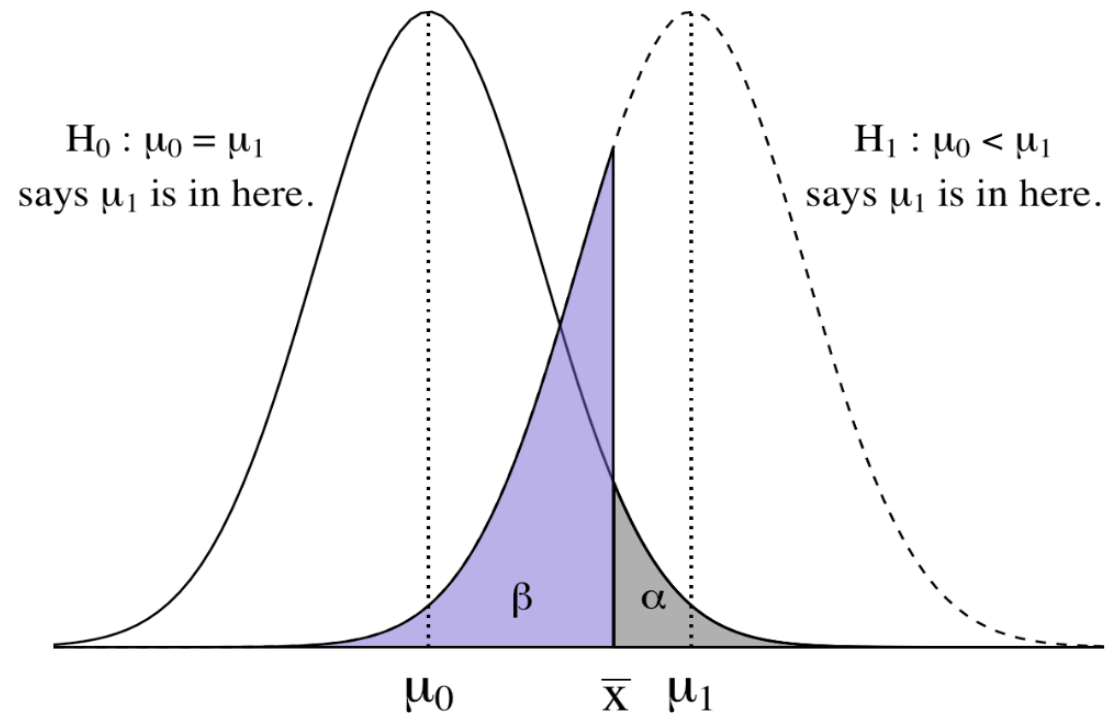


- $P(\text{false positive}) = \alpha = \text{Type I error rate} = .05$
- $P(\text{false negative}) = \beta = \text{Type II error} = ?$
- $P(\text{true positive}) = 1 - \beta = \text{power} = ?$



# Dueling distributions

- In a one-tailed test, we have two dueling hypothetical distributions:
  - The null distribution, centered at  $\mu_0$
  - The alternative distribution, centered around some specific or unspecified other mean (higher or lower?)  $\mu_1$

