Additional Testing Topics

Stat 120

Section 4.4 – 4.5

Days 13

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).

$$\text{p-value} < \alpha \quad \Rightarrow \quad \text{Reject } H_0$$
$$\text{p-value} \geq \alpha \quad \Rightarrow \quad \text{Do not Reject } H_0$$

- 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

Decision

Truth	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 suppose to work).
  $H_0: defendant\ is\ innocent \quad vs. \quad H_A: defendant\ is\ guilty$
- 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 α=0.05; reported p-value of 0.46
- Decision?
  - Do not reject 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 α=0.05; data gives p-value of 0.048
- Decision?
  - Reject 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.

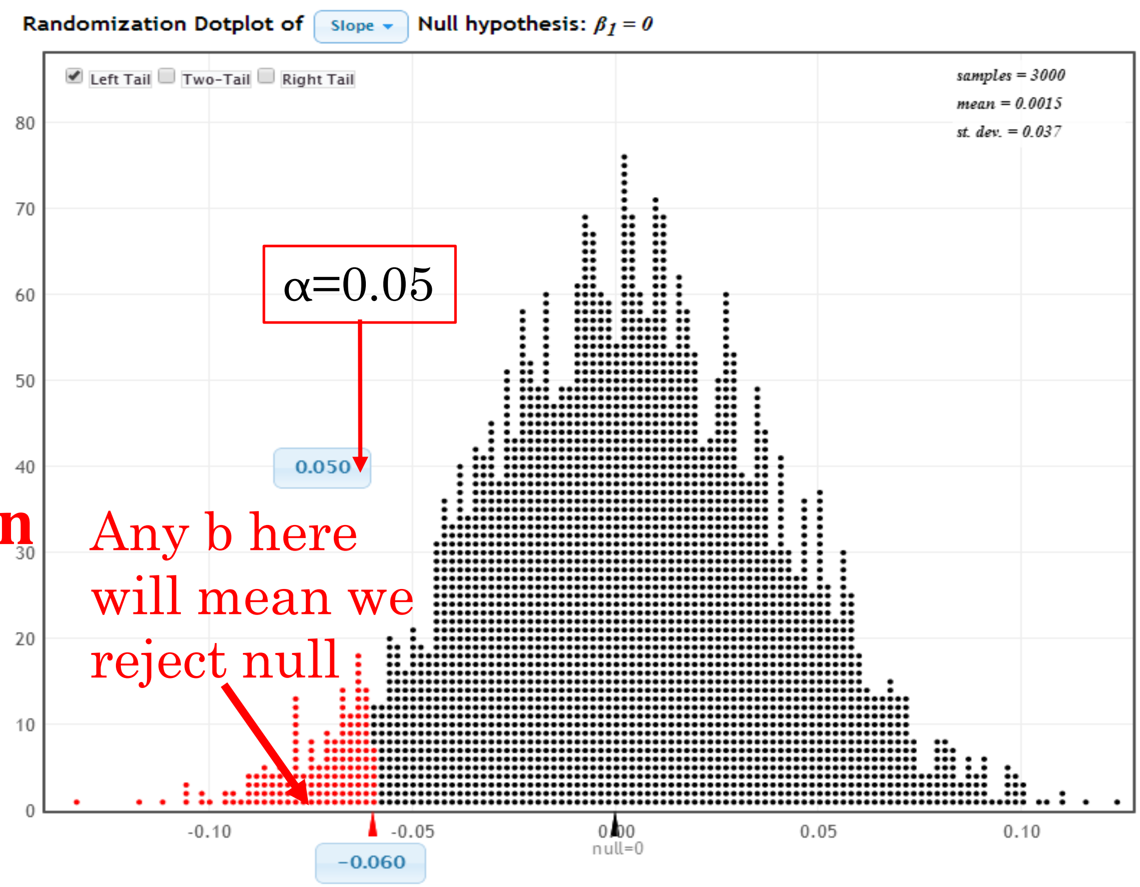
- The nice thing about type I errors is that we can control the chance of such an error…

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<-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)!!!



