Resampling methods

Readings for today

- Chapter 5: Resampling methods. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R (Vol. 6). New York: Springer.

Topics

1. Permutation tests

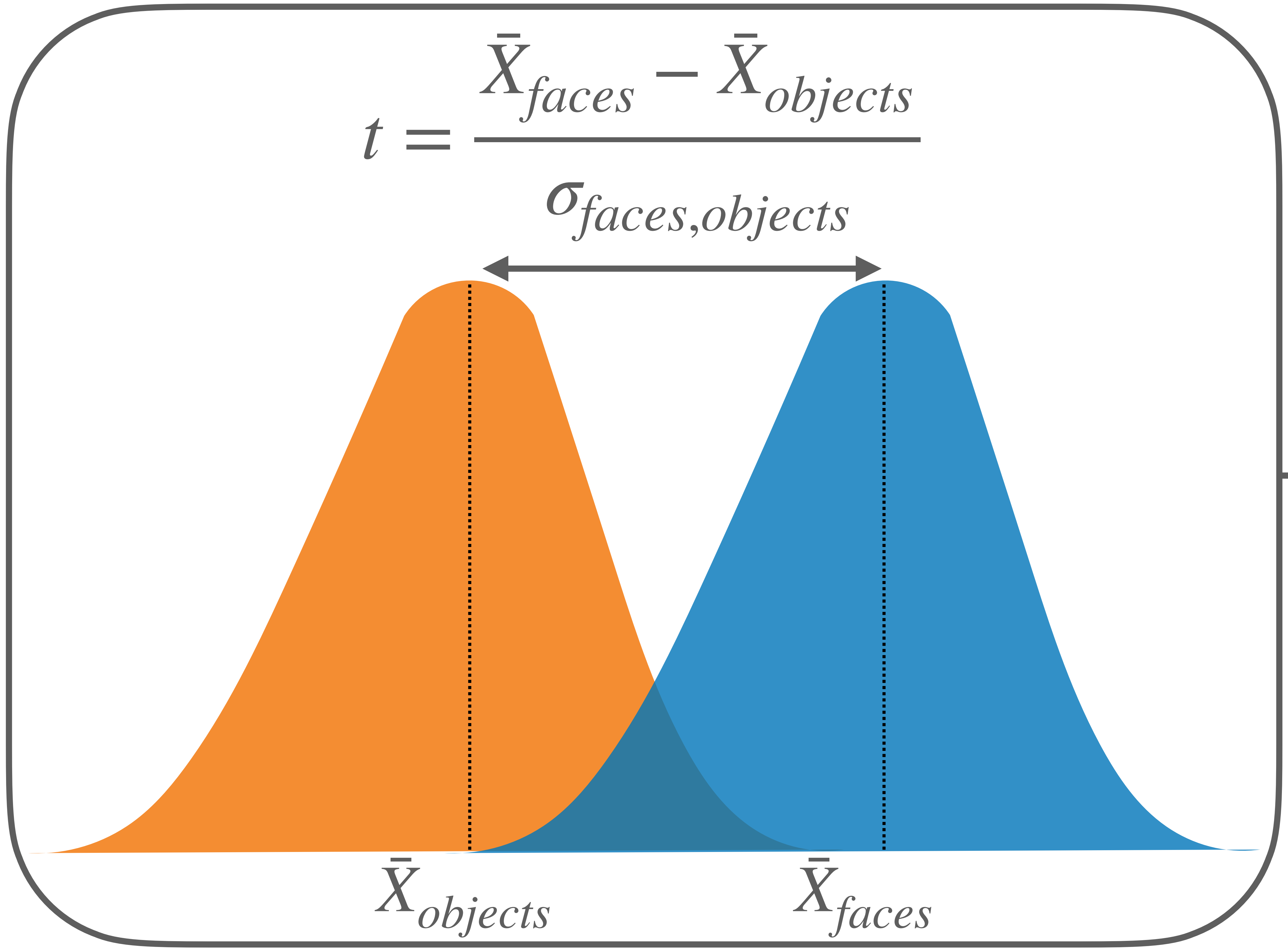
2. Bootstrapping

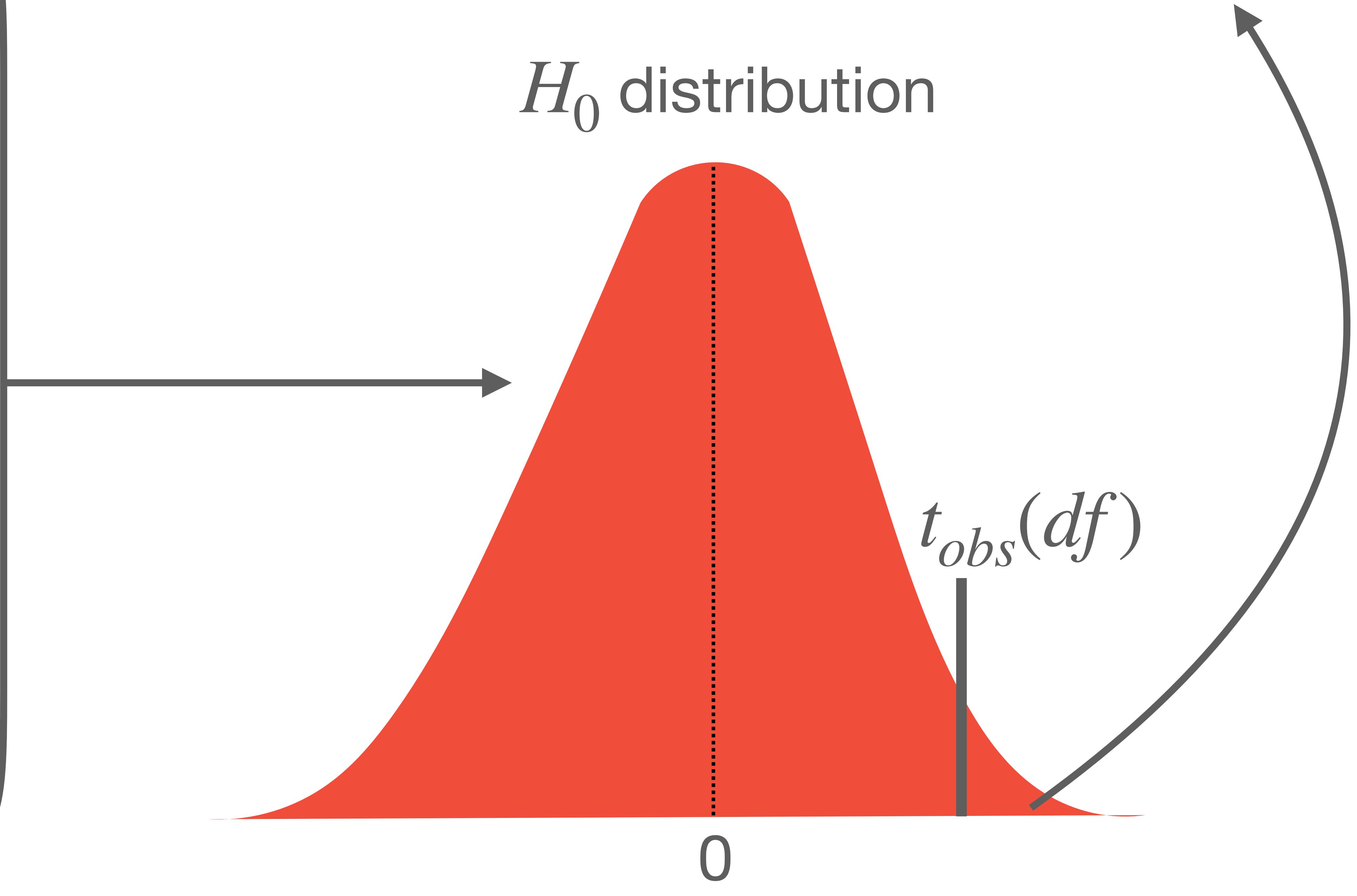
Permutation tests

