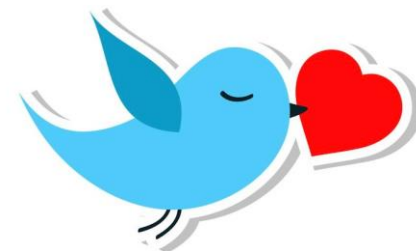# Math 530/630: CM 4.4
Classical Inference:
Hypothesis Testing and Confidence Intervals

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

General idea for an estimator

- Sample estimates are our "best guesses" for population parameters
- Such a best guess is called a **point estimate**. We know, however, that point estimates — which are sample statistics, like the sample mean x — vary from sample to sample.
- Interval estimate, i.e. **confidence interval**, provides a set of possible values for that parameter value that are "compatible" with the data
- Construction of confidence interval requires knowledge of the estimator's distribution
  - Known distribution
  - Simulated distribution (e.g., "bootstrapping")

Bootstrapping

- **Idea:**
  Simulate the population sampling distribution by repeatedly selecting samples (with replacement) and computing the desired statistic.

- **Motivation:**
  Since the sample came from the population of interest, it can be used to estimate that population, and determine the Confidence Interval.

Confidence Intervals

Typically, a confidence interval takes the form:

point estimate ± a margin of error

We'll come back to this…

General idea for hypothesis testing

- Partition the *parameter* sample space into two regions
- One region is defined by the **null hypothesis**: H_0
  - Usually what you wish to see evidence against
  - If the parameter estimate based on my sample is **null/dull**, then it will fall in this region (given some reasonable amount of random variation)
- Other region defined by the **alternative hypothesis**: H_1
  - Usually what you wish to see evidence in support of
  - If the parameter estimate based on my sample is **interesting**, then it will fall in this region (and thus, outside of the null/dull region)
- This is traditional null hypothesis significant testing (**NHST**)
  - Also called reject-support hypothesis testing

General idea for a test statistic

- A numerical summary used to collapse sample data into a single number
- When "big" or "extreme", suggests that the observed data is very unexpected under the null hypothesis H_0
- A p-value quantifies this incompatibility between the data and H_0 – specifically, it's a tail probability
- So, to get a p-value, you must know or approximate the probability distribution of the test statistic under the null H_0
- This means we need to know what the *null* sampling distribution looks like

Therefore…

- To complete a hypothesis test or convey the precision of a point estimate, we need the statistic's or estimator's *sampling distribution*

- "sampling" here should invoke "long-run", "hypothetical repeats of the experiment"

- For an estimator, the standard deviation of the sampling distribution is called the *standard error*

Process

1) Write down **null** and **alternative** hypotheses
2) Figure out good test statistic that estimates the parameter you are interested in (i.e., what numeric summary based on your sample captures what you want)
3) Work out null distribution for that statistic (known or simulated)
4) Calculate p-value (or critical value) by comparing actual value to null distribution
5) Reject H_0 if probability (p-value) $< \alpha$
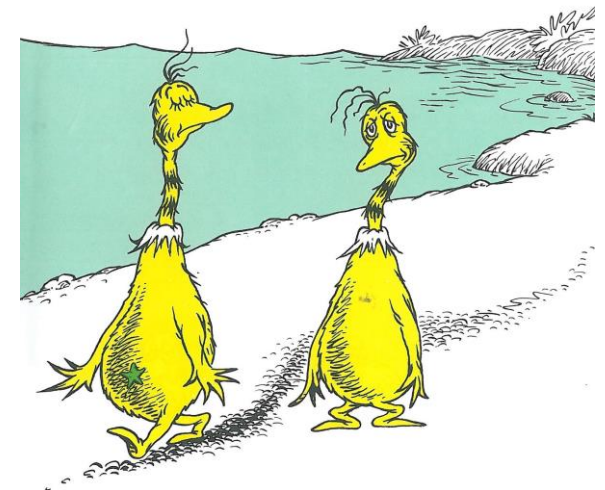
Hypotheses

- Null hypothesis = H_0

- Alternative hypothesis = H_a / H_1

HINT: we have to know the null distribution, so H_0 will always be of the form:

Some parameter = some value as in…

$\mu = 0$

The null distribution

- If we know the population distribution, we can "construct" the null distribution: the distribution of a test statistic if the null hypothesis is true
- If we don't know the population distribution, we can simulate it using resampling
  - In the *two sample* scenario, this means we resample observations from our sample data without replacement, each time reshuffling the "sample" or "group" assignment (a la Sneetches).

Permutation testing

- **Idea:**
  Simulate the sampling distribution under the null hypothesis by shuffling the group labels repeatedly and computing the desired statistic.

- **Motivation:**
  If the labels really don't matter, then switching them shouldn't change the result! We do this when we want to test whether observations from two groups follows the same distribution, without making any assumptions about the shape of that distribution (e.g., normality).

Building Intuition – 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.
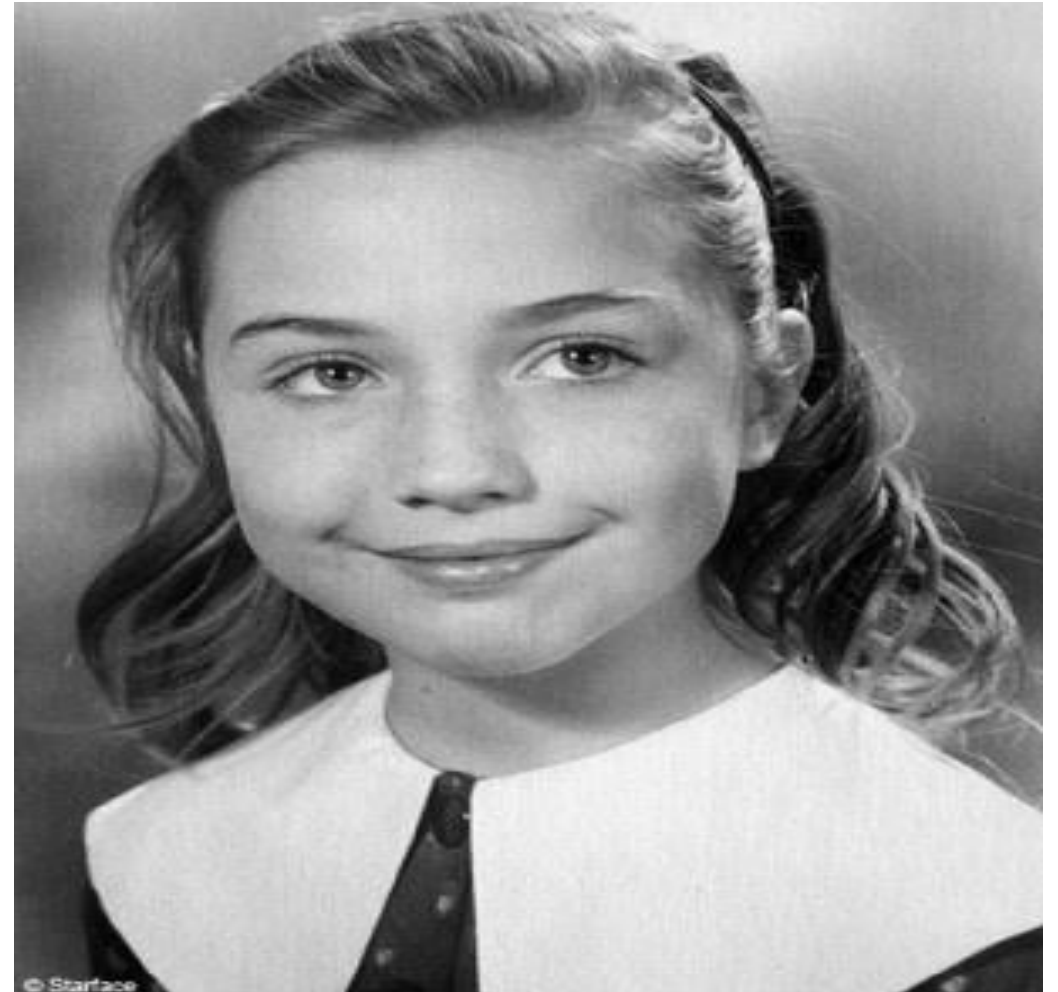
Let's pretend

- 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
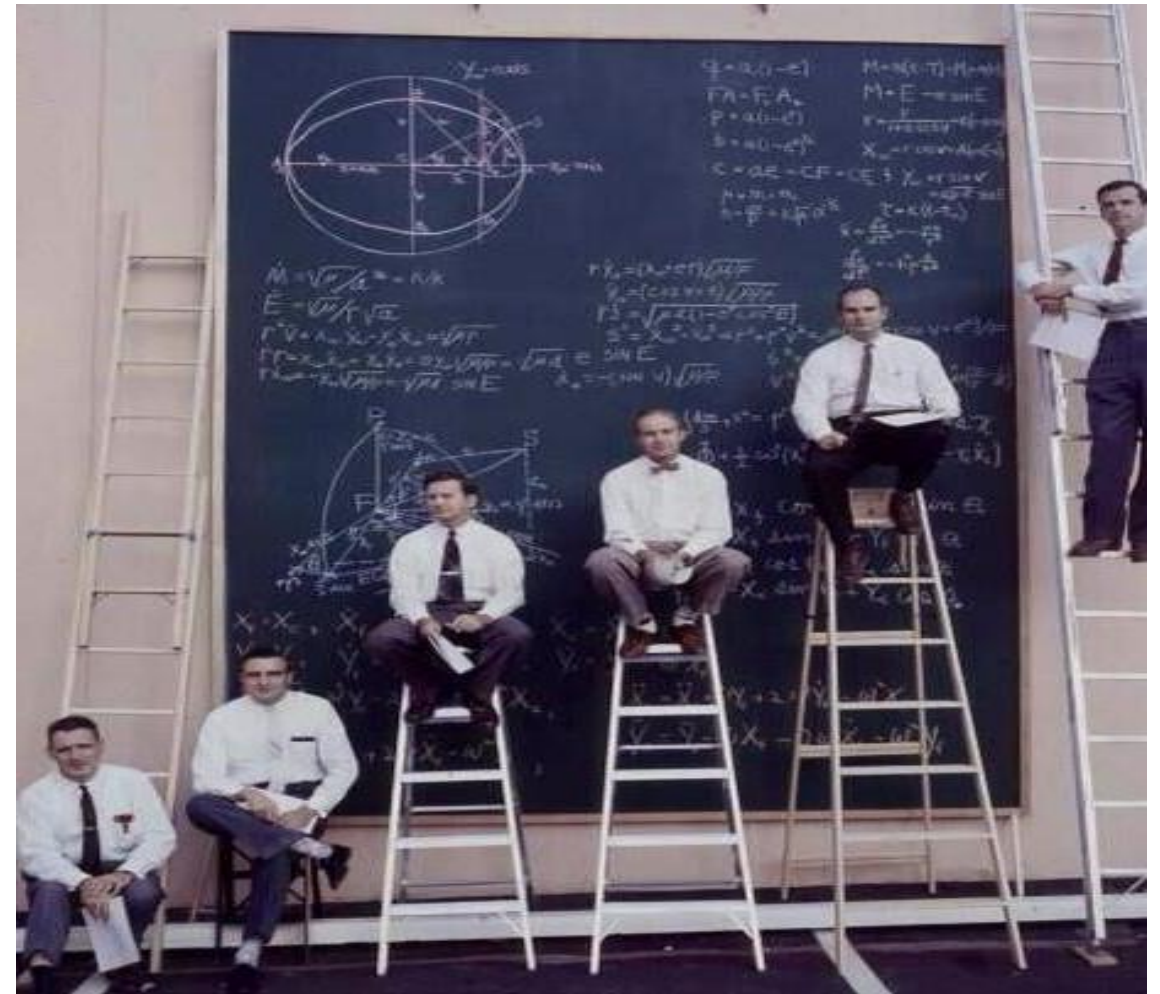
The (fictitious) study

- 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

NASA Budget, in millions of dollars, adjusted to 2007 values.
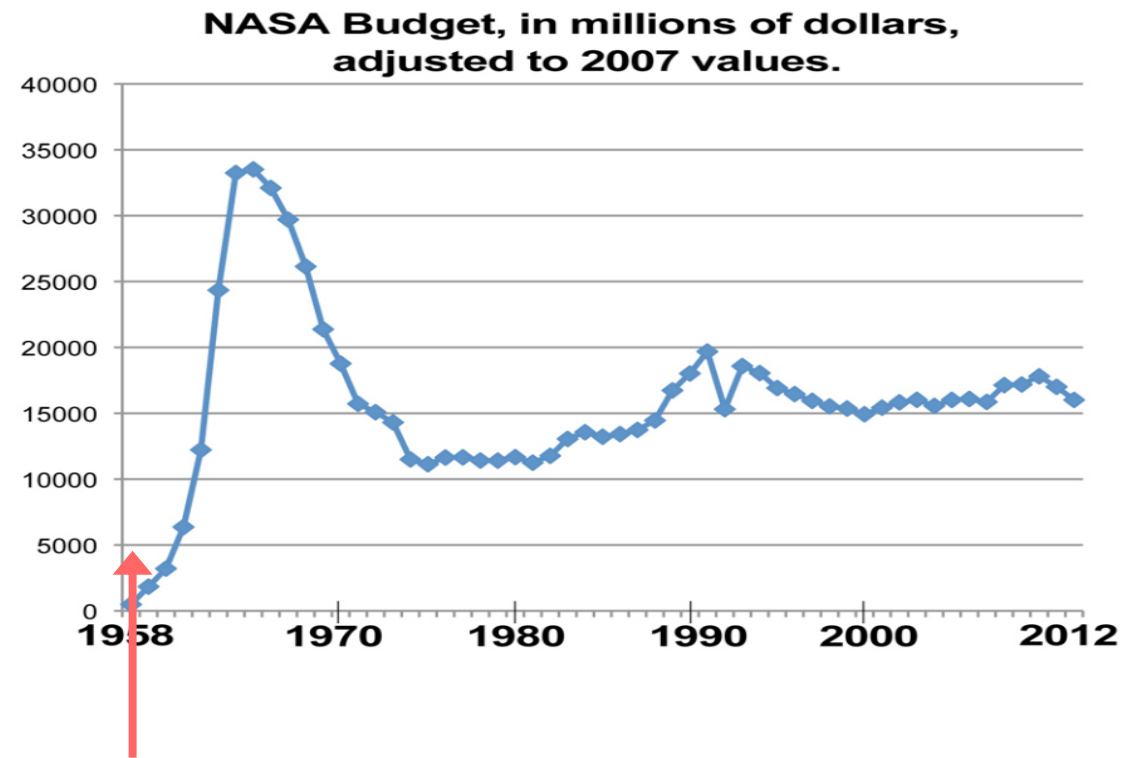
A (directional) alternative hypothesis

- H_0:

  AA will have lower or comparable IQs;

  $$\mu_{aa} \leq \mu_0$$

  $$\mu_{aa} \leq 100$$

- H_1:

  AA will have higher IQs;

  $$\mu_{aa} > \mu_0$$

  $$\mu_{aa} > 100$$

Rearrange…

- H_0:

AA will have lower or comparable IQs;

$$\mu_{aa} \leq \mu_0$$

$$\mu_{aa} - \mu_0 \leq 0$$

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

- H_1:

AA will have higher IQs;

$$\mu_{aa} > \mu_0$$

$$\mu_{aa} - \mu_0 > 0$$

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

Obtaining a test statistic

We know the population mean and standard deviation.

