Confidence Intervals and Hypothesis Testing

Given an estimator $\widehat{\theta}$ and estimated standard error $\widehat{\text{se}}$, the $(1-\alpha)\%$ *confidence interval*

$$\widehat{\theta} \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\widehat{\text{se}}$$

is commonly used, where Φ^{-1} is the standard normal inverse cdf.

For example,

$$\alpha = 0.10 \ : \ \Phi^{-1}(0.95) = 1.645 \qquad \alpha = 0.05 \ : \ \Phi^{-1}(0.975) = 1.960.$$

This formula uses the asymptotic normal approximation to the distribution of $\widehat{\theta}$.

Confidence Intervals and Hypothesis Testing

Recall the definition of a confidence interval in a simple situation where a *pivotal quantity* is available. Suppose the sampling distribution of estimator $\widehat{\theta}$ is given **exactly** by

$$\widehat{\theta} \sim \mathcal{N}(\theta, \text{se}^2).$$

Recall that this is a **frequentist** statement, conditional on the **known** true values θ and se. Then

$$Z = \frac{\widehat{\theta} - \theta}{\text{se}} \sim \mathcal{N}(0, 1)$$

so that, for $c > 0$,

$$\mathbb{P}_\theta[-c \leq Z \leq c] = \Phi(c) - \Phi(-c) = 1 - 2\Phi(-c)$$

Choosing c such that $\Phi(c) = \alpha/2$, say $c = z_{\alpha/2}$, we have

$$\mathbb{P}_\theta \left[-z_{\alpha/2} \leq \frac{\widehat{\theta} - \theta}{\text{se}} \leq z_{\alpha/2} \right] = 1 - \alpha$$

which can be rewritten

$$\mathbb{P}_\theta \left[\widehat{\theta} - \text{se} \times z_{\alpha/2} \leq \theta \leq \widehat{\theta} + \text{se} \times z_{\alpha/2} \right] = 1 - \alpha$$

This is a probability statement concerning the **random interval**

$$[\widehat{\theta} - \text{se} \times z_{\alpha/2}, \widehat{\theta} + \text{se} \times z_{\alpha/2}]$$

given the true value of θ, and the true standard error se.

Confidence Intervals and Hypothesis Testing

Recalling the equivalence with hypothesis testing, the $(1-\alpha)\%$ confidence interval is the set of values of θ_0 that **do not** lead to a rejection of the null hypothesis

$$H_0 \ : \ \theta = \theta_0$$

at significance level α, in a two-tailed test, in light of the observed data and the estimate $\widehat{\theta}$.

Typically, α is chosen to be 0.05, or 0.01.

The *Normal* interval

$$[\widehat{\theta} - \text{se} \times z_{\alpha/2}, \widehat{\theta} + \text{se} \times z_{\alpha/2}]$$

is justified asymptotically, but for finite n, the *Student-t* interval

$$[\widehat{\theta} - \text{se} \times t_{\alpha/2}(n-1), \widehat{\theta} + \text{se} \times t_{\alpha/2}(n-1)]$$

is preferable, where

$$t_\alpha(\nu)$$

is the α percentile of the Student-t density with ν degrees of freedom.

The standard error se is typically not known, so it is replaced by the usual estimate $\widehat{\text{se}}$ obtained from the plug-in procedure.

For example, for an unknown mean

$$\widehat{\text{se}}(\widehat{\theta}) = \frac{\widehat{\sigma}}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

that is, using the plug-in estimator of the population variance based on $\widehat{F}_n$.

The corresponding bootstrap interval is based on the bootstrap resampled quantities

$$Z^{(b)} = \frac{\widehat{\theta}^{(b)} - \widehat{\theta}}{\widehat{\text{se}}^{(b)}}$$

where

- $\widehat{\theta}$ is the estimator from the original data
- $\widehat{\theta}^{(b)} = t_n^{(b)}$ is the estimator derived from the bth bootstrap resample.
- $\widehat{\text{se}}^{(b)}$ is the estimator of the standard error for $\widehat{\theta}^{(b)}$ from the bth bootstrap resample.