Null Hypothesis (H_0)

Q: Do IT cells fire more to images of faces than objects?

- The probability that you see this difference in means if the real difference is zero
- Assumed distribution of differences (t distribution)





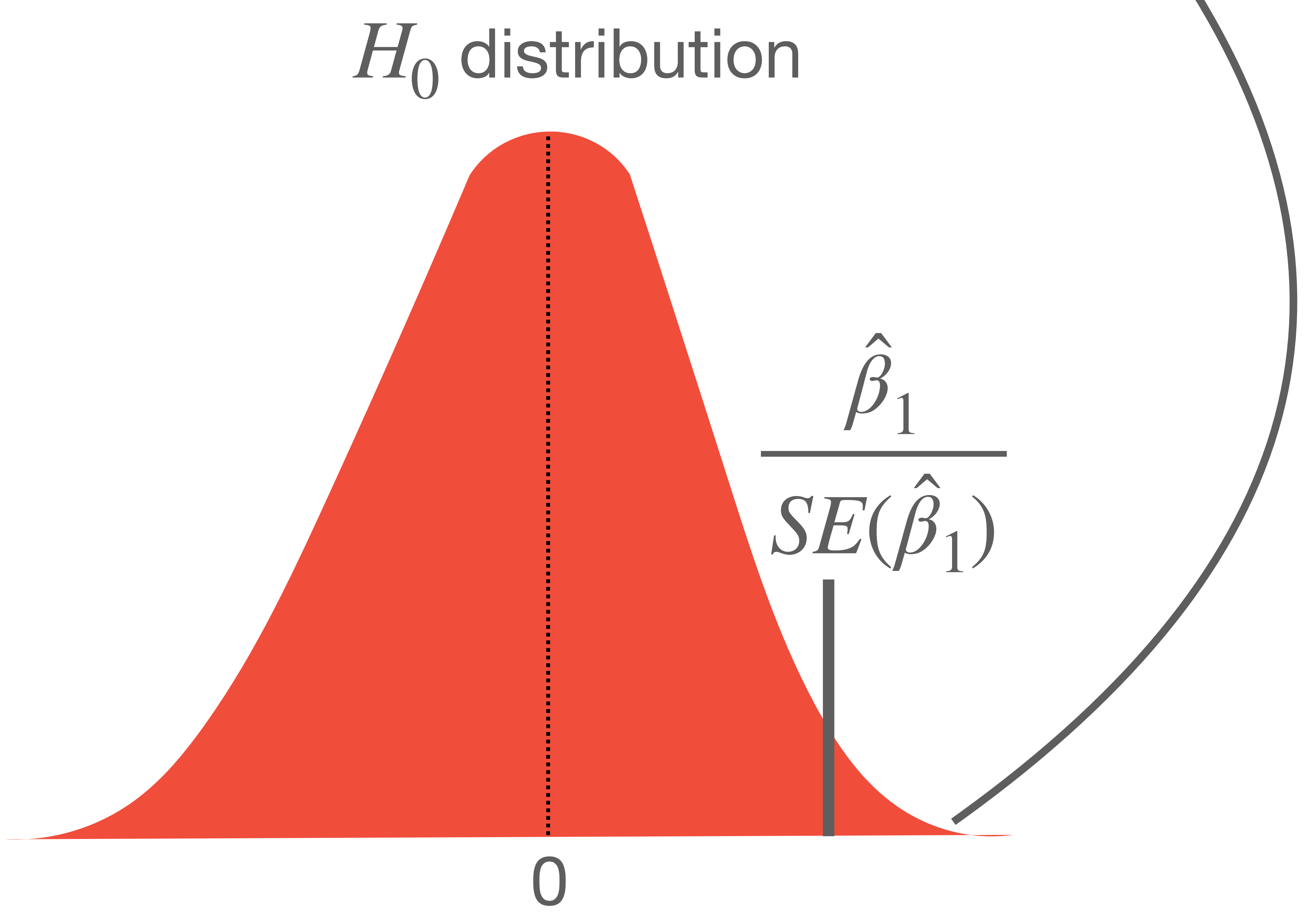
Null Hypothesis (H_0)

