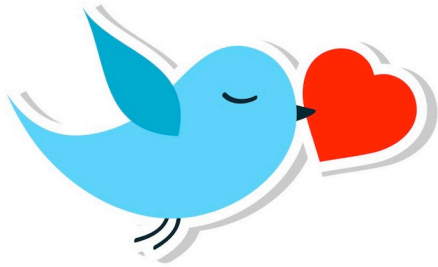
Class 11
Classical Inference: Hypothesis Testing

Alison Presmanes Hill

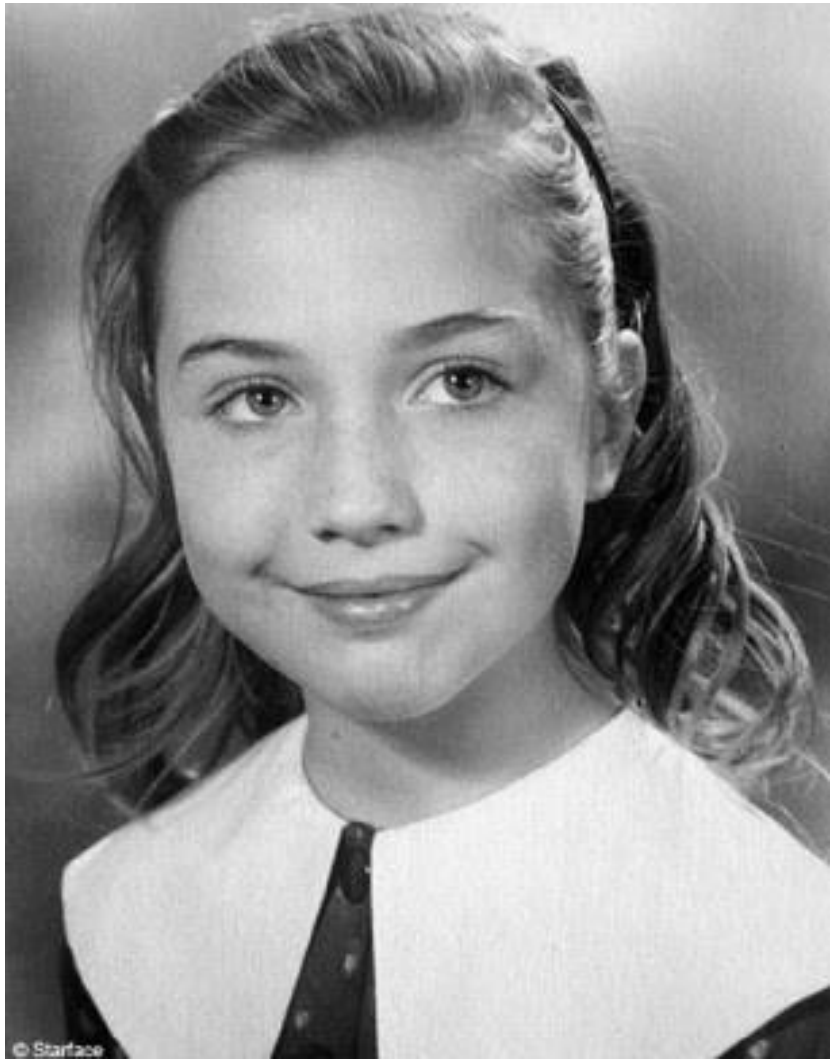
Remember, statistics…

- 2 main reasons we love them:
  - Sometimes they are **estimators** for population parameters we care about
  - Sometimes they are **test statistics**, i.e. the basis for a hypothesis test



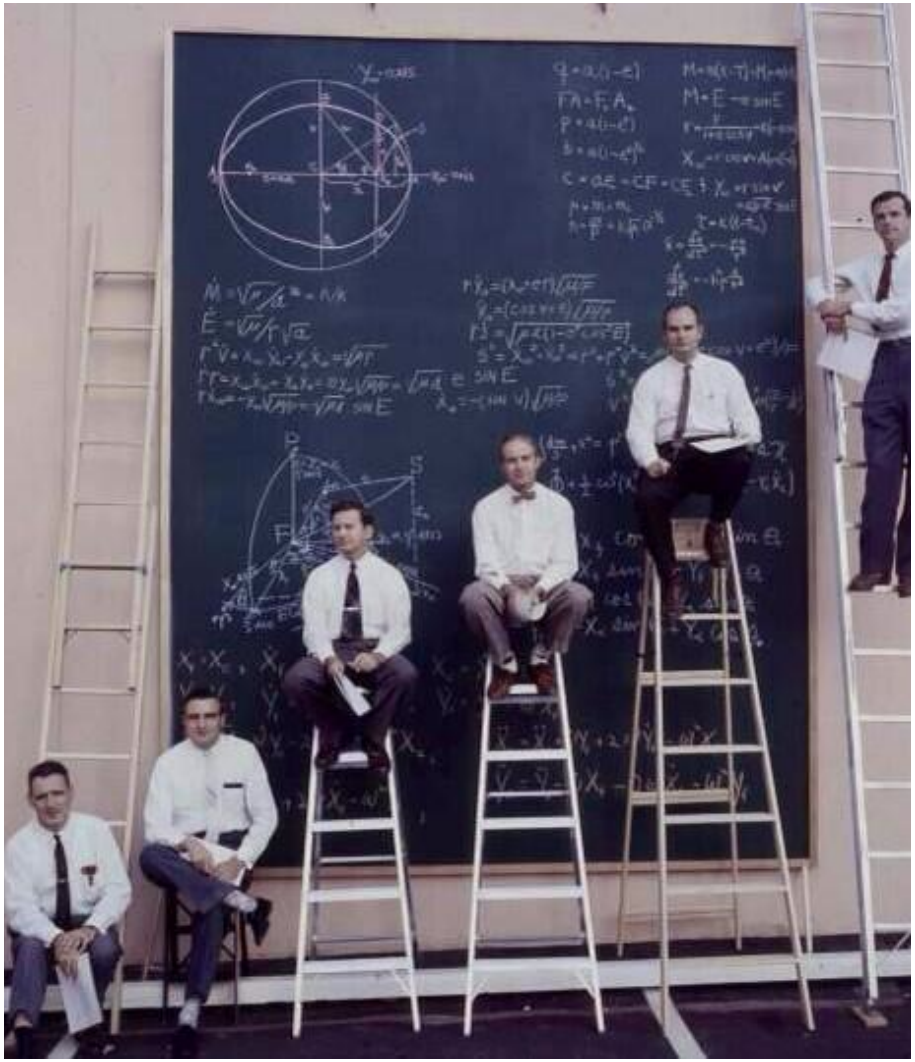
Aspiring astronauts

- 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.



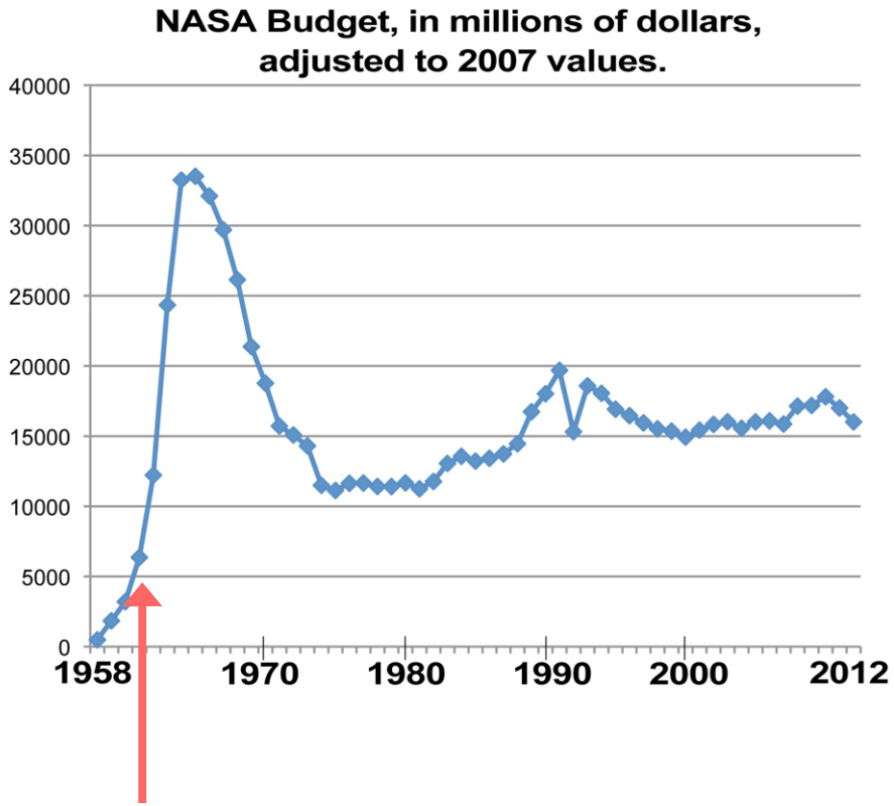
Let's pretend

- 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

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

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

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

Two-tailed test

And the sample mean is…

One-sample z-test

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

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

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

$$= \blacksquare$$

One-sample z-test

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

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

$$= \frac{105 - 100}{15/\sqrt{25}}$$

$$= \frac{5}{3} = 1.667$$

From wikipedia…

- The p-value is defined as the probability, under the assumption of the null hypothesis, of obtaining a result equal to or more extreme than what was actually observed.
- Depending on how it is looked at, the "more extreme than what was actually observed" can mean
  - $X \geq x$ (right-tail event) or
  - $X \leq x$ (left-tail event) or
  - the "smaller" of $X \leq x$ and $X \geq x$ (double-tailed event).
- Thus, the p-value is given by
  - $Pr(X \geq x|H)$ for right tail event,
  - $Pr(X \leq x|H)$ for left tail event,
  - $2*\min\{Pr(X \leq x|H), Pr(X \geq x|H)\}$ for double tail event.

- http://math.stackexchange.com/questions/1493880/two-tailed-hypothesis-test-why-do-we-multiply-p-value-by-two

One-sample *z*-test

- Does sample perform differently than general population (known: μ & σ)?

$$z_{obs} = \frac{\bar{Y}_i - \mu}{\sigma / \sqrt{n}}$$

- Now what?
  - Determine $z_{critical}$ values for your α
  - Must "beat" ±1.64 for α = .10, 2-tailed, to reject null
  - Must "beat" ±1.96 for α = .05, 2-tailed, to reject null
  - Must "beat" ±2.32 for α = .01, 2-tailed, to reject null

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

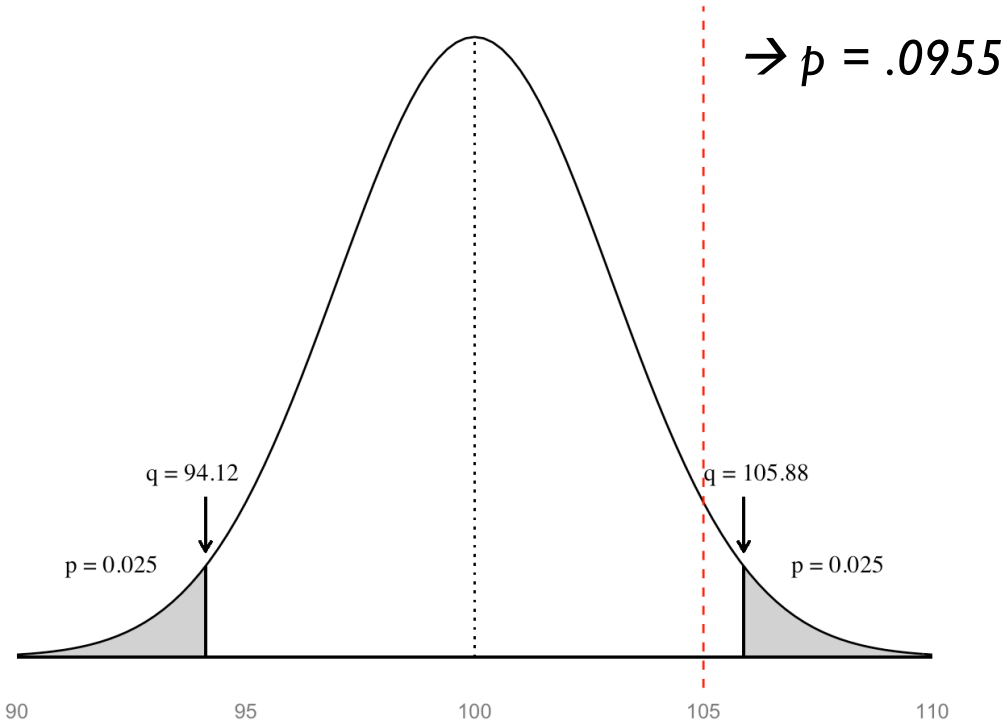
```
pz_2 <- 2*pz_up
```

```
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
```