The empirical distribution of the observed bootstrap resampled quantities $z^{(1)}, \ldots, z^{(B)}$ yields the lower and upper percentiles

$$\widehat{l}_{\alpha/2} \quad : \quad \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\{z^{(b)} \leq \widehat{l}_{\alpha/2}\}} \doteq \alpha/2$$

$$\widehat{u}_{\alpha/2} \quad : \quad \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\{z^{(b)} \geq \widehat{u}_{\alpha/2}\}} \doteq \alpha/2$$

and hence the bootstrap interval

$$[\widehat{\theta} - \widehat{\text{se}} \times \widehat{l}_{\alpha/2}, \widehat{\theta} + \widehat{\text{se}} \times \widehat{u}_{\alpha/2}]$$

Here $\widehat{\theta}$ and $\widehat{\text{se}}$ are estimated from the original data.

Example: Mouse data

Example (Mouse data)

The following data are the survival times for mice in a randomized experiment comparing a treated group with a control group.

Control C	52	104	146	10	50	31	40	27	46
Treatment T	94	197	16	38	99	141	23		

Example: Mouse data

Example (Mouse data)

Bootstrap percentiles used for reporting CI for $\mathbb{E}[Y_C]$:

B	0.05	0.1	0.5	0.9	0.95
100	-4.6258	-4.1015	0.0382	1.2886	1.4606
200	-4.5560	-2.4595	-0.2170	1.0194	1.4346
500	-4.6962	-3.3398	-0.0873	1.2819	1.5853
1000	-4.4927	-2.3528	-0.0752	1.1252	1.5343
2000	-4.5014	-3.3380	-0.0562	1.1478	1.4993
5000	-4.6618	-3.5709	-0.0495	1.1760	1.5561
10000	-4.5987	-3.4164	-0.0828	1.1706	1.4984
20000	-4.6032	-3.3034	-0.0762	1.1406	1.4817
50000	-4.6337	-3.2448	-0.0710	1.1709	1.5248
100000	-4.6083	-3.2710	-0.0720	1.1603	1.5032
Normal	-1.6449	-1.2816	0.0000	1.2816	1.6449
Student	-1.8595	-1.3968	0.0000	1.3968	1.8595

Example: Mouse data

Example (Mouse data)

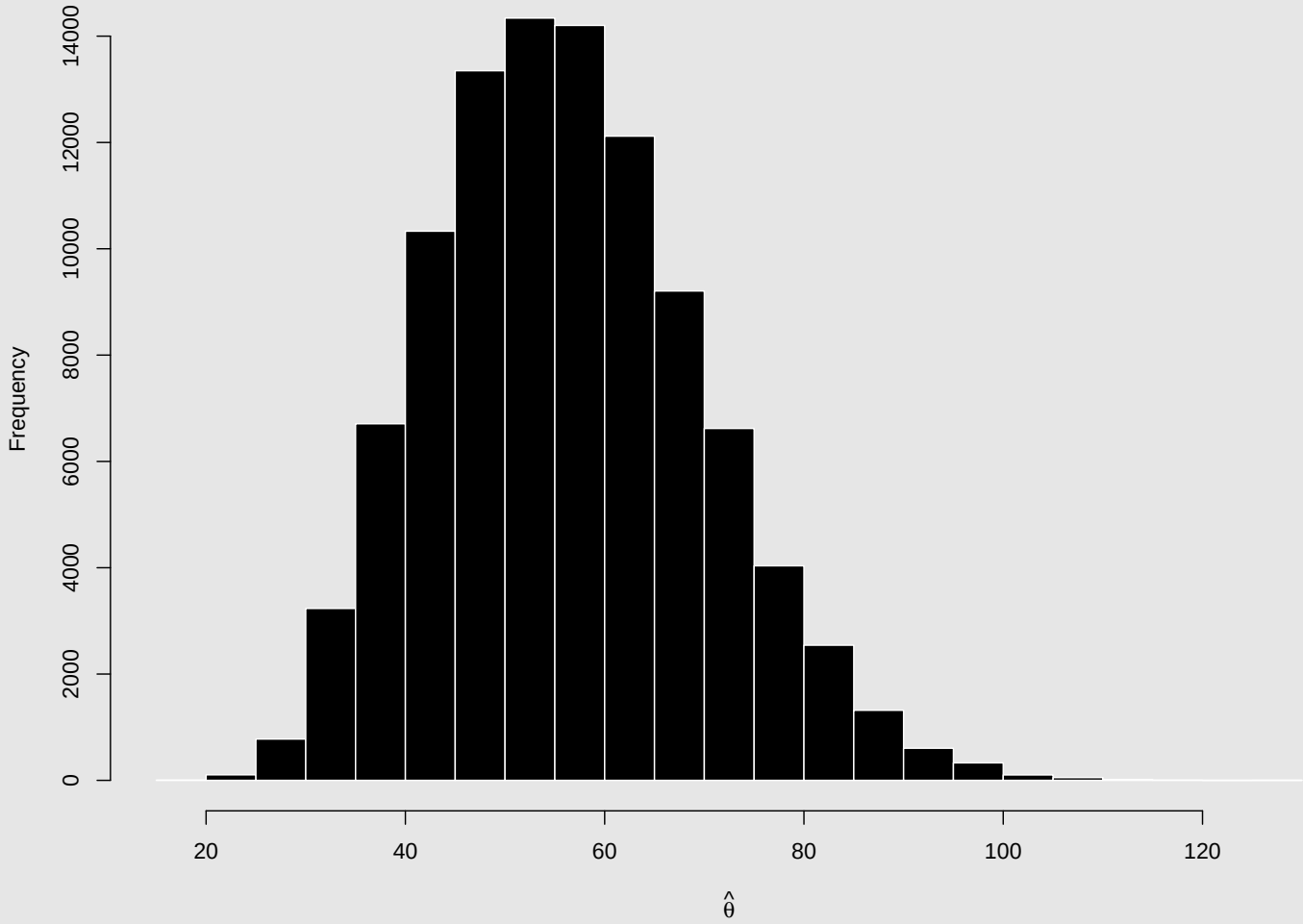
CI for $\mathbb{E}[Y_C]$:

B	Lower	Upper
100	-5.4411	75.6928
200	-4.5104	75.3462
500	-6.3801	77.3555
1000	-3.6664	76.6748
2000	-3.7825	76.2090
5000	-5.9217	76.9659
10000	-5.0803	76.1960
20000	-5.1401	75.9739
50000	-5.5468	76.5478
100000	-5.2086	76.2598
Normal	31.4338	81.0106
Student	34.2957	78.1487

Example: Mouse data

Example (Mouse data)

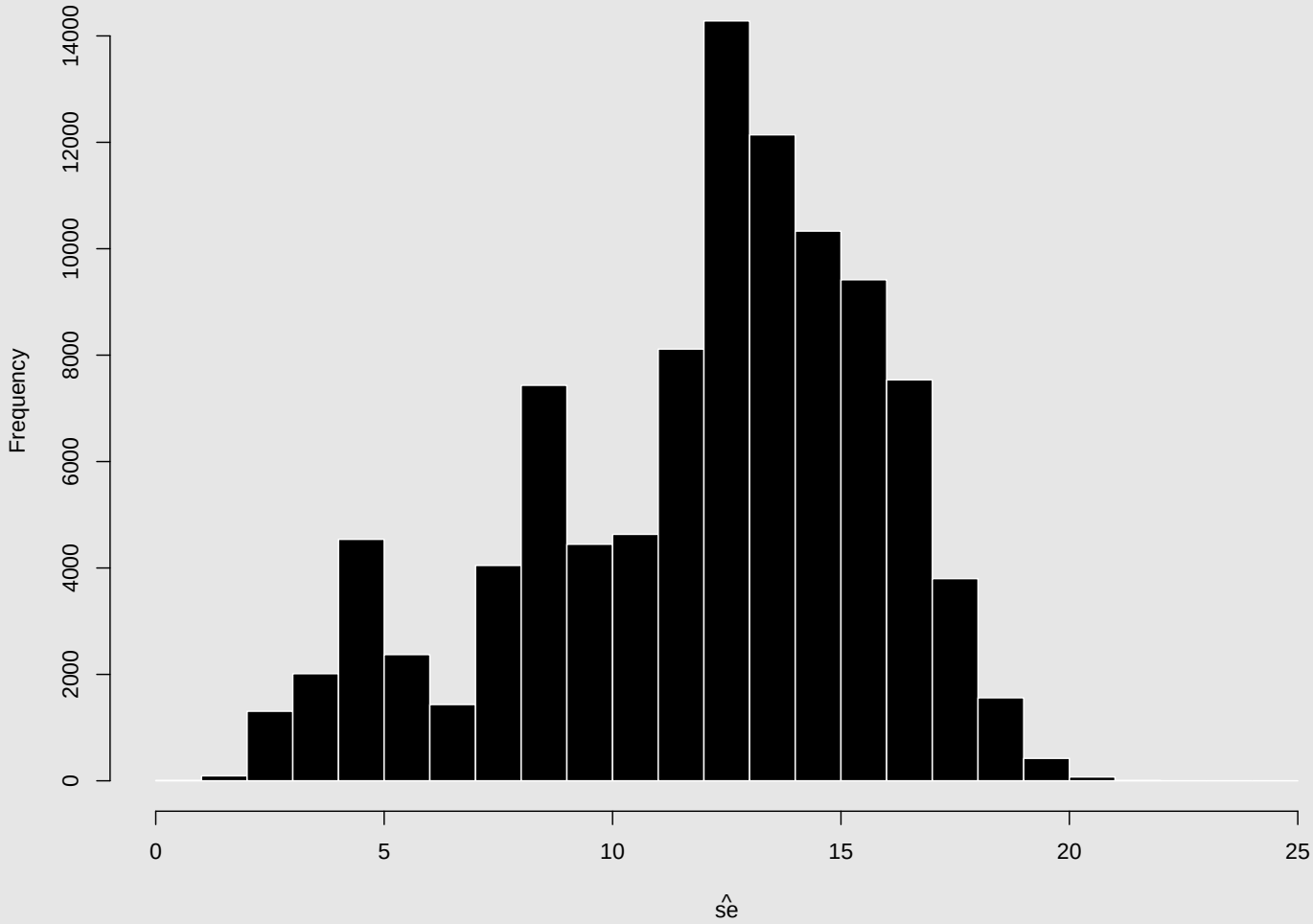
$\widehat{\theta}^{(b)}$:



