Additional topics in testing

Stat 120

February 03 2023

Significance Level & Formal Decisions

The significance level, α is the threshold below which the p-value is deemed small enough to reject the null hypothesis (evidence is statistically significant).

$$\text{p-value} < \alpha \implies \text{Reject } H_0$$

$$\text{p-value} \geq \alpha \implies \text{Do not Reject } H_0$$

Common levels:

- 10% : need some evidence to reject the null
- 5% : need moderate evidence to reject the null
- 1% : need strong evidence to reject the null

Errors

	Reject H_0	Do not reject H_0
H_0 true	TYPE I ERROR	😄
H_0 false	😄	TYPE II ERROR

- A Type I Error is rejecting a true null (false positive)
- A Type II Error is not rejecting a false null (false negative)

Statistical Significance

Hypothesis testing is similar to how our justice system works (or is supposed to work!).

$$H_0 : \text{defendant is innocent}$$
$$H_A : \text{defendant is guilty}$$

Assumption: Defendant is innocent (H_0)

Verdicts:

Guilty: evidence (data) "beyond a reasonable doubt" points to guilt (Statistically significant)

- Type I error possible: convict an innocent person

Not Guilty: evidence (data) not beyond a reasonable doubt, but we don't know if they are truly innocent (H_0)

- Type II error possible: release a guilty person

Examples

Science study of gender stereotypes:

- Comparing interest between 5-year-old boys and girls in a game for "really, really smart kids"
- test using $\alpha = 0.05$; reported p-value of 0.46

Decision?

- Do not reject $\mathbf{H}_0$: no evidence of a difference in mean interest level

Possible error?

- Type II: if a difference in mean interest level exists, then we would have made an error when not finding evidence of a gender difference in interest levels.

Consequence of making this error?

- Mislead the public about when gender stereotypes start emerging in young children

Examples

Memory: test using $\alpha = 0.05$; data gives p-value of 0.048

Decision?

- Reject $\mathbf{H}_0$

Possible error?

- Type I: if there is no difference in treatments, then we would have made an error in claiming that there was.

Consequences of making this error?

- Mislead the public about the benefits of sleep over caffeine.
- The nice thing about type I errors is that we can control the chance of such an error...

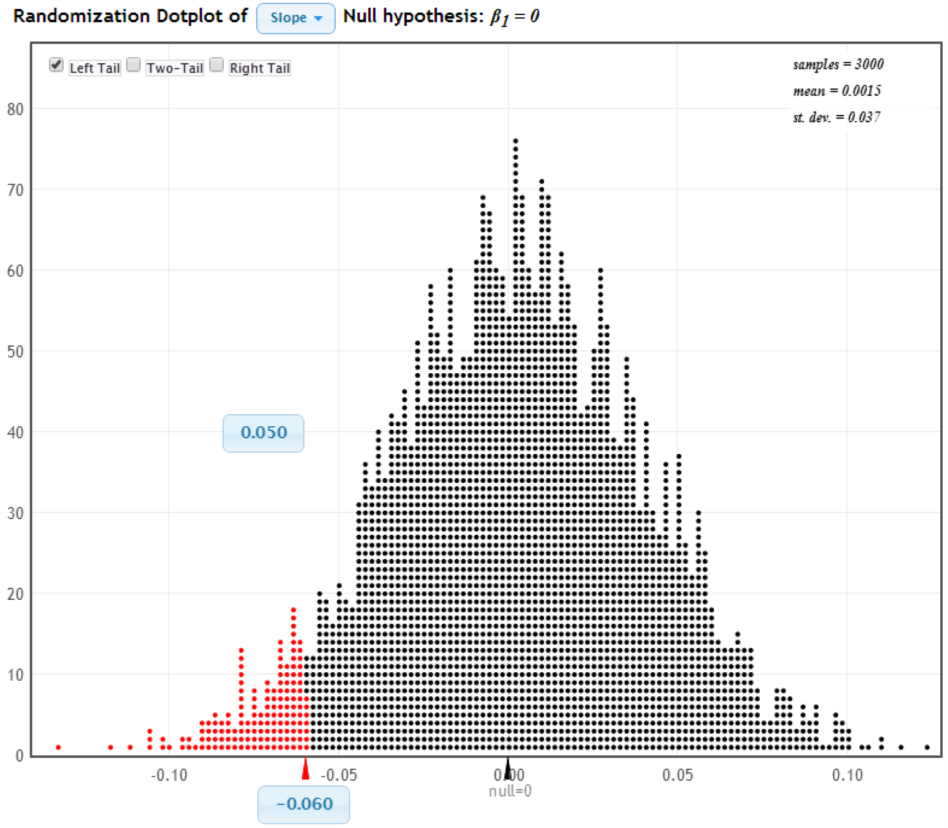
α = Probability of Type I Error

The significance level α controls the type I error rate.

- Recall the Florida Lakes slope test:

$$H_0 : \beta = 0 \quad H_a : \beta < 0$$

- If H_0 is true and $\alpha = 0.05$, then 5% of sample slopes will be lower red tail ($b \leq 0.06$).
- 5% of the sample slopes will give p-values less than 0.05, so 5% of statistics will lead to rejecting H_0 if it is true (Type I error)!!!