Q: Do IT cells fire more to images of faces than objects?

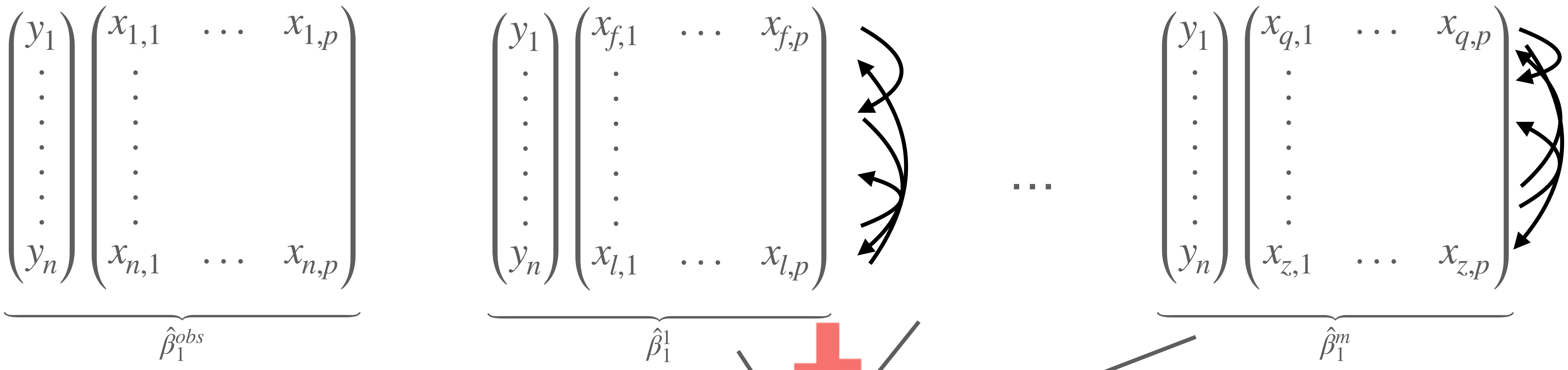
$$Y_{FR} = \hat{\beta}_0 + \hat{\beta}_1 X_{faces}$$

$$X_{faces} : \begin{cases} 1, & \text{face image} \\ 0, & \text{object image} \end{cases}$$

- The probability that you see the effect of faces if the real difference is zero.
- Assumed distribution if $\hat{\beta}_1 = 0$.