Example: Mouse data

Example (Mouse data)

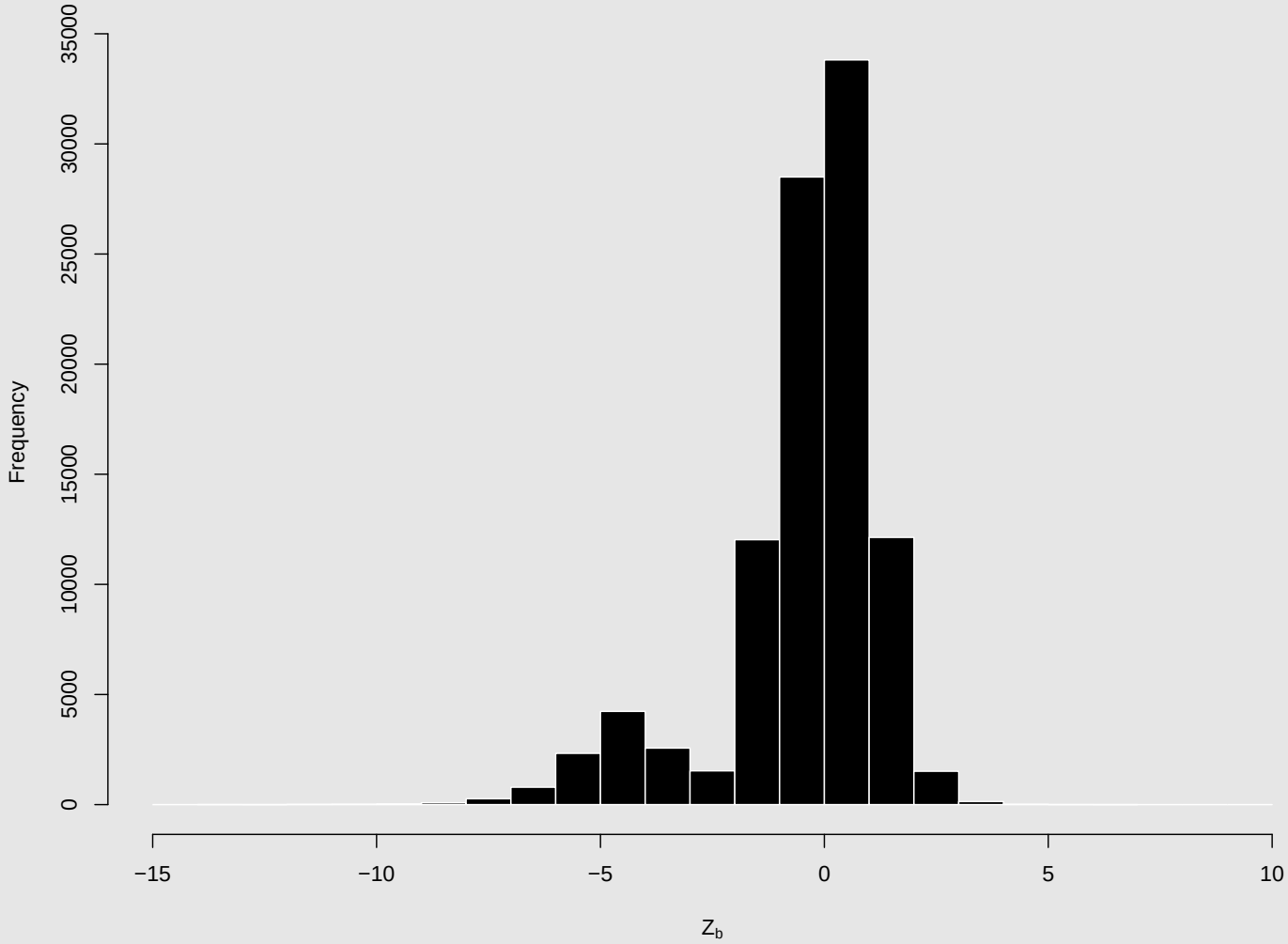
$\widehat{\text{se}}^{(b)}$:



Example: Mouse data

Example (Mouse data)

$Z^{(b)} = (\widehat{\theta}^{(b)} - \widehat{\theta})/\widehat{\text{se}}^{(b)}$:



Issues

Some issues:

- In general, a formula for $\widehat{\text{se}}$ is not available
  - Need to do a **nested** bootstrap: within each bootstrap iteration, need an internal bootstrap calculation of the standard error estimate, $\widehat{\text{se}}_B$.
  - Can be computationally demanding
- The symmetric interval of the form

$$\widehat{\theta} \pm \text{se} \times q_{\alpha/2}$$

  may include negative values, even if the quantity of interest is positive.
  - A transformation may be necessary
- The bootstrap interval is not invariant to transformations (specifically, the computation of standard errors needs care)

Variance stabilizing transformations

The bootstrap-t interval is constructed by appealing to asymptotic normality, so therefore we seek a transformation $\phi = g(\theta)$ such that the distribution of $\widehat{\phi}$ is better represented by a normal distribution for finite n.

Suppose that, by the Central Limit Theorem, we have that

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$$

where the asymptotic variance $\sigma^2(\theta)$ in general depends on the true θ. We seek a transform $\phi = g(\theta)$ such that

$$\sqrt{n}(\widehat{\phi} - \phi) \xrightarrow{d} \mathcal{N}(0, \varsigma^2)$$

where ς^2 does not depend on the true θ (or ϕ).

Variance stabilizing transformations

For example, for the Pearson correlation coefficient R in a bivariate normal problem with true correlation ρ, we saw that if

$$Z = \frac{1}{2}\log\left(\frac{1+R}{1-R}\right) \qquad\qquad \varrho = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$$

then

$$\sqrt{n-3}Z \xrightarrow{d} \mathcal{N}(\varrho, 1)$$

suggesting that we should construct the bootstrap-t interval for Z, and then back-transform to the R scale to get the confidence interval for ρ.

Variance stabilizing transformations

Suppose that X is a random variable with

$$\mathbb{E}[X] = \theta \qquad \mathbb{V}ar[X] = s(\theta)^2$$

Let transformation g be defined by the relation

$$\dot{g}(x) = \frac{1}{s(x)} \qquad \text{or, equivalently} \qquad g(x) = \int^x \frac{1}{s(u)}\, du$$

Then if $Y = g(X)$, by a Taylor series approximation

$$\mathbb{V}ar[Y] = \{\dot{g}(\theta)\}^2 s(\theta)^2 = 1$$

This *variance-stabilizing transformation* can be used to assist in the construction of the bootstrap-t intervals.

Variance stabilizing transformations

In statistical applications, the function $s(.)$ is the standard error, a function of the true θ; often this is not known. However, it can be estimated from the data.

Let

$$s(u) = \text{se}(\widehat{\theta}|\theta = u).$$

This function can be estimated using a nested bootstrap for a collection of fixed values, then integrated numerically to yield an estimate of the transform function g.