Null distribution

α = Probability of Type I Error

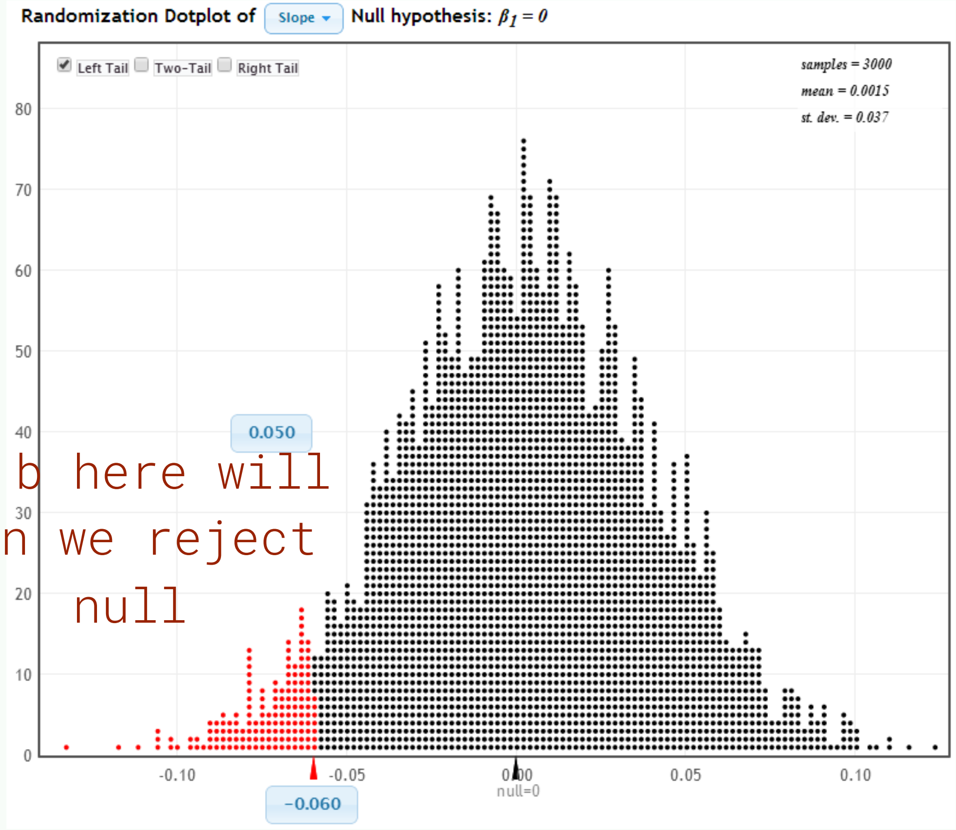
The significance level α controls the type I error rate.

- Recall the Florida Lakes slope test:

$$\mathrm{H}_0 : \beta = 0 \quad \mathrm{H_a} : \beta < 0$$

- If H_0 is true and $\alpha = 0.05$, then 5% of sample slopes will be lower red tail ($b \leq 0.06$).
- 5% of the sample slopes will give p-values less than 0.05, so 5% of statistics will lead to rejecting H_0 if it is true (Type I error)!!!

Any b here will mean we reject null



Null distribution

Selecting a significance level

Decreasing α will lower your Type I error rate (makes it harder to reject the null)

- but it will also increase your type II error rate (makes it harder to accept a true alternative)

Selecting a significance level

If a Type I error (rejecting a true null) is much worse than a Type II error, we may choose a smaller α, like $\alpha = 0.01$ (need lots of evidence to reject null).

- E.g. sending an innocent person to jail

Selecting a significance level

If a Type II error (not rejecting a false null) is much worse than a Type I error, we may choose a larger α, like $\alpha = 0.10$

- E.g. a false negative test for a serious disease

Probability of Type II Error

Not as simple to compute since the alternative is assumed to be true

- E.g. which value in $H_a: \beta < 0$ do we select to create an "alternative" randomization distribution?

The probability of making a Type II Error (not rejecting a false null) depends on

- Effect size (how far the truth is from the null)
- Sample size (bigger n means less uncertainty)
- Variability of measurements
- Significance level (bigger α means more false positives but fewer false negatives)