- $\mu = 100$
- $\sigma = 15$

A general formula for any test statistic about some generic parameter, θ:

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

?What are these values?

Let's pretend we live in a crazy 1960's world, and NASA interns calculated just the sample mean and not the standard deviation before they "lost" the data. Luckily, we know μ *and σ*.

What is the NULL distribution of this test statistic?

Before we even look at the data:

What is the mean and standard deviation (the "standard error") of the null distribution?
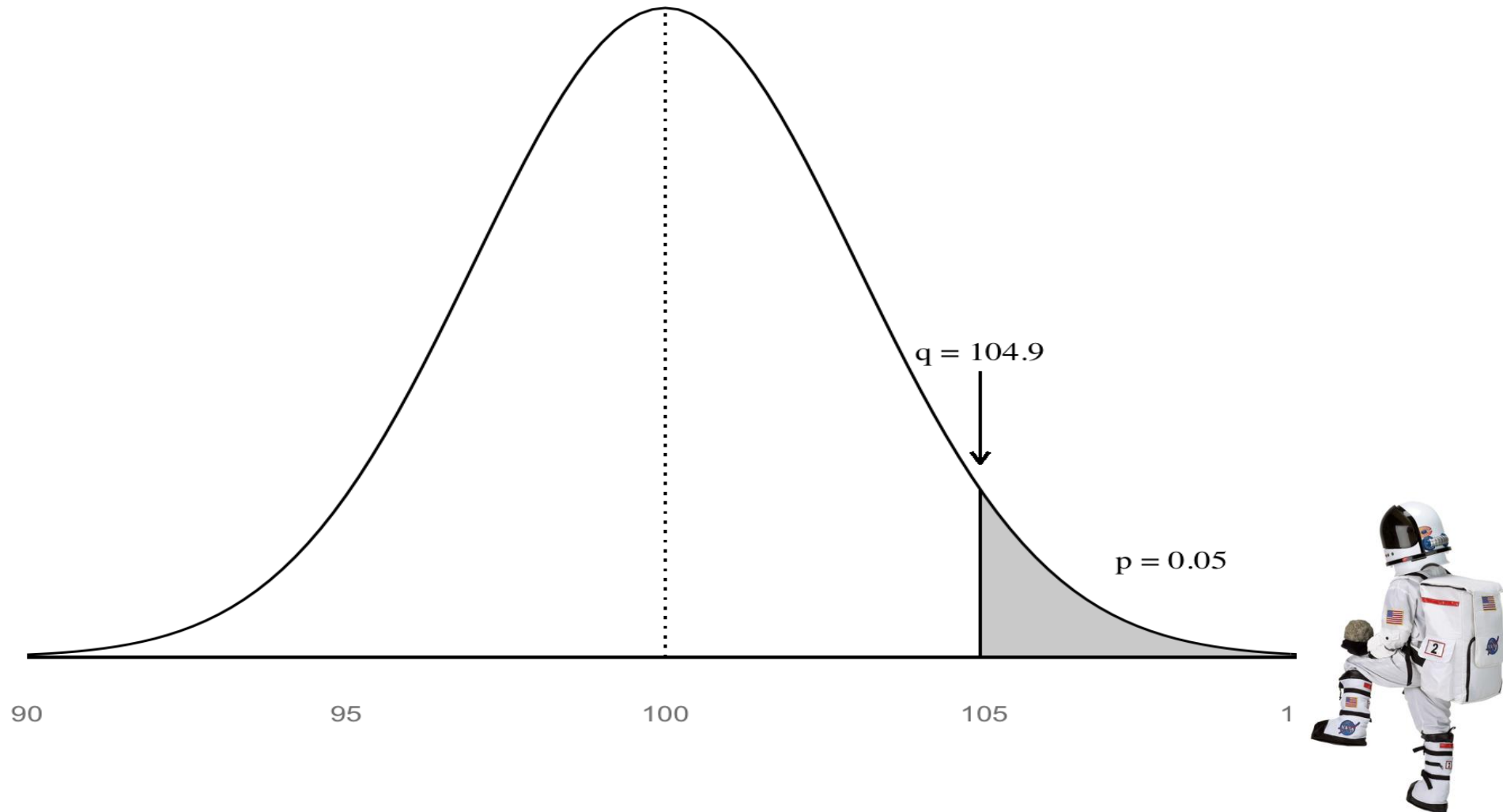
What shall we assume this distribution looks like? Draw it!

Let α = .05, 1- tailed: What value does our sample mean need to "beat" in order for us to conclude that it is higher than the population mean (given random variation present in our sample)? Draw this!

What would you conclude if our sample mean is 104? What if it is 108?

Your Turn

The null distribution for the z-statistic (normal)

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;

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

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

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

Before we even look at the data:

What is the mean and standard deviation (the "standard error") of the null distribution?

What shall we assume this distribution looks like? Draw it!

Let α = .05, 2- tailed: What value does our sample mean need to "beat" in order for us to conclude that it is different than the population mean (given random variation present in our sample)? Draw this!
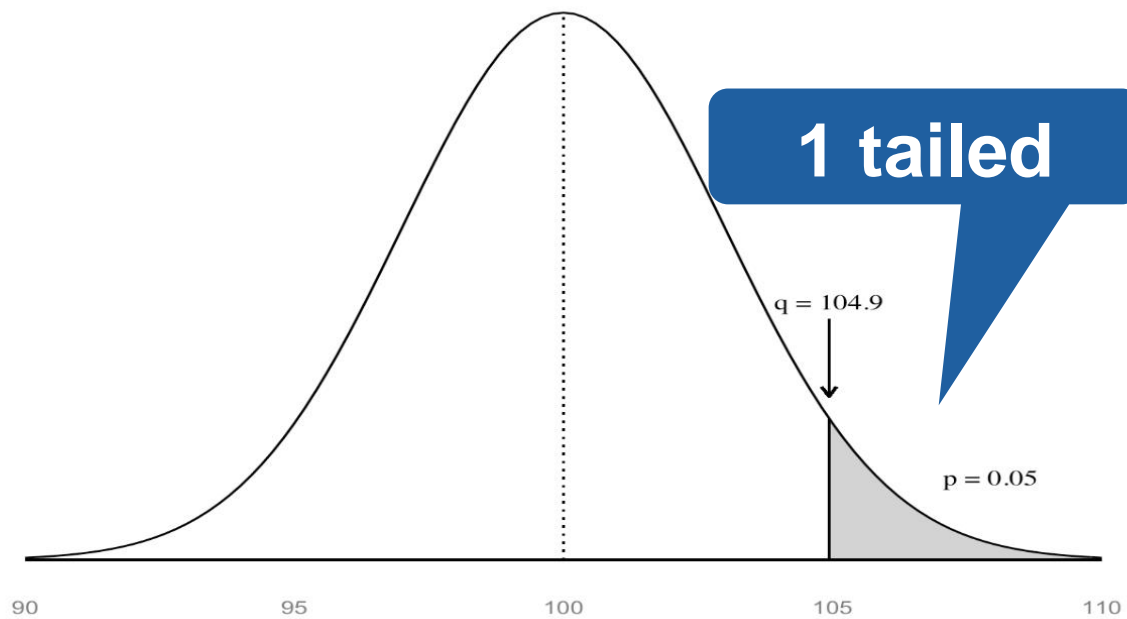
What would you conclude if our sample mean is 105?

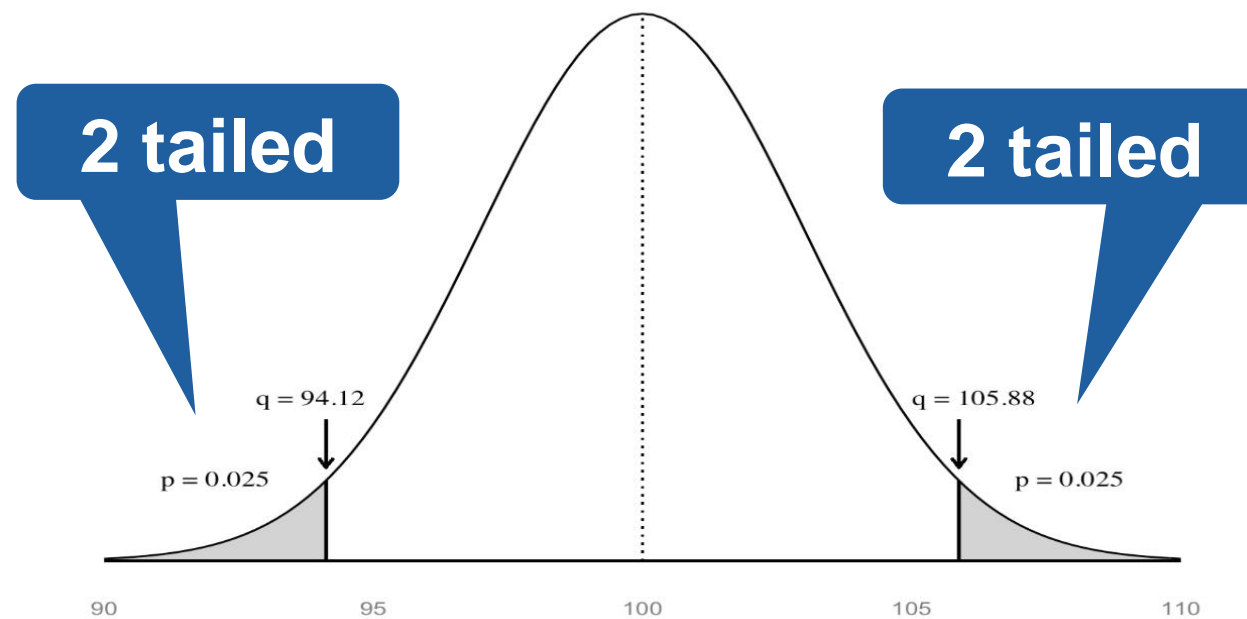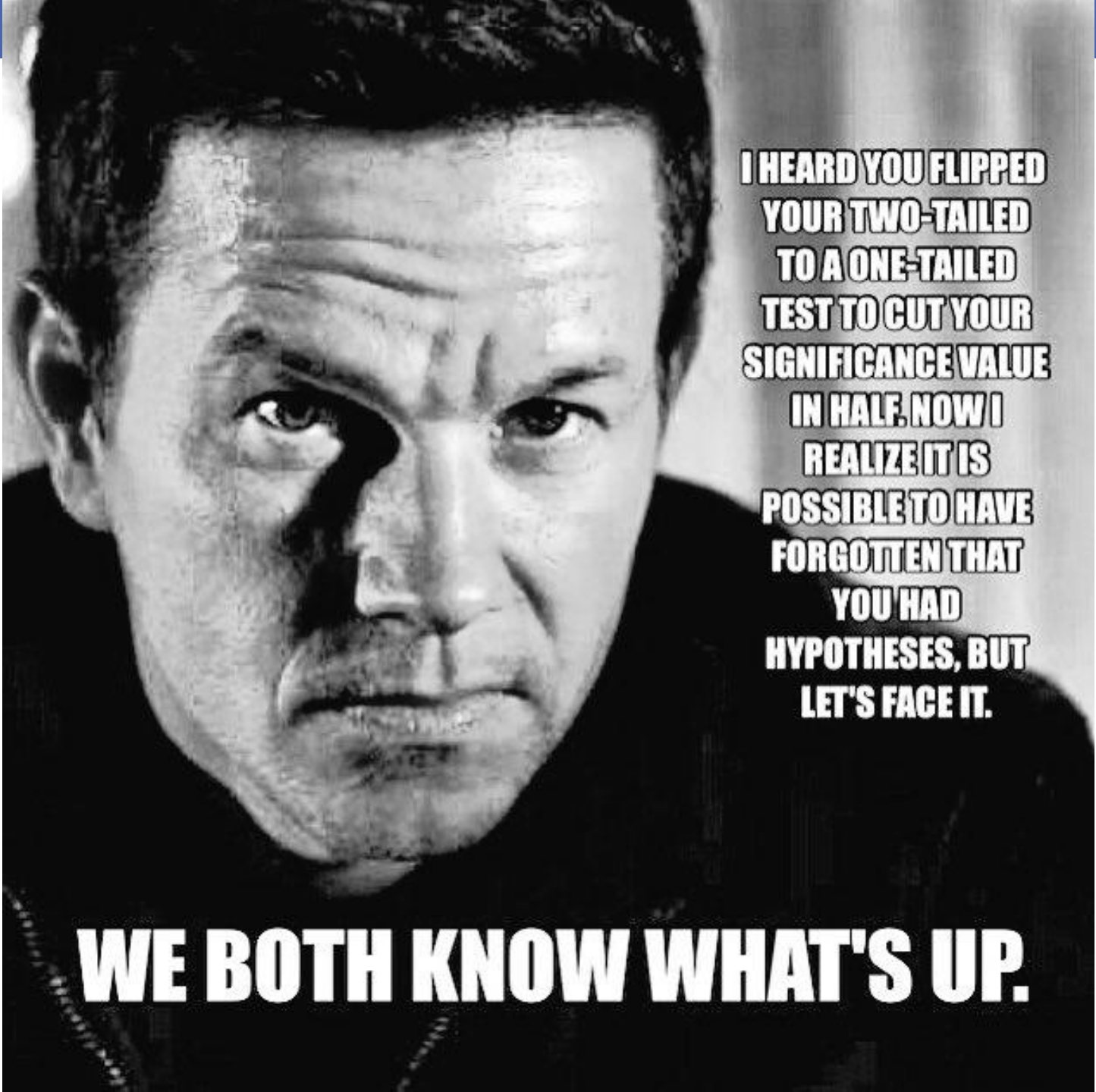Your Turn