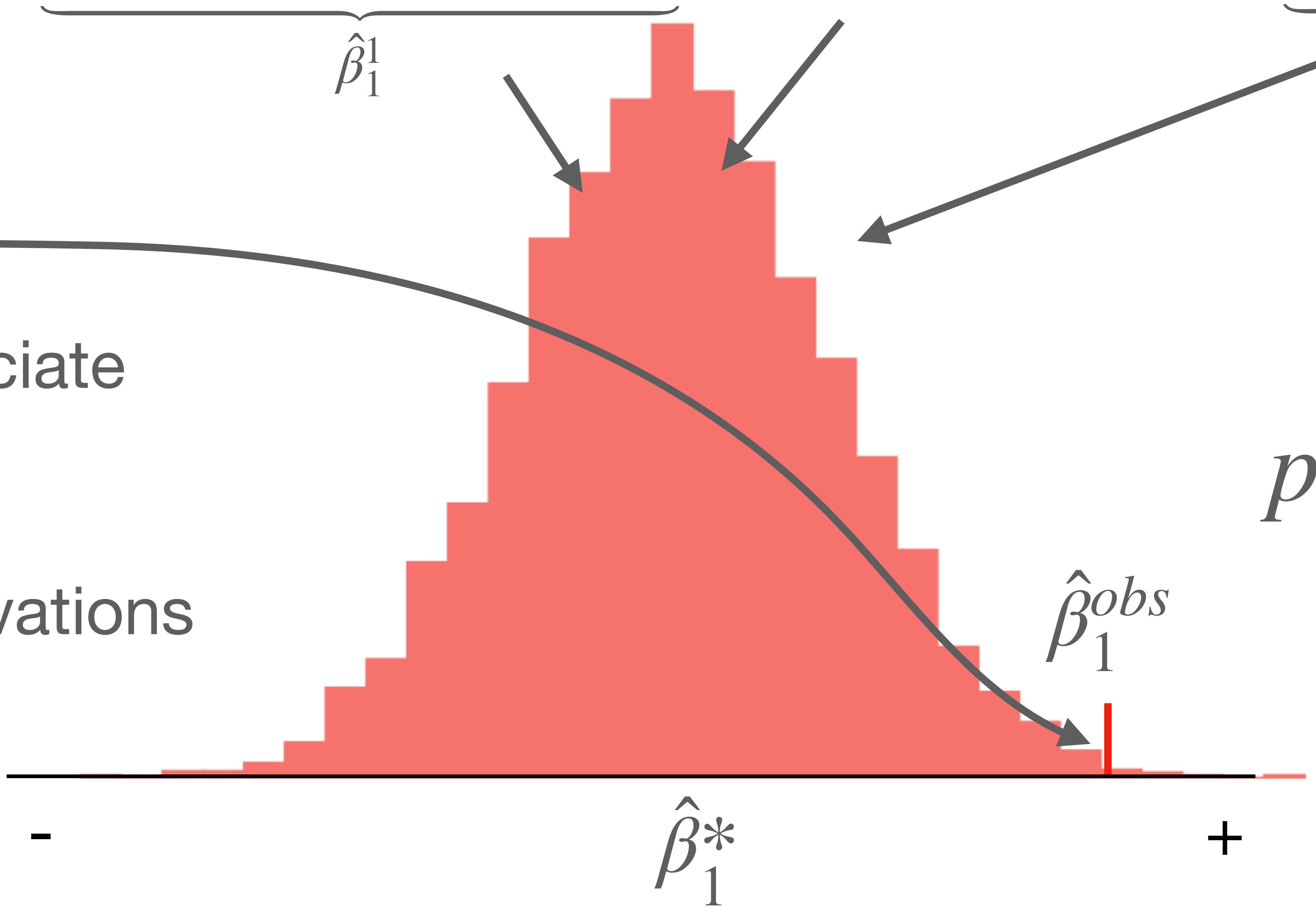


Directly generating H_0



H_0: Variation in X does not associate with variation Y.

mapping across observations is arbitrary



$$p = \frac{\sum_{i=1}^{m} I(\hat{\beta}_1^i > \hat{\beta}_1^{obs})}{m}$$

The permutation algorithm

Goal: Directly calculate a H_0 distribution from your data.

Step 1: Calculate the observed effects in your data $\hat{f}^{obs}(X)$ & set aside the relevant parameters for evaluating your hypothesis (e.g., $\hat{\beta}_i^{obs}$).

Step 2: Run m iterations where, on each iteration:

- A new variable set $X*$ is generated by randomly reassigning (permuting) observations *within variables relevant to your hypothesis*.
- A new model $\hat{f}*$ is calculated from $X*$.
- All relevant parameter (e.g., $\hat{\beta}_i^*$) are stored.

Step 3: Compare the observed (unpermuted) parameters (e.g., $\hat{\beta}_i^{obs}$) against the distribution of parameters generated from the set of $\hat{f}*$.

Example: regression

Q: After controlling for age, does income level negatively associate with subjective stress level?

$$Y_{stress} = \hat{\beta}_0 + \underbrace{\hat{\beta}_1 X_{income}}_{\text{target}} + \underbrace{\hat{\beta}_2 X_{age}}_{\text{nuisance}}$$