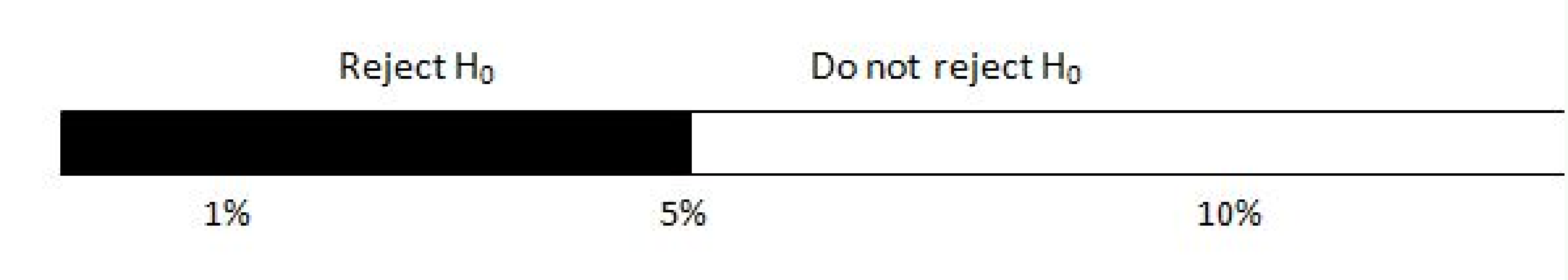
Power of a test

The power of a test is the chance that it will correctly reject the null, or

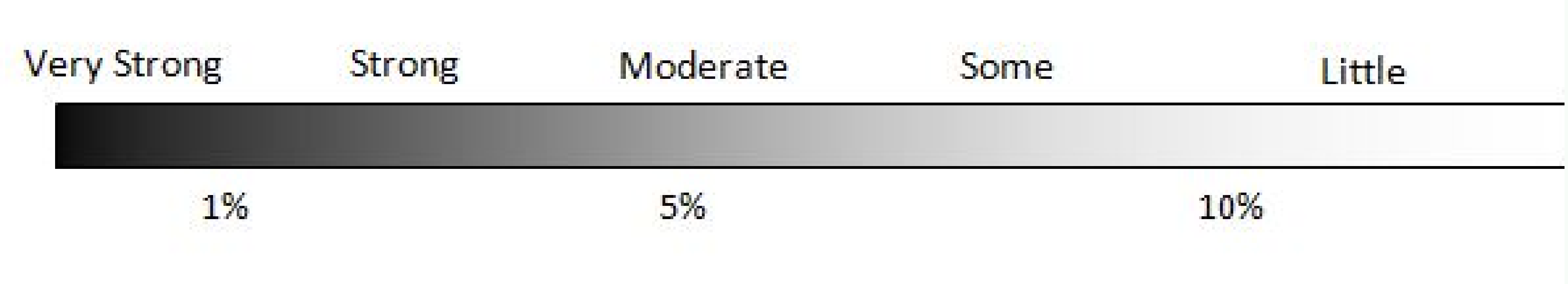
$$1 - \text{Prob}(\text{Type II error})$$

Statistical Conclusions

Formal decision of hypothesis test, based on a = 0.05 :



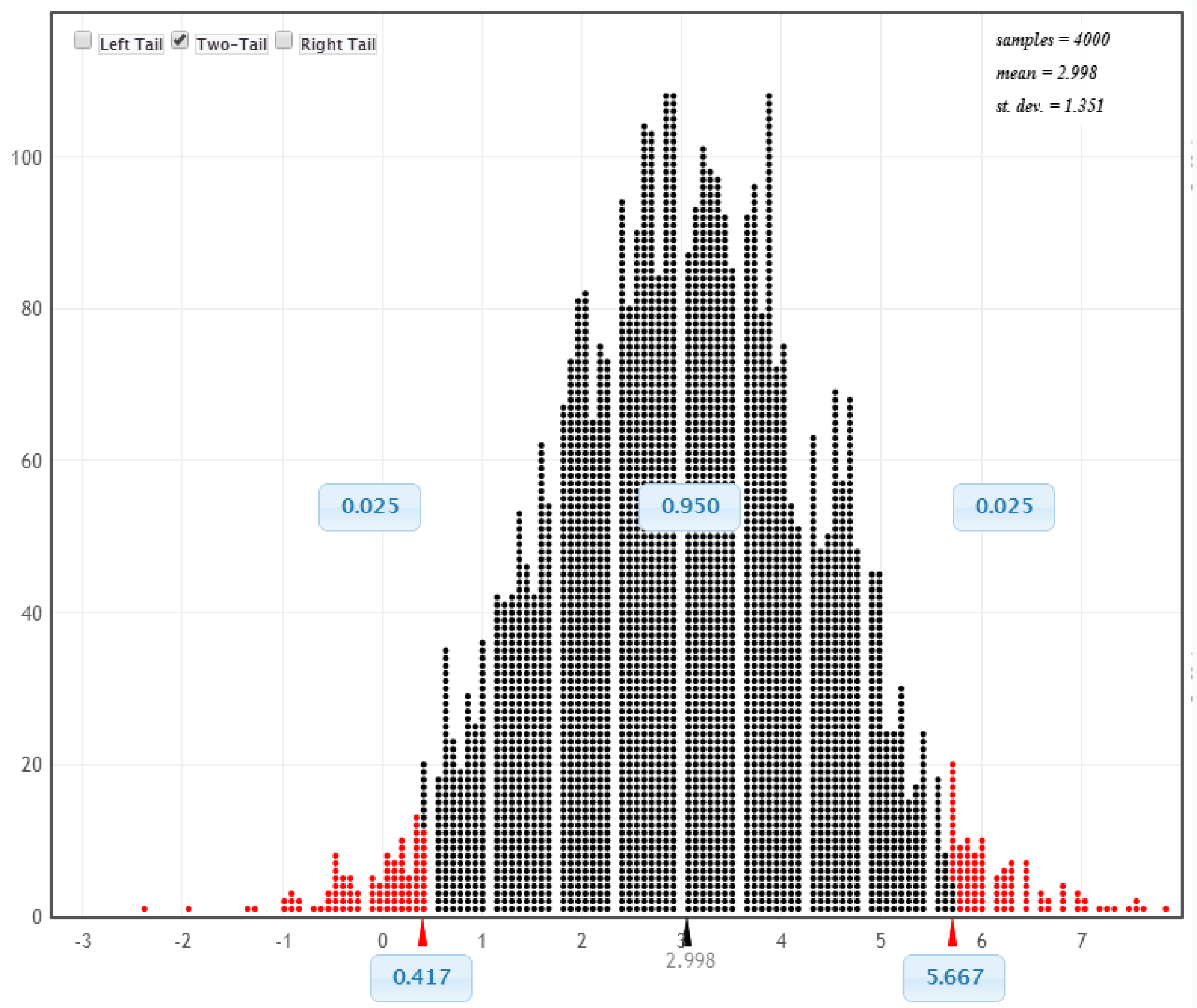
Informal strength of evidence against H0:



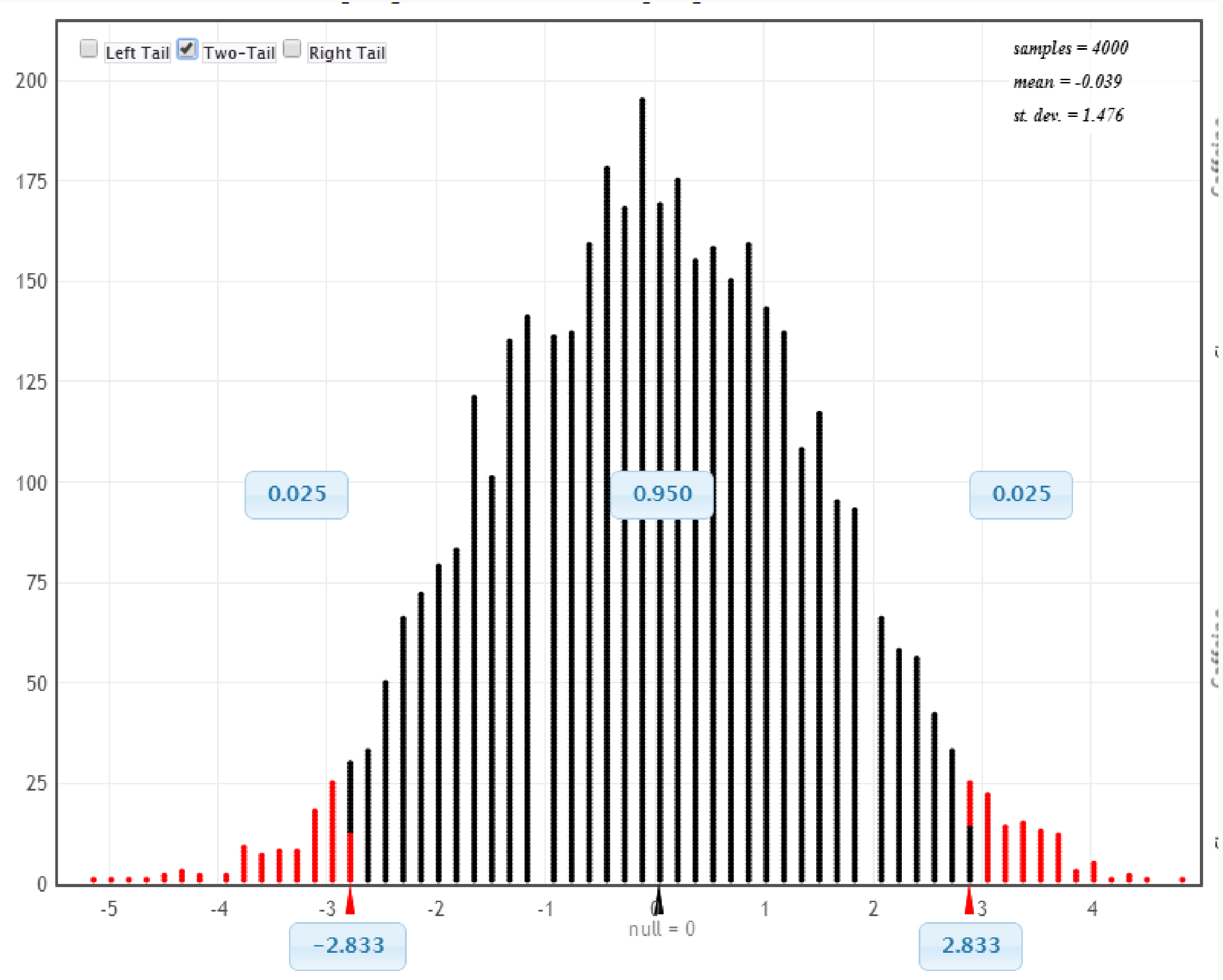
Strength of evidence

- *Smaller p-values give us stronger and stronger evidence for the alternative hypothesis.*
- *Larger p-values indicate little evidence for the alternative hypothesis.*

Example: bootstrap or randomization?