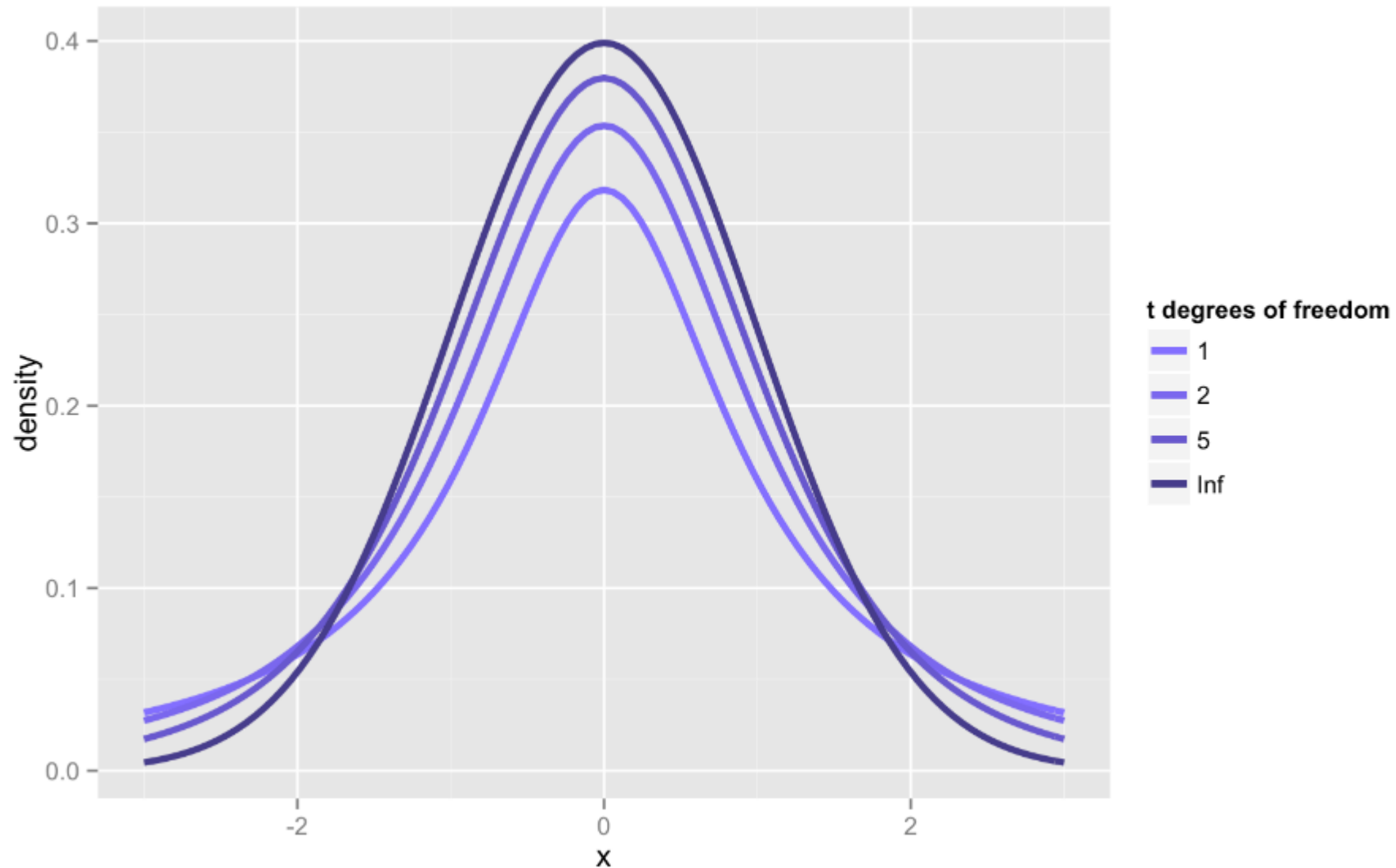


# Finding $\beta$ (and $1 - \beta$ )

- 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)
```

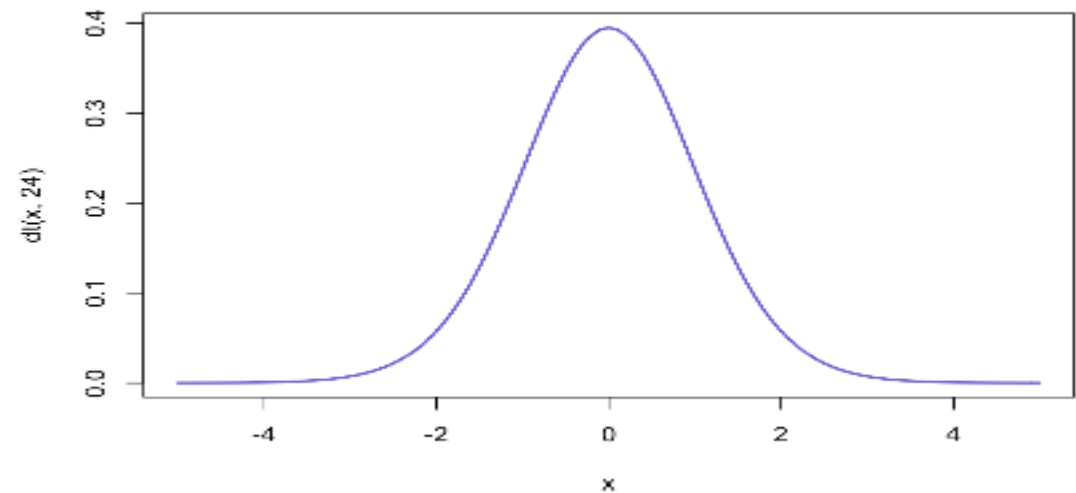
## Arguments

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



# Central t-distribution ( $v$ = degrees of freedom)

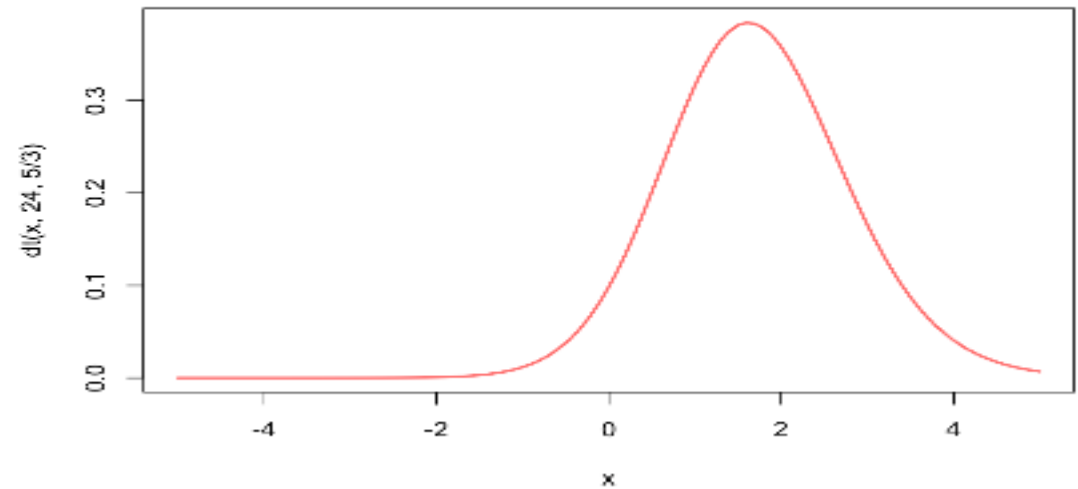
- $\delta = 0$
- Mean = 0
- Variance slightly  $>$  than  $N(0, 1)$
- Kurtosis (biased) is  $> 3$
- Symmetric