The null distribution for the z-statistic (normal)
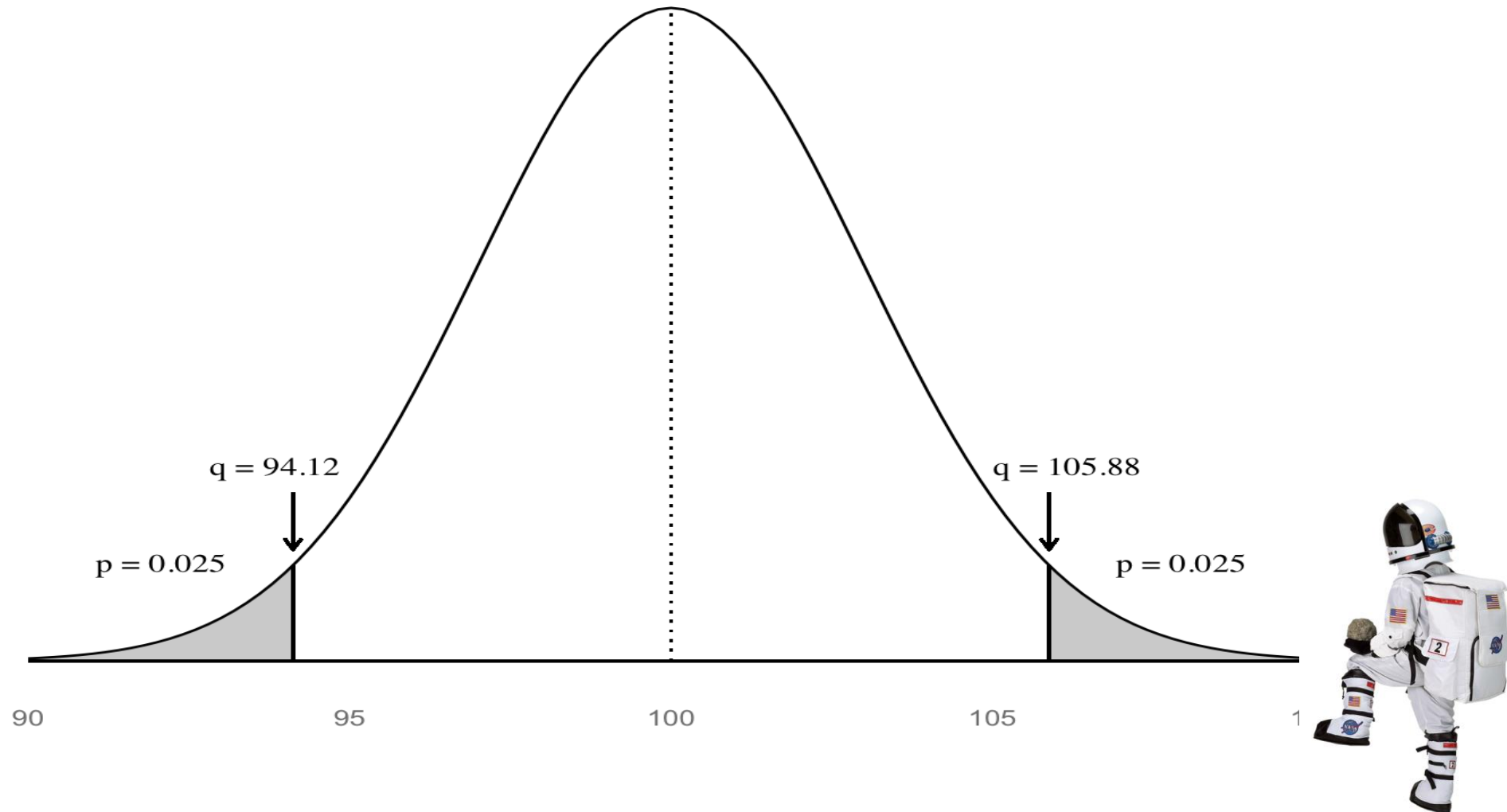
The alternative hypothesis

I HEARD YOU FLIPPED YOUR TWO-TAILED TO A ONE-TAILED TEST TO CUT YOUR SIGNIFICANCE VALUE IN HALF. NOW I REALIZE IT IS POSSIBLE TO HAVE FORGOTTEN THAT YOU HAD HYPOTHESES, BUT LET'S FACE IT.

WE BOTH KNOW WHAT'S UP.

One sample means z-test

And the sample mean is…

$$\overline{x} = 105$$

One-sample *z*-statistic

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

$$z_{obs} = \frac{\overline{Y}_i - m}{s / \sqrt{n}}$$

One-sample z-statistic

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

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

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

$$= \boxed{}$$

One-sample z-statistic

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

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

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

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

Z-statistic – Now what? Critical Values
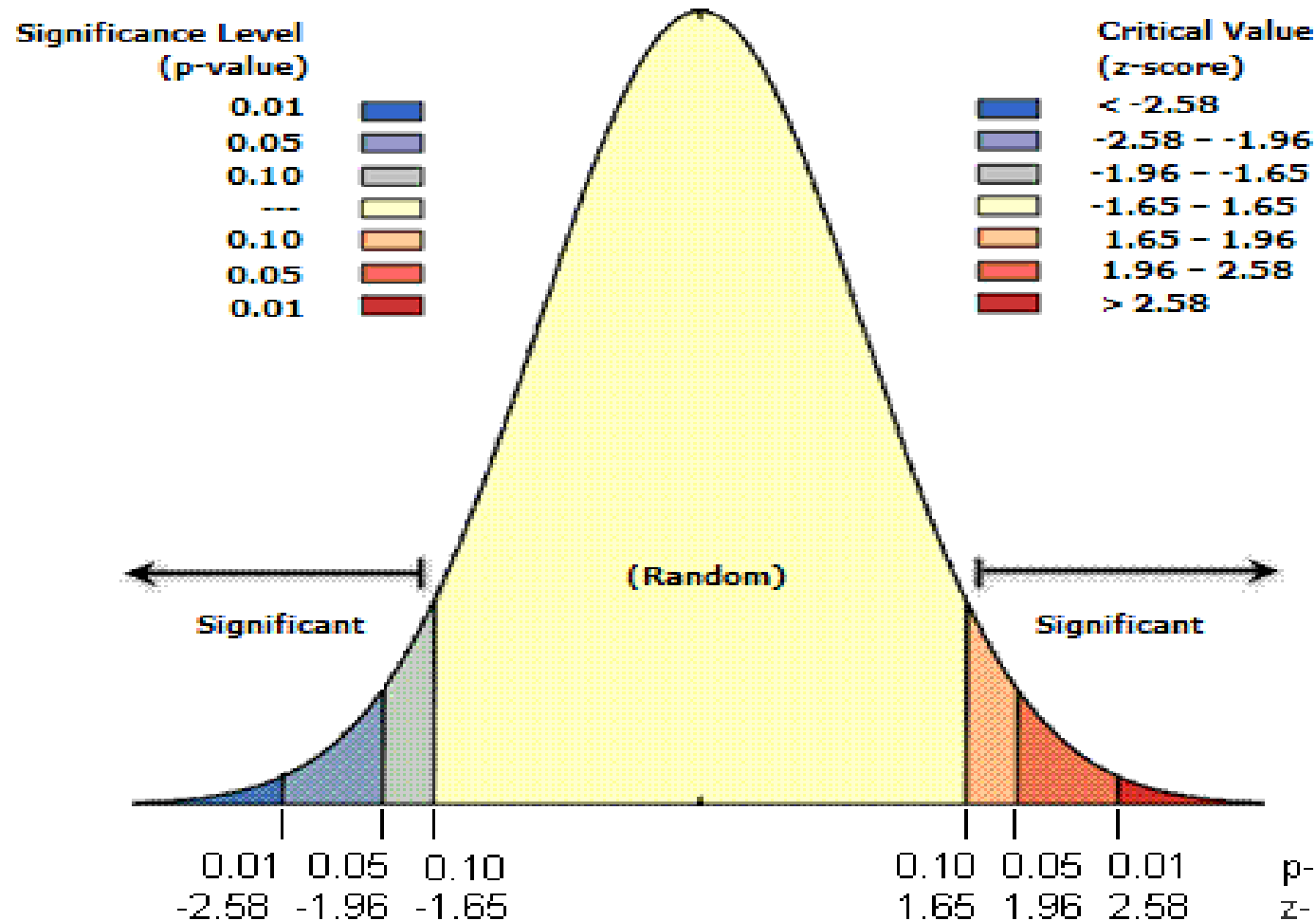
Determine $z_{critical}$ values for your α…

Must "beat" ±1.65 for α = .10, 2-tailed, to reject null
Must "beat" ±1.96 for α = .05, 2-tailed, to reject null
Must "beat" ±2.58 for α = .01, 2-tailed, to reject null

Not shown…
Must "beat" ±1.28 for α = .10, 1-tailed, to reject null
Must "beat" ±1.65 for α = .05, 1-tailed, to reject null
Must "beat" ±2.32 for α = .01, 1-tailed, to reject null

General idea for a p-value

- Probability under the null H_0 of observing a test statistic value as or more extreme than that computed from the observed data (your sample data)
- The p-value is **not** the probability that the null hypothesis is true
- Example in a two-sided test, i.e. when both very small and very large values of test stat are "extreme":

$$\text{p-value(obs. test statistic)} = P(|\text{test statistic rv}| \geq |\text{obs. test statistic}|)$$

"A *p*-value is a measure of how embarrassing
the data are to the null hypothesis"

--Nicholas Maxwell

Maxwell (2004). Data Matters: Conceptual Statistics for a Random World. Emeryville, CA: Key College Publishing.

Cut-offs

- Whether we like to admit it or not, p-values ultimately are cut-offs:

p-value $< \alpha$	p-value $\geq \alpha$
Hit	Not hit
Statistically significant	Not statistically significant
Fame and glory!	?
Reject H_0	Accept H_0 (cringe) Fail to reject H_0 (eye-roll)

P-value

- Standardized measurement of evidence
- Measure of probabilistic significance, not conceptual.
- **LOW P-VALUE:** low probability of sample looking like the null population
  - Decision → reject the null
- **HIGH P-VALUE:** high probability of sample looking unlike the null
  - Decision → do not reject null