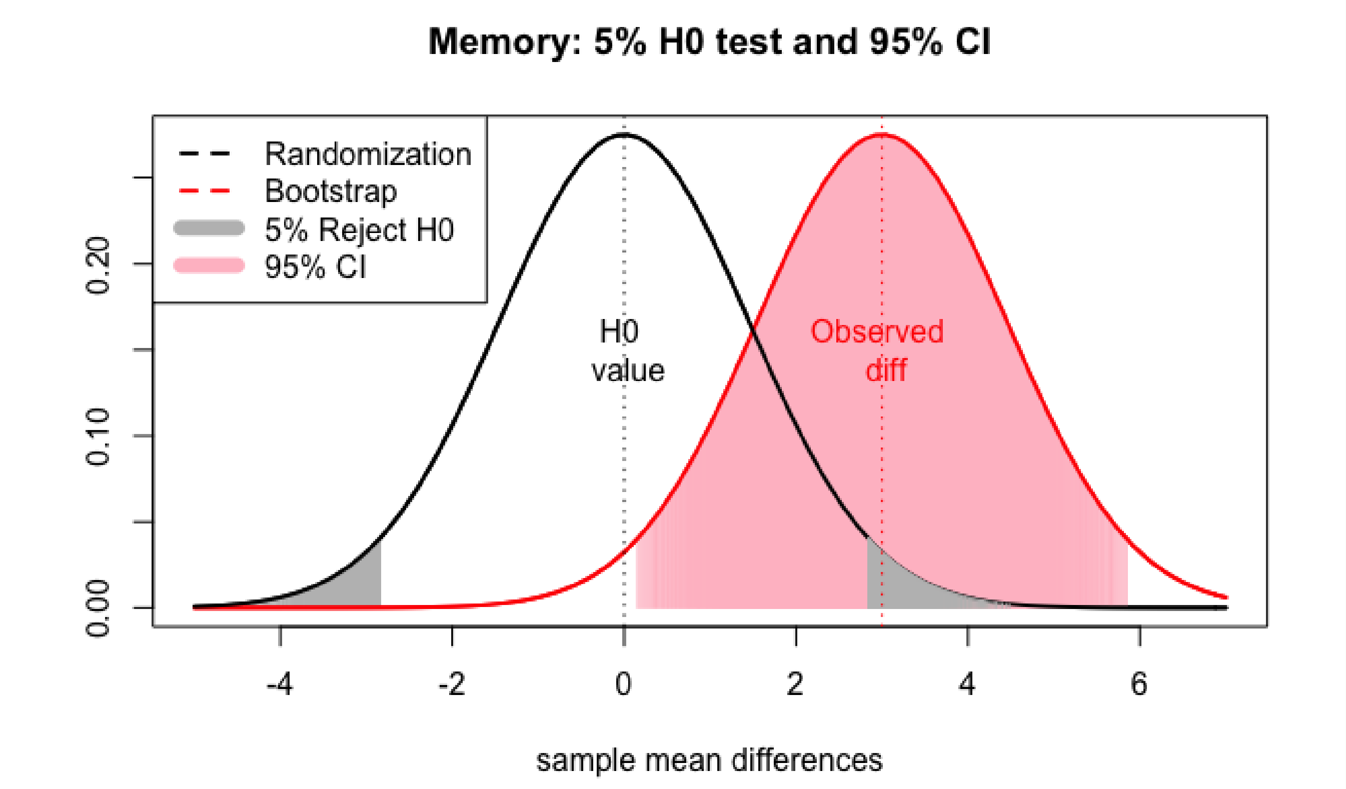
Example: bootstrap or randomization?