Selecting a significance level

- Common level is $\alpha = 0.05$, but…
- 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)
- 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
- 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

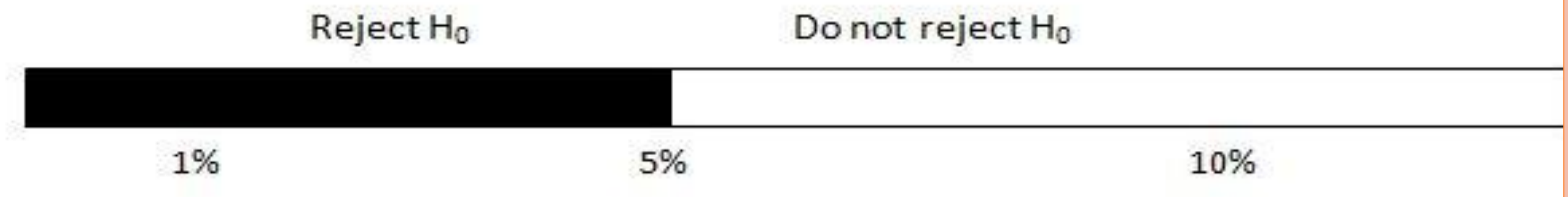
Probability of Type II Error

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

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

Statistical Conclusions

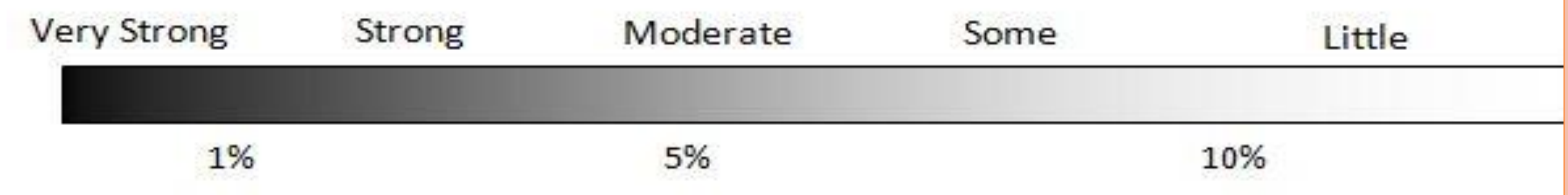
Formal decision of hypothesis test, based on $\alpha = 0.05$:



statistically significant

not statistically significant

Informal strength of evidence against H_0:



Less Formal Decisions

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

- For Formal and Informal decisions:
  - Consider how sample size may play a role in your test results
  - **If an "effect" is found (evidence for your research hypothesis), quantify the effect with a confidence interval.**

Beware of the Prosecutor's fallacy

- Misinterpreting evidence:
  - **Correct**: If a **defendant is innocent (assumption)**, the **probability** that his **DNA will match** the crime scene DNA is 1 in a million. (1 in a million people have this profile)
  - **Incorrect**: Since the **DNA matched (assumption)**, there is a 1 in a million **probability** that the **defendant is innocent**.
- Same issues can arise with p-values
  - **Correct**: If the null is true, the probability of observing a stat as, or more, extreme than the one observed is 0.01.
  - **Incorrect**: Given this stat, the probability of our null being true is 0.01.

EXAMPLE :

- Hypotheses: $\boldsymbol{p}$ = true proportion of users who relapse
- H_0: $\boldsymbol{p_D}$ - $\boldsymbol{p_L} = 0$ vs H_a: $\boldsymbol{p_D}$ - $\boldsymbol{p_L} < 0$
- Sample statistic used to test claims: $\hat{p}_D - \hat{p}_L$
- Observed statistic:

$$\hat{p}_D - \hat{p}_L = \frac{10}{24} - \frac{18}{24} \approx -0.333$$

- With the difference D-L, we have a left-tail test.
- The proportion of resamples with a sample difference of -0.333 or less is about 2%.
- The difference is statistically significant at 5% level.
- We have evidence that Despramine leads to a lower relapse rate than Lithium.

EXAMPLE : P-VALUE = 2%, OBSERVED DIFFERENCE = -33% (D – L)