Obtaining the p-value for our z-statistic

- One-tailed

`pz_up <-` ▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮

- Two-tailed

`pz_2 <- 2` ▮▮▮▮▮▮▮▮▮▮▮

`c(pz_up, pz_2)`
`[1] 0.04` ▮▮▮▮▮▮▮▮▮▮

Obtaining the p-value for our z-statistic

- 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


Cut-off values
Lower <- qnorm(.025, 100, 15/sqrt(25))
Upper <- qnorm(.975,100,15/sqrt(25))
```
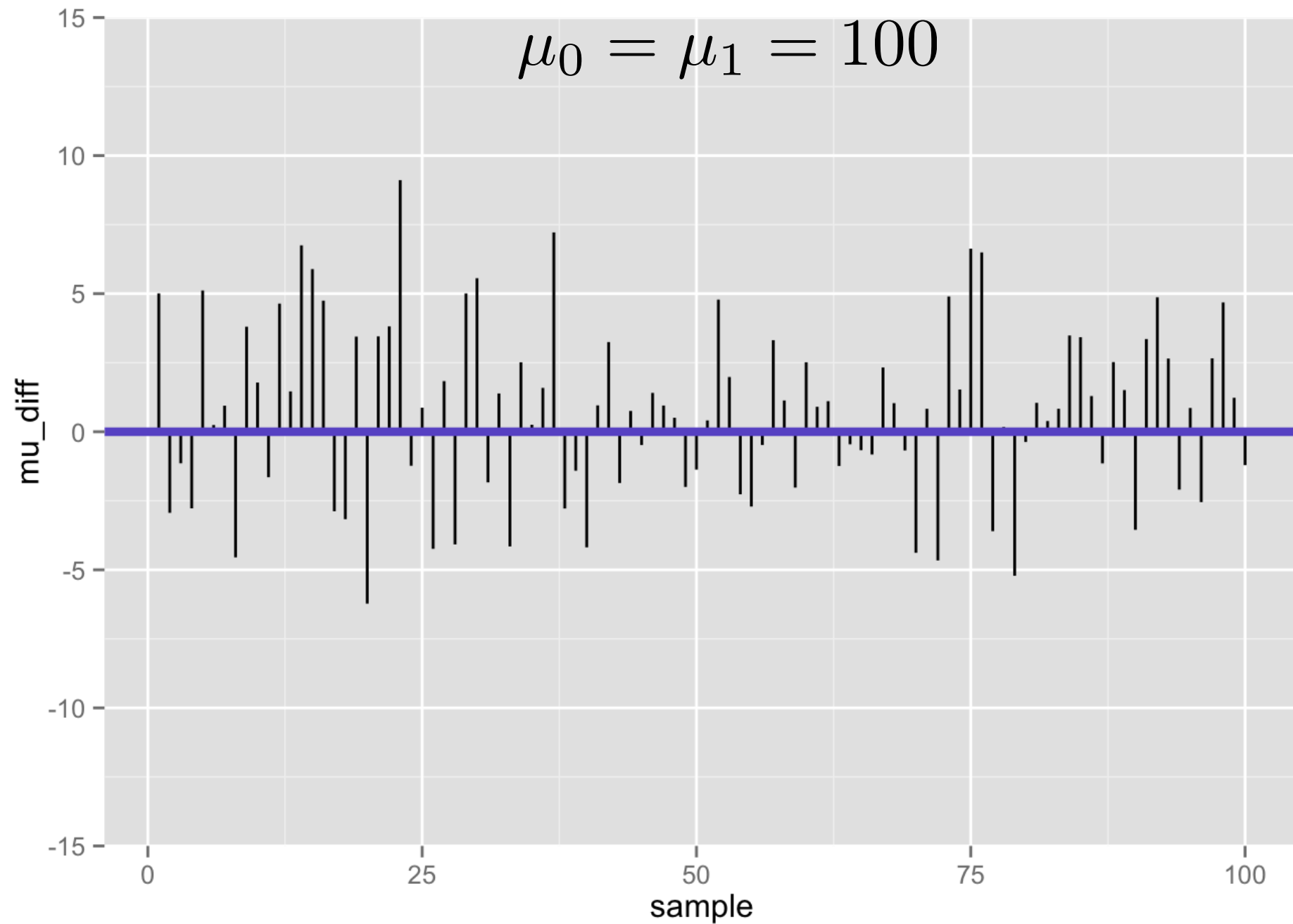
→ *p = .0955*

q = 94.12

q = 105.88

p = 0.025

p = 0.025

90   95   100   105   110
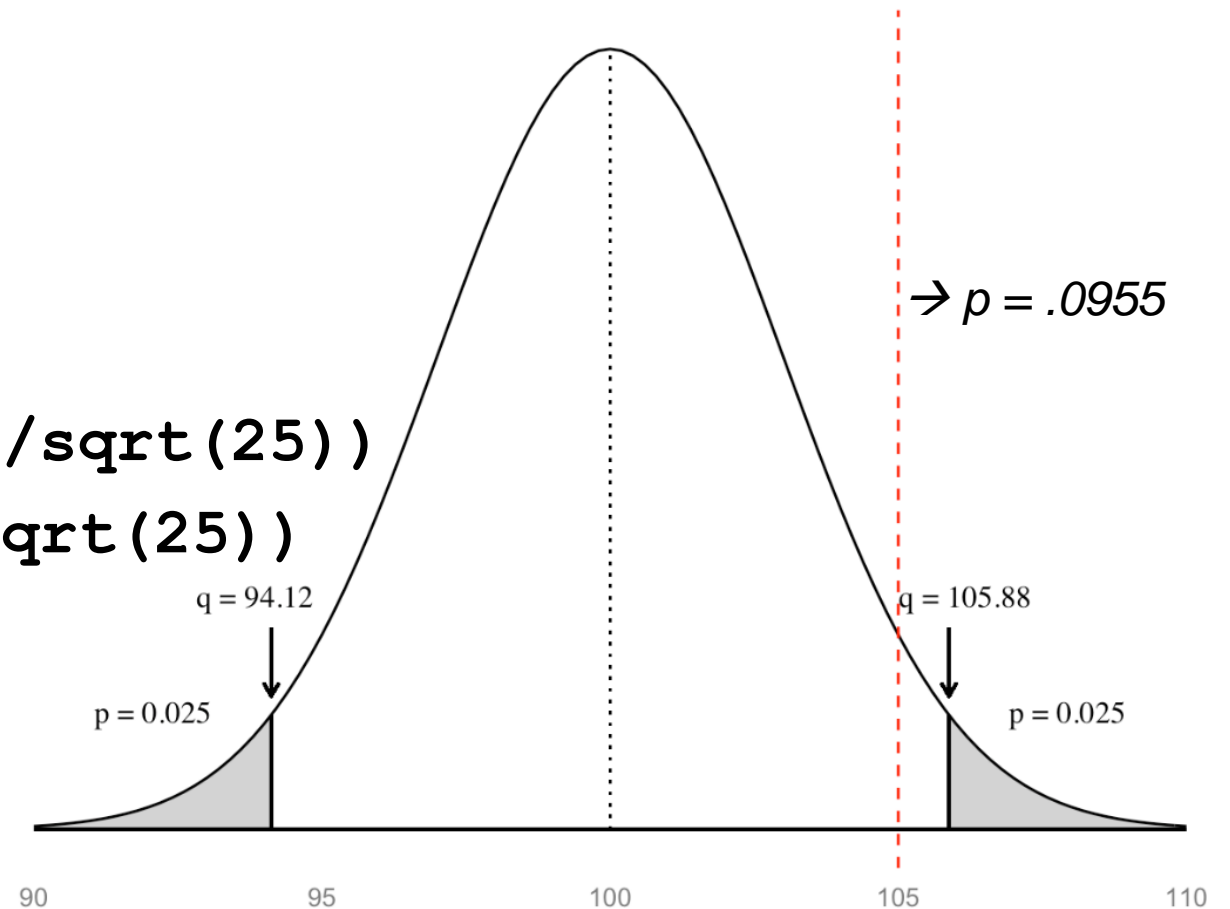
When the null hypothesis is true

$$\mu_0 = \mu_1 = 100$$

1-sample z-test
n = 25

$$\mu_0 = \mu_1 = 100$$

$$\mu_0 = \mu_1 = 100$$
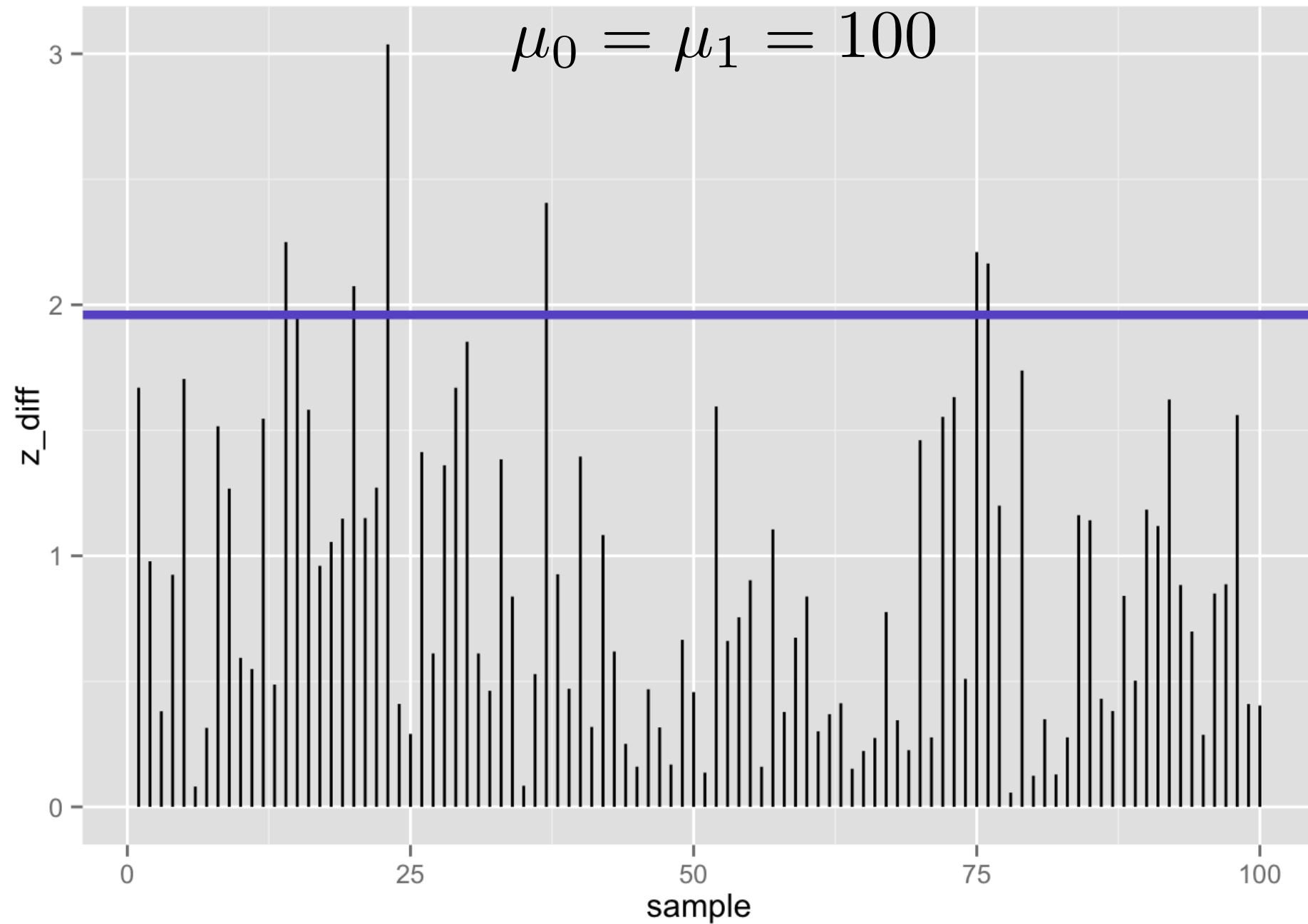
100 p-values when null is **true**: 7% false positives using z-test ($\alpha = .05/2$)
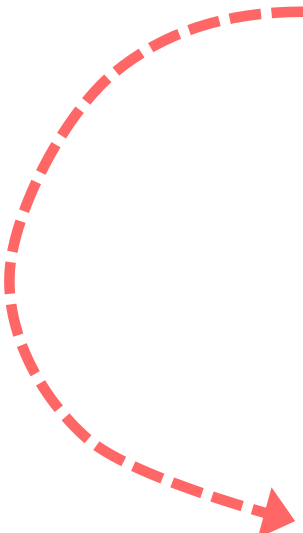
$$\mu_0 = \mu_1 = 100$$
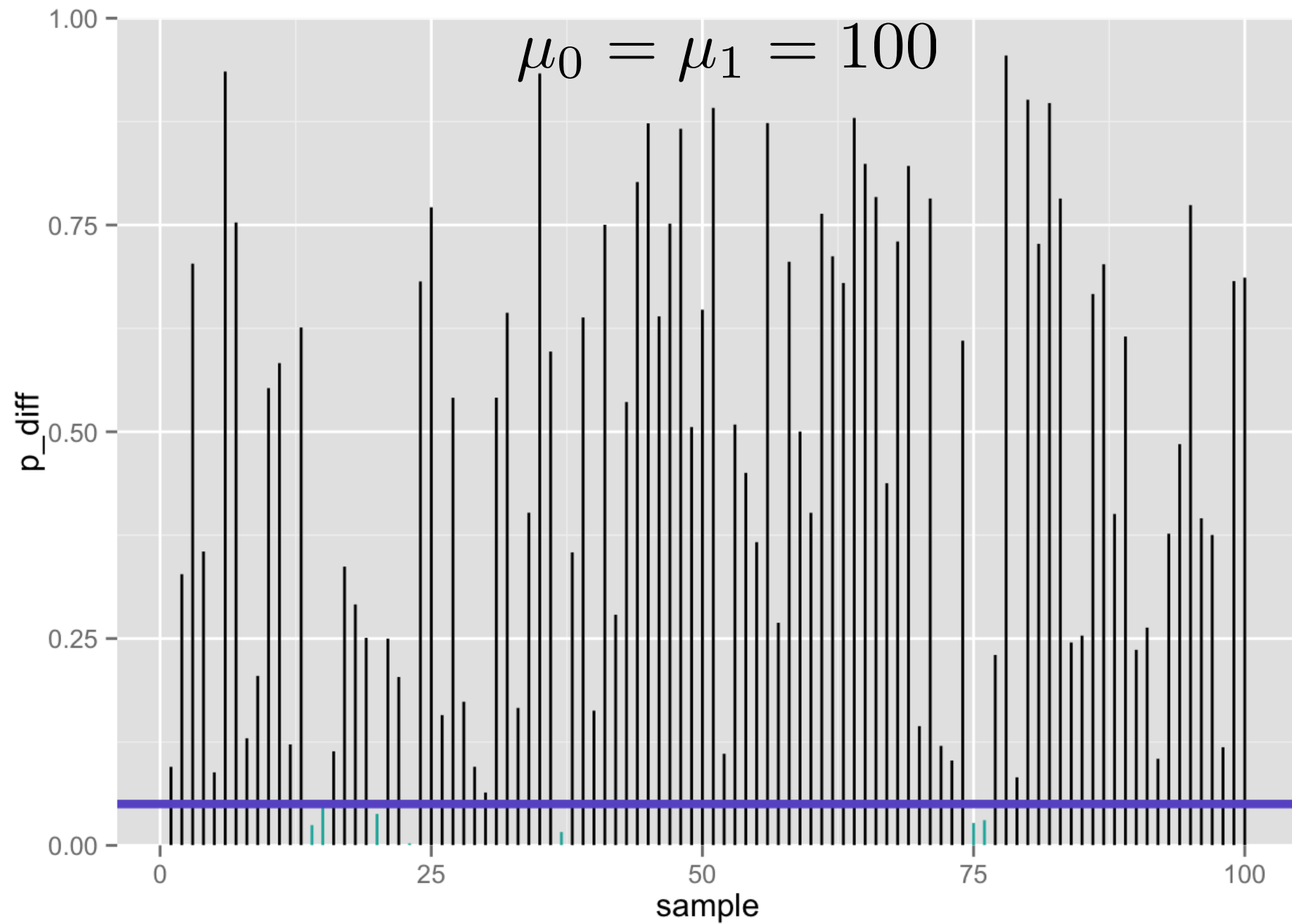
One sample means T-test

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:

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

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

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:

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