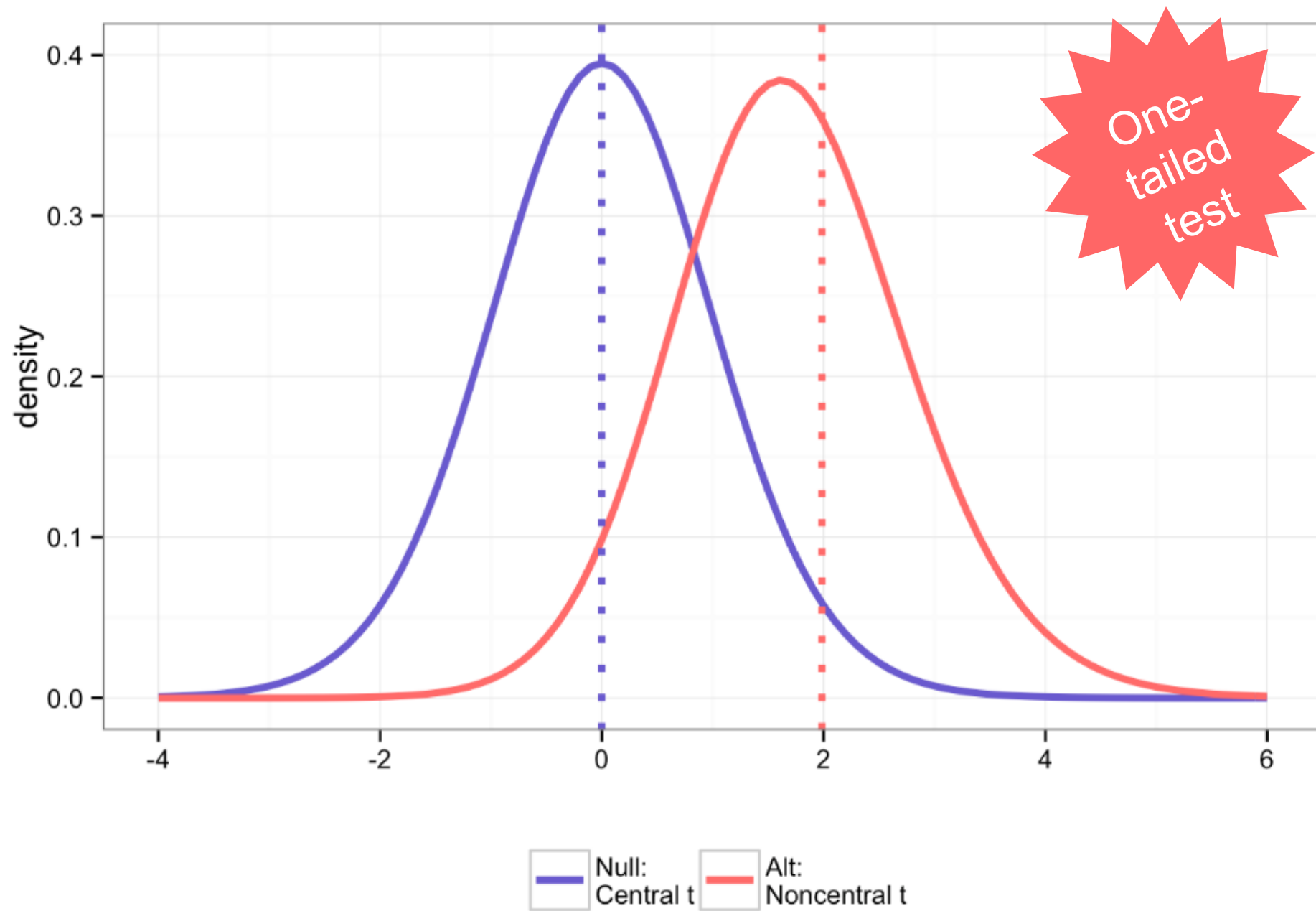


# Noncentral t distribution ( $\nu$ = degrees of freedom)

- $\delta \neq 0$
- Asymmetric: skewed in the direction of  $\delta$

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





# 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

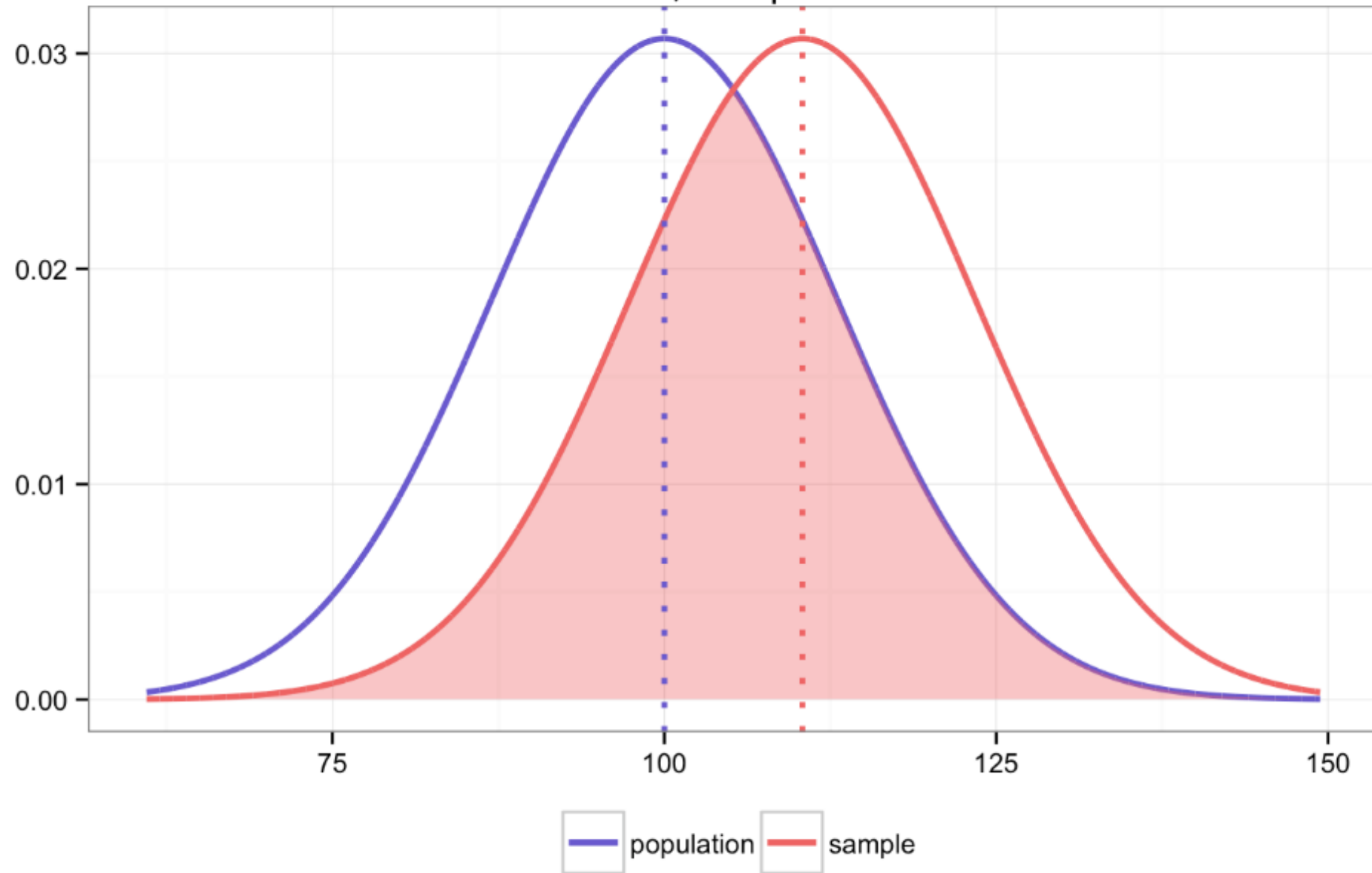
$$t_{\nu} = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$$

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

What is **not** in this formula?