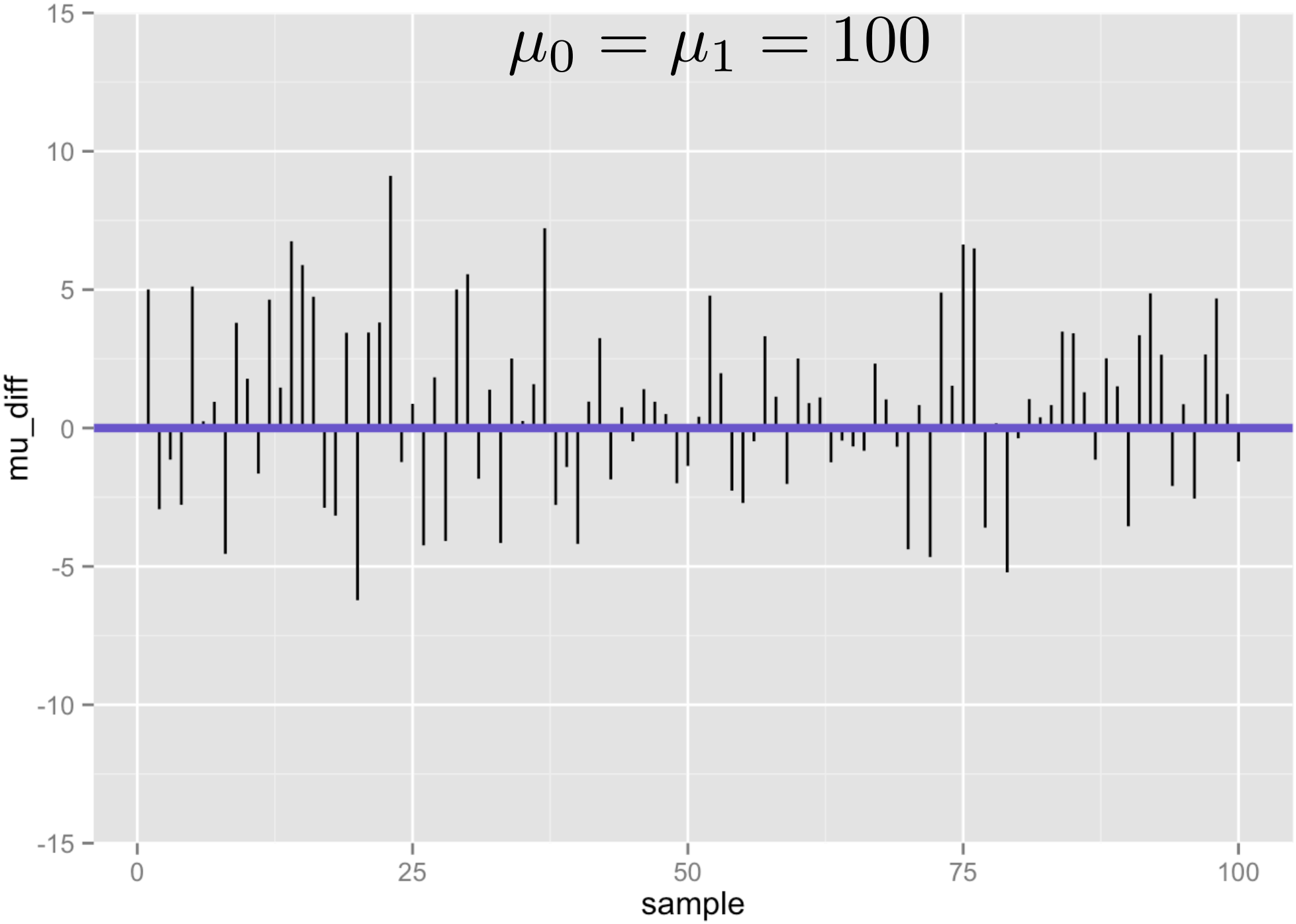
When the null hypothesis is true

$$\mu_0 = \mu_1 = 100$$

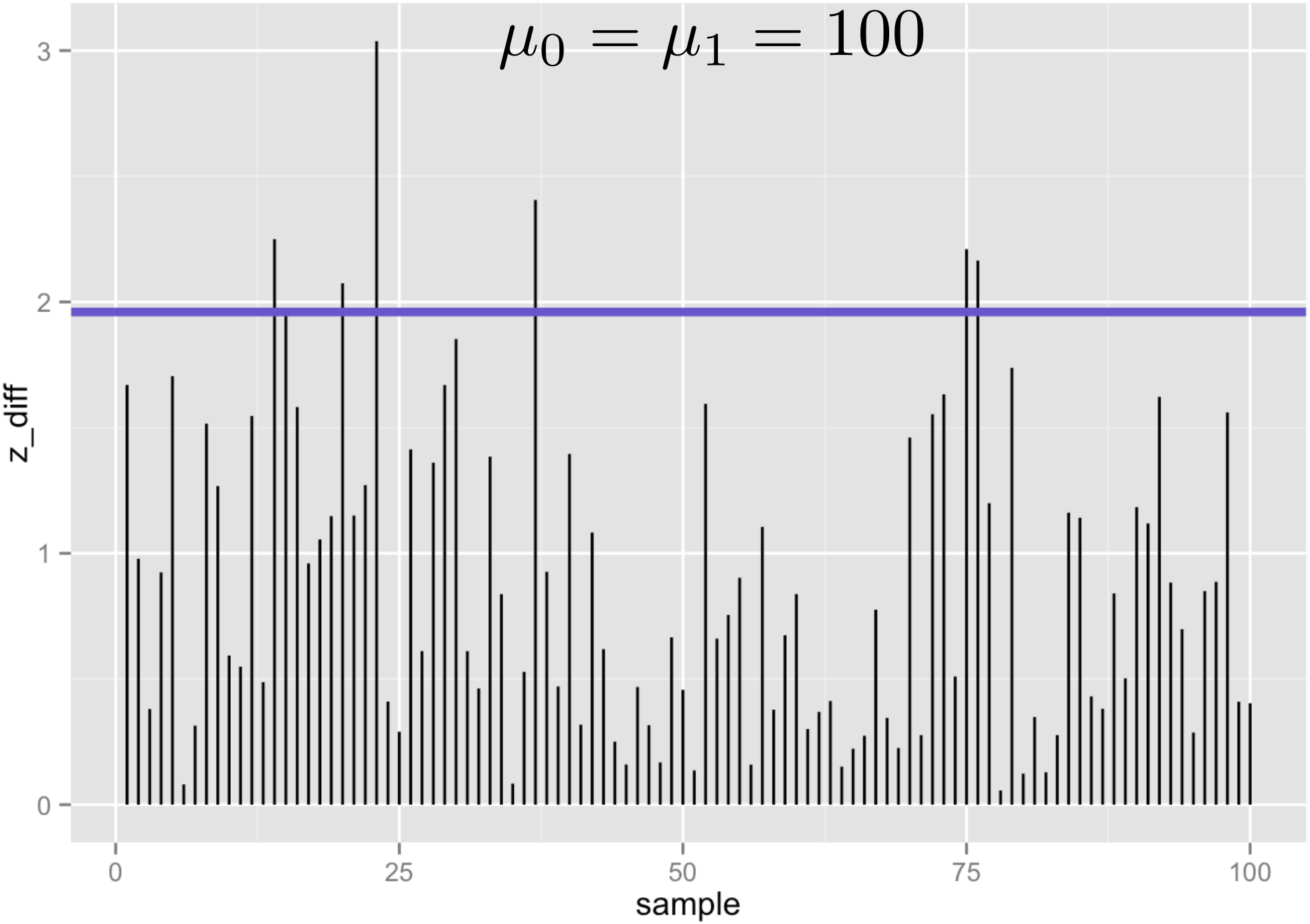
1-sample z-test

n = 25

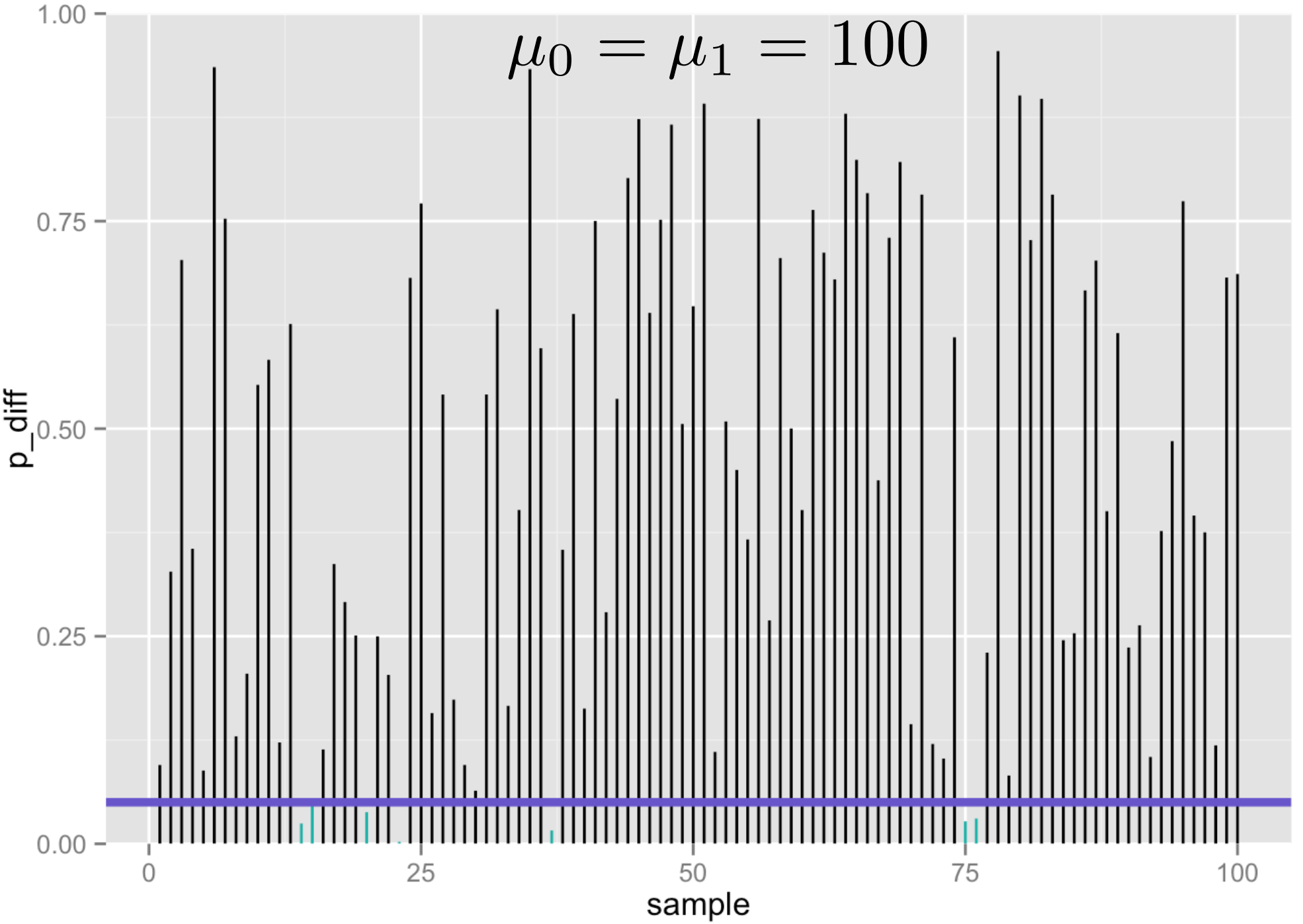
Differences between population mean and sample mean (100 samples) when null is **true**