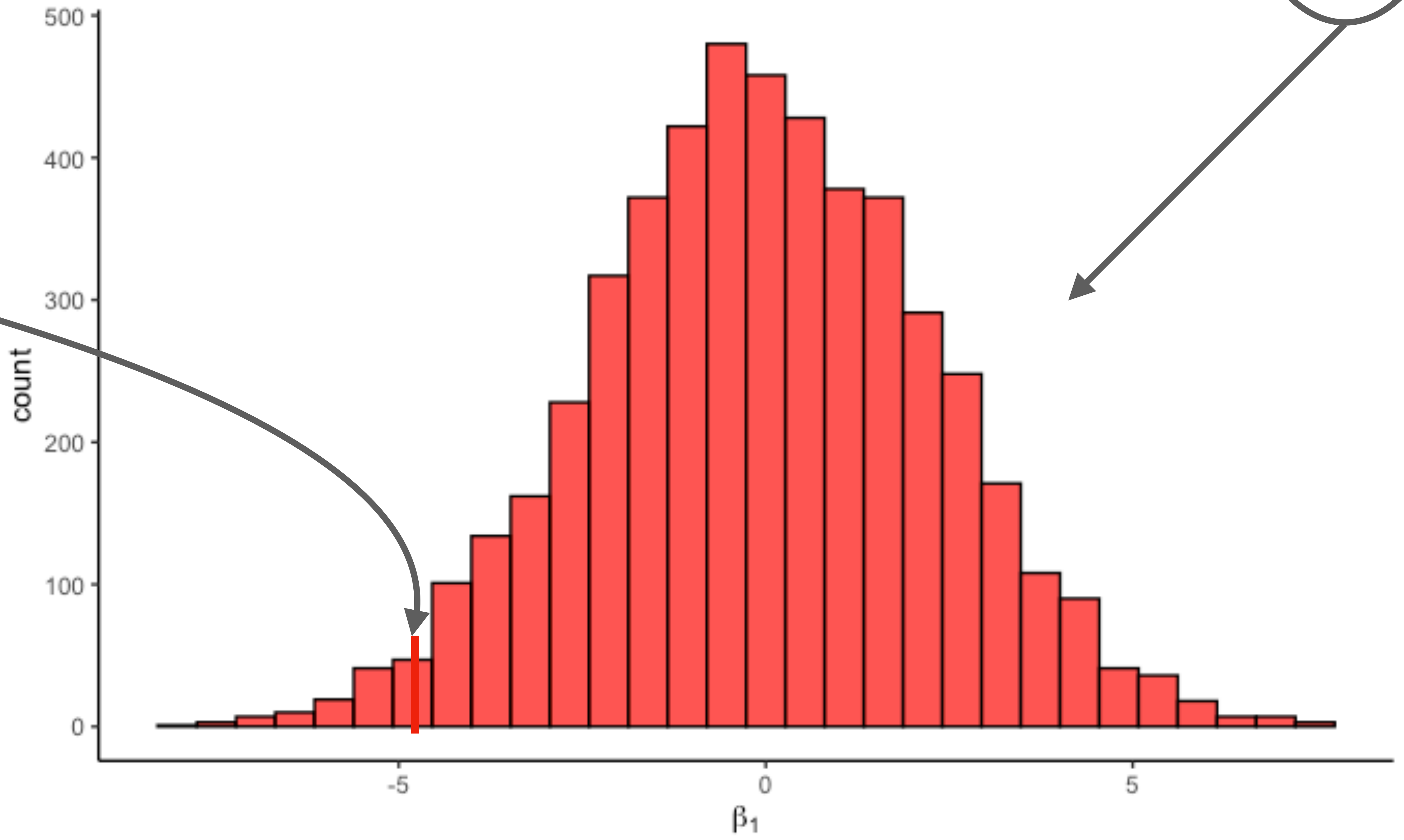
$$\hat{\beta}_1 = -4.2$$

$$p = \frac{\sum_{i=1}^{m} (I(\hat{\beta}_1^i < \hat{\beta}_1^{obs})}{m} = \frac{164}{5000} = 0.0328$$

$$Y_{stress} = \hat{\beta}_0 + \hat{\beta}_1^1 X^*_{income} + \hat{\beta}_2 X_{age} \rightarrow \hat{\beta}_1^1$$

$$\vdots$$

$$Y_{stress} = \hat{\beta}_0 + \hat{\beta}_1^m X^*_{income} + \hat{\beta}_2 X_{age} \rightarrow \hat{\beta}_1^m$$



Thinking carefully about H_0

- Need to specify your H_0 models very carefully. Know which variables are relevant to your hypothesis and which are not.

- Permutation tests have no assumptions on the shape of the H_0 distribution, but scrambling assumes *completely random links across observations*, which may be inconsistent with your hypotheses (e.g., time-series data).

- Permutation nulls may not necessarily have zero means. Beware assuming that zero is your expected H_0 effect.

Bootstrapping

Confidence of your parameter estimates

Confidence: What is the certainty of the value of a particular metric or measure?

Parametric: $$\sigma^2_{model} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

$$SE(\hat{\beta}_i) = \frac{\sigma^2_{model}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Assumptions:

1. $y \rightarrow RSS \rightarrow N(\mu, \sigma)$
2. $X_i \perp X_j$

Confidence of your regression coefficient *assuming* normality and independence.

(James et al. 2013)

The bootstrap

Q: Does hypothalamic stimulation increase cortisol levels?

$$\overbrace{\begin{pmatrix} 4.3 \\ 2.4 \\ 11.1 \\ 7.7 \\ 1.2 \\ 33.3 \\ 5.9 \\ 2.8 \\ 5.7 \end{pmatrix}}^{Y} \overbrace{\begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}}^{X}$$

Sample *with* replacement

$$\overbrace{\begin{pmatrix} 33.3 \\ 33.3 \\ 2.4 \\ 4.3 \\ 1.2 \\ 5.9 \\ 11.1 \\ 7.7 \\ 2.8 \end{pmatrix}}^{Y*} \overbrace{\begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}}^{X*} \quad \cdots \quad \overbrace{\begin{pmatrix} 7.7 \\ 33.3 \\ 2.4 \\ 11.1 \\ 1.2 \\ 4.3 \\ 4.3 \\ 2.8 \\ 5.9 \end{pmatrix}}^{Y*} \overbrace{\begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}}^{X*}$$

$$Y_{cort} = \hat{\beta}_0^1 + \hat{\beta}_1^1 X_{stim} \qquad Y_{cort} = \hat{\beta}_0^m + \hat{\beta}_1^m X_{stim}$$