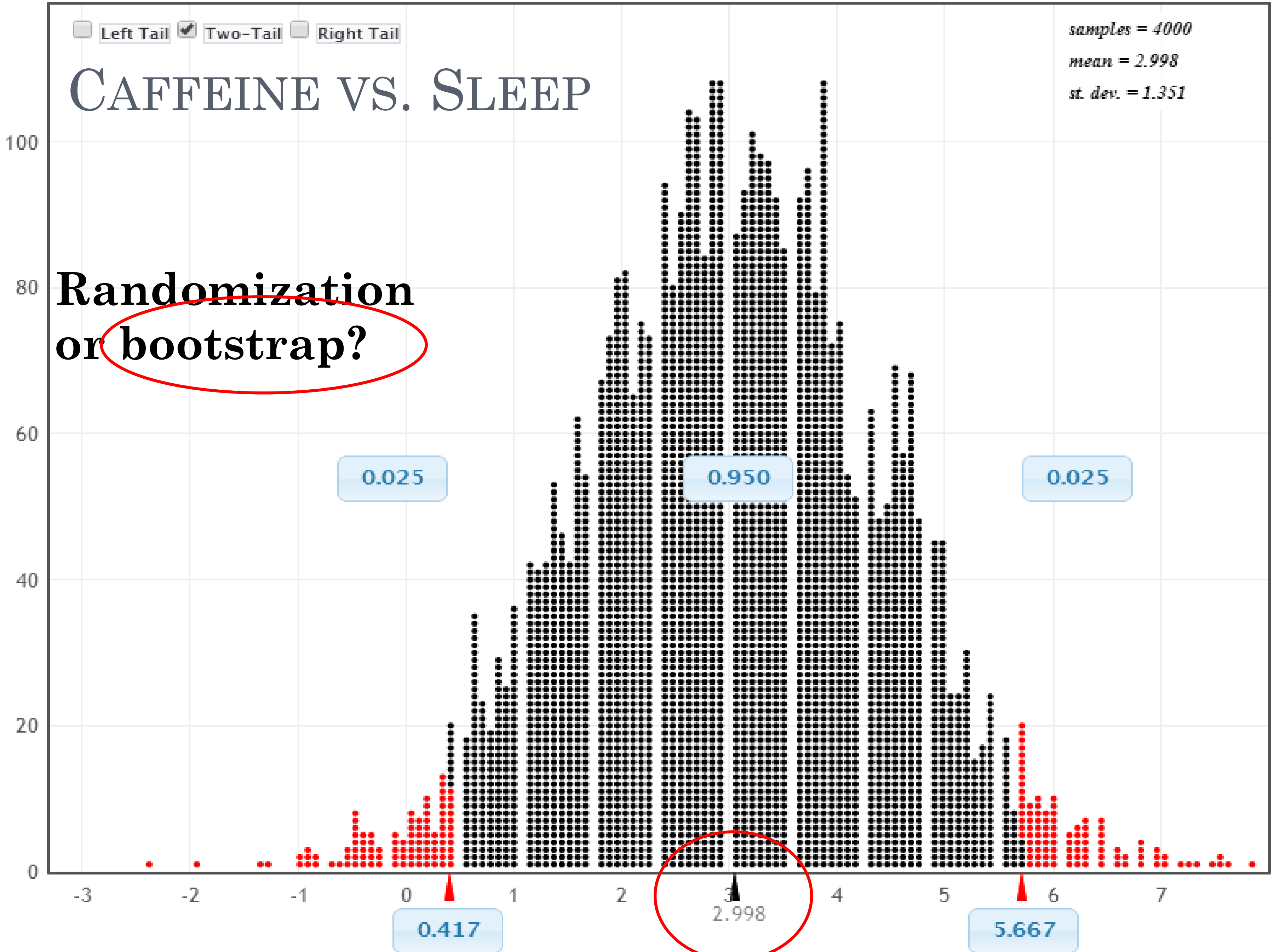
- True or False:
- "There is a 2% chance that the true relapse rates are equal."
- False!
  - The parameters $\boldsymbol{p_D}$ and $\boldsymbol{p_L}$ are fixed (unknown) values.
  - The null is either true or not true (no chance involved)
  - Only the sample proportions are random – depend on the treatment randomization

Wed. example : p-value = 2%, observed difference = -33% (D – L)

- True or False:
- "There is a 2% chance of seeing at least 33% fewer people relapse using Despramine just by chance."
- True!
  - "Just by chance" implies that that chance is the only reason for the difference in relapse rates – i.e. that the true rates of relapse are the same (null assumption).
  - The p-value measures the likelihood of seeing a difference as extreme, or more extreme, than the observed difference.