Variance stabilizing transformations: Algorithm

1. Generate B_1 bootstrap samples, and bootstrap estimates $\widehat{\theta}^{(b)}$. For **each** bootstrap sample $\boldsymbol{y}^{(b)}$, generate B_2 nested resamples, and estimate the standard error $\widehat{\text{se}}(\widehat{\theta}^{(b)})$ for this bootstrap sample.
2. Fit a curve through the collection of points $\{(\widehat{\theta}^{(b)}, \widehat{\text{se}}(\widehat{\theta}^{(b)}),$ to get an estimate of $s(u) = \text{se}(\widehat{\theta}|\theta = u)$.
3. Estimate the variance-stabilizing transform g by integrating $s(u)$ numerically
4. Use B_3 new bootstrap samples to compute a bootstrap-t interval for $\phi = g(\theta)$ by inspecting

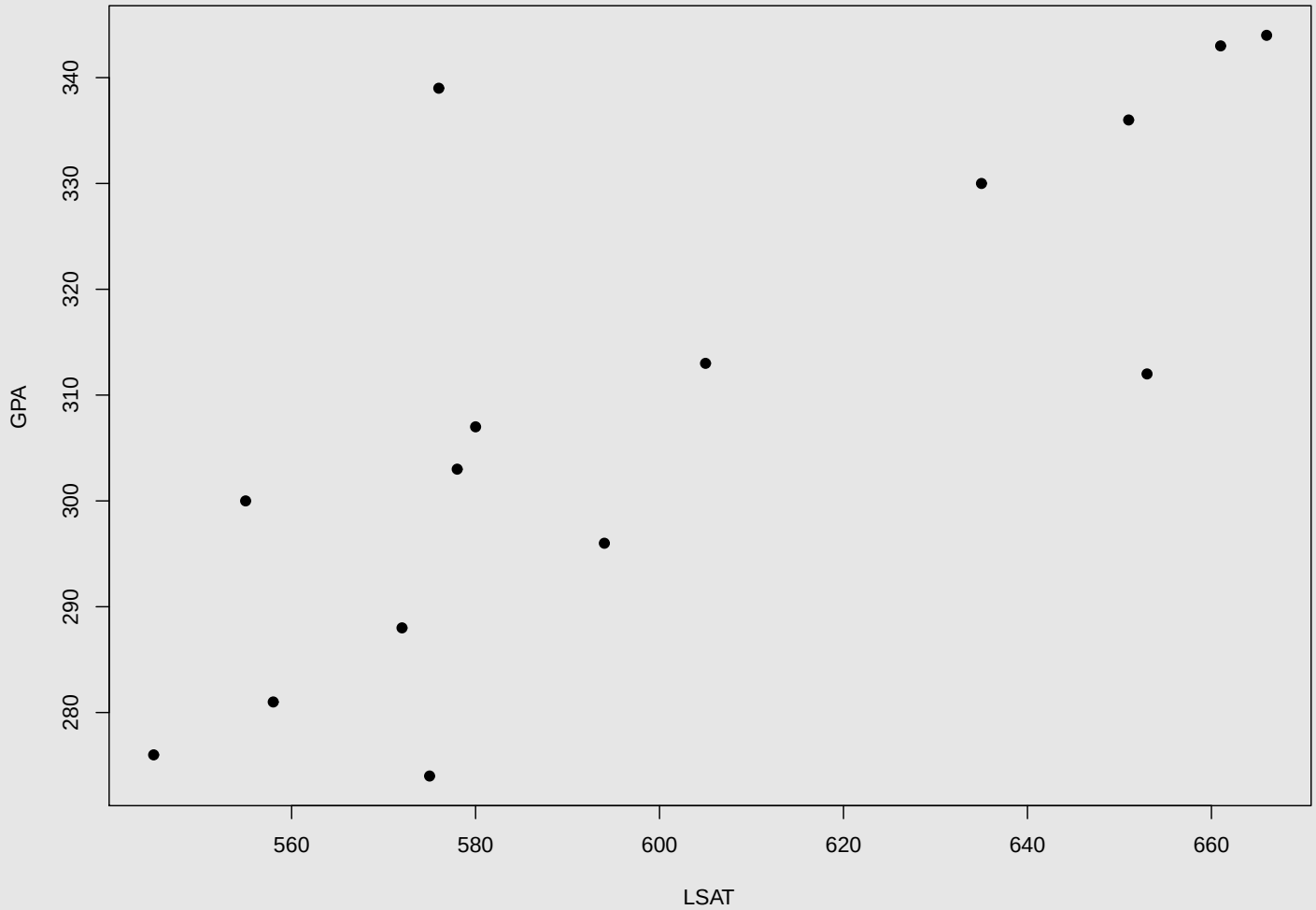
$$\widehat{\phi}^{(b)} - \widehat{\phi}$$

   as the variance has been stabilized to 1.
5. Back-transform to the θ scale using the numerically defined g^{-1}.