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 (α = .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 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…

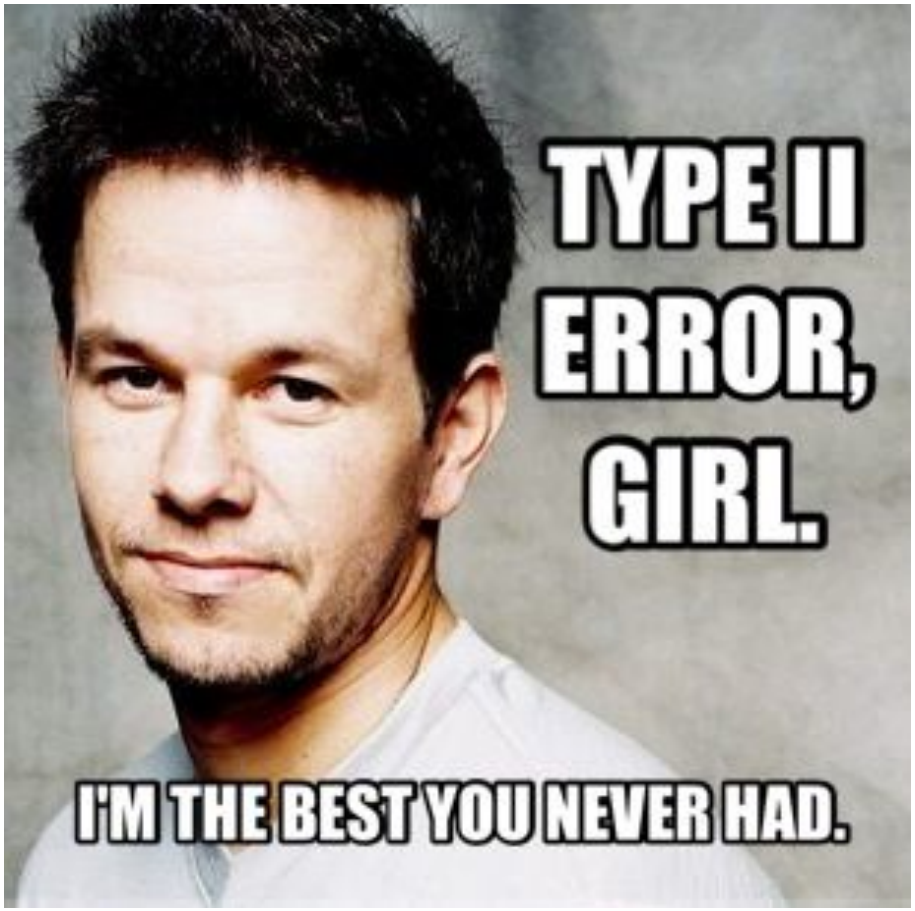
Call: fail to reject H_0

- 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

Type II errors

- The one(s) that got away…



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

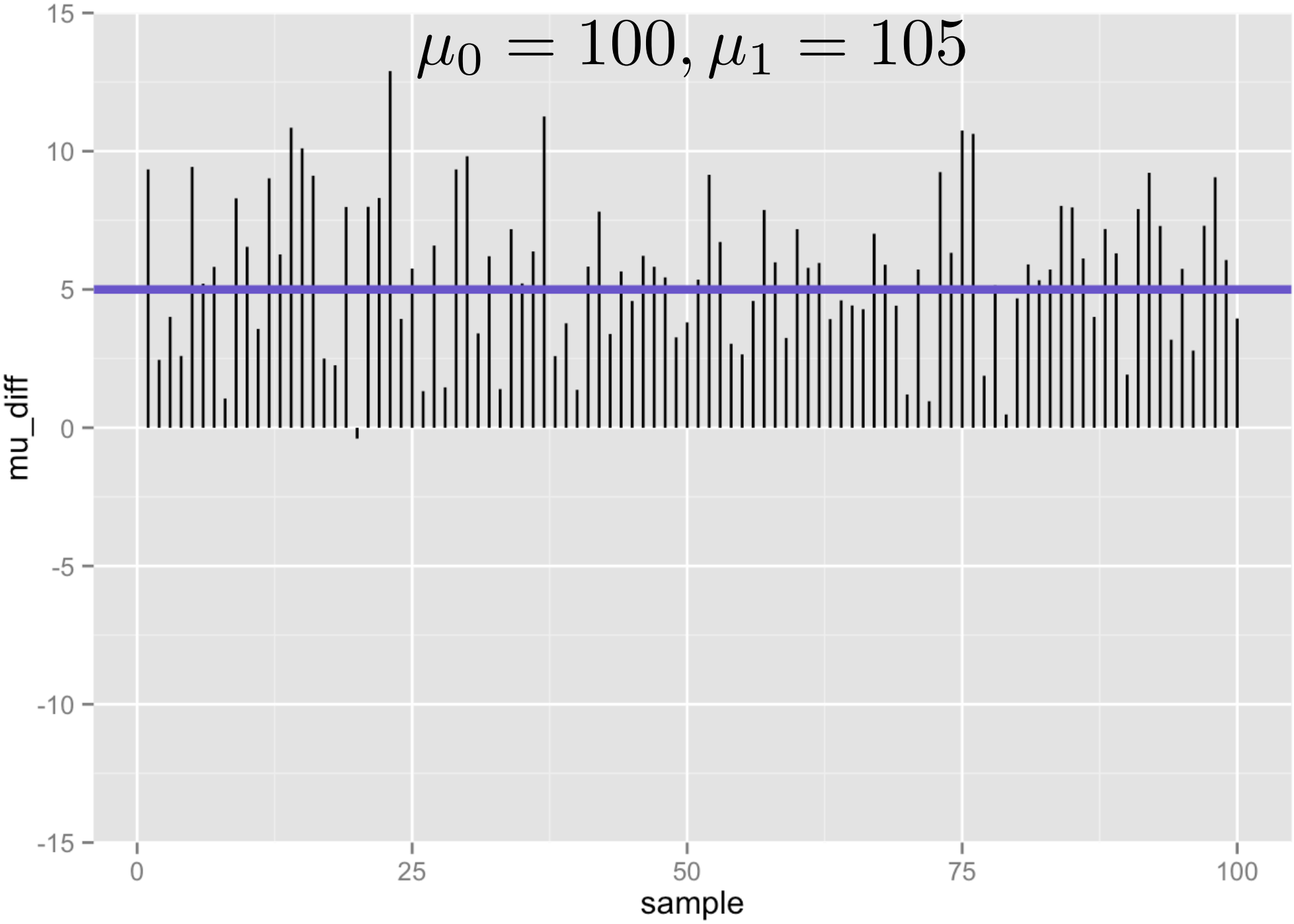
Decide:
"You're **not** smarter than average!"
Reality:
but you **are** actually

What if:
the null hypothesis is FALSE?