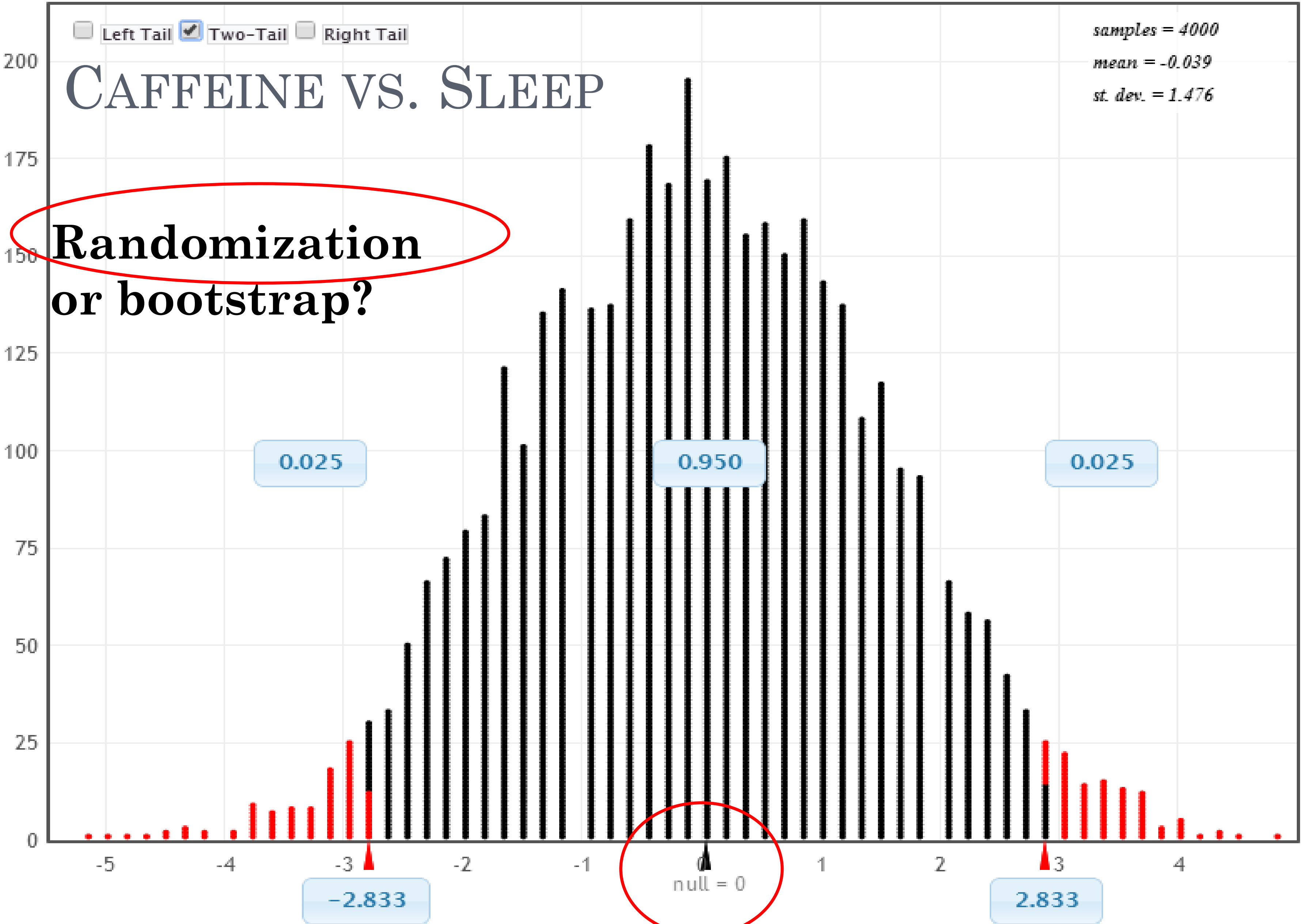
Caffeine vs. Sleep

Randomization or bootstrap?



Caffeine vs. Sleep

Randomization or bootstrap?