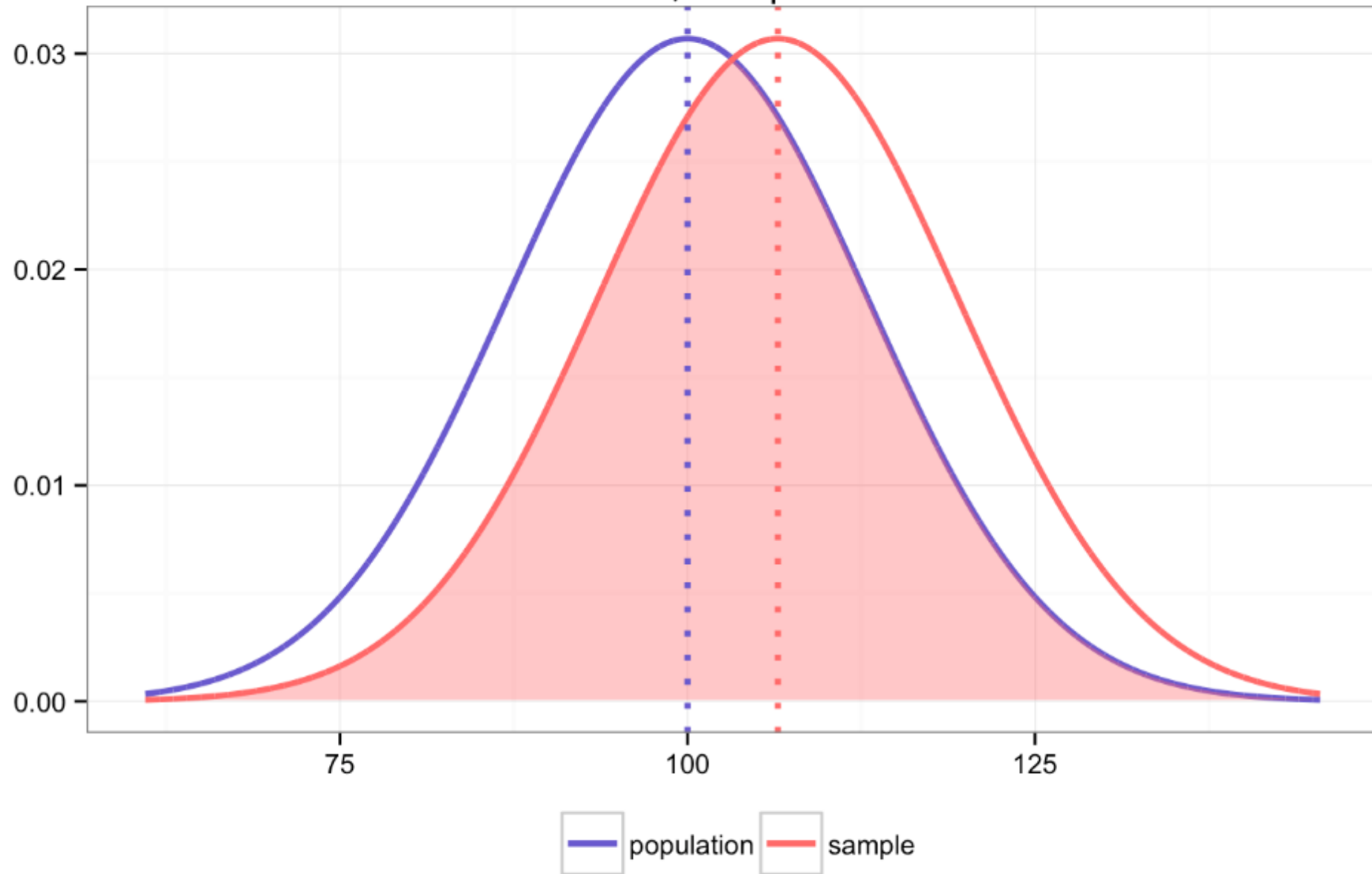


Effect size = 0.8; sample mean = 110.4

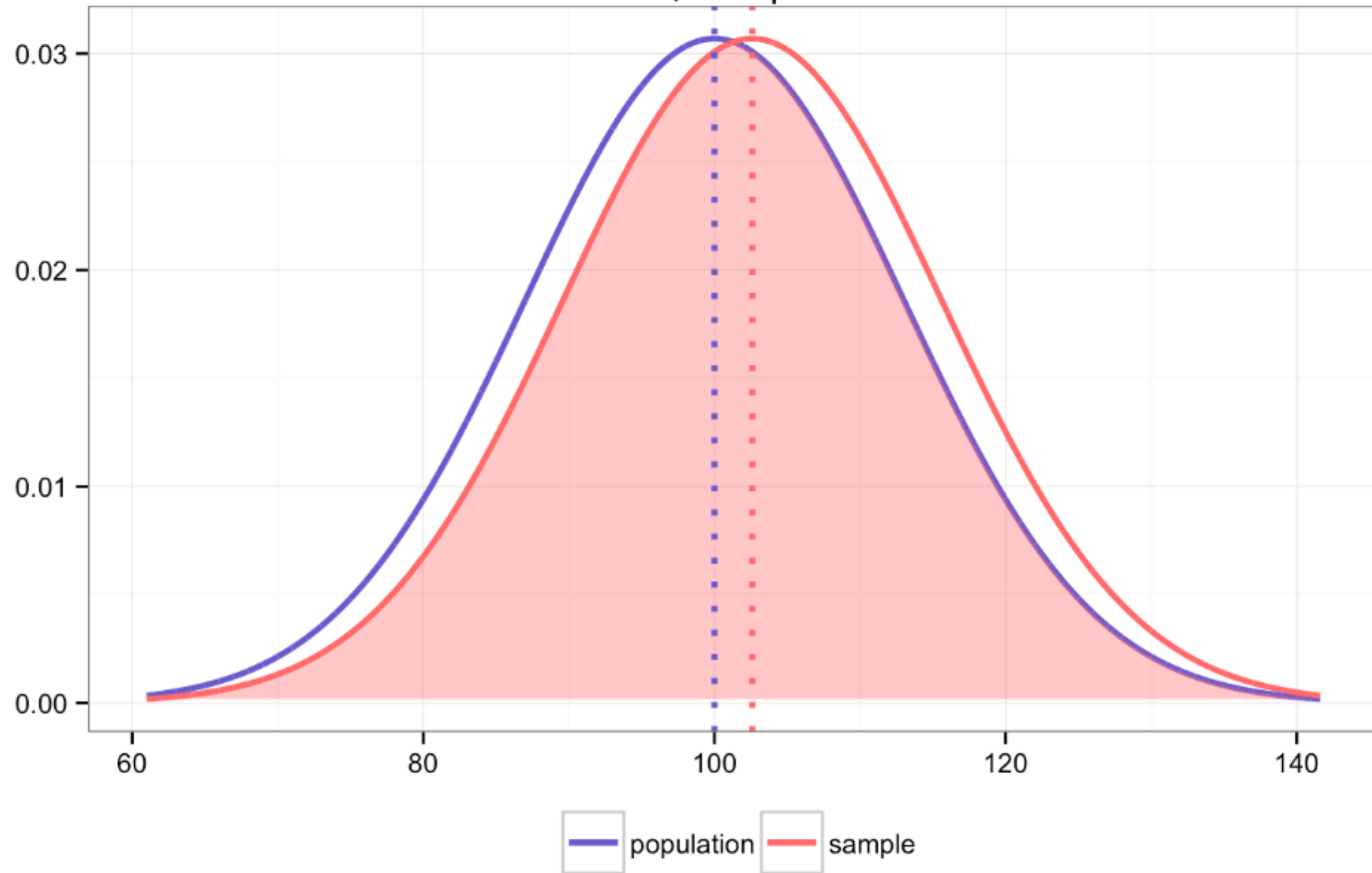




Effect size = 0.5; sample mean = 106.5



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

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  $\beta$



# POWER: $p(\text{true positive}) = 1 - \beta$

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