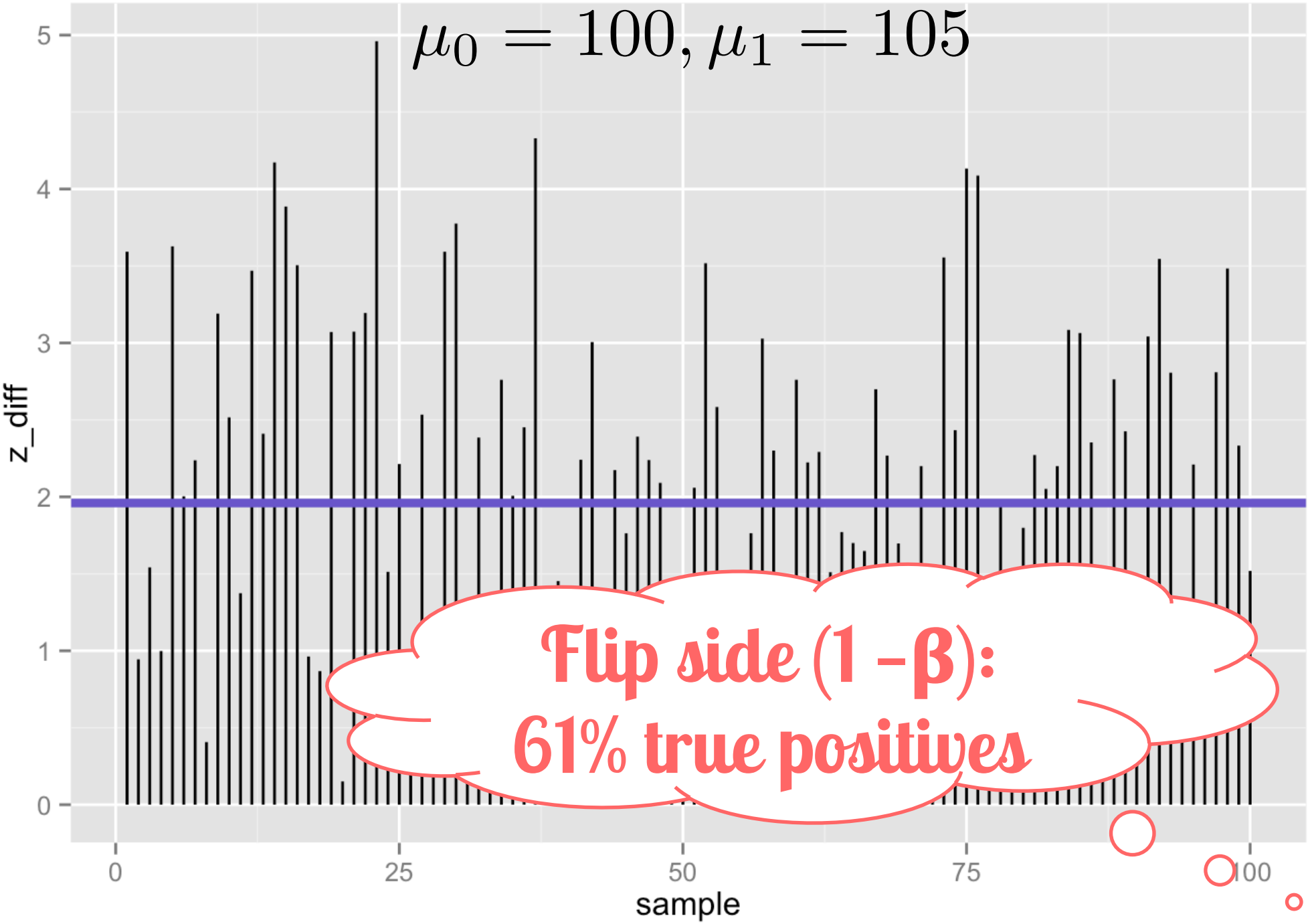
$$\mu_0 = 100, \mu_1 = 105$$

1-sample z-test

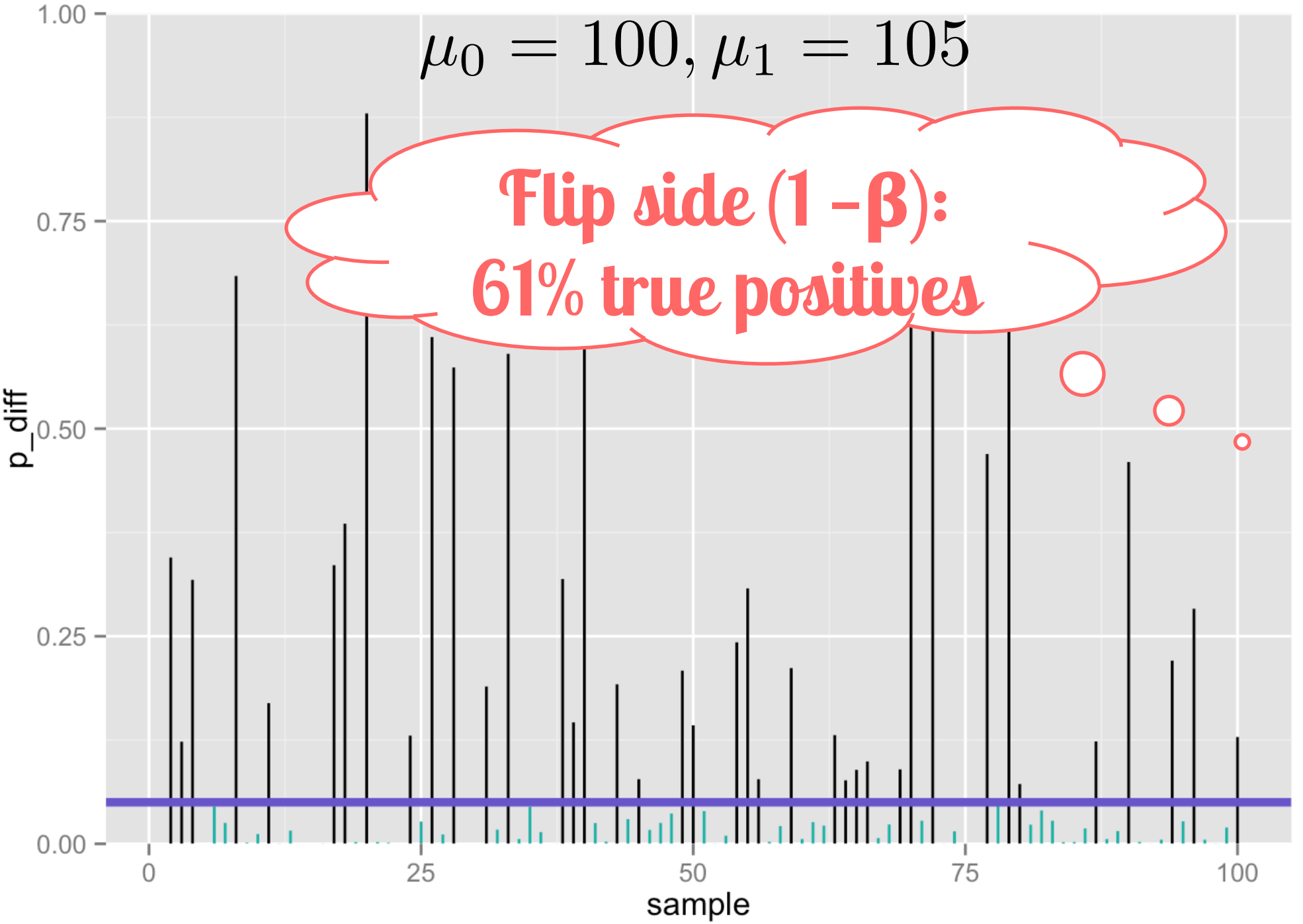
n = 25


$\mu_0 = 100, \mu_1 = 105$
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 (α = .05/2)



100 p-values when null is false: 39% false negatives using z-test (α = .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. ($\boldsymbol{\sigma}$).

What is the NULL distribution of this new test statistic?

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 means T test

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

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

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

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

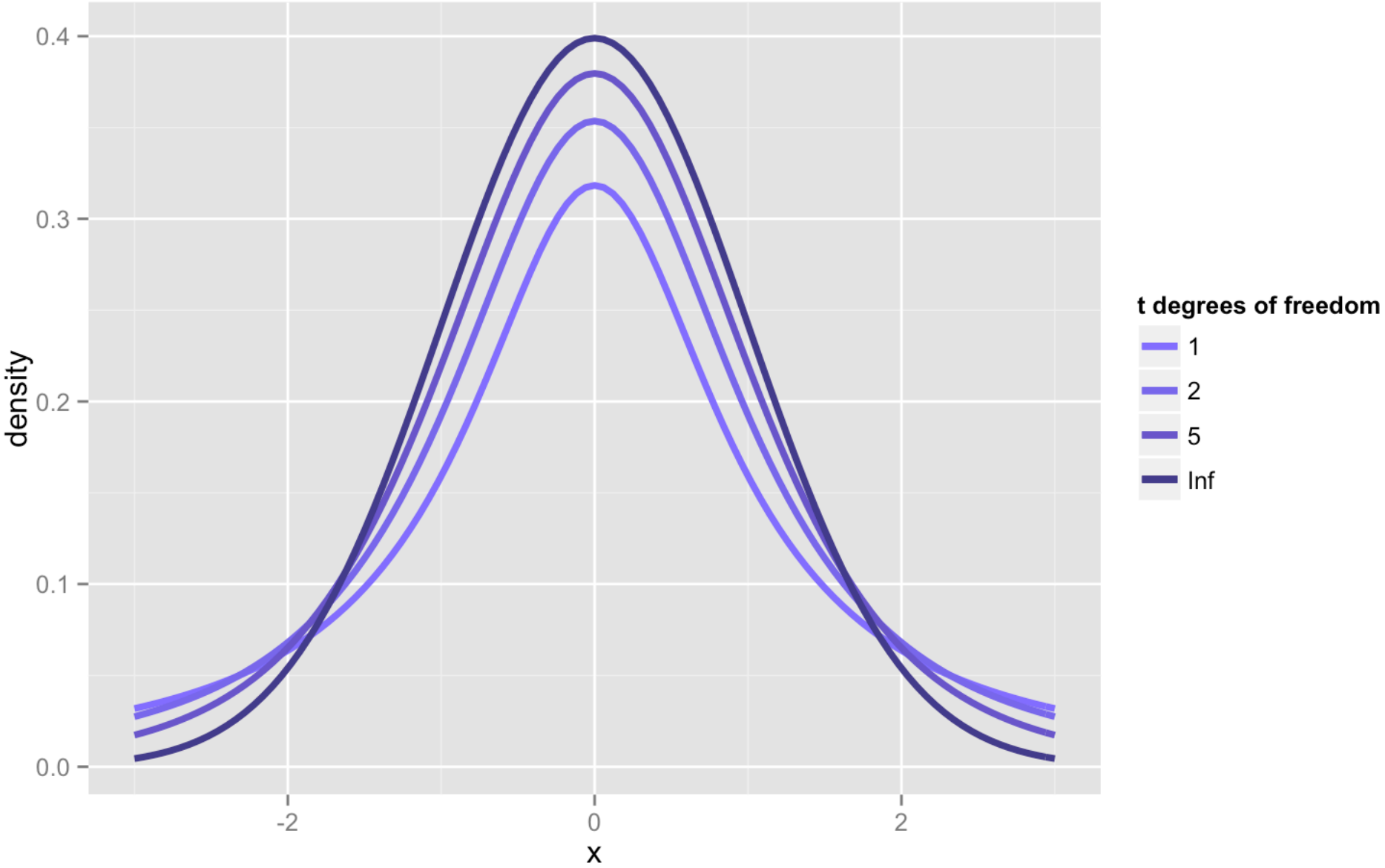
One-sample *t*-test

- Does sample perform differently than general population (known: μ, unknown: σ)?

$$t_{obs} = \frac{\bar{Y}_i - \mu}{s / \sqrt{n}}$$

- Now what?
  - Determine $t_{critical}$ values for your α and *df*
  - Must "beat" that value to reject null

New distribution family: student's t

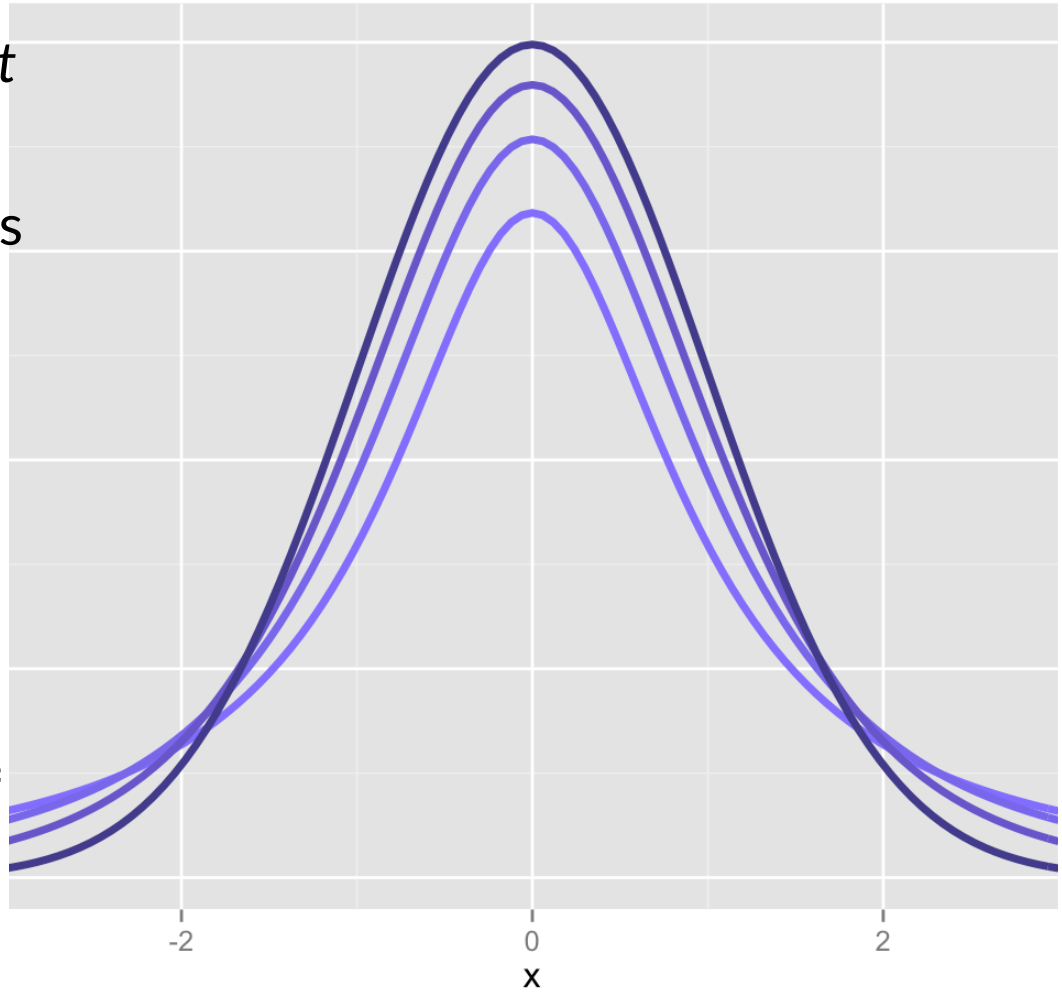


The *t*-distributions: PDFs

- Symmetric (skewness = 0)
- Bell-shaped, but notice that the *t* always has always has relatively more AUC in the tails vs. the unit-normal, and unit-normal has relatively more scores in the center; thus t-distribution is leptokurtic