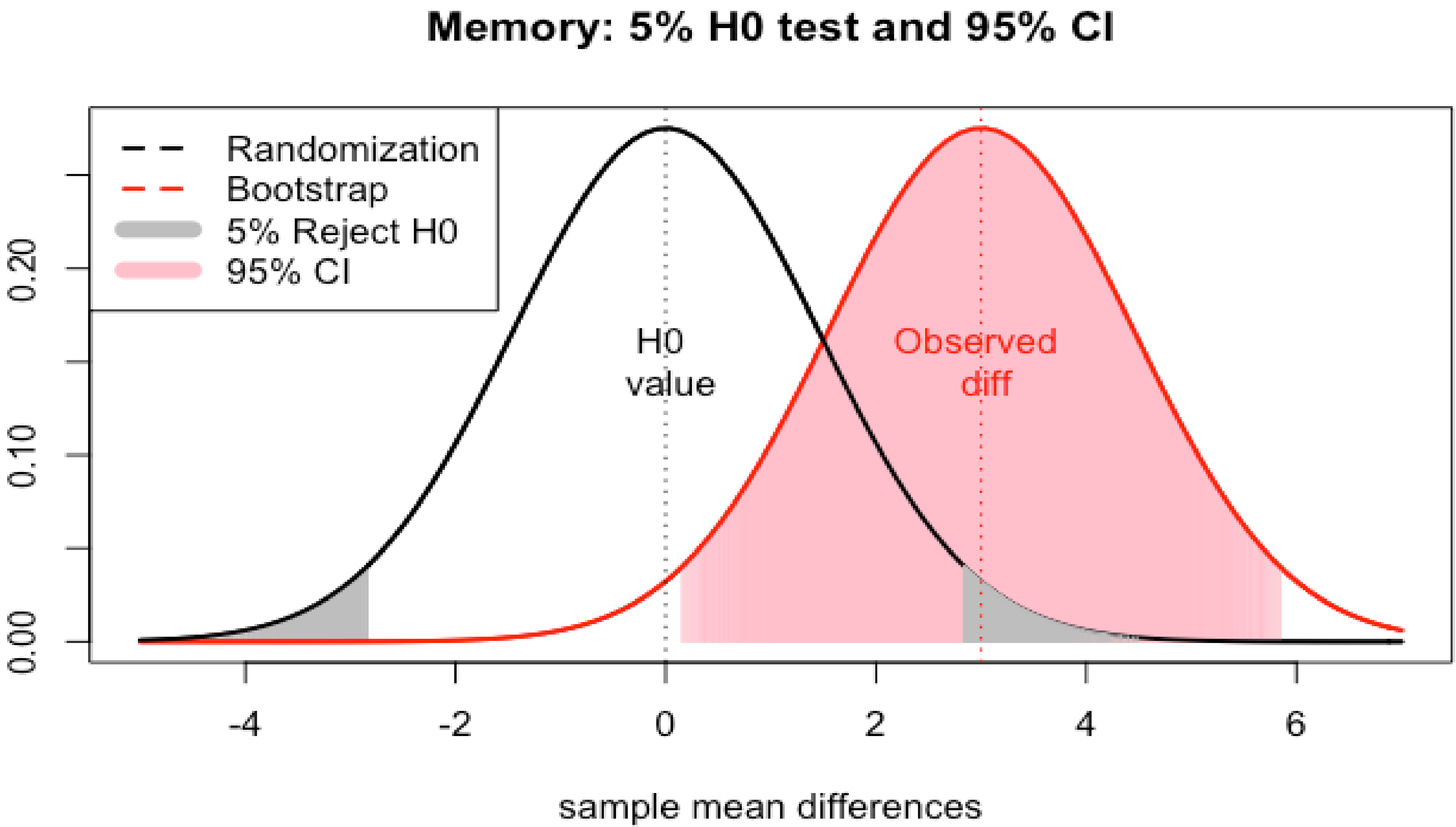
Intervals and Tests

If a 95% CI *contains* the parameter in H_0, then a two-tailed test should *not reject* H_0 at a 5% significance level.

If a 95% CI *misses* the parameter in H_0, then a two-tailed test should *reject* H_0 at a 5% significance level.