```
qt(.95, 24) # tcritical, null dist
```

```
[1] 1.710882
```

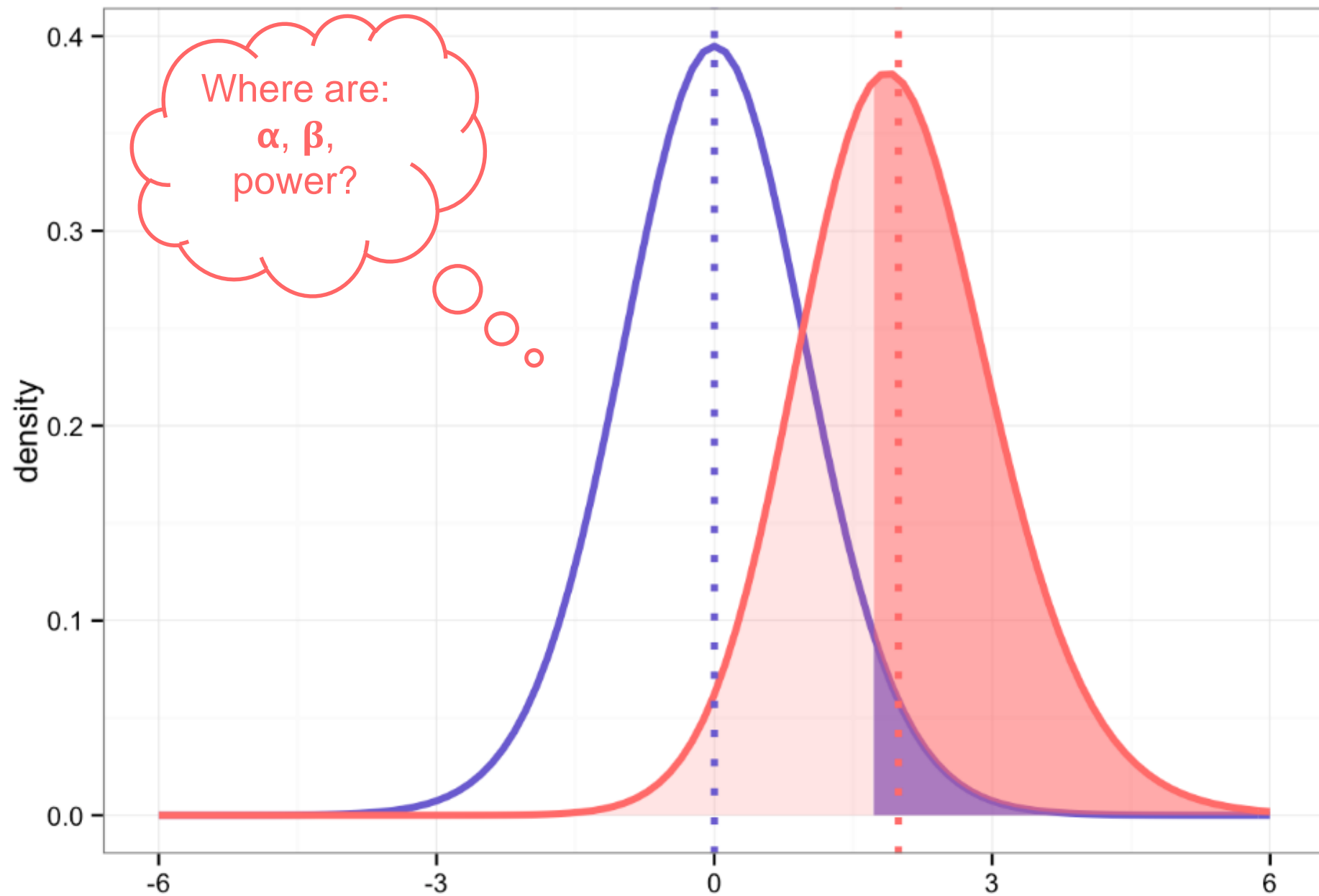
```
pt(qt(0.95, 24), 24, 25/13) # beta
```

```
[1] 0.4115342
```

```
1 - pt(qt(0.95, 24), 24, 25/13) # power
```

```
[1] 0.5884658
```





# Power

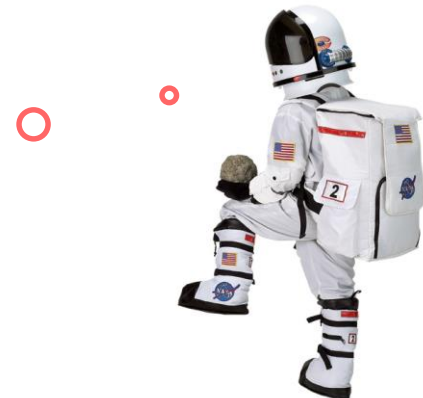
```