- 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

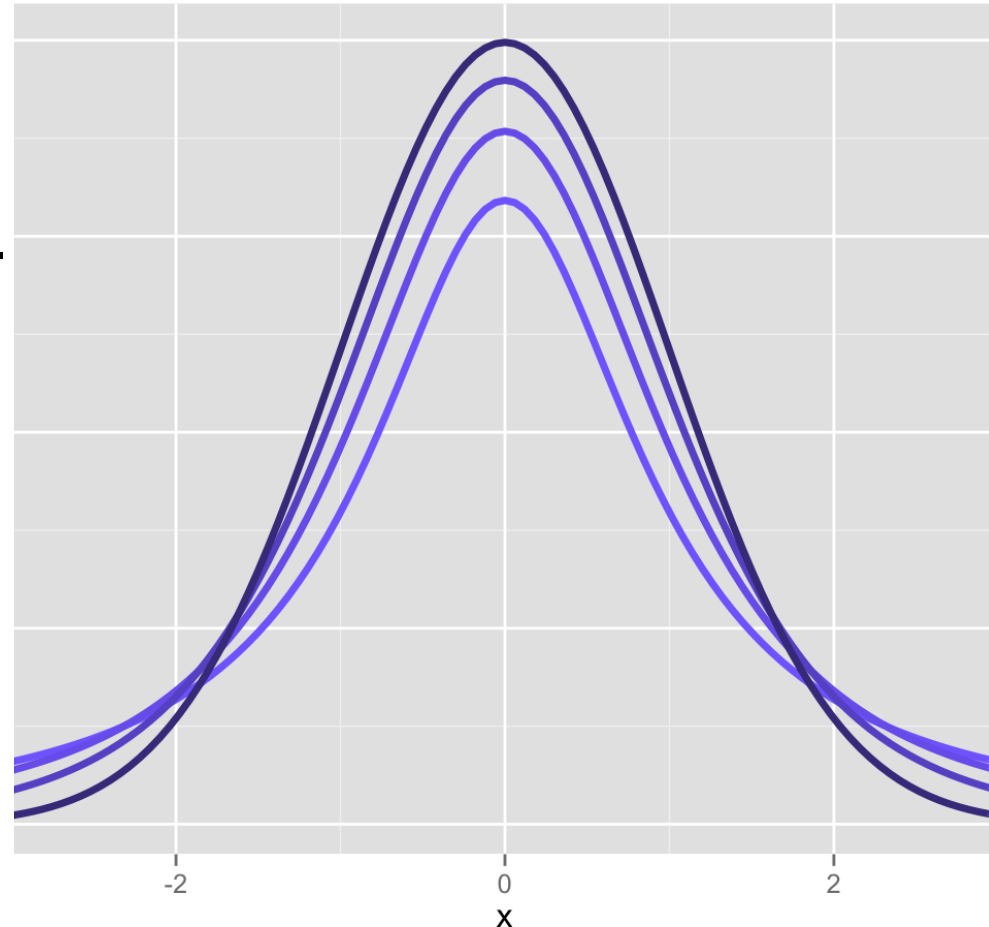# New distribution family: student's t
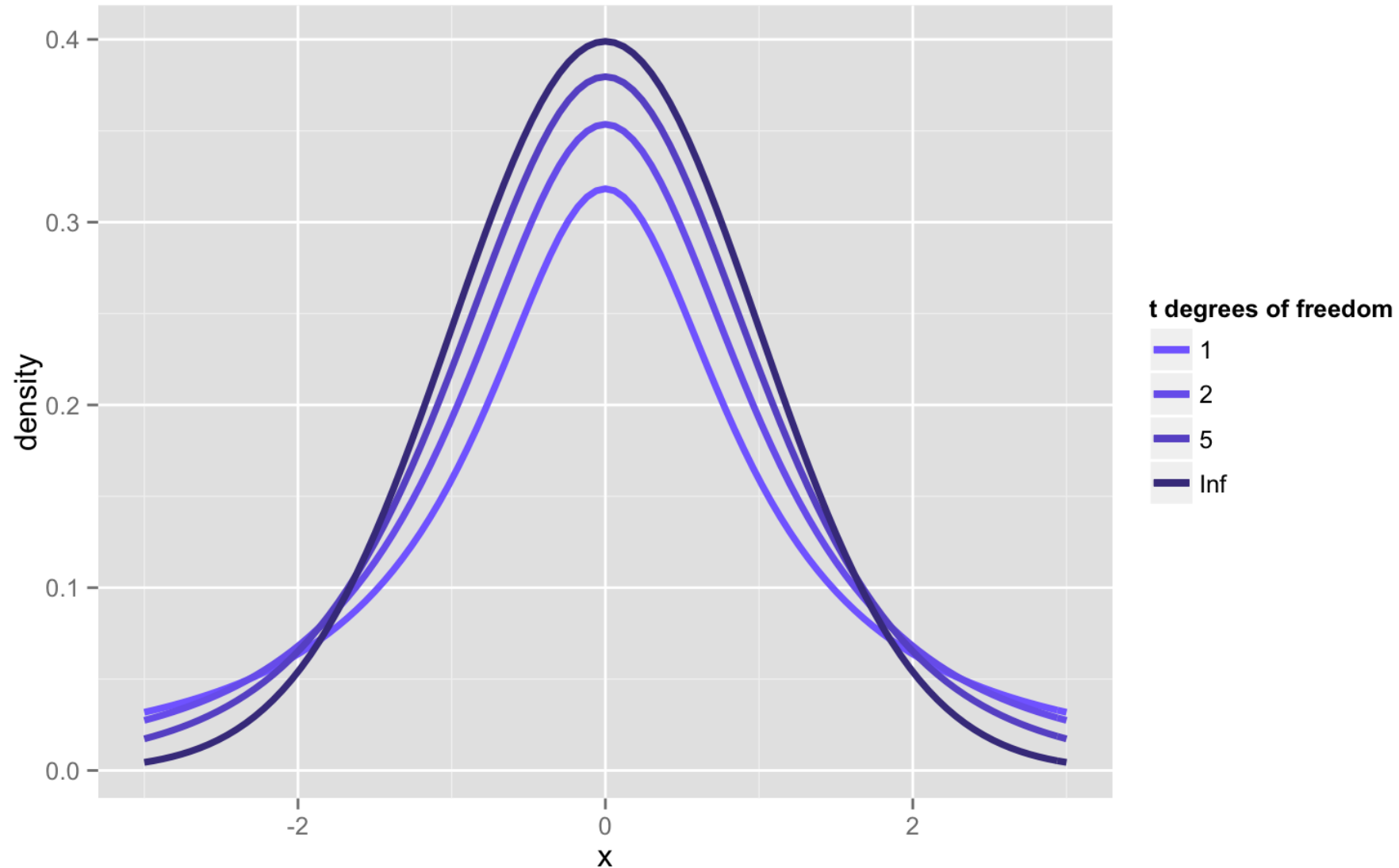
The *t*-distributions: PDFs

- 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

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.
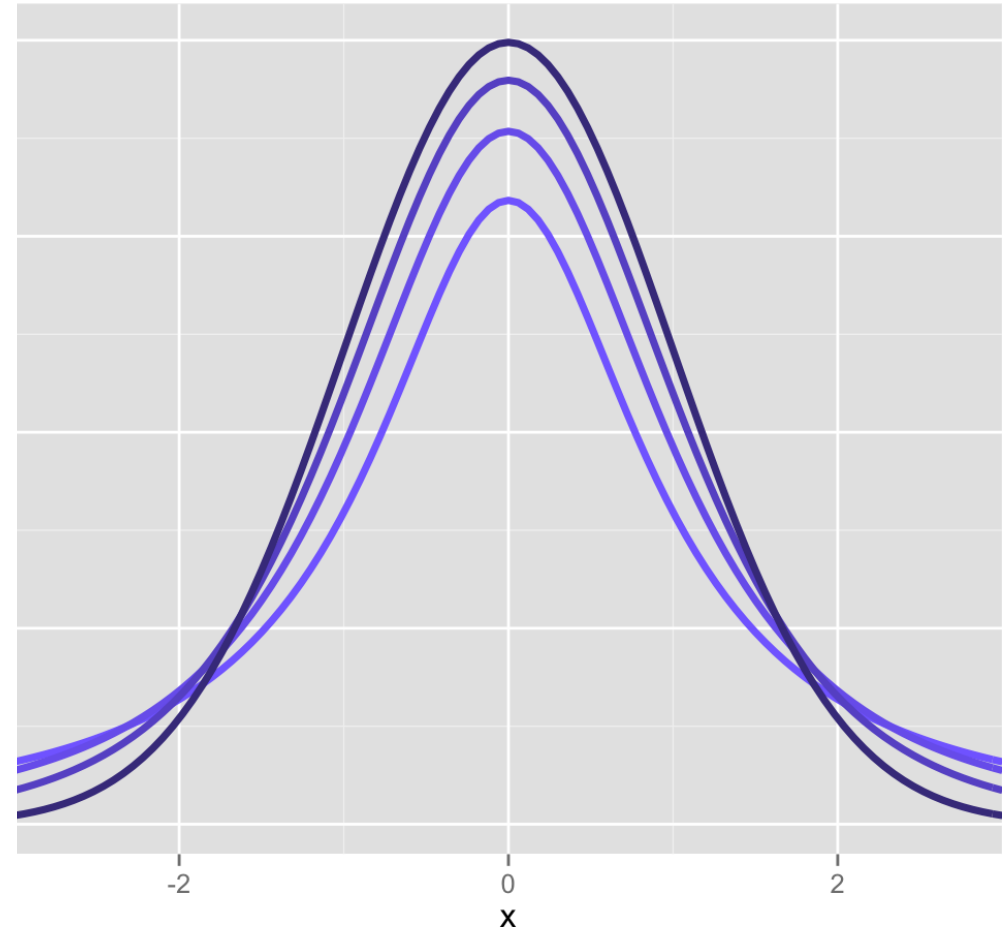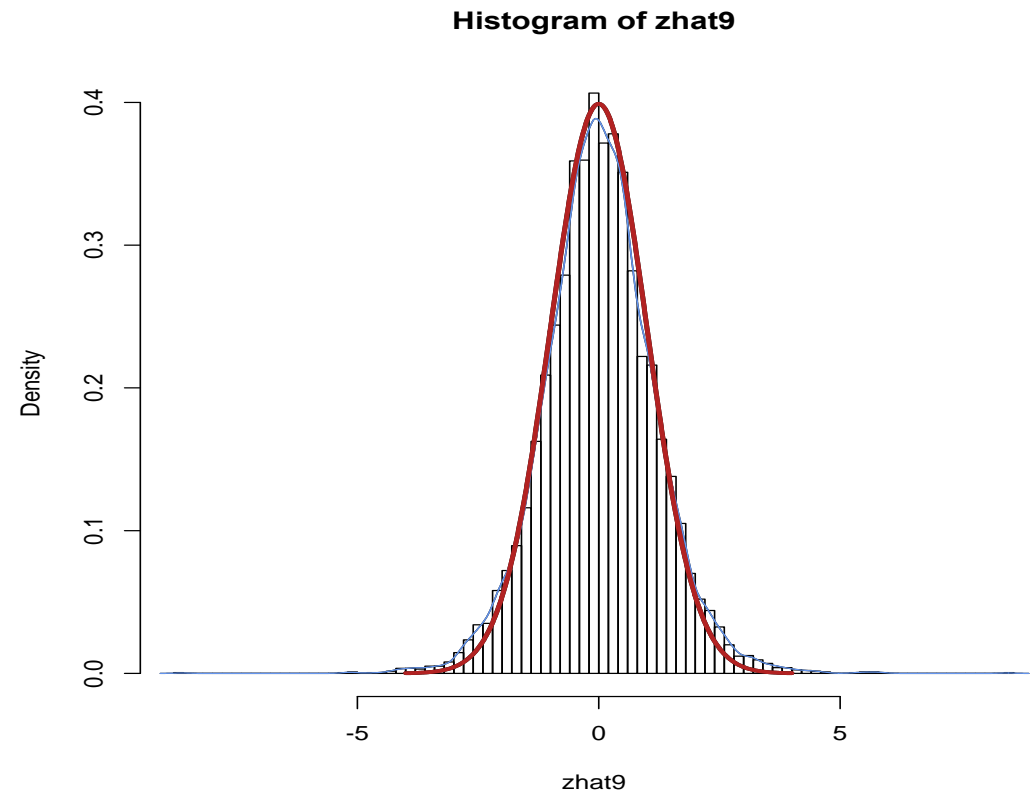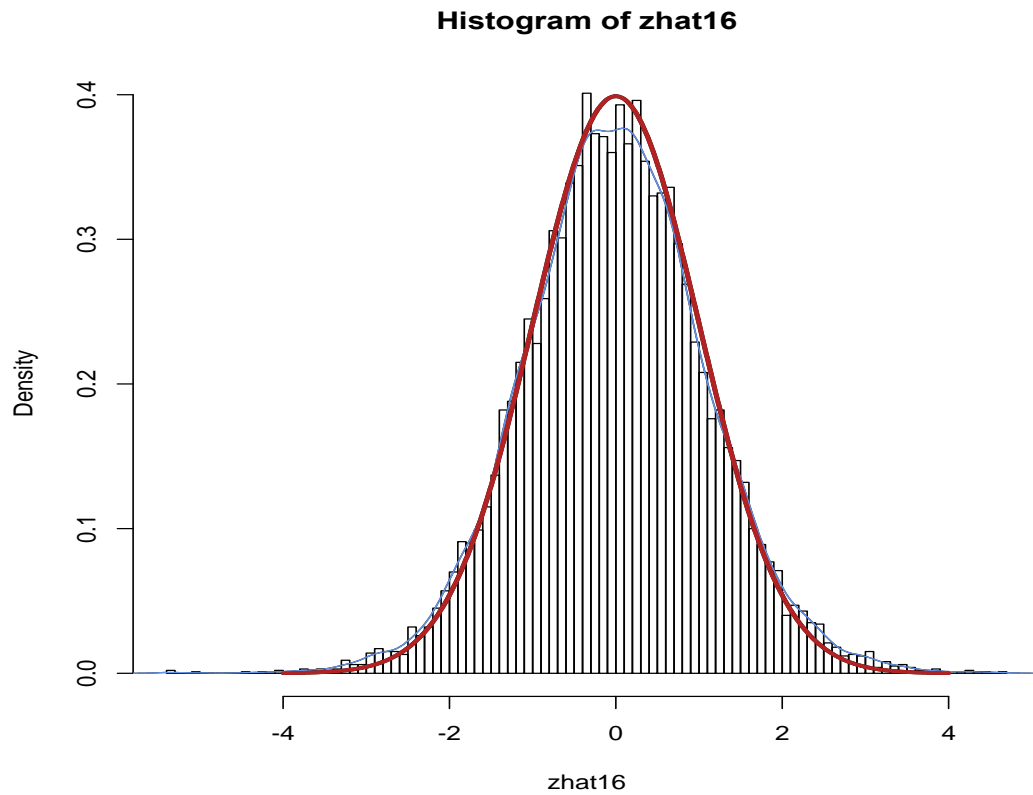
Red = z, blue = t

n = 4 (×10,000 simulations)
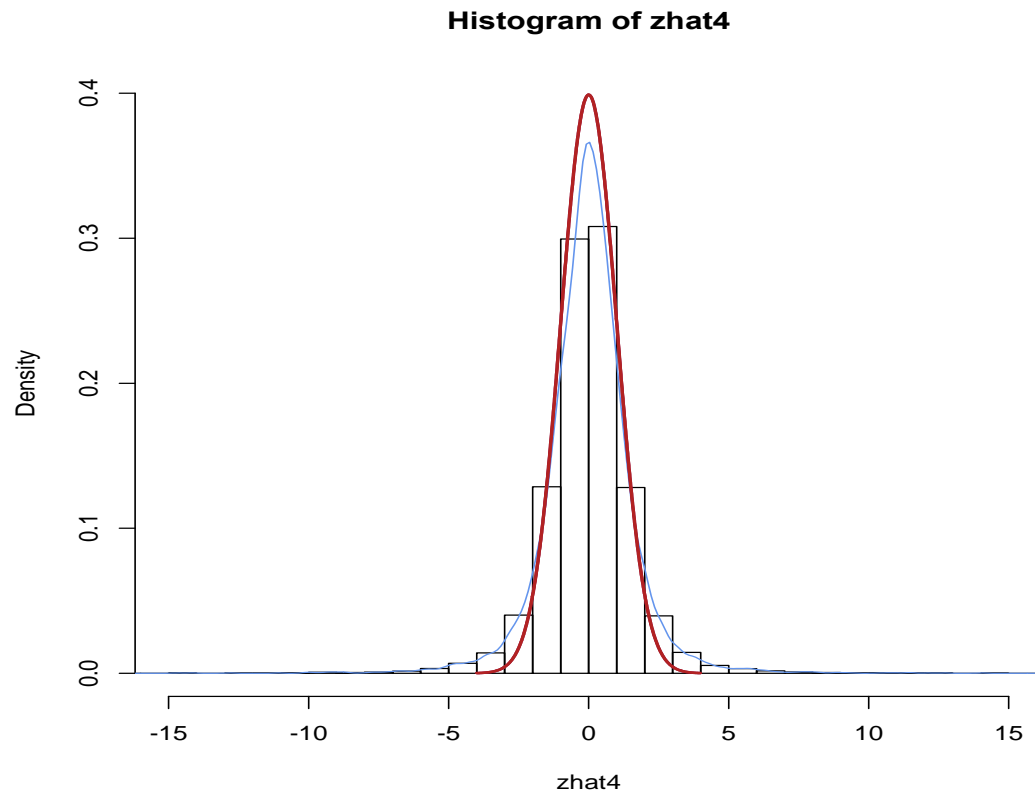
n = 9 (×10,000 simulations)

Histogram of zhat4

Histogram of zhat9

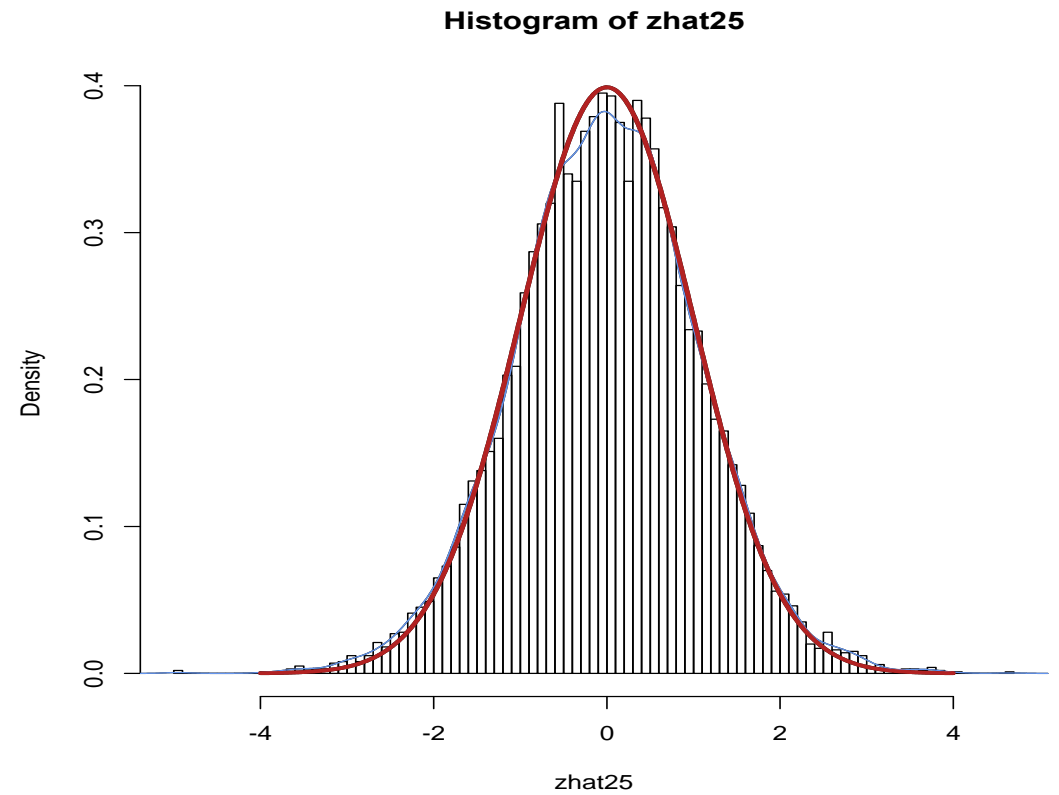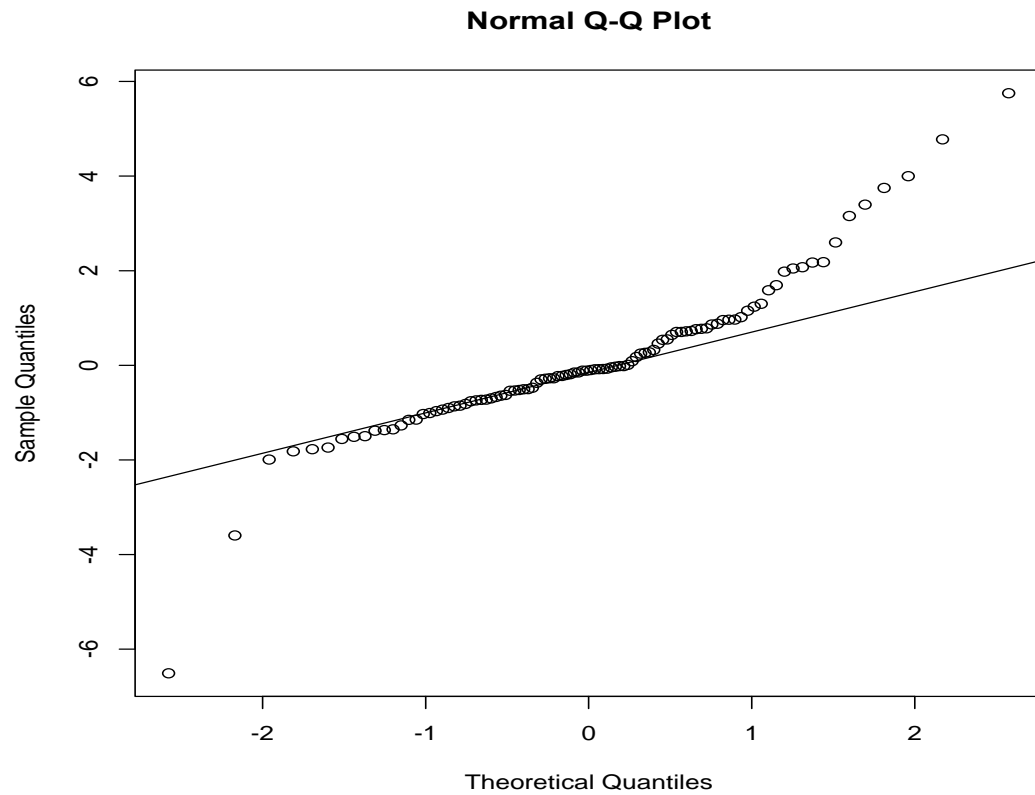# Red = z, blue = t

n = 16 (×10,000 simulations)

n = 25 (×10,000 simulations)

Q-Q Plots of rt(100,df)

t with *df*=4

t with *df*=30
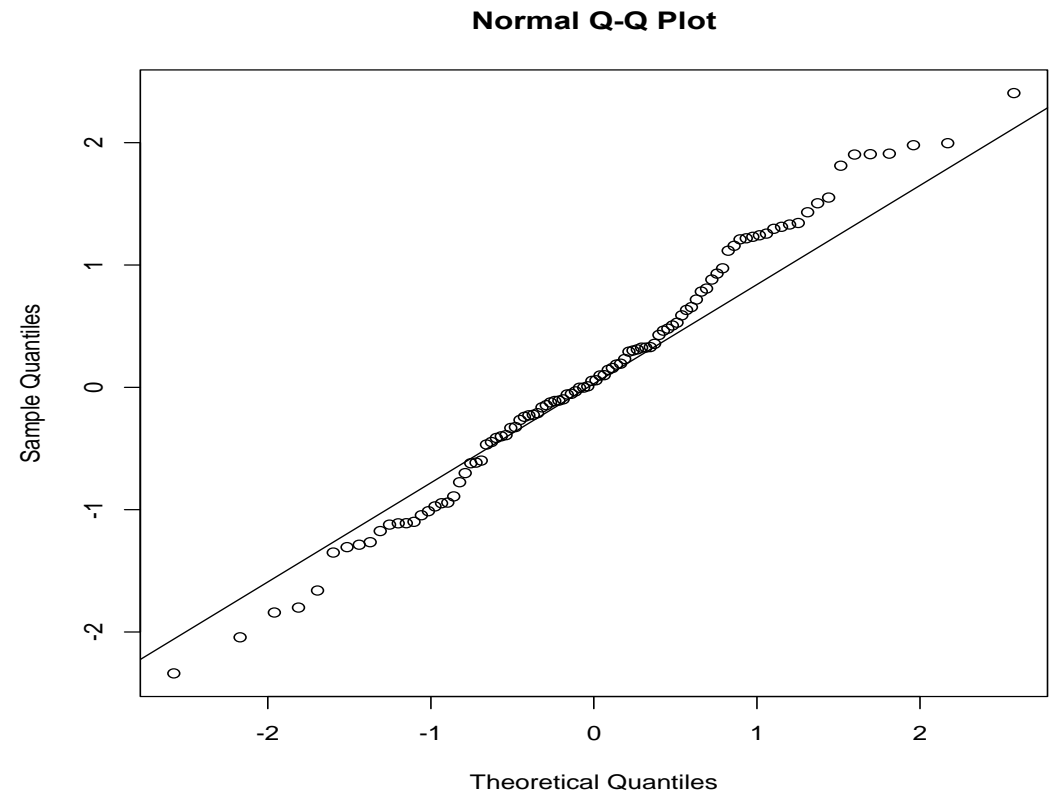
One-sample *t*-test

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

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)?

$$t_{obs} = \frac{\overline{Y}_i - m}{s / \sqrt{n}}$$

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

α = ?

df = ?

$t_{critical}$ = ?   Is it directional?

p?

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

- One-tailed, upper