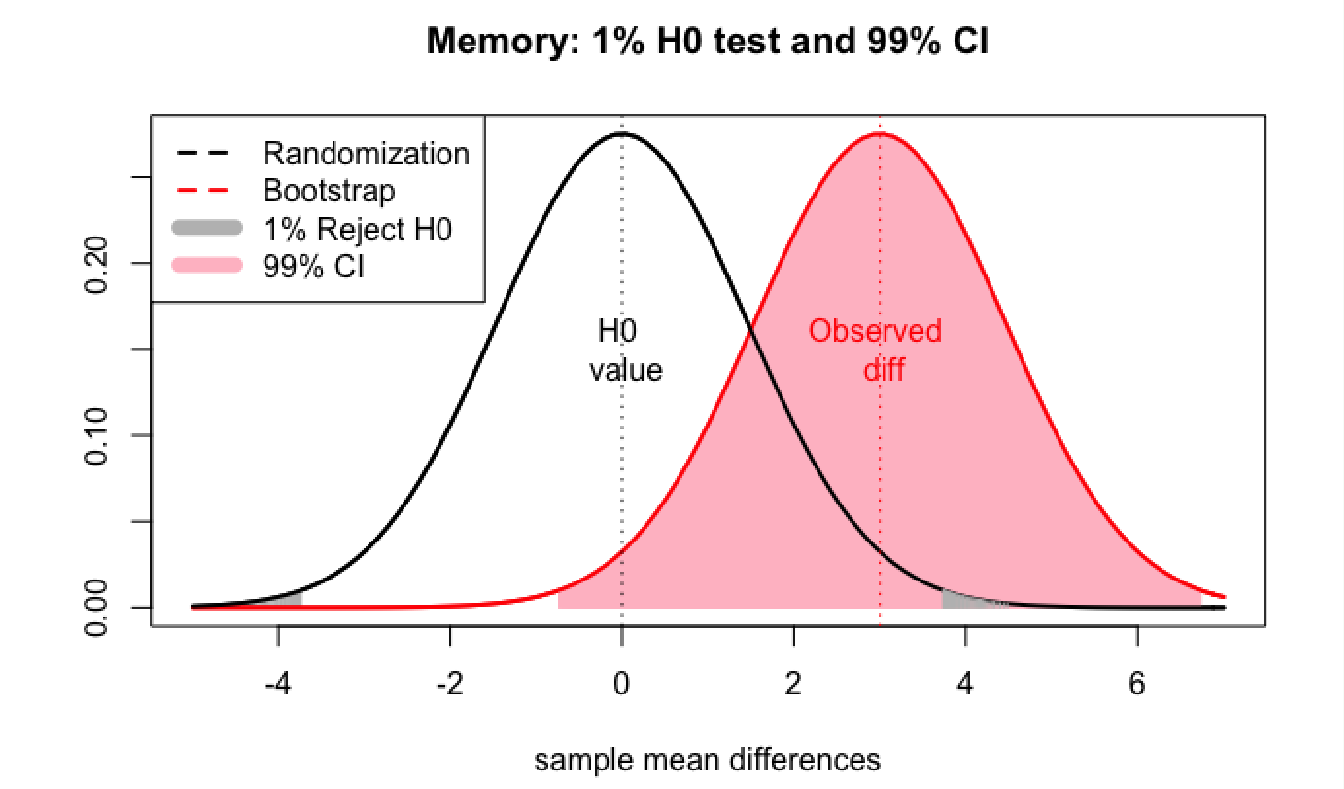
Using confidence intervals for hypothesis testing

If a 95% CI contains the parameter in $\mathbf{H}_0$, then a two-tailed test should not reject $\mathbf{H}_0$ at a 5% significance level.

If a 95% CI misses the parameter in $\mathbf{H}_0$, then a two-tailed test should reject $\mathbf{H}_0$ at a 5% significance level.



The 95% confidence interval misses null difference of 0. Reject the null at 5% level