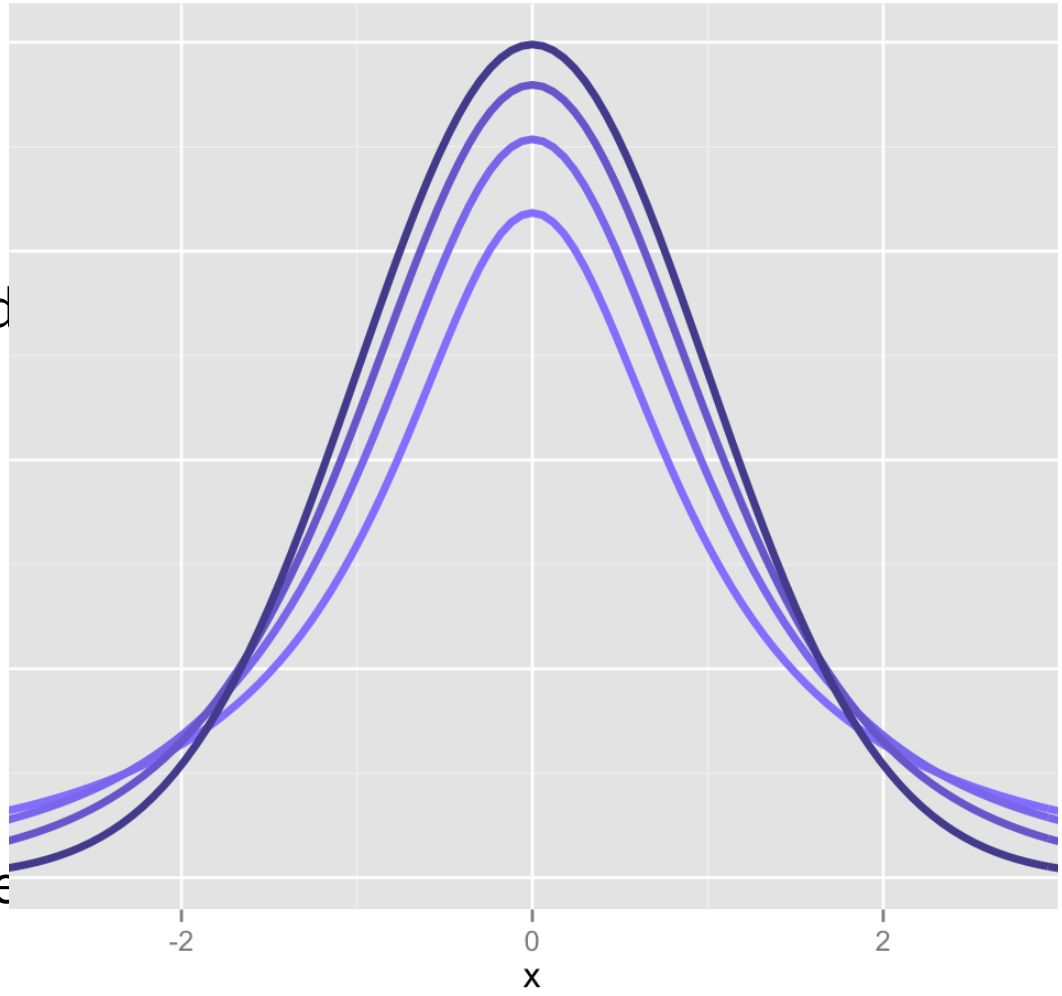
$$= \frac{3(df-2)}{df-4}$$

- Kurtosis is undefined for *t*-variables with $df < 4$
- On your own, at what *N* and *df* is the kurtosis of a *t*-variable = 3 (= to kurtosis of normal distribution)?



The *t*-distributions

- Let T denote a random variable with a *t*-distribution with *df* degrees of freedom. Then:
  - *E*(*T*) = 0; same as unit-normal
  - Var(*T*) = *df*/(*df*-2); more spread out than unit-normal (↑ variance)
- As *df* increases, the *t*-distributions converge to the unit normal.
- *t*-distributions will be useful for statistical inference for one or more populations of quantitative variables.



Using the sample estimate of the variance

- Recall the sample estimate of the population variance:

$$s_Y^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y}.)^2}{n-1}$$

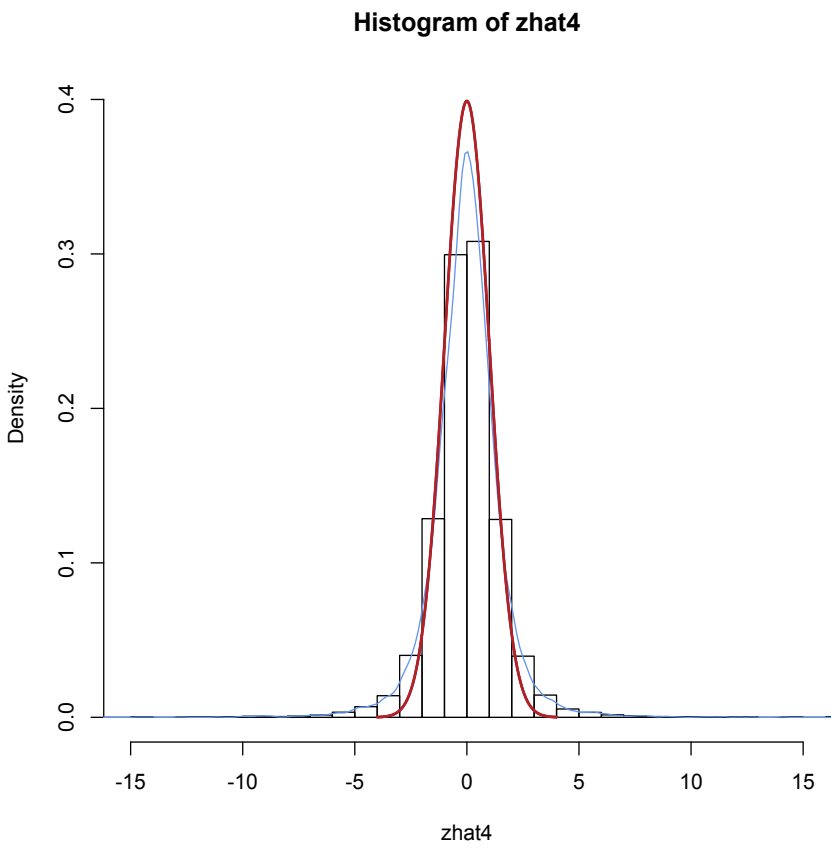
- The t-statistic formula is:

$$t_{obs} = \frac{\bar{Y}_i - \mu}{s / \sqrt{n}}$$

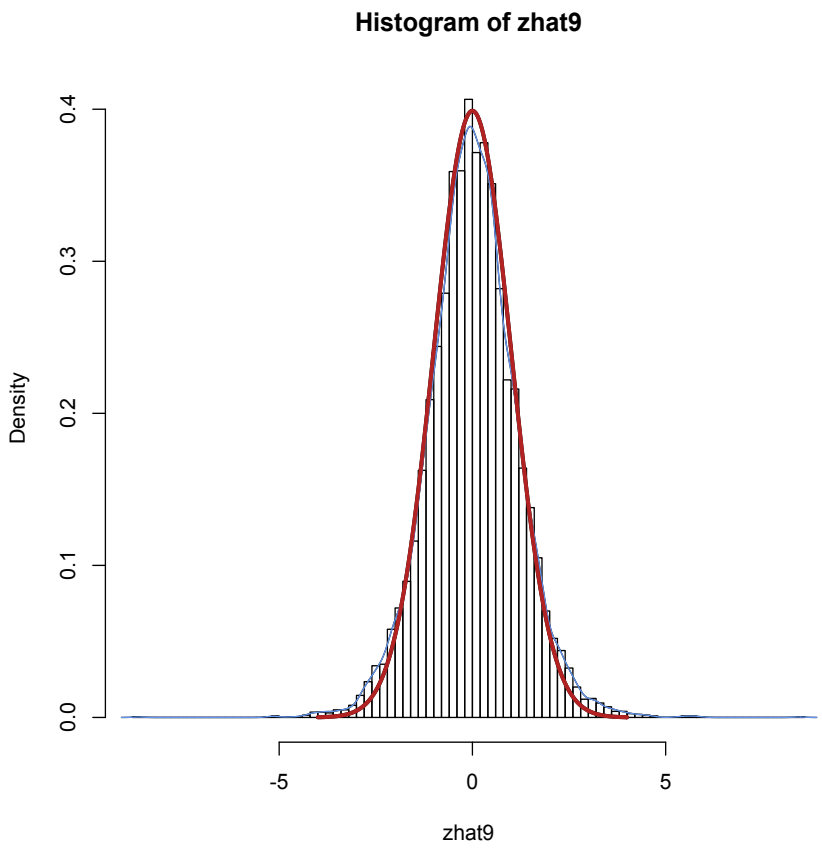
- At this point, you may be worried about the effect of variability of the variance estimate
- You would be right: the random variable, *t*, is no longer normally distributed
  - It has a unique distribution: the *t*-distribution, with *n-1* degrees of freedom
  - The *t*-distribution is *asymptotically* normal

Red = z, blue = t

n = 4 (×10,000 simulations)

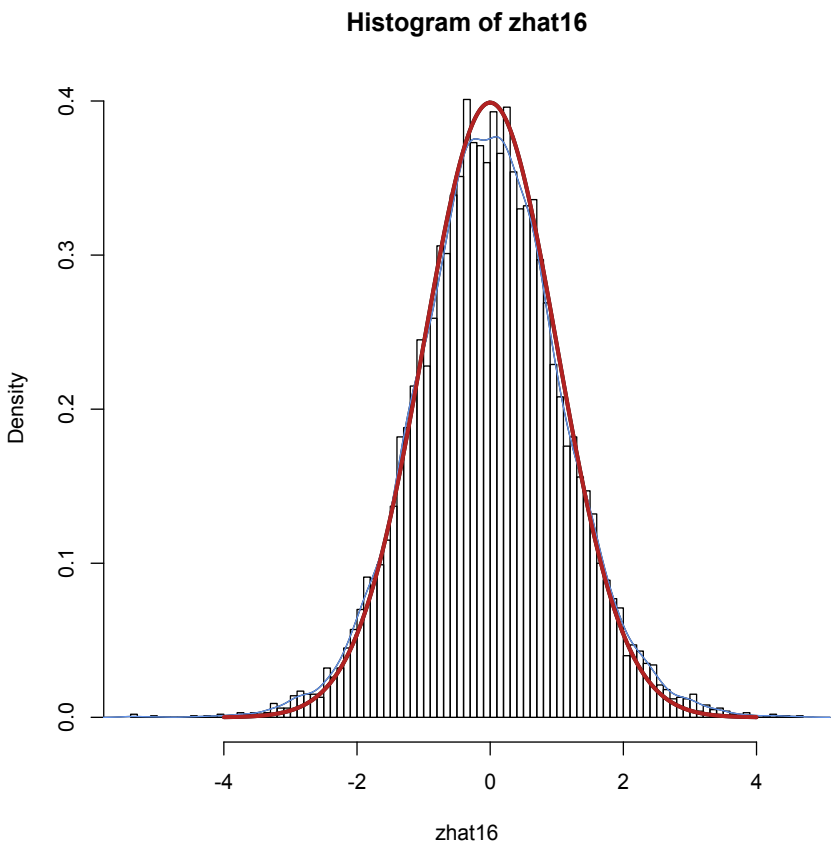


n = 9 (×10,000 simulations)