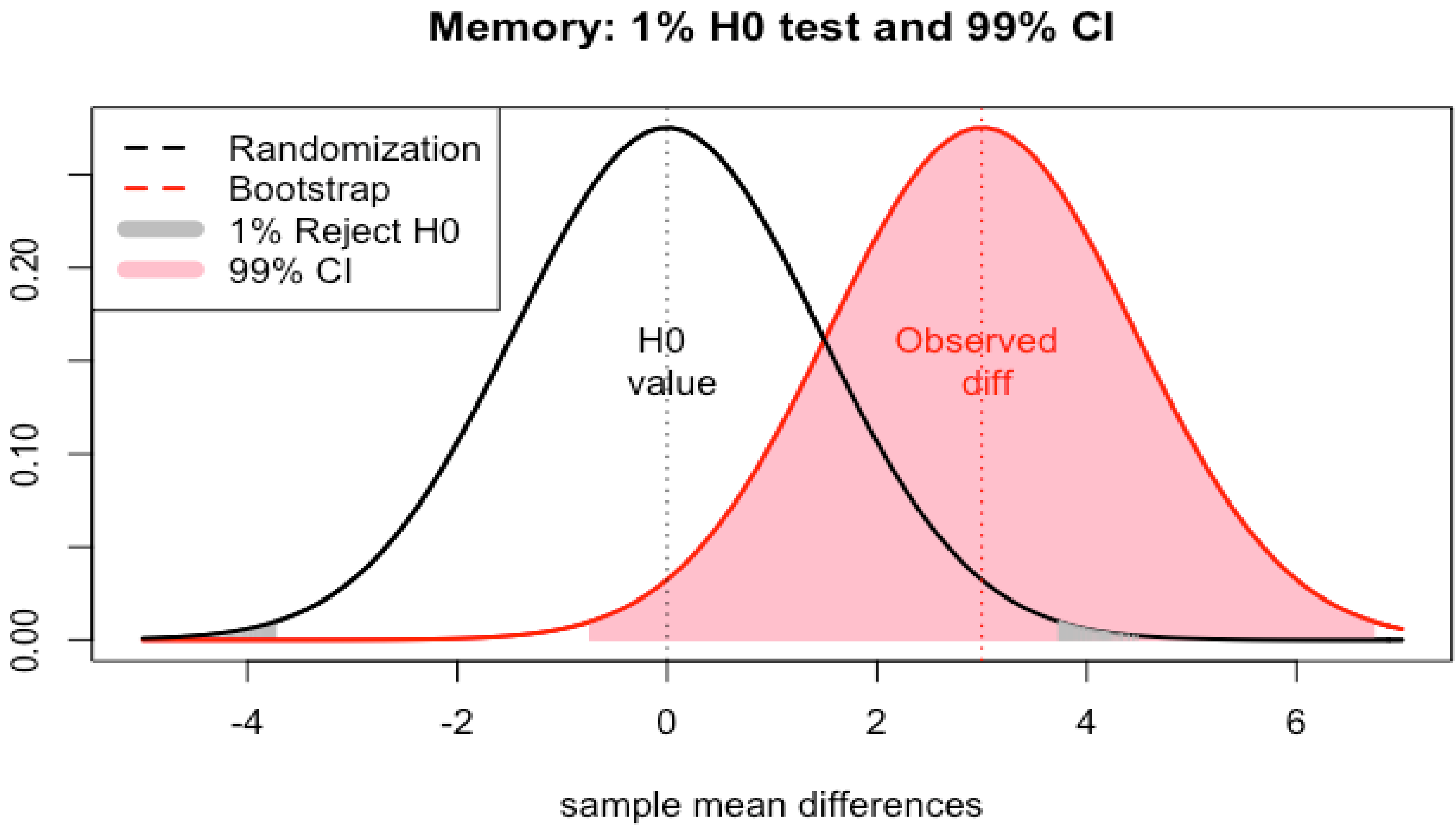
Caffeine vs. Sleep

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



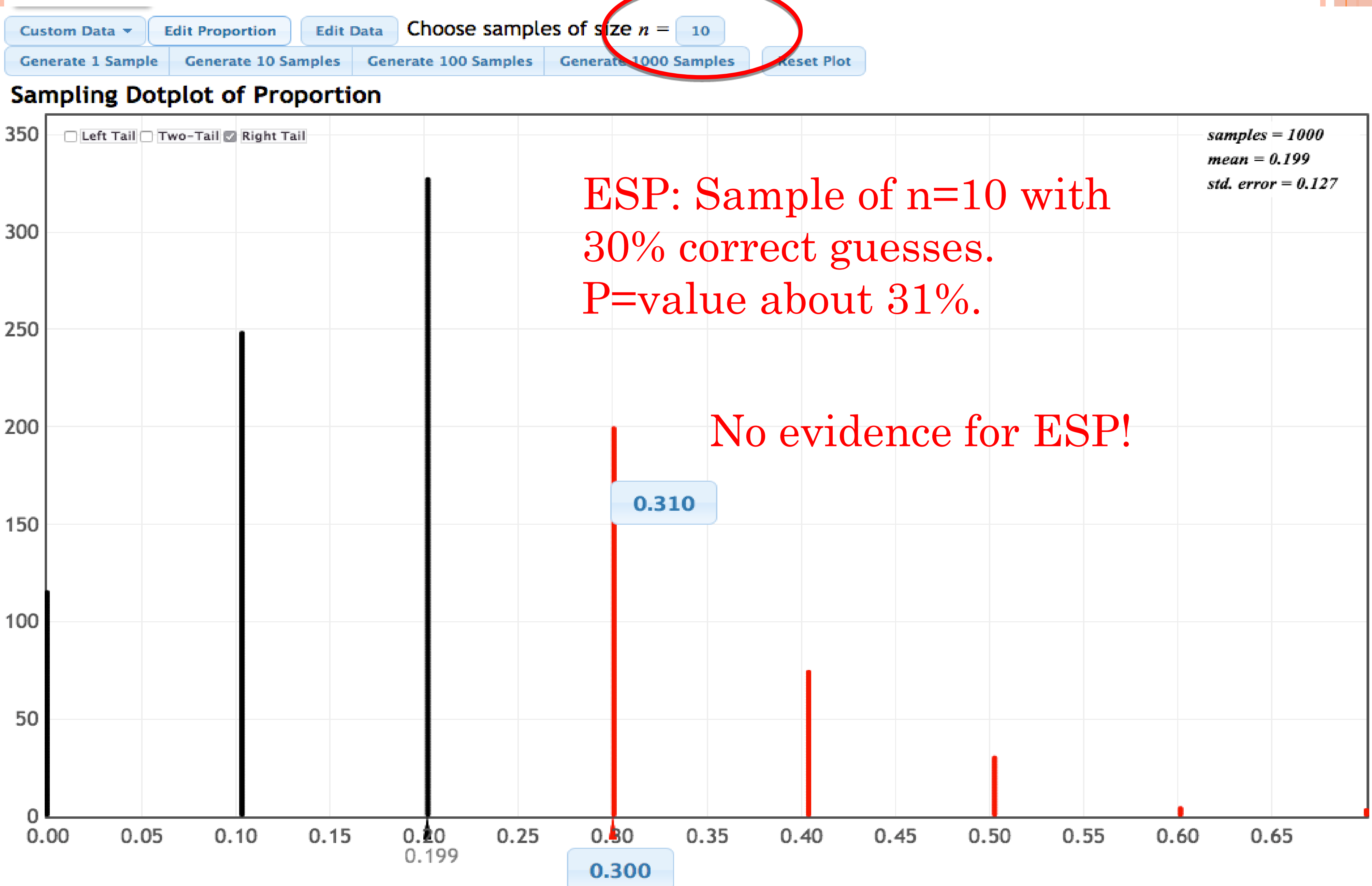
Caffeine vs. Sleep

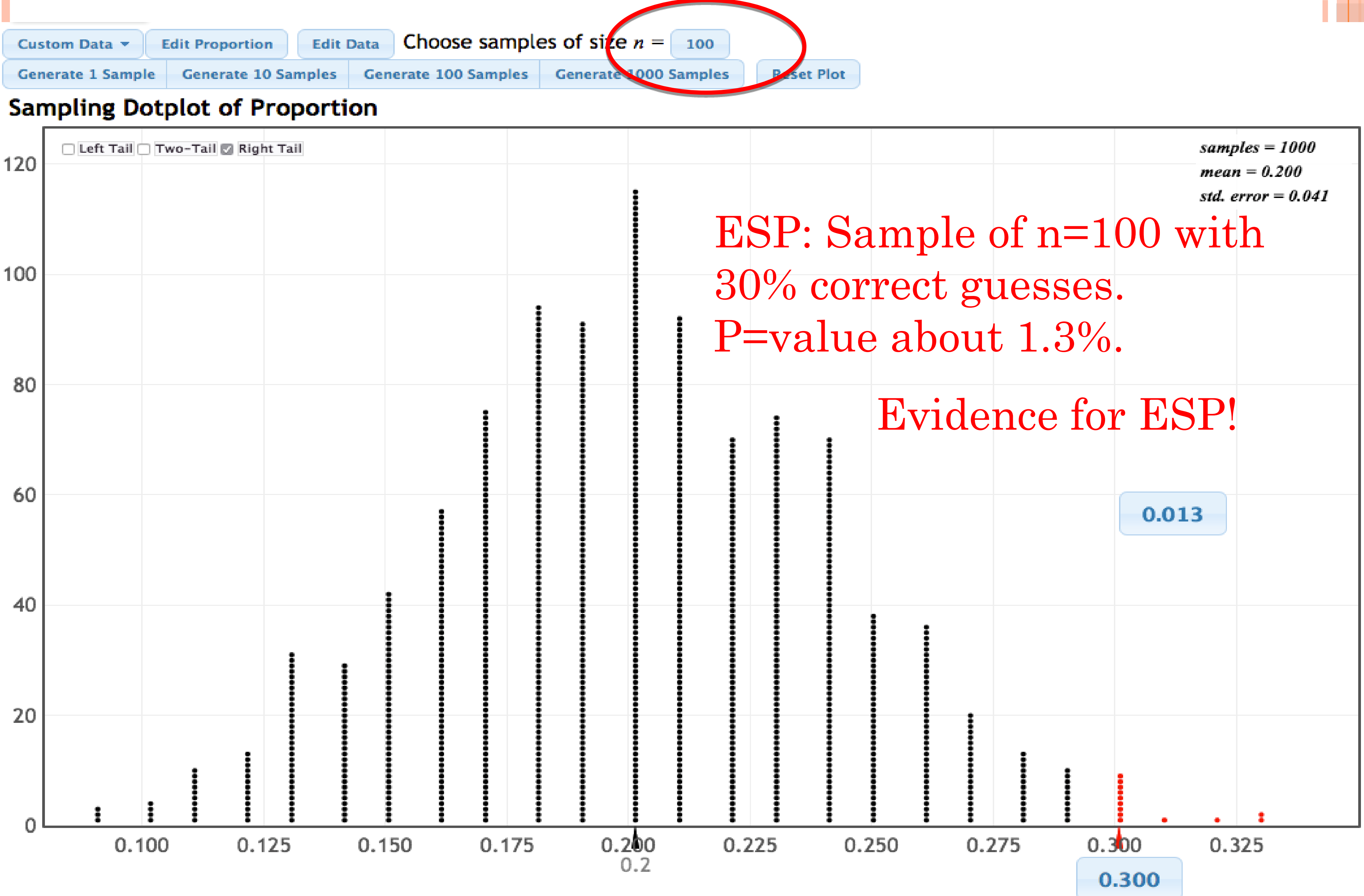
The 99% confidence interval contains null difference of 0. Do not reject the null at 1% level



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


Custom Data
Edit Proportion
Edit Data
Choose samples of size n = 10
Generate 1 Sample
Generate 10 Samples
Generate 100 Samples
Generate 1000 Samples
Reset Plot
Sampling Dotplot of Proportion
Left Tail
Two-Tail
Right Tail
samples = 1000
mean = 0.199
std. error = 0.127
ESP: Sample of n=10 with
30% correct guesses.
P=value about 31%.
No evidence for ESP!
0.310
0.199
0.300
0, 50, 100, 150, 200, 250, 300, 350