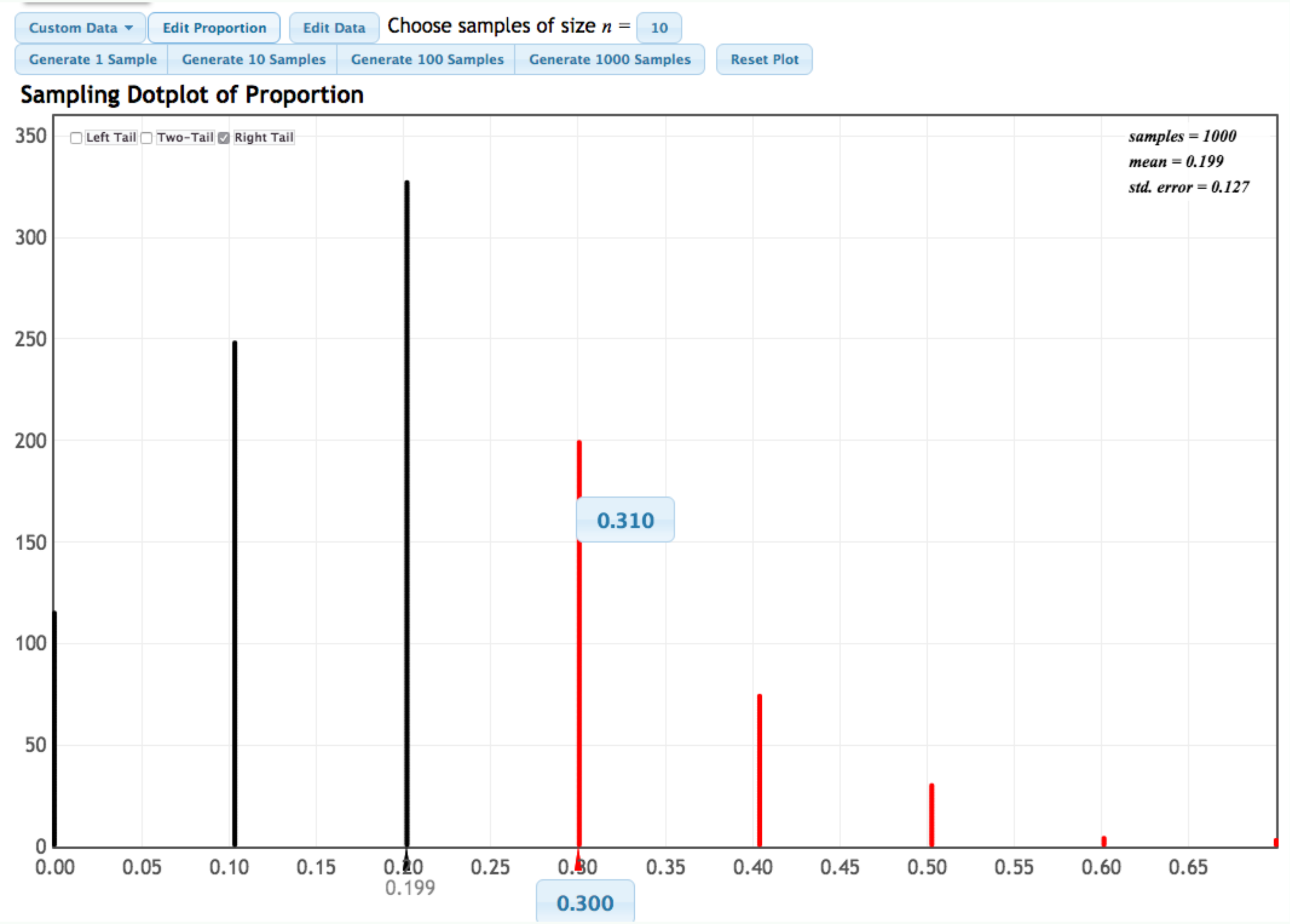


The 99% confidence interval contains null difference of 0. Do not reject the null at 1% level

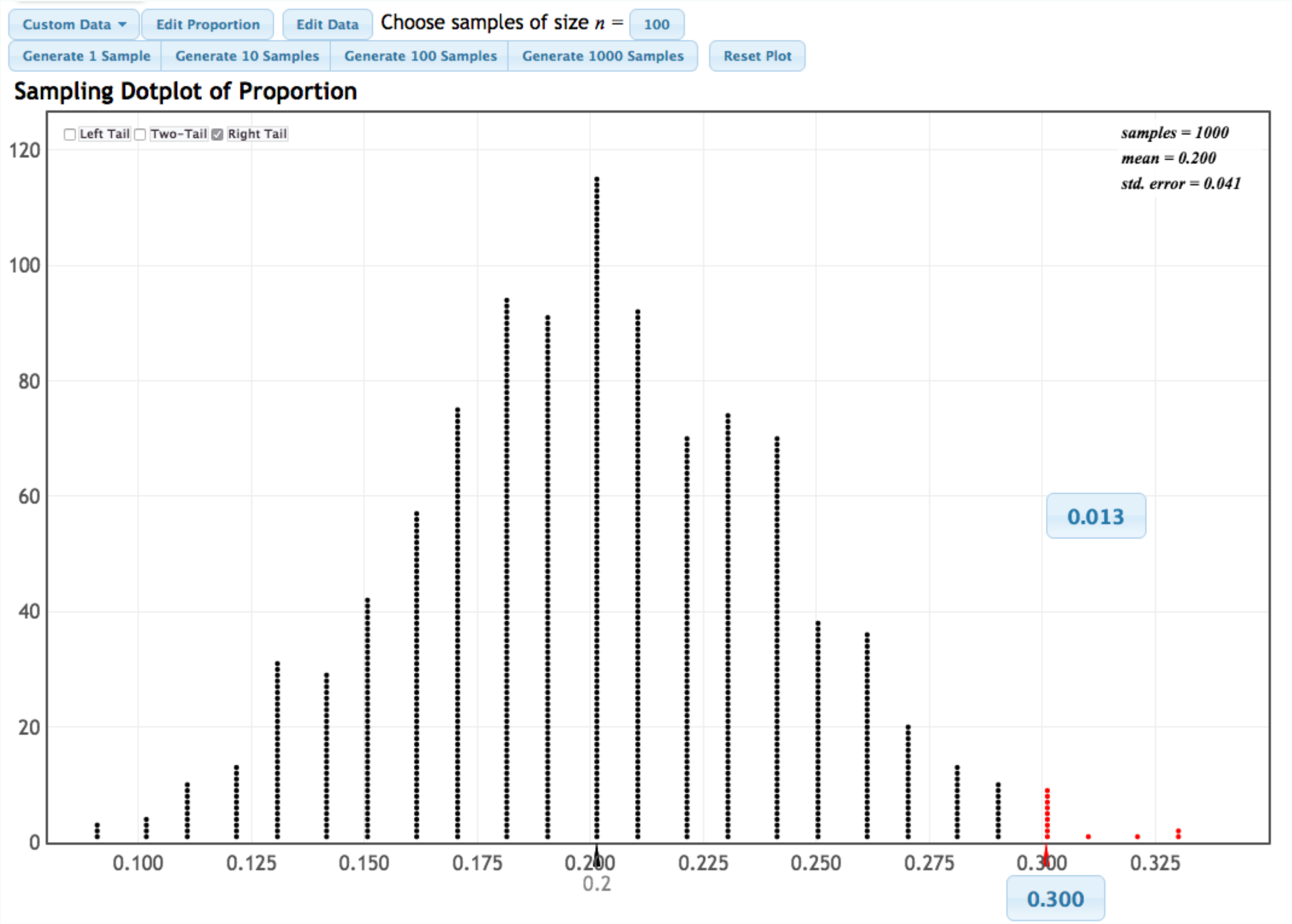
Sample Size and Statistical Significance

- ***With small sample sizes, even large differences or effects may not be significant.***
- ***With large sample sizes, even a very small difference or effect can be significant***

Randomization distribution with $n = 10$



Randomization distribution with $n = 100$



Multiple Testing/Comparison

When multiple hypothesis tests are conducted, the chance that at least one test incorrectly rejects a true null hypothesis increases with the number of tests.