power.t.test(n = 25, delta = 5, sd = 13, type = "one.sample", alternative =  
c("one.sided"))
```

One-sample t test power calculation

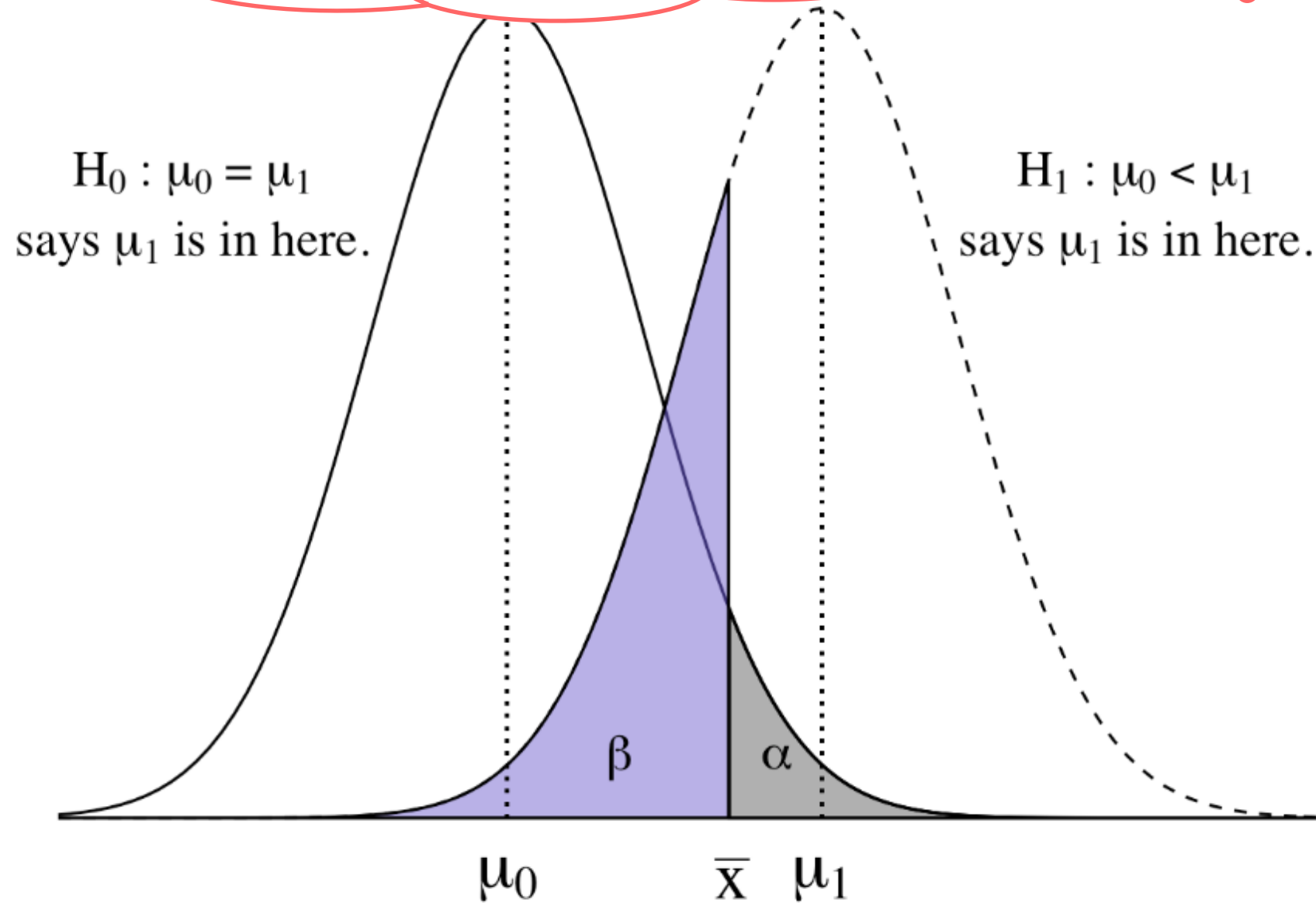
```
      n = 25  
    delta = 5  
      sd = 13  
sig.level = 0.05  
  power = 0.5884658  
alternative = one.sided
```

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  $\beta$  if I  
make  $\alpha$  smaller?