$$[\hat{\beta}_0^1, \ldots, \hat{\beta}_0^m]$$

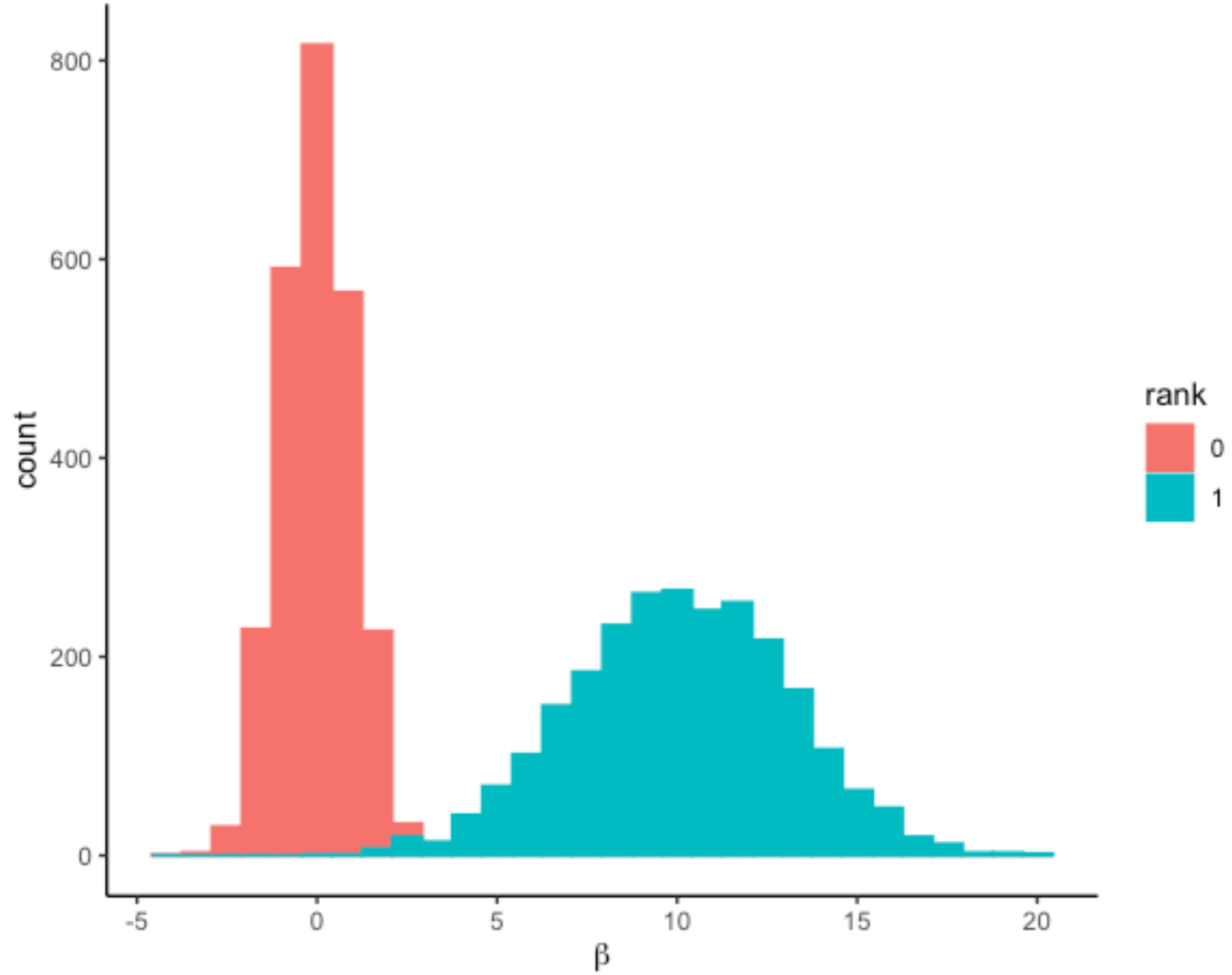
$$[\hat{\beta}_1^1, \ldots, \hat{\beta}_1^m]$$

$$Y_{cort} = \hat{\beta}_0 + \hat{\beta}_1 X_{stim}$$

(James et al. 2013)

The bootstrap

Q: Does hypothalamic stimulation increase cortisol levels?



$$\overbrace{\begin{pmatrix} 33.3 \\ 33.3 \\ 2.4 \\ 4.3 \\ 1.2 \\ 5.9 \\ 11.1 \\ 7.7 \\ 2.8 \end{pmatrix}}^{Y^*} \overbrace{\begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}}^{X^*} \quad \cdots \quad \overbrace{\begin{pmatrix} 7.7 \\ 33.3 \\ 2.4 \\ 11.1 \\ 1.2 \\ 4.3 \\ 4.3 \\ 2.8 \\ 5.9 \end{pmatrix}}^{Y^*} \overbrace{\begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}}^{X^*}$$

$$Y_{cort} = \hat{\beta}_0^1 + \hat{\beta}_1^1 X_{stim} \qquad Y_{cort} = \hat{\beta}_0^m + \hat{\beta}_1^m X_{stim}$$