```
pt_up <-
```

- Two-tailed

```
pt_2 <- 2
```

```
c(pt_up,
[1] 0.03
```

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
```

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
```

Two-tailed test

Mansplain it to me…

```
devtools::install_github(c("hilarypark
er/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?
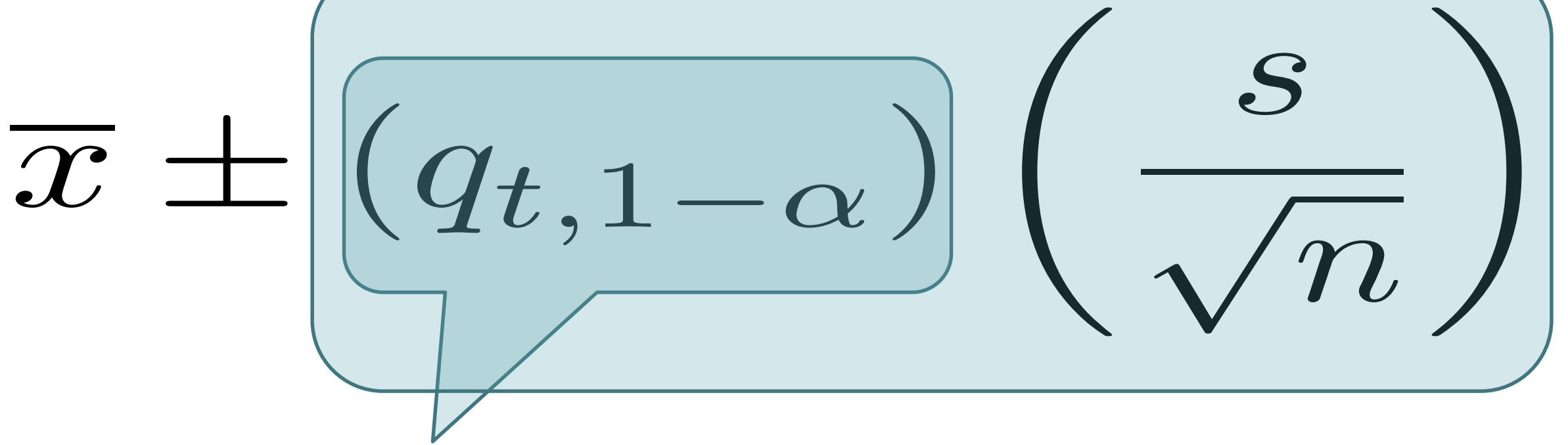
Family of *t*-tests

- 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

Confidence Intervals

Confidence interval for μ

Margin of Error

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

Critical Value!