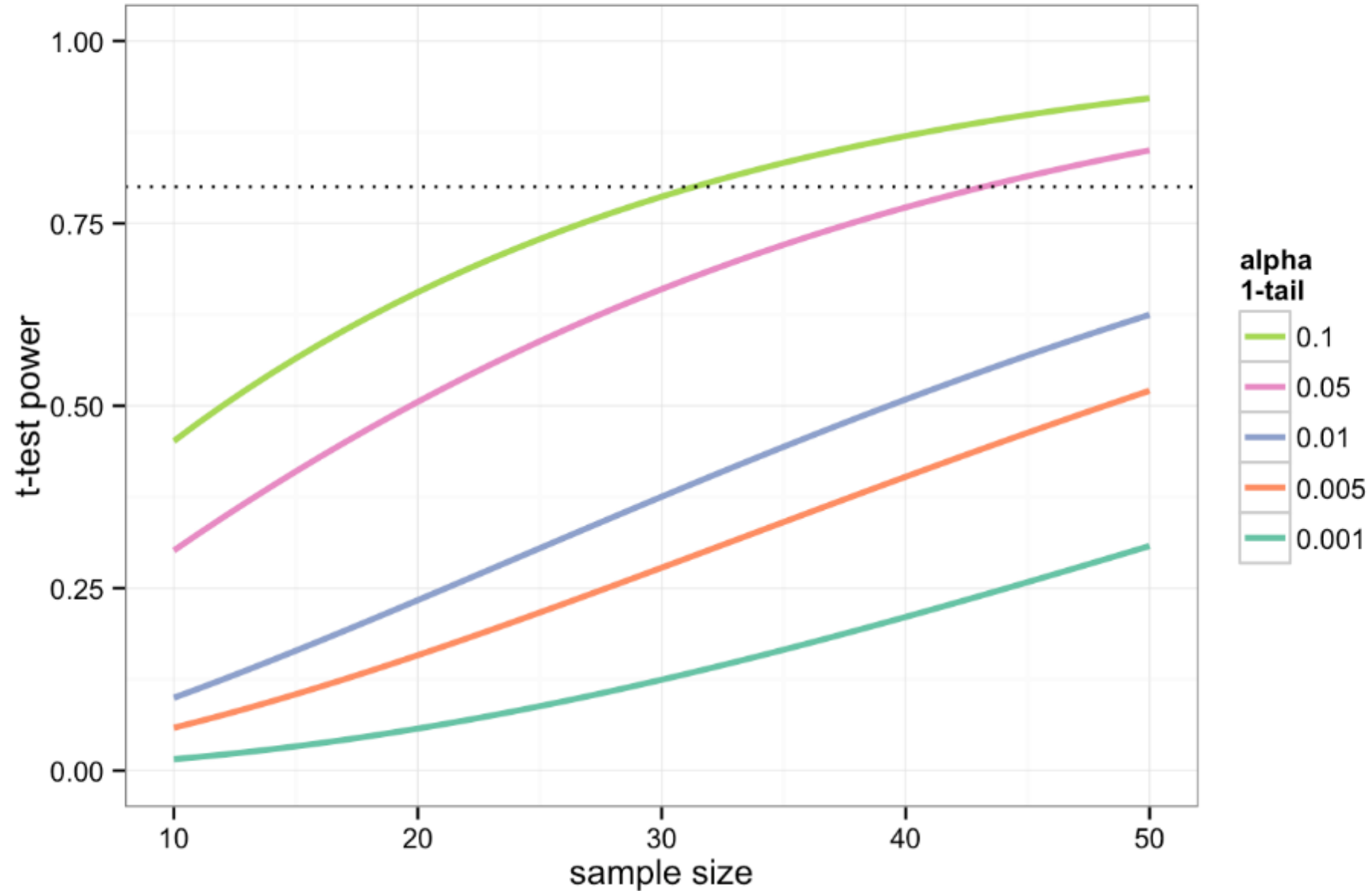
$\alpha \uparrow \rightsquigarrow \beta \downarrow$

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

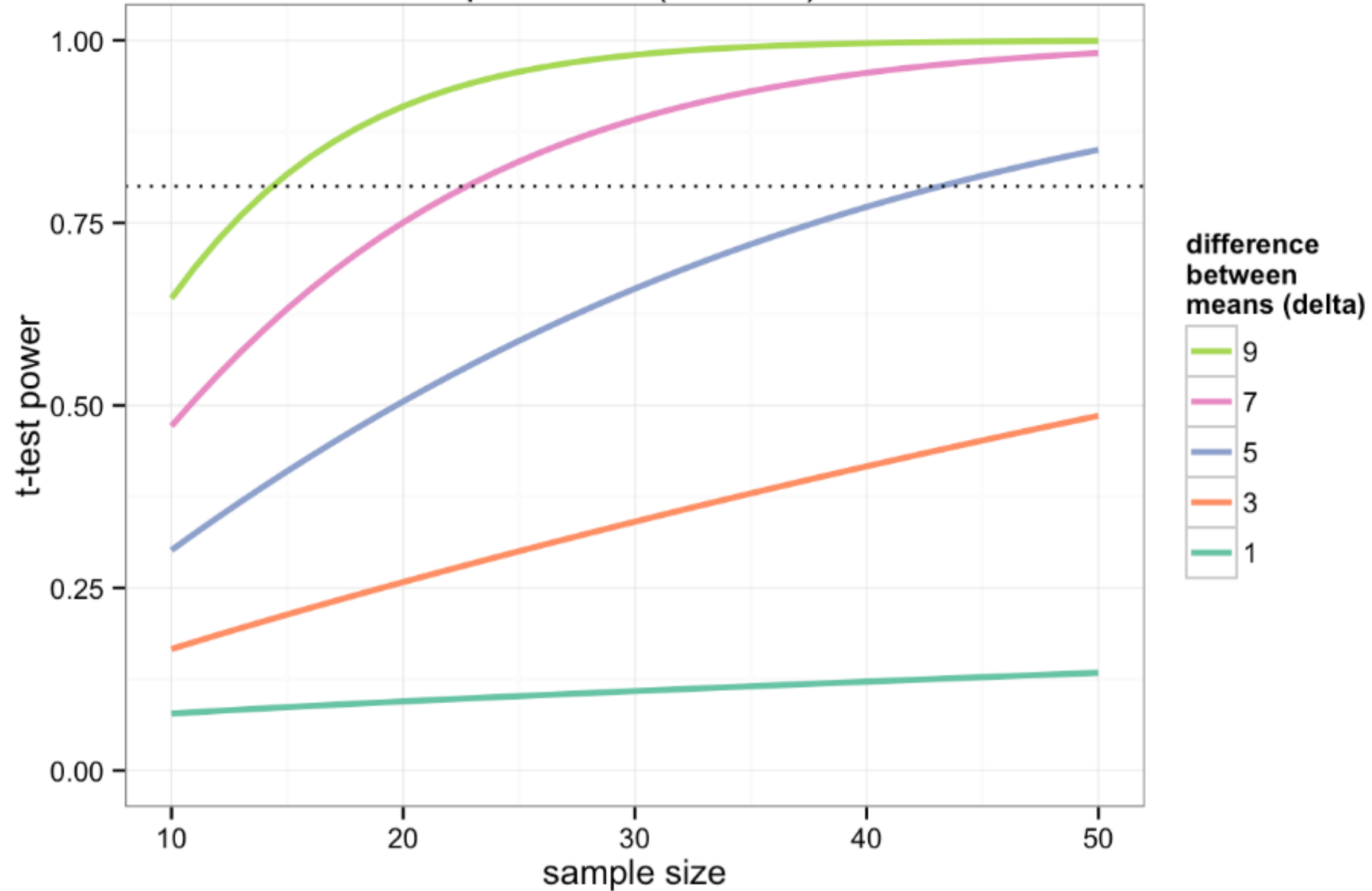
# Factors that affect Power ( $1 - \beta$ )

- Sample size
  - Increased  $n$  reduces  $SE_{\text{mean}}$
- Level of significance
  - Power increases as  $\alpha$  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  $SE_{\text{mean}}$

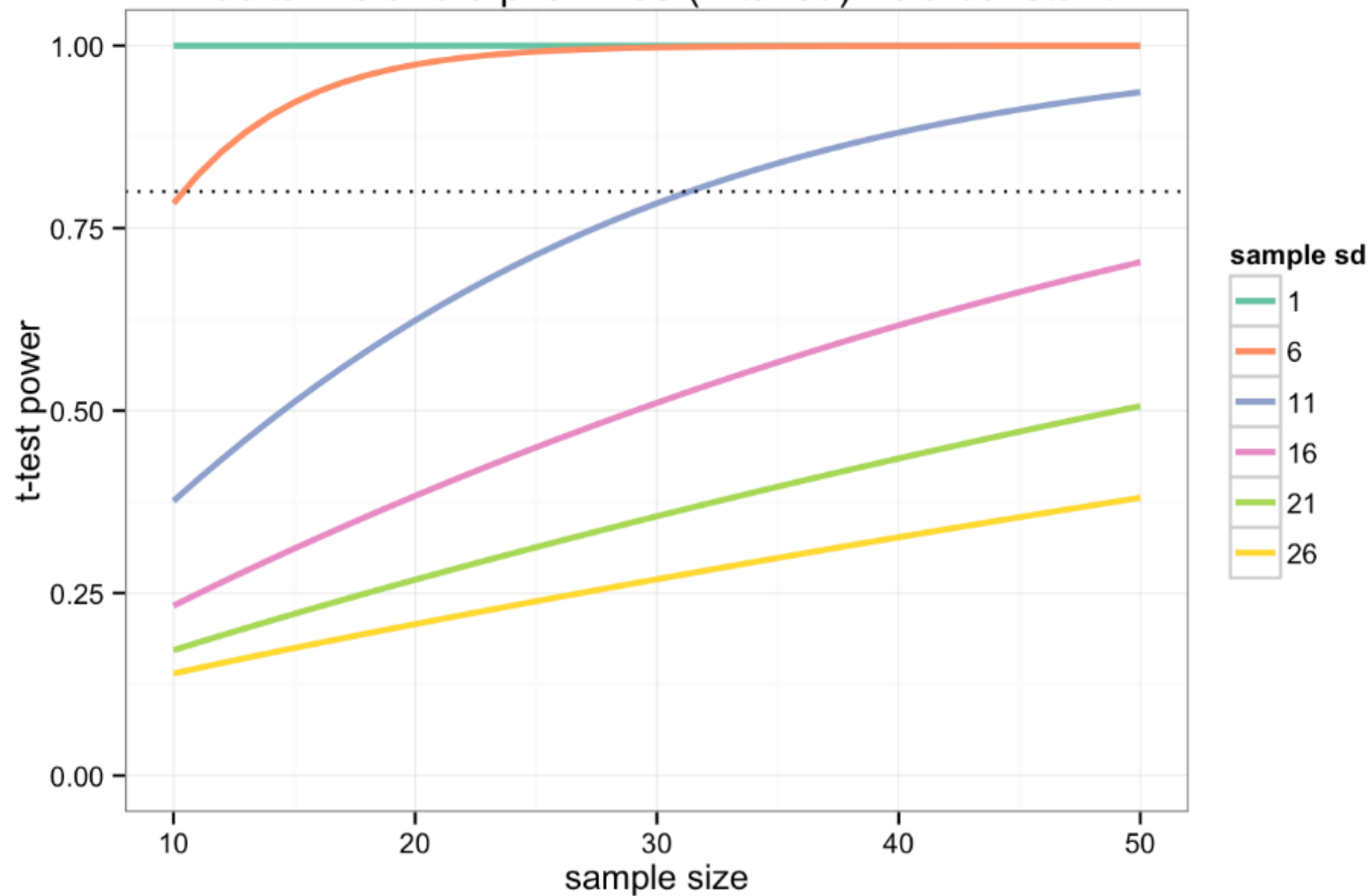
Power increases as  $n$  and  $\alpha$  (1-tailed) increase  
 $\delta = 5$  and  $sd = 13$  held constant



Power increases as  $n$  and  $\delta$  increase  
 $sd = 13$  and  $\alpha = .05$  (1-tailed) held constant



Power increases as  $n$  increases and sample sd decreases  
 $\delta = 5$  and  $\alpha = .05$  (1-tailed) held constant



Let's play. Maybe...

# How large would our “n” have to be?

To detect:

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



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 = )
```

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
```



# 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:

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

With s.d. = 13



# Effect size determination

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

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.**