$$[\hat{\beta}_0^1, \ldots, \hat{\beta}_0^m]$$

$$[\hat{\beta}_1^1, \ldots, \hat{\beta}_1^m]$$

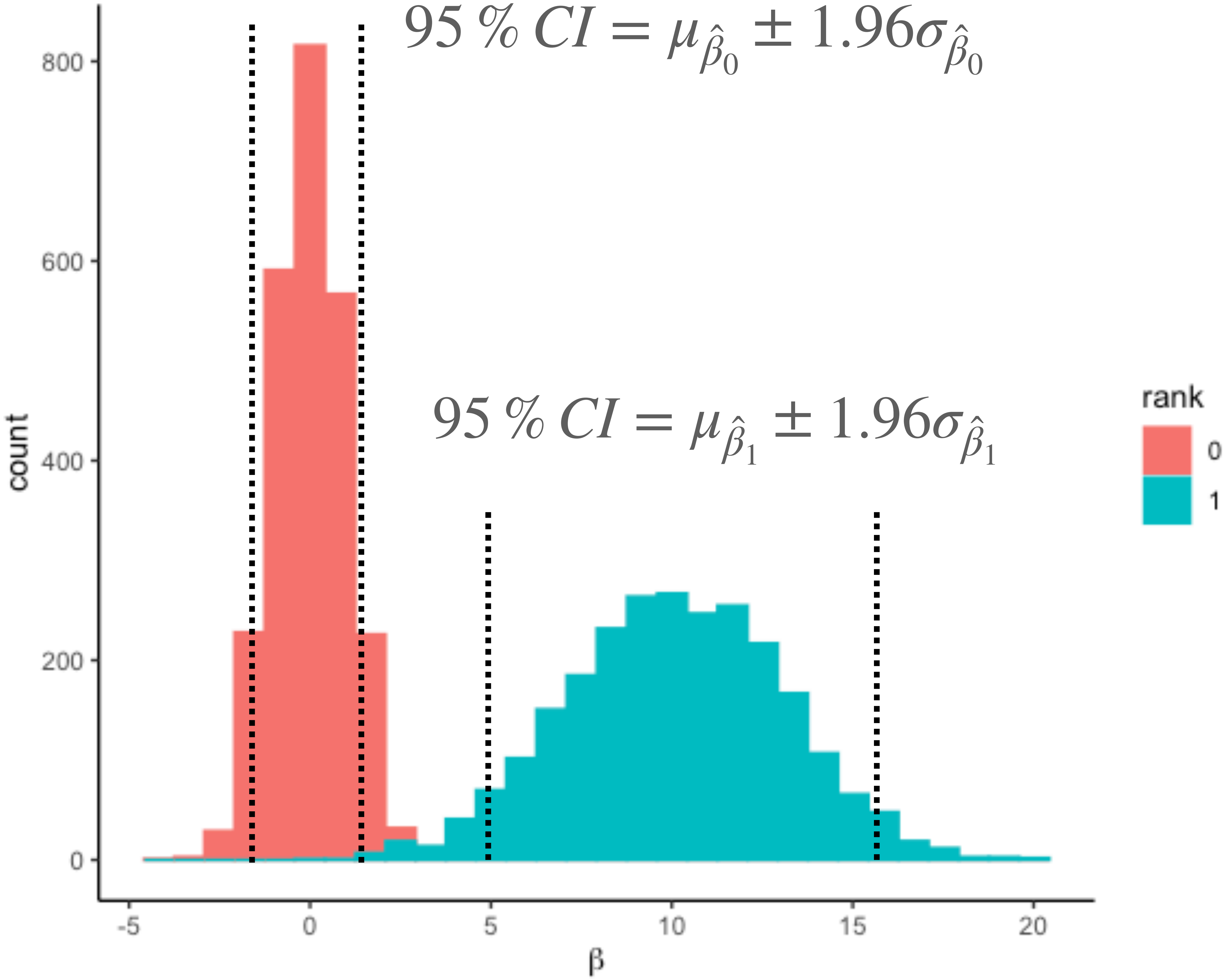
(James et al. 2013)

The bootstrap

Q: Does hypothalamic stimulation increase cortisol levels?

$$95\ \%\ CI = \mu_{\hat{\beta}_0} \pm 1.96\sigma_{\hat{\beta}_0}$$

$$95\ \%\ CI = \mu_{\hat{\beta}_1} \pm 1.96\sigma_{\hat{\beta}_1}$$



$$\overbrace{\begin{pmatrix} 33.3 \\ 33.3 \\ 2.4 \\ 4.3 \\ 1.2 \\ 5.9 \\ 11.1 \\ 7.7 \\ 2.8 \end{pmatrix}}^{Y*} \overbrace{\begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}}^{X*} \quad \cdots \quad \overbrace{\begin{pmatrix} 7.7 \\ 33.3 \\ 2.4 \\ 11.1 \\ 1.2 \\ 4.3 \\ 4.3 \\ 2.8 \\ 5.9 \end{pmatrix}}^{Y*} \overbrace{\begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}}^{X*}$$

$$Y_{cort} = \hat{\beta}_0^1 + \hat{\beta}_1^1 X_{stim} \qquad Y_{cort} = \hat{\beta}_0^m + \hat{\beta}_1^m X_{stim}$$

$$[\hat{\beta}_0^1, \ldots, \hat{\beta}_0^m]$$

$$[\hat{\beta}_1^1, \ldots, \hat{\beta}_1^m]$$

(James et al. 2013)

The bootstrap algorithm

Goal: Simulate m many replications of your dataset to estimate confidence of specific effects.