- *If the null hypotheses are all true,* α *of the tests will yield statistically significant results just by random chance.*

Diet and Sex of Baby

- Are certain foods in your diet associated with whether or not you conceive a boy or a girl?
- To study this, researchers asked women about their eating habits, including asking whether or not they ate 133 different foods regularly

http://www.newscientist.com/article/dn13754-breakfast-cereals-boost-chances-of-conceiving-boys.html

Diet and Sex of Baby

For each of the 133 foods studied, a hypothesis test was conducted for a difference between mothers who conceived boys and girls in the proportion who consume each food

What are the null and alternative hypotheses?

Compare two populations: mothers who have boys vs. mothers who have girls

$\boldsymbol{p_b}$: proportion of mothers who have boys that consume the food regularly

$\boldsymbol{p_g}$: proportion of mothers who have girls that consume the food regularly

$$\mathrm{H}_0 : p_b = p_g$$
$$\mathrm{H}_a : p_b \neq p_g$$

Diet and Sex of Baby

A significant difference was found for breakfast cereal (mothers of boys eat more), prompting the headline

“Breakfast Cereal Boosts Chances of Conceiving Boys”

How might you explain this?

Random chance; several tests (about 6 or 7) are going to be significant, even if no differences exist

Diet and Sex of Baby

If there are NO differences (all 133 null hypotheses are true), about how many significant differences would be found using a = 0.05?

$$133 \pm 0.05 = 6.65$$

Expect about 6-7 statistically significant foods even if the rate of food consumption is equal for women who have boys and women who have girls

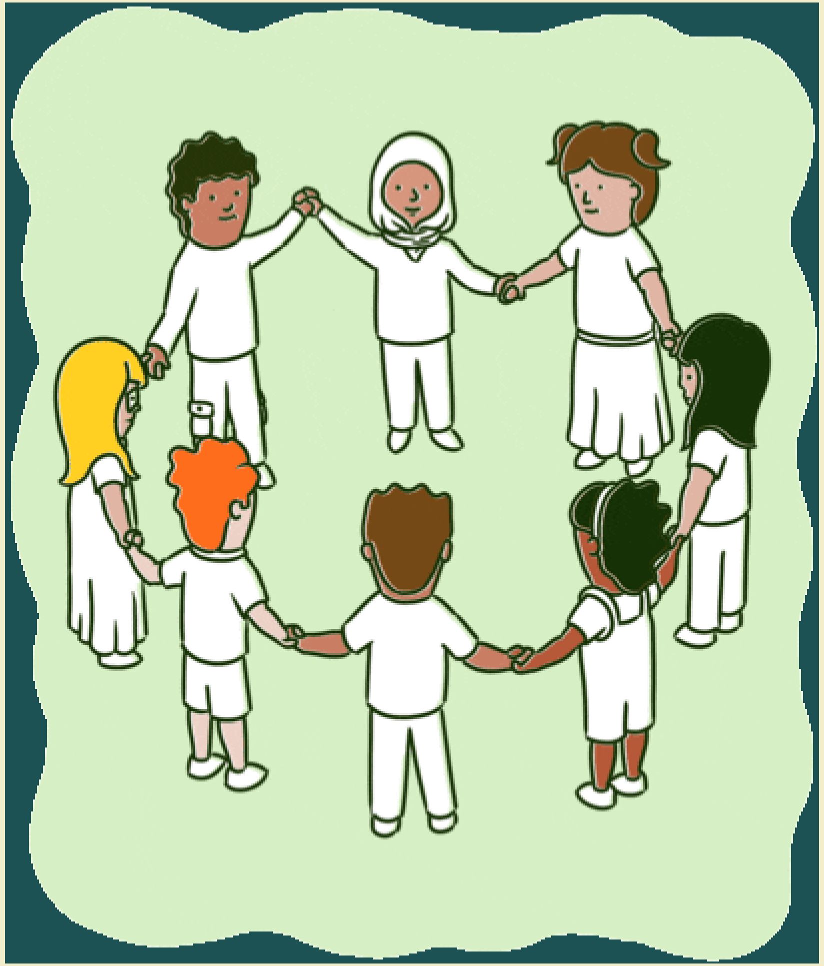
Multiple Comparisons

The most important thing is to be aware of this issue, and not to trust claims that are obviously one of many tests (unless they specifically mention an adjustment for multiple testing)

There are ways to account for this (e.g. Bonferroni's Correction), but these are beyond the scope of this class

Your Turn 1

10:00



Please go over the class activity for today and let me know if you have any questions.