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
```

Is μ = 100 in there??

Two-tailed CI

Let's attempt to summarize…

We are 95% confident that the IQ for AA HS girls in this sample is between 99.12 and 110.88.

This is not correct:
A confidence interval is for a population parameter, and cannot be applied to individuals in the sample.

Let's attempt to summarize…

95% of all samples of AA HS girls will give an average IQ between 99.12 and 110.88.

This is not correct:
Each sample will give rise to a *different* confidence interval- 95% of *those* intervals will contain the true mean.

Let's attempt to summarize…

There is a 95% chance that the true mean IQ for AA HS girls is between 99.12 and 110.88.

This is not correct:
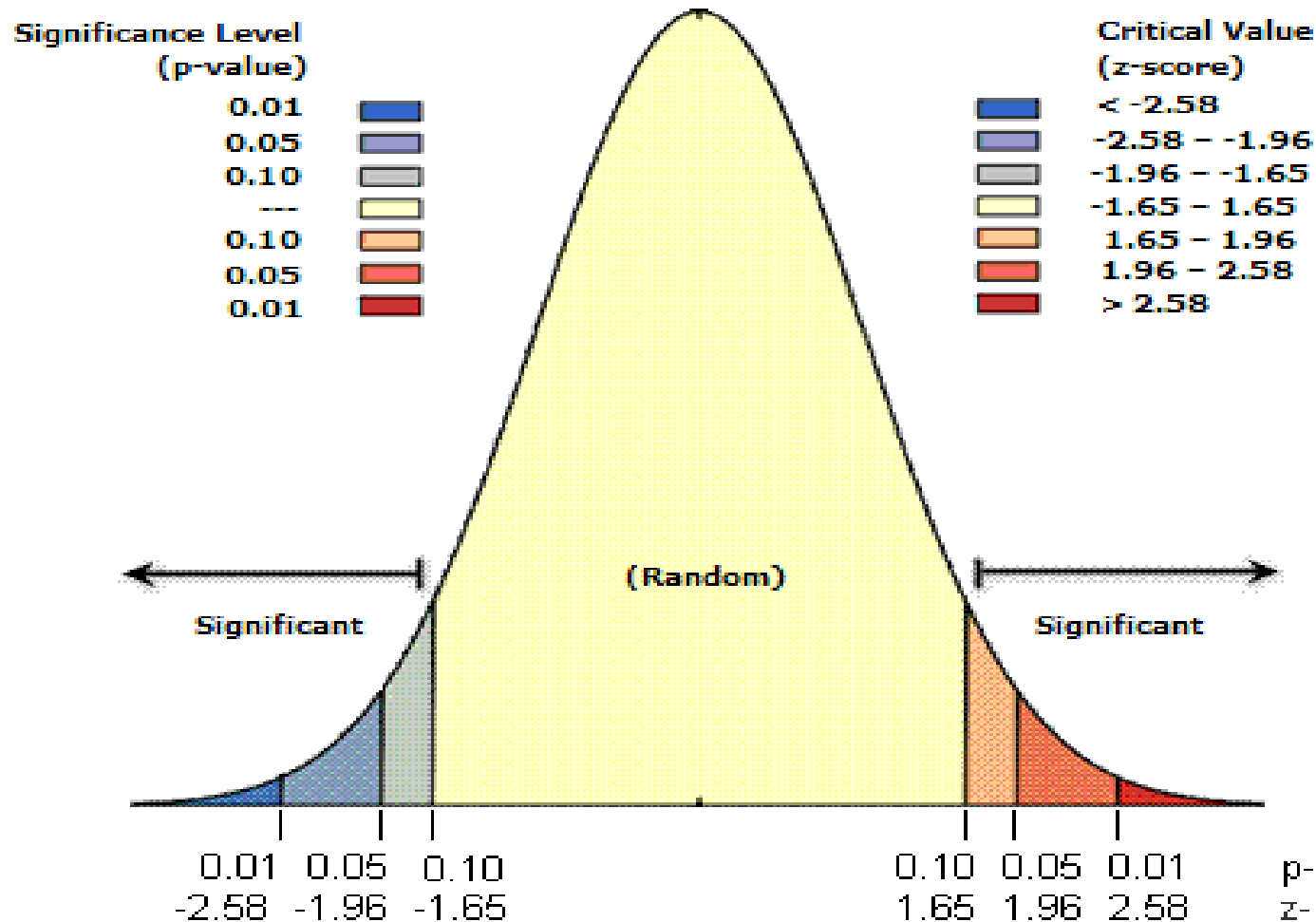μ is not random. The probability that μ is between 99.12 and 110.88 is either 0 or 1.

You got it…

We are 95% confident that the population mean IQ of AA HS girls is between 99.12 and 110.88.

This is correct!:
The confidence interval is based on the sample, and hence is random. There is a 95% probability that the interval (99.12, 110.88) contains μ.

The "Other" Confidence Interval

Determine $z_{critical}$ values for your α…

Must "beat" ±1.65 for α = .10, 2-tailed, to reject null
Must "beat" ±1.96 for α = .05, 2-tailed, to reject null
Must "beat" ±2.58 for α = .01, 2-tailed, to reject null

Not shown…
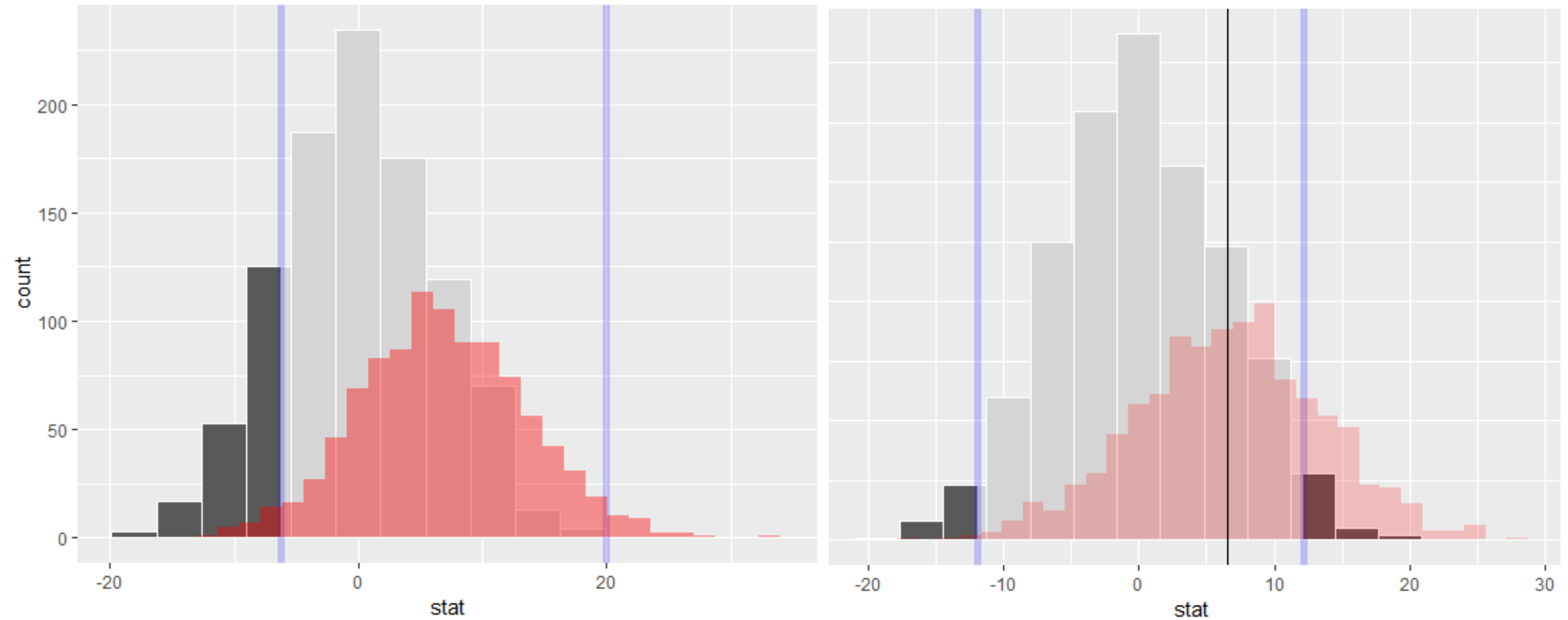Must "beat" ±1.28 for α = .10, 1-tailed, to reject null
Must "beat" ±1.65 for α = .05, 1-tailed, to reject null
Must "beat" ±2.32 for α = .01, 1-tailed, to reject null

Dueling Confidence Intervals - Sneetches

Hypothesis Tests and Confidence Intervals

- Suppose: z-test with $\alpha = .05$, 2-tailed
- H_0: $\quad \bar{Y}_. = m$
- H_1: $\quad \bar{Y}_. \neq m$

$$z_{obs} = \frac{\bar{Y}_. - m}{S / \sqrt{n}}$$

- **I will reject H_0** if:

$$\frac{\bar{Y}_. - m}{S / \sqrt{n}} < -1.96 \qquad\qquad \frac{\bar{Y}_. - m}{S / \sqrt{n}} > 1.96$$

- Some algebra to re-arrange, **I will reject H_0** if:

$$m < \bar{Y}_. - 1.96 \frac{S}{\sqrt{n}} \qquad\qquad m > \bar{Y}_. + 1.96 \frac{S}{\sqrt{n}}$$

- And **I will not reject H_0** if μ lies between those 2 rejection points:

$$\left[\bar{Y}_. - 1.96 \frac{S}{\sqrt{n}} \leq m \leq \bar{Y}_. + 1.96 \frac{S}{\sqrt{n}} \right]$$

This is my 95% confidence interval for μ

Hypothesis Tests and Confidence Intervals

- The confidence interval also tells when to reject H_0

  I will reject H_0 if μ is outside of this interval:

  $$\left(\overline{Y} - 1.96\frac{S}{\sqrt{n}} \leq \mu \leq \overline{Y} + 1.96\frac{S}{\sqrt{n}} \right)$$

- But what if instead of solving for μ, I solve for the sample mean?

  I will reject H_0 if:

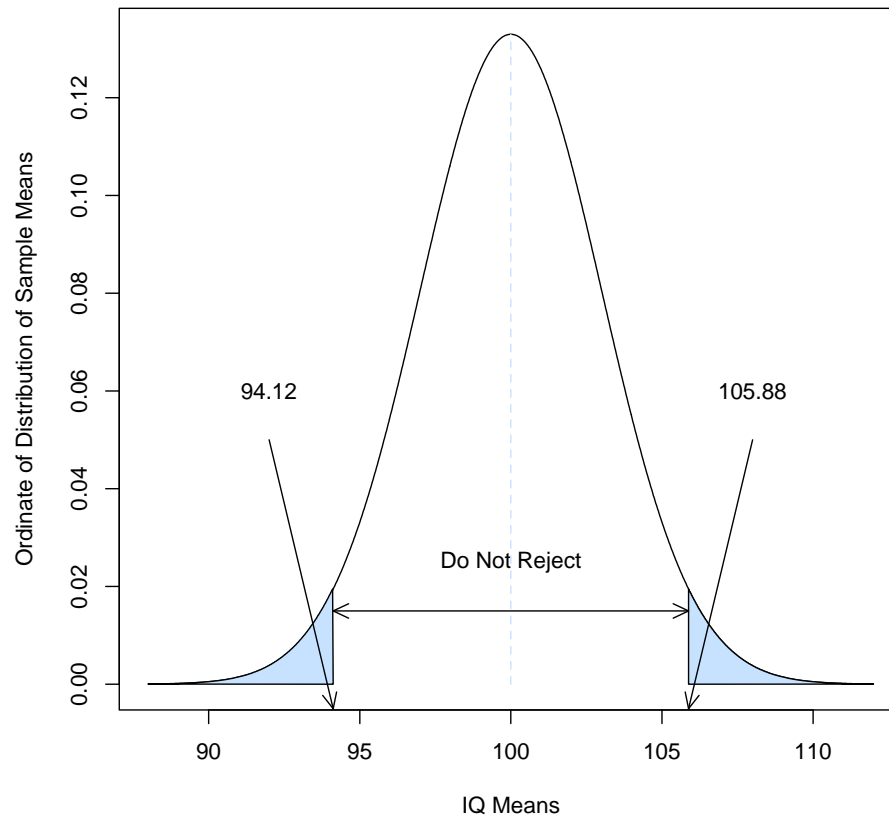 $$\overline{Y} < \mu - 1.96\frac{S}{\sqrt{n}} \qquad \overline{Y} > \mu + 1.96\frac{S}{\sqrt{n}}$$

- These two values define both of my regions of rejection (2-tailed).

- My region of non-rejection is all values in between those 2 rejection points

  I will not reject H_0 if my sample mean is within:

  $$\left(\mu - 1.96\frac{S}{\sqrt{n}} \leq \overline{Y} \leq \mu + 1.96\frac{S}{\sqrt{n}} \right)$$

Regions of rejection/non-rejection:
What if I observe a sample mean = 106?

- **Rejection regions**
  I reject if:
  $$\bar{Y}_{.} < 100 - 1.96\frac{15}{\sqrt{25}} = 94.12$$
  or if:
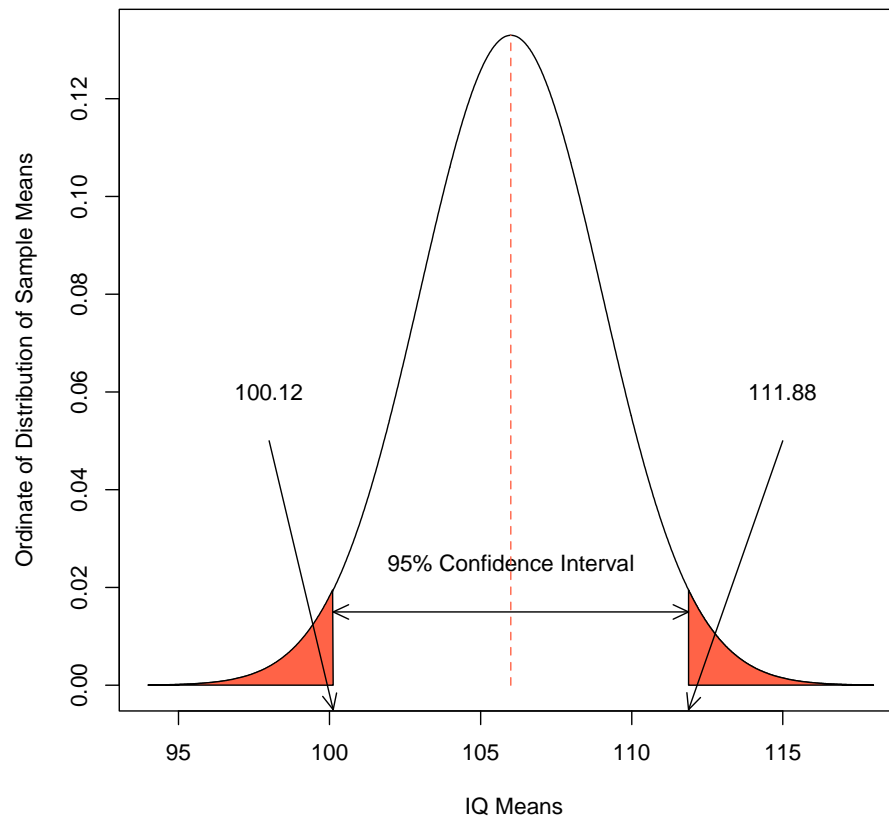 $$\bar{Y}_{.} > 100 + 1.96\frac{15}{\sqrt{25}} = 105.88$$

- **Non-rejection region**
  I will not reject if:
  $$\left(94.12 \le \bar{Y}_{.} \le 105.88\right)$$

- With observed sample mean=106, I reject H_0