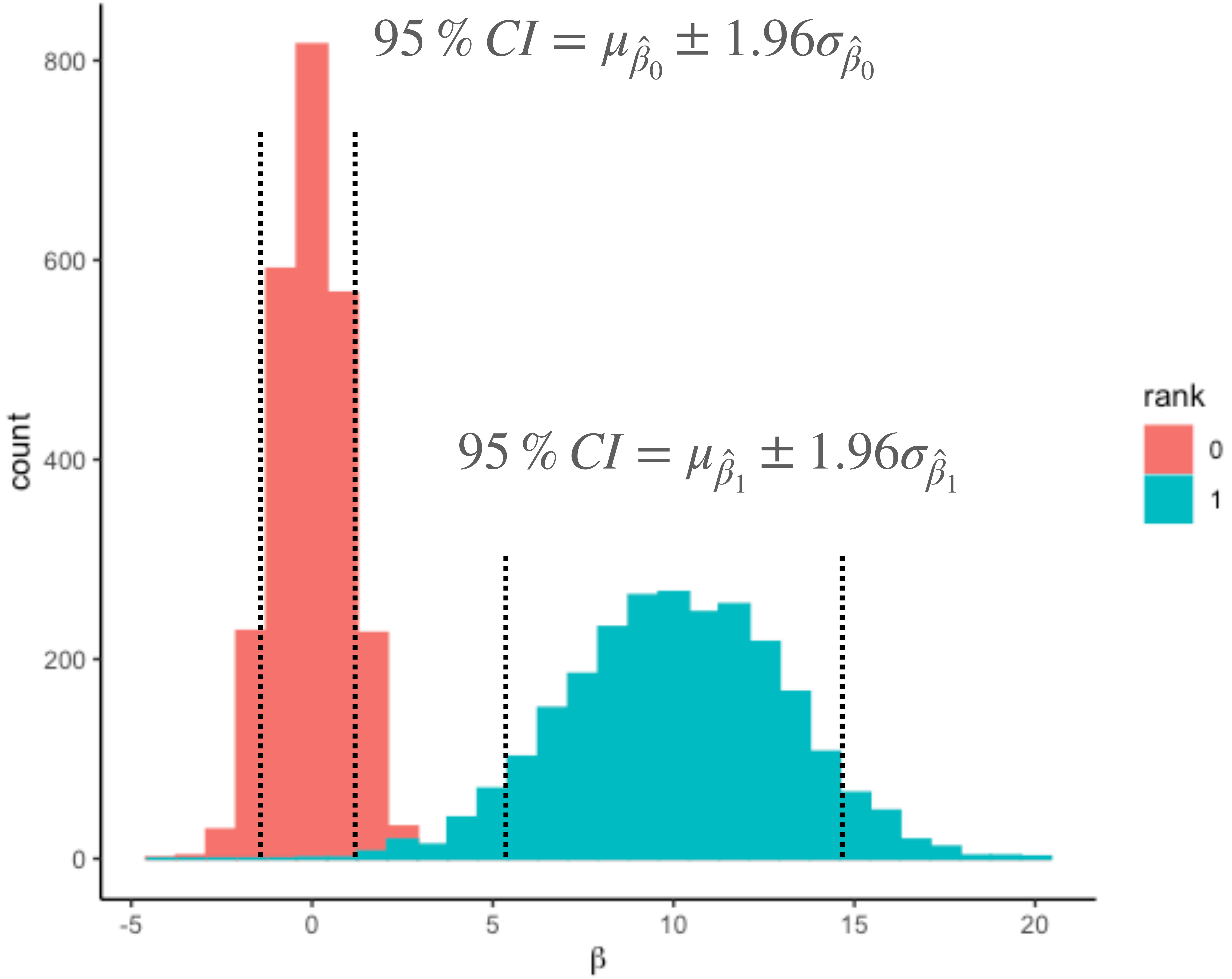
Step 1: Run m iterations where, on each iteration:

- A new variable set Y^* & X^* is generated by randomly sampling observations (rows) from the original data set, with replacement, keeping values across variables (columns) intact.
- A new model $\hat{f}^*$ is calculated from Y^* & X^*.
- All relevant parameter (e.g., $\hat{\beta}_i^*$) are stored.

Step 2: Calculate confidence estimates (e.g., $95\,\%\,CI$, SE) on target parameters (e.g., $\hat{\beta}_i^*$) from bootstrapped results.

(James et al. 2013)

Inference from bootstrapping



<u>Inference:</u>

If the confidence intervals on your bootstrapped distribution include the expectation of H_0 (most often 0), then you cannot reject the H_0 for that effect.

(James et al. 2013)

Things to consider

- As n and m increase, the bootstrapped distribution approaches the real distribution (Law of Large Numbers).
- Can estimate bias in the bootstrap by comparing the mean of the bootstrapped distribution to the original effect size.
- Use of standard error to estimate confidence on the bootstrapped distribution ($\sigma * \sqrt{m}^{-1}$) can lead to artificially high certainty with large m.

(James et al. 2013)

Take home message

- Resampling methods offer a data-driven way of estimating confidence on observed effects (bootstrap) and on inferences with regards to your hypotheses (permutation tests).