0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65


Custom Data
Edit Proportion
Edit Data
Choose samples of size n = 100
Generate 1 Sample
Generate 10 Samples
Generate 100 Samples
Generate 1000 Samples
Reset Plot
Sampling Dotplot of Proportion
Left Tail
Two-Tail
Right Tail
samples = 1000
mean = 0.200
std. error = 0.041
ESP: Sample of n=100 with
30% correct guesses.
P=value about 1.3%.
Evidence for ESP!
0.013
0.300
0.2
0.100
0.125
0.150
0.175
0.200
0.225
0.250
0.275
0.300
0.325
0
20
40
60
80
100
120

Statistical vs Practical Significance

- A statistically significant result is not always **practically significant**, especially with large sample sizes
    - Ask yourself (or an expert) if the difference or effect size is meaningful.

Publication Bias

- ***publication bias*** refers to the fact that usually only the significant results get published
- The one study that turns out significant gets published, and no one knows about all the insignificant results
- This combined with the problem of ***multiple comparisons*** can yield very misleading results

Multiple Testing/Comparisons

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

p_b: proportion of mothers who have boys that consume the food regularly
p_g: proportion of mothers who have girls that consume the food regularly

$$H_0: p_b = p_g$$
$$H_a: p_b \neq p_g$$

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 $\alpha = 0.05$?

$133 \times 0.05 = 6.65$

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

- ***This is a serious problem***
- 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