Confidence interval:
What if I had hypothesized μ = 100?

- **95% confidence interval**
  I will not reject H_0 if μ is within:

$$\left(106 - 1.96\frac{15}{\sqrt{25}} \leq m \leq 106 + 1.96\frac{15}{\sqrt{25}}\right)$$

$$\left(100.12 \leq m \leq 111.88\right)$$

- Because μ=100 is not contained in this interval, I reject H_0 as it does not contain my population value
- Note that this is the same decision I reached with my regions of rejection
- I also know more than I did before: I know that I am 95% confident that the true value of μ is between 100.12 and 111.88

Reject H$_0$...

ROR:

If the sample mean is outside the region of rejection.
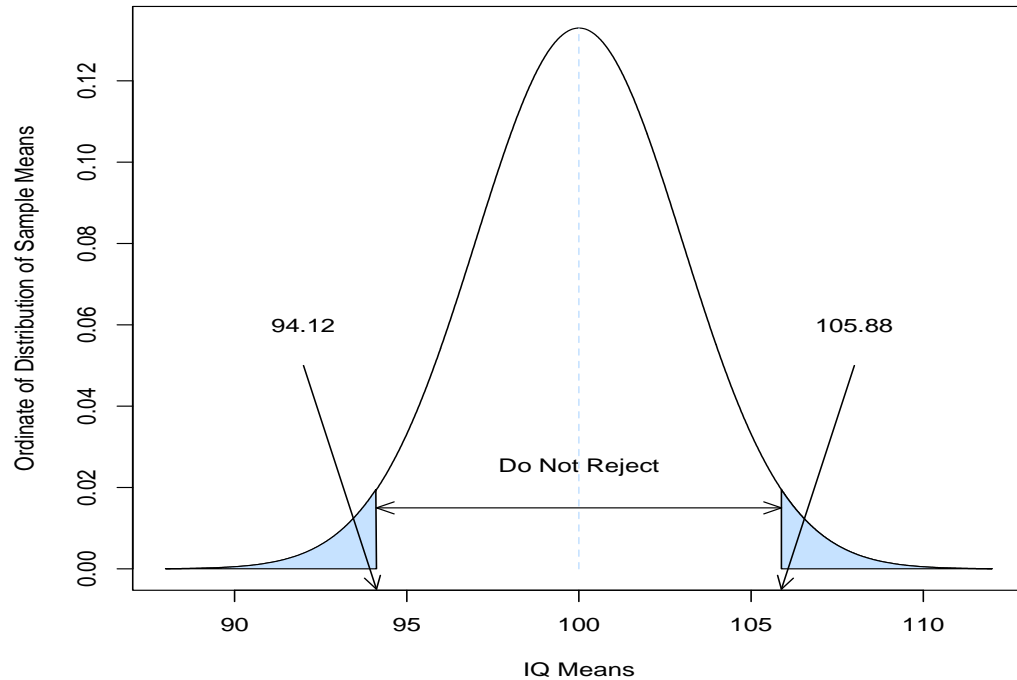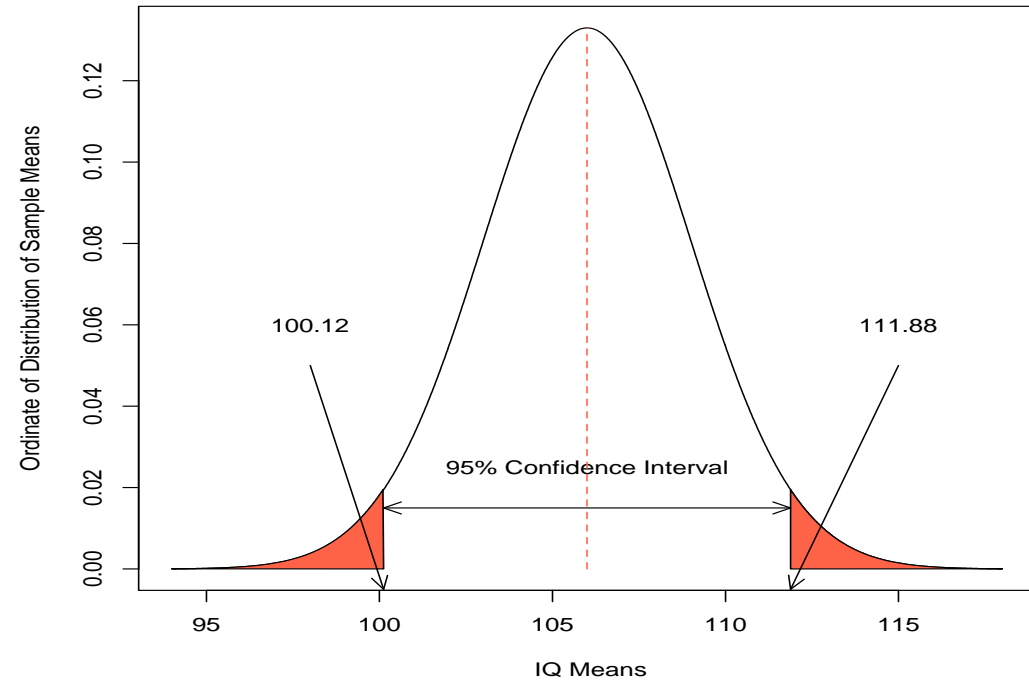

RoR: μ = 100 (population parameter / hypothesized value)

Given that μ = 100, is my sample mean a reasonable value from that population?

Confidence Interval:

If the null hypothesis mean is outside the Confidence Interval.

Confidence interval: μ = 106 (sample statistic)

Given that my sample mean = 106, is μ a reasonable value for the population mean?

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

Confidence intervals provide several advantages over NHST alone

- Provides bounded estimate of the population parameter

- Permits tests of all possible null hypotheses simultaneously
  - You can reject/not reject H_0 not only for value hypothesized in H_0 but also for other hypothesized values

- Interval width provides information about precision
  - Significance plus very narrow interval suggests power so great as to enable us to reject even a trivial effect
  - Non-significance with wide intervals suggests that our measures/procedures/experiment lacks precision
  - *p*-values do not give you this information, heavily influenced by sample size (*n*)

3 new distributions…

If you see the ratio of an independent normal random variable (numerator) to the square-root of a chi-squared variable (denominator; any time you have an SE on the bottom!),

- think **Student's t**.

If you see squares of normal random variables (or of their differences!),
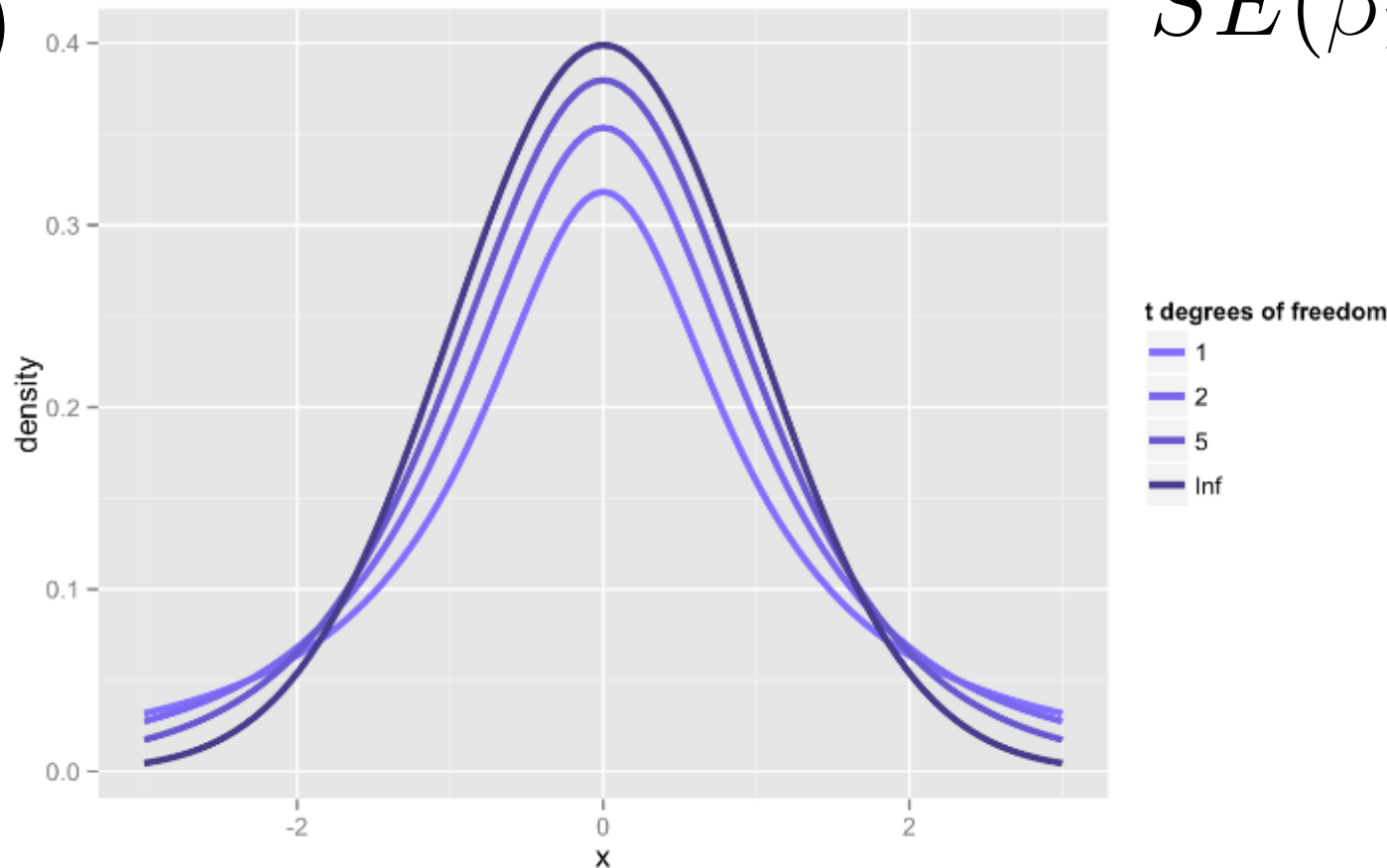
- think **chi-squared**.

If you see the ratio of two chi-squared random variables (i.e., sample variances from a normal distribution),

- think **F distribution**.

If you see the ratio of an independent normal random variable to the square-root of a chi-squared, think Student's t

$$T = \frac{\hat{\theta} - \theta}{SE(\hat{\theta})}$$

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$$

If you see squares of normal random variables, think chi-squared

If you see the ratio of two chi-squared random variables (i.e., sample variances from a normal distribution), think *F* distribution
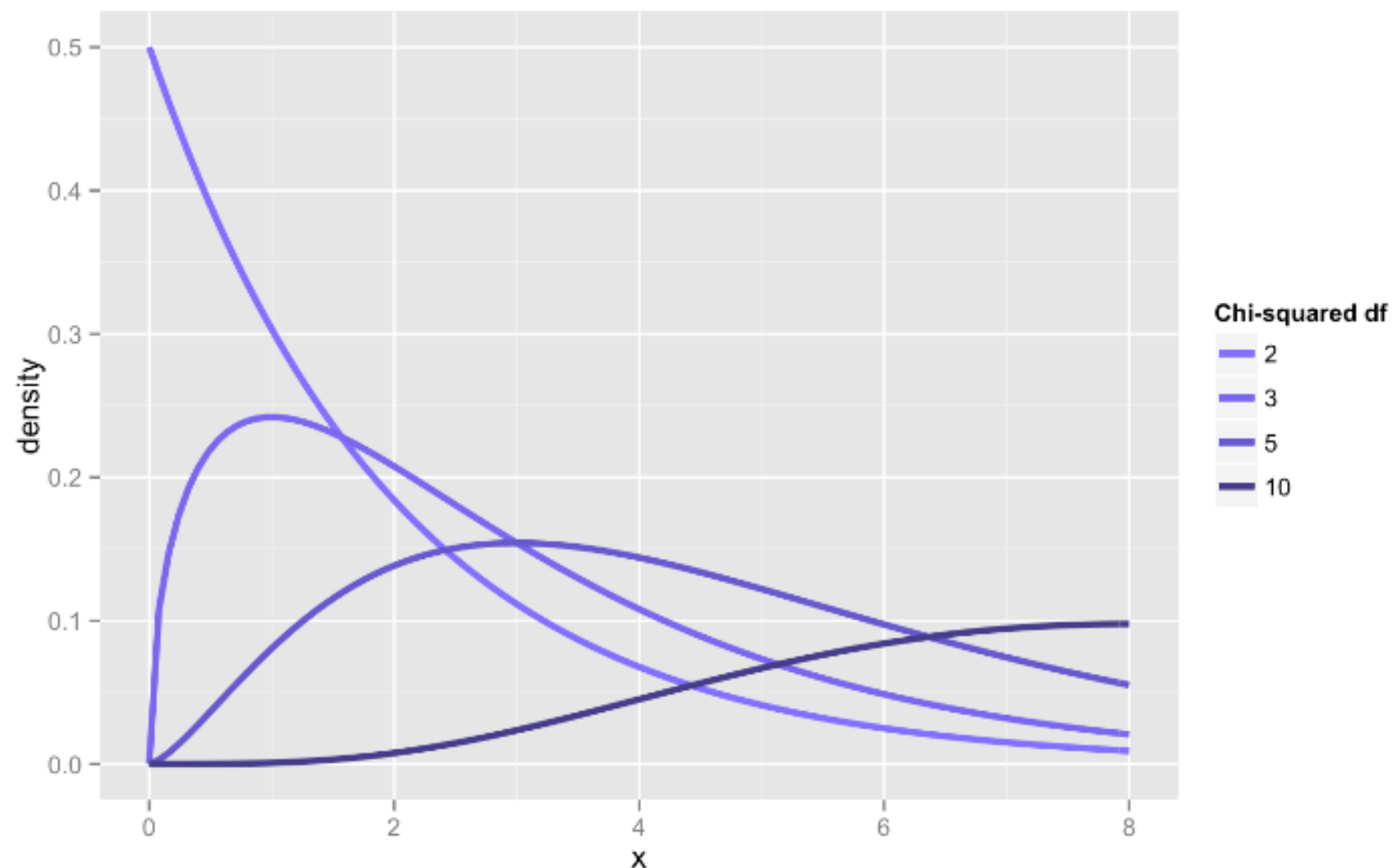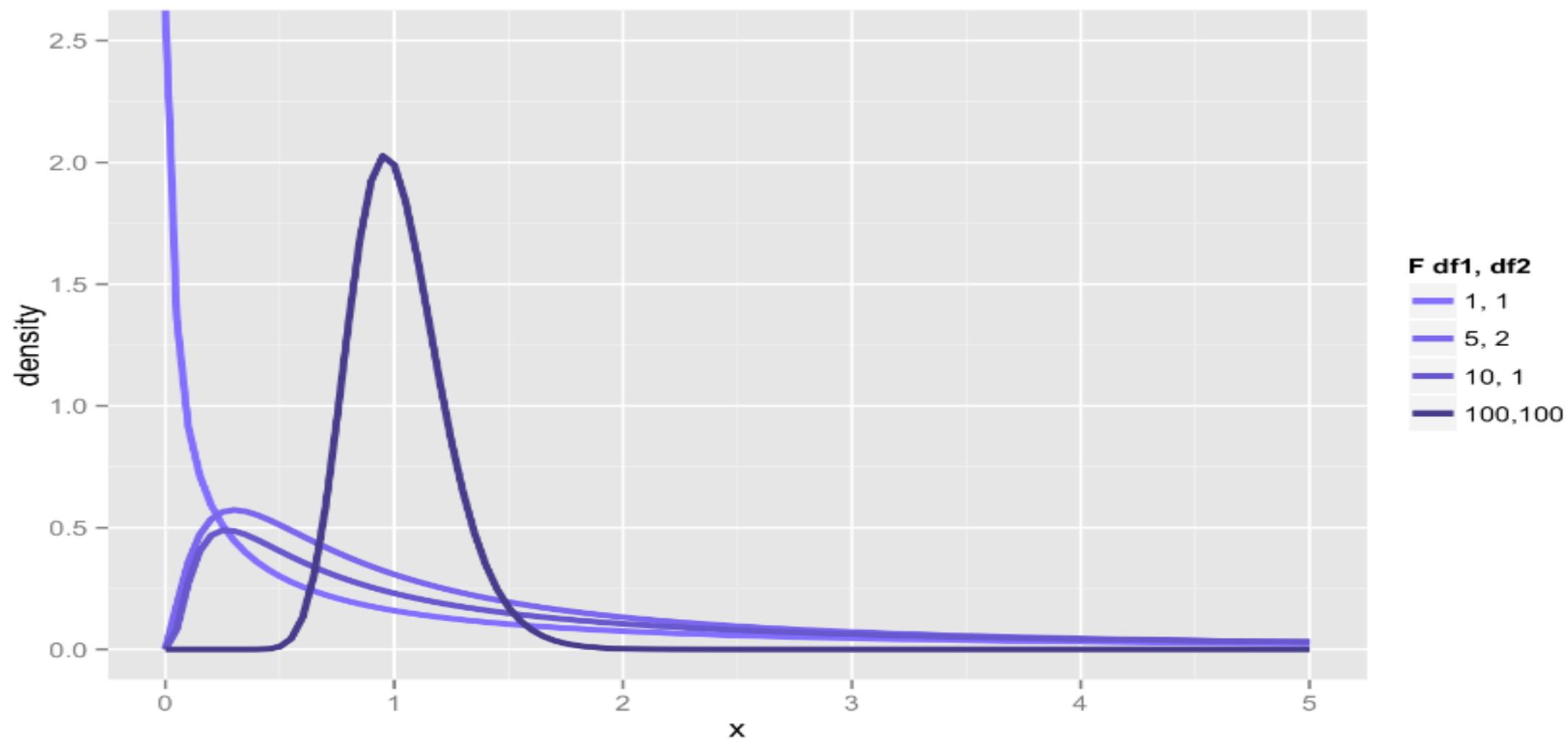
Done-ish…

- References
  - Assumptions!!
  - Perfect T-test

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.

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

From Chihara & Hesterberg - permutation:

- Say that you have one sample with *m* observations, and a second sample with *n* observations
1. Pool the m + n values
2. **Repeat…**
   - Draw a resample of size m without replacement
   - Use the remaining n observations for the other sample
   - Calculate the difference in means or another statistic that compares samples
3. **…until you have enough samples**
4. Calculate the p-value as the fraction of times the random statistics exceed the original statistic (note: "Exceeds" generally means ≥ rather than >)
5. Multiple by 2 for a two-sided test, or use abs()