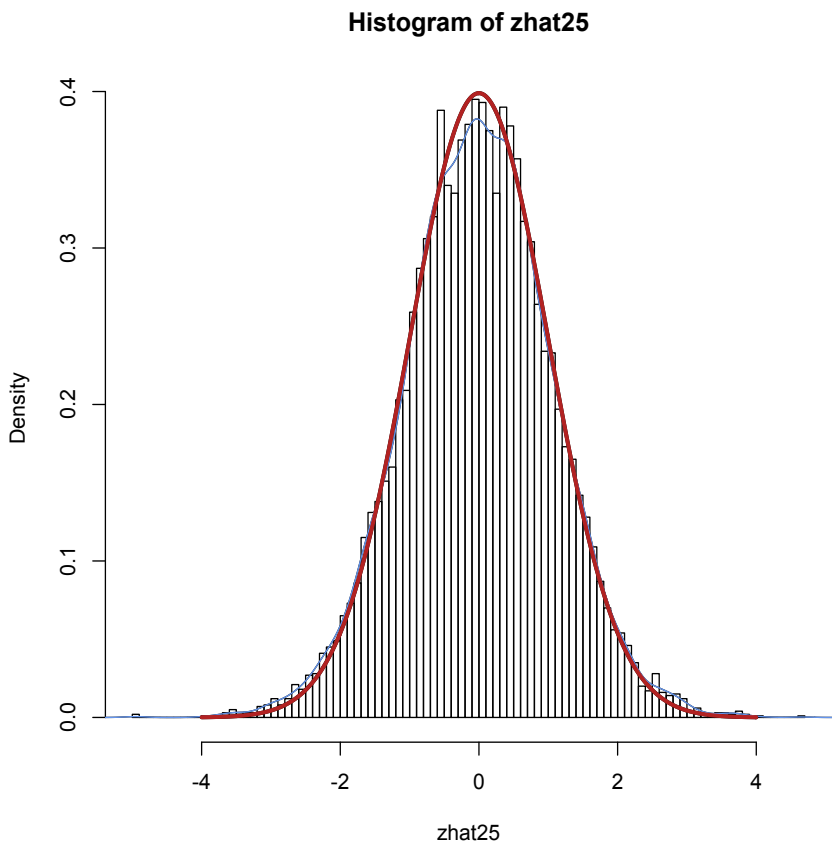


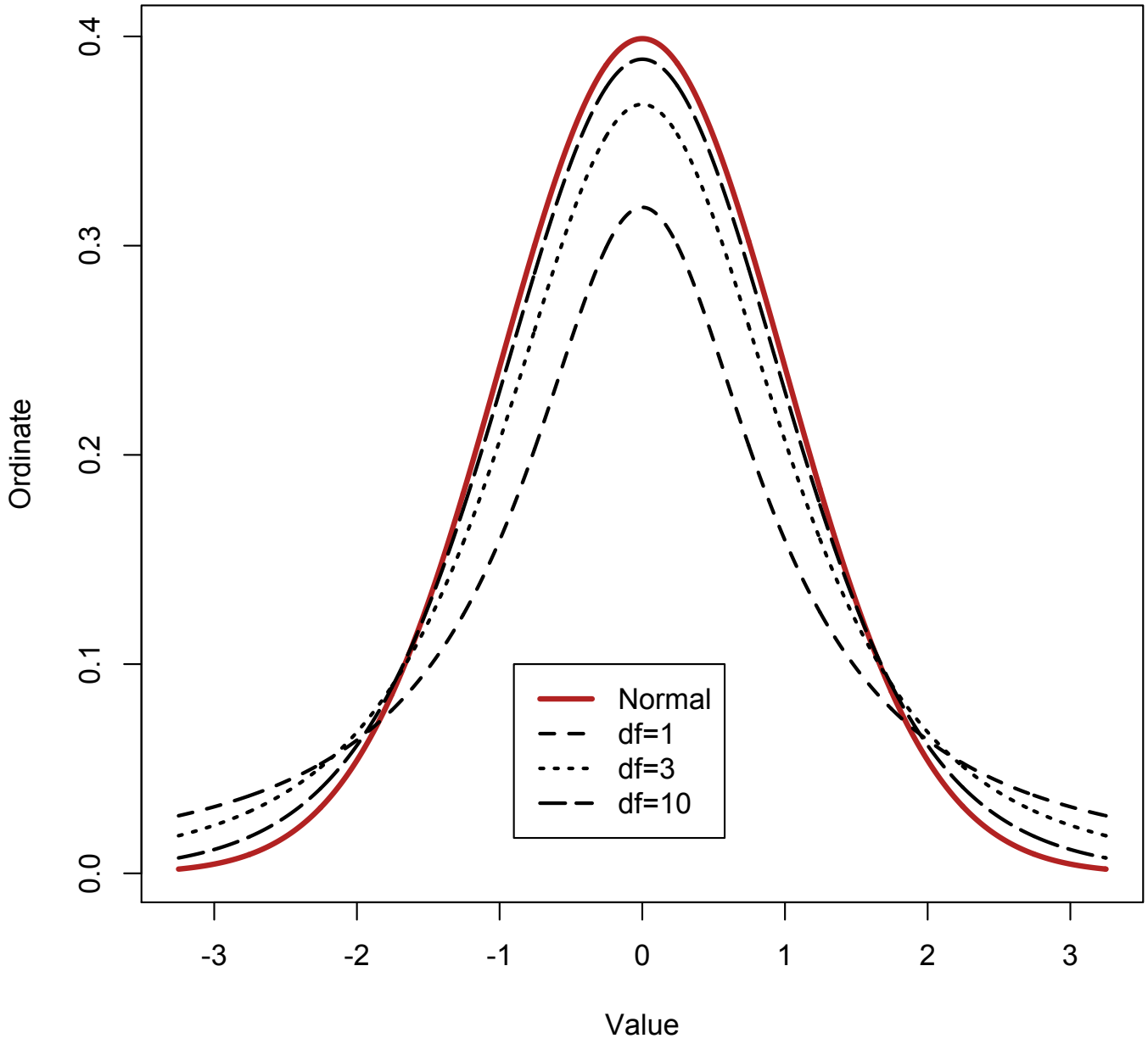
Red = z, blue = t

n = 16 (×10,000 simulations)



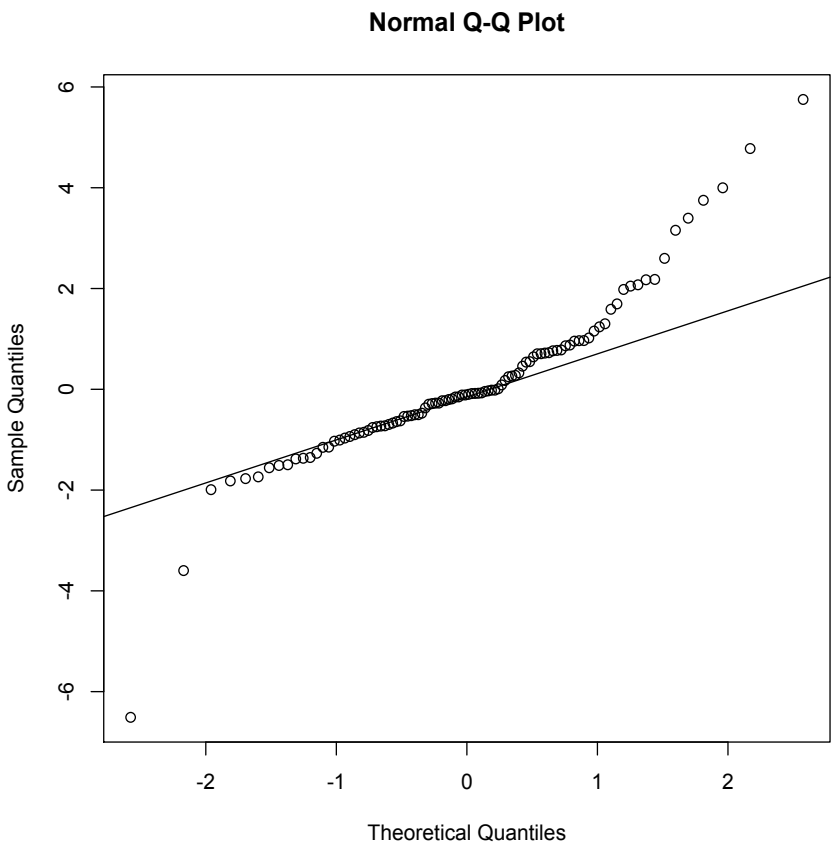
n = 25 (×10,000 simulations)



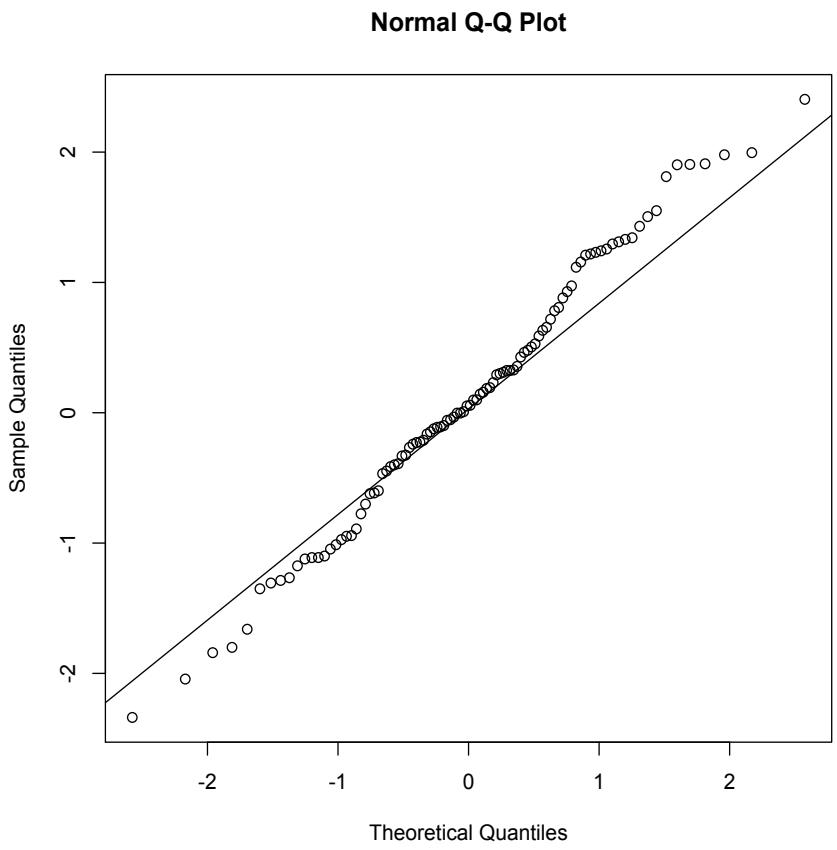

Ordinate
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
Value
Normal
df=1
df=3
df=10

Q-Q Plots of rt(100,df)

t with *df*=4



t with *df*=30



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
```

Family of *t*-tests

- One-sample t-test for a single mean
- Two-sample t-test (independent samples) for comparing 2 means
- Two-sample t-test (correlated samples; dependent samples) for comparing 2 means with correlated or repeated measures

Interpreting significance in NHST

- Non-significance ≠ true null hypothesis
  - Suppose we fail to reject the null because $p > .05$
  - We cannot assume then that null is true, but merely that we lack sufficient evidence to reject it
- It is all too easy to find non-significant results by conducting…
  - A poor study…
  - With poor measures…
  - And low power.
- Consider: "Not guilty" verdict in a jury trial- prosecutor may have failed to present a strong case
  - Not guilty ≠ innocent
- If you truly want to support the null, look into equivalence testing/tests of close fit

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??

Two-tailed CI

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?
```

Rationale for one sample T test

- We collect sample data and calculate a sample mean.
- If this sample come from the population we think it came from, then we would expect the sample mean to be approximately equal to the population mean.
- Although they may differ by chance, we would expect large differences between sample means to happen infrequently.
- Under the null hypothesis, we assume no difference in means.
- We compare the the sample mean to the population mean to see if the difference is more than we would expect to get by chance under the null hypothesis.

Rationale for one sample T test

- At alpha=.05, we therefore seek evidence that there is only a 5% chance that the magnitude of the difference we observe is consistent with what we would expect based on chance alone.
- **Remember:**
  The *p*-value does not tell you if the result was due to chance. It tells you whether the results are consistent with being due to chance. These two things are not the same.
- We use the standard error as the gauge of variability of the sample mean. If it is small, we expect most samples to have very similar means. If it is large, then large differences in sample means are more likely.

Rationale for one sample T test

- If the difference between the sample means is larger than we would expect based on the standard error, then we know that one of two things has happened:
  - **We made a mistake (boo!).** There is no difference between the population and our sample means fluctuate a lot. By chance, we have collected a sample that is atypical of the population we drew from. Here, the difference is a fluke and the null is true; we incorrectly reject the true null hypothesis and thus commit a Type I error.
  - **We made a discovery (yay!).** The sample comes from a different population, that may be typical of THAT population. Here, the difference is genuine, and we correctly reject the null hypothesis.

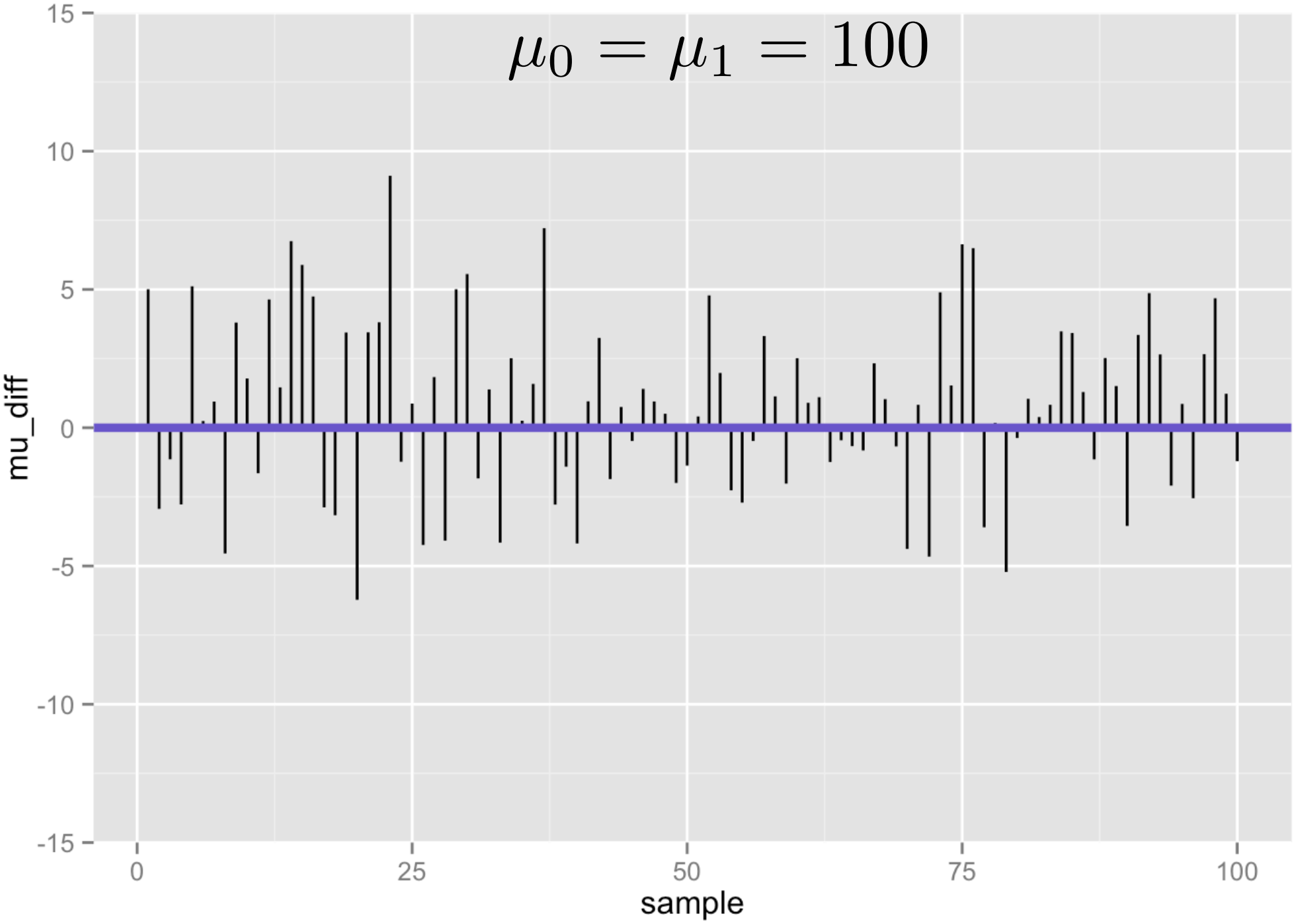
When the null hypothesis is true

$$\mu_0 = \mu_1 = 100$$

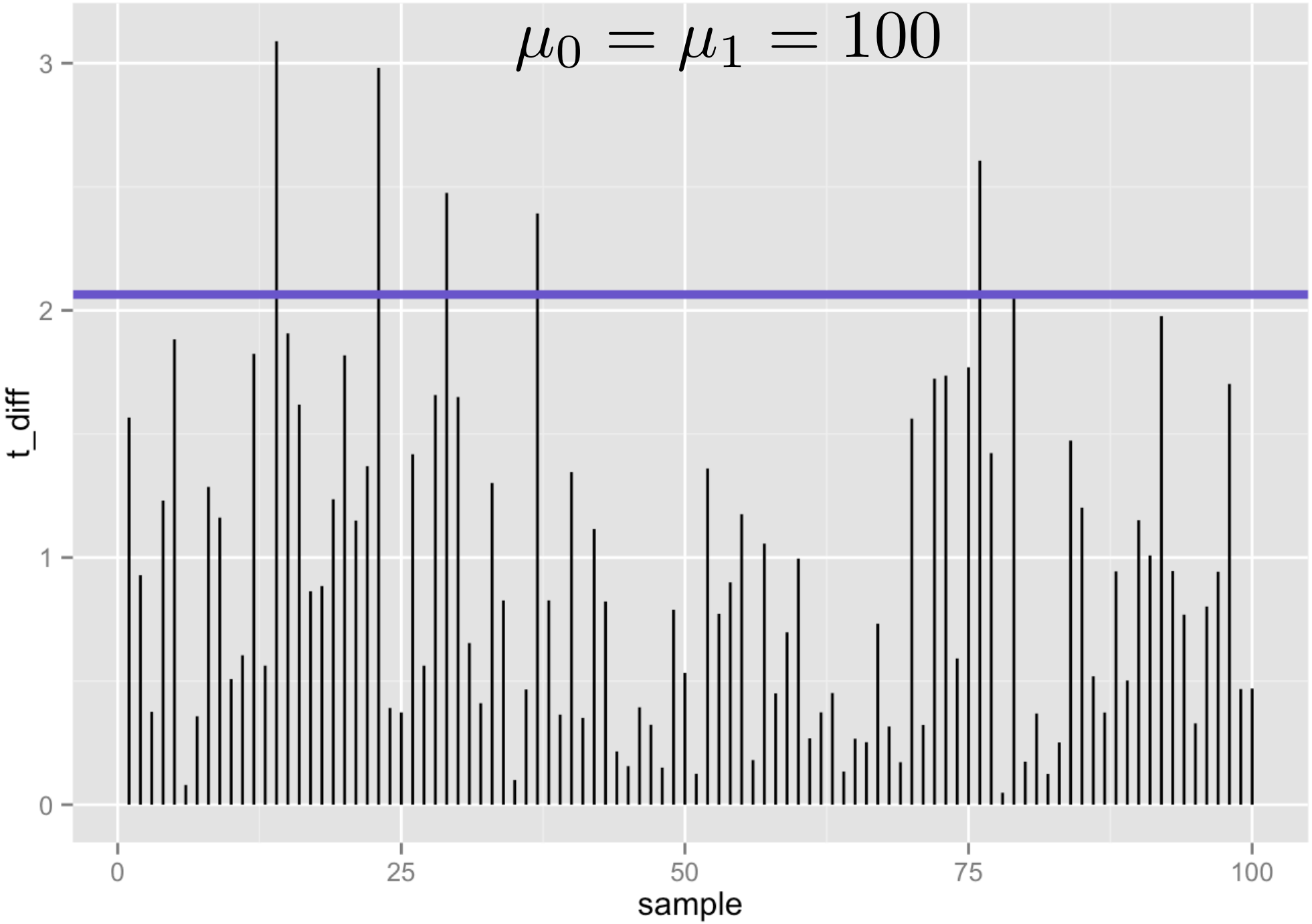
1-sample t-test

n = 25

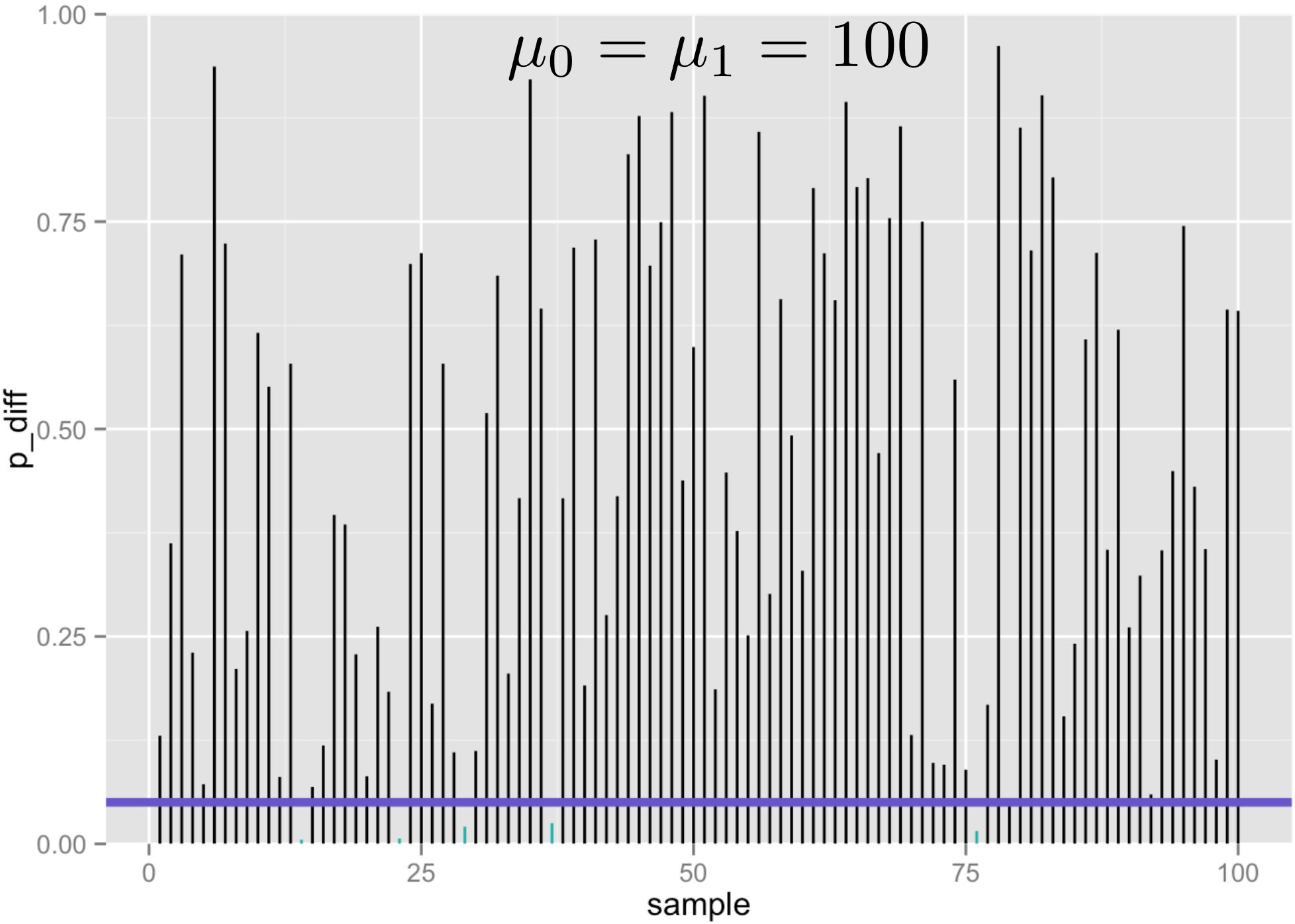
Differences between population mean and sample mean (100 samples) when null is **true**



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)



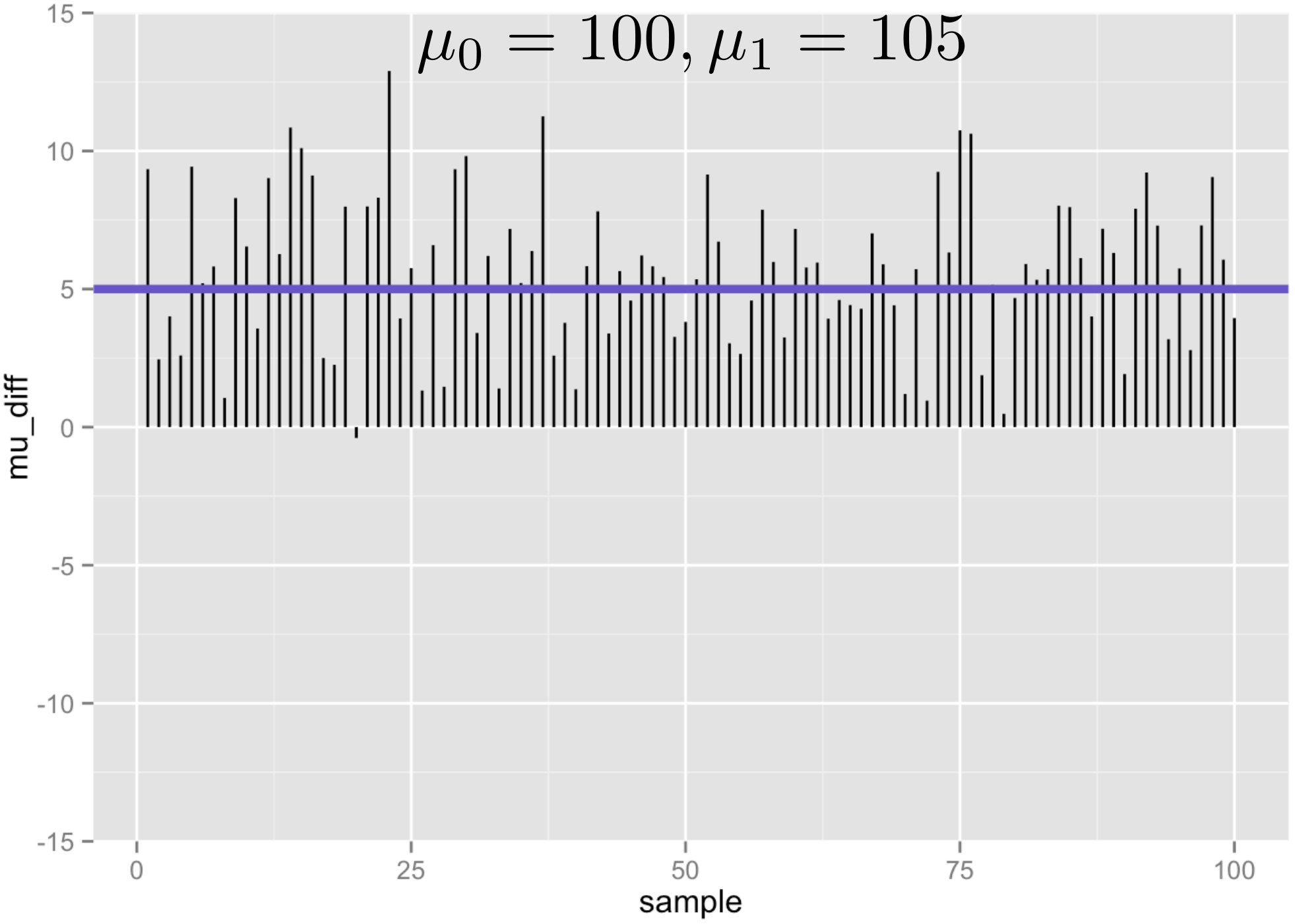
What if:
the null hypothesis is FALSE?

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

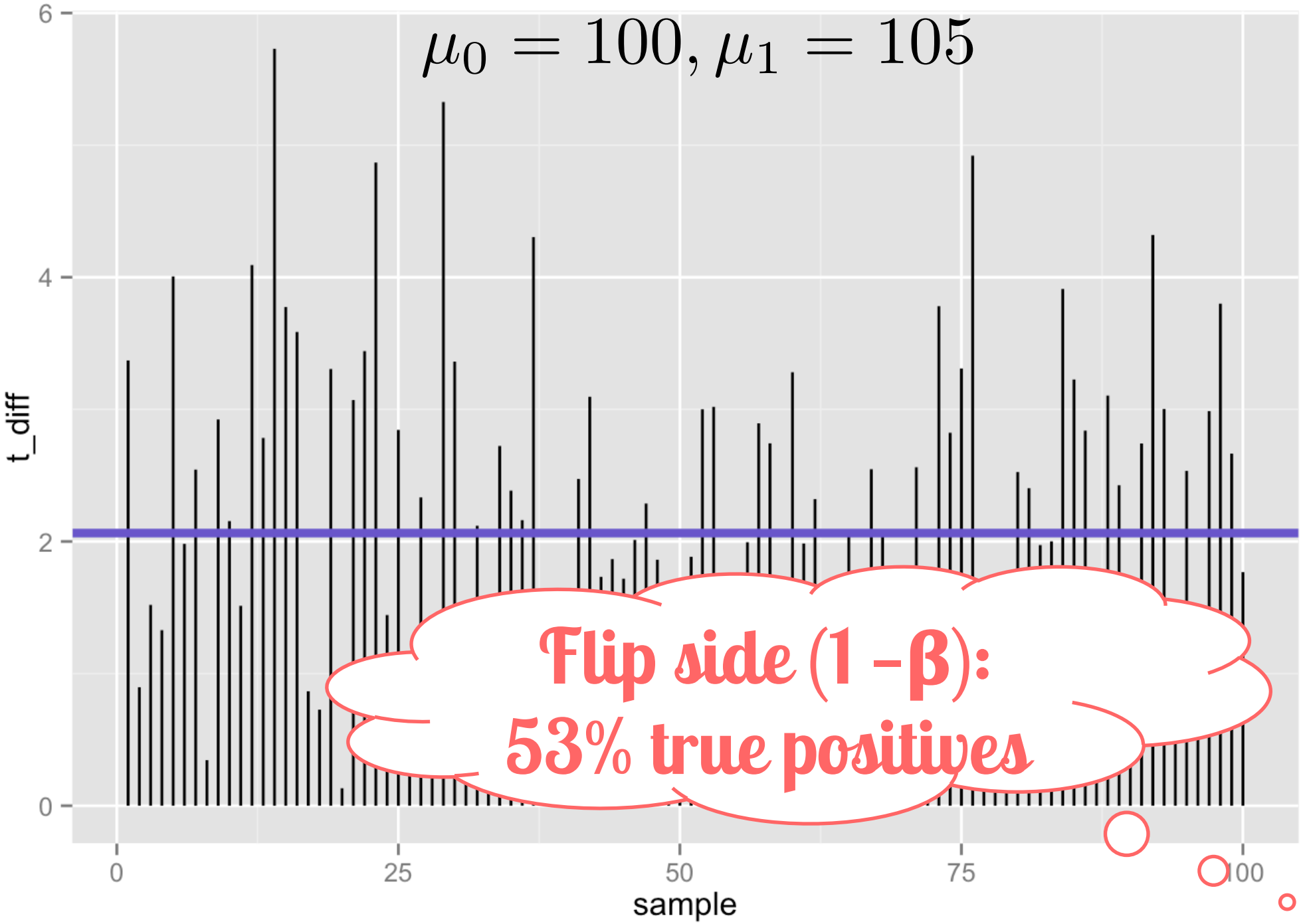
1-sample t-test

n = 25

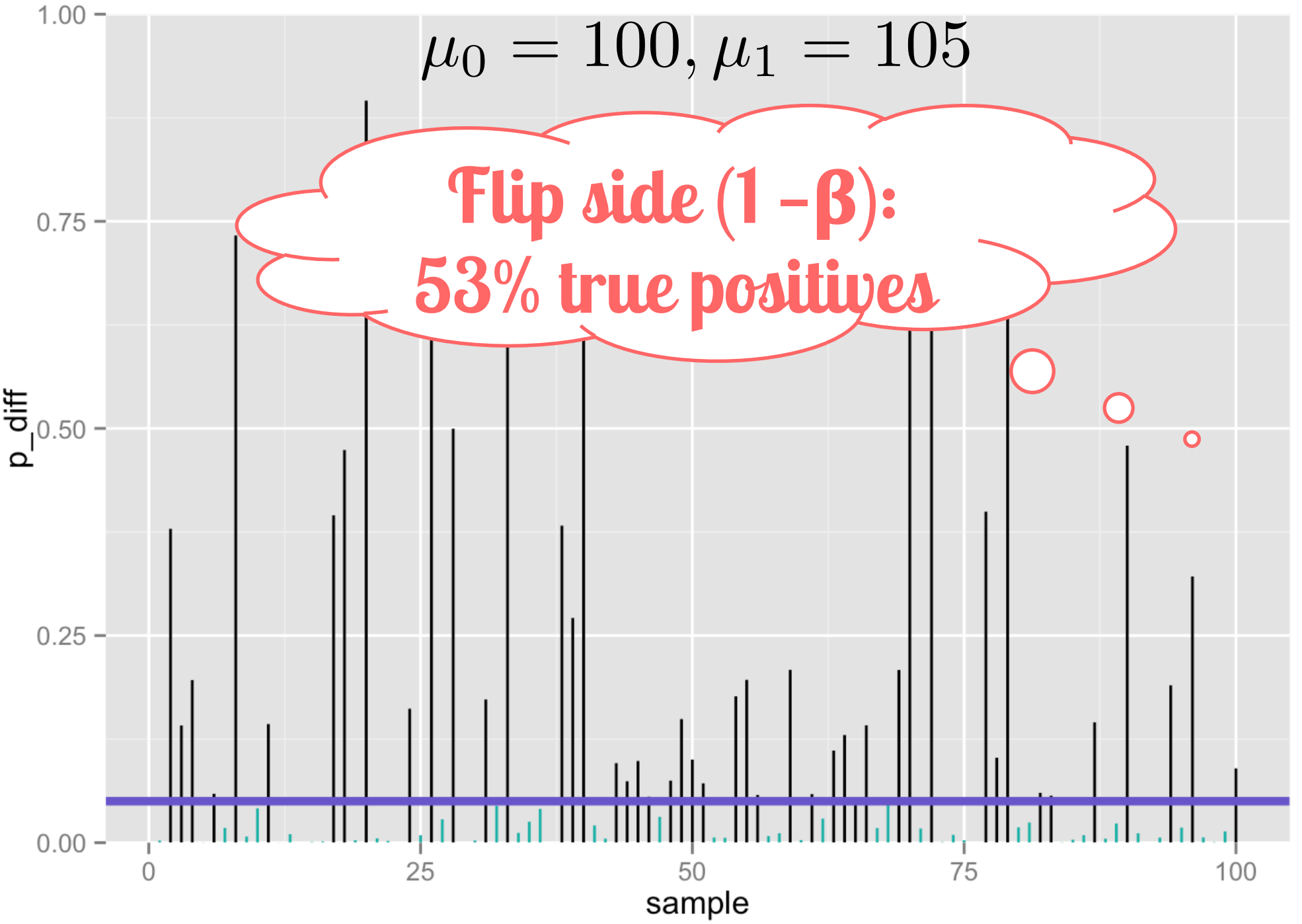
Differences between population mean and sample mean (100 samples) when null is false



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



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