Variance stabilizing transformations: Example

Example (LSAT vs GPA data)

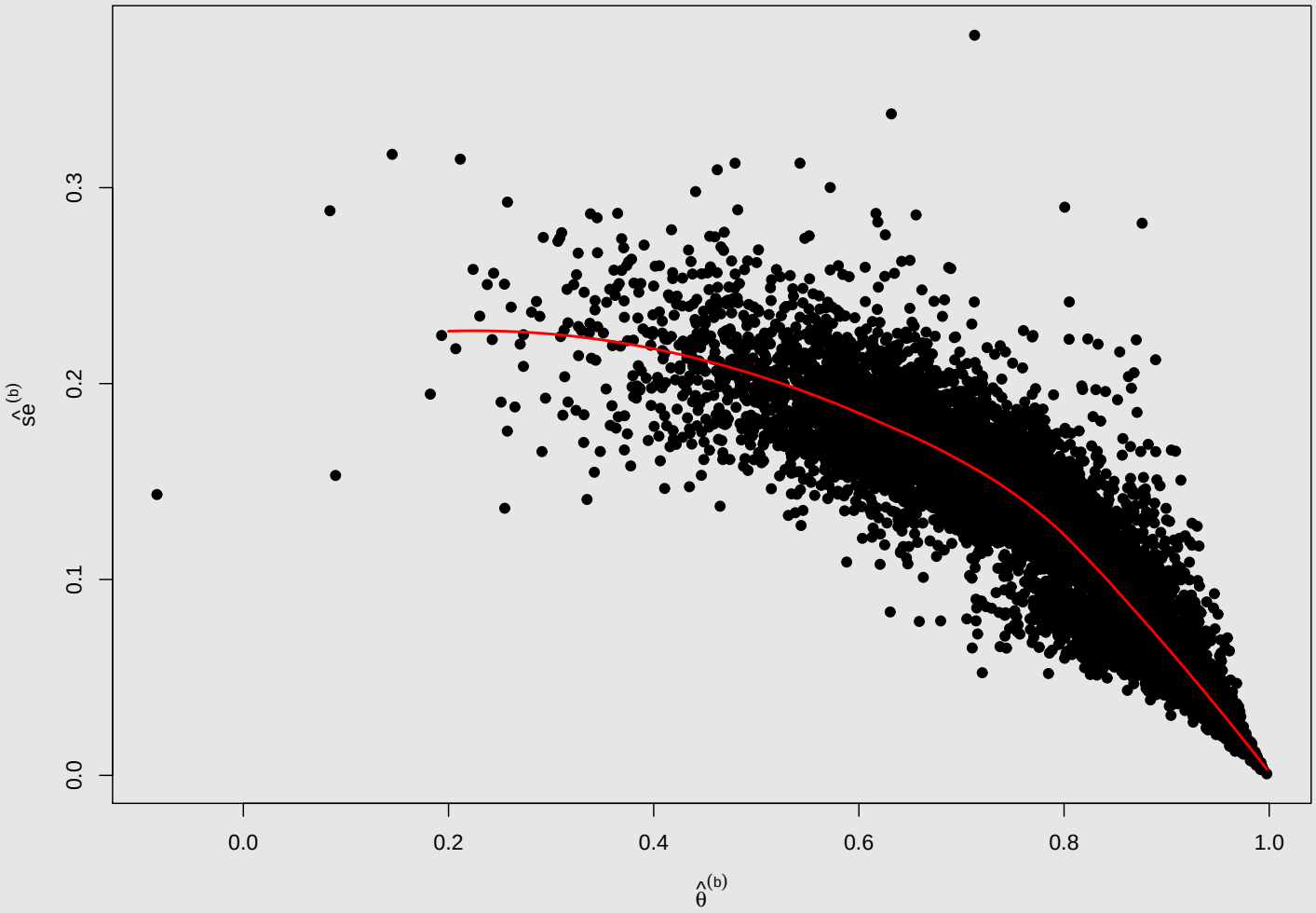
Raw data



Variance stabilizing transformations: Example

Example (LSAT vs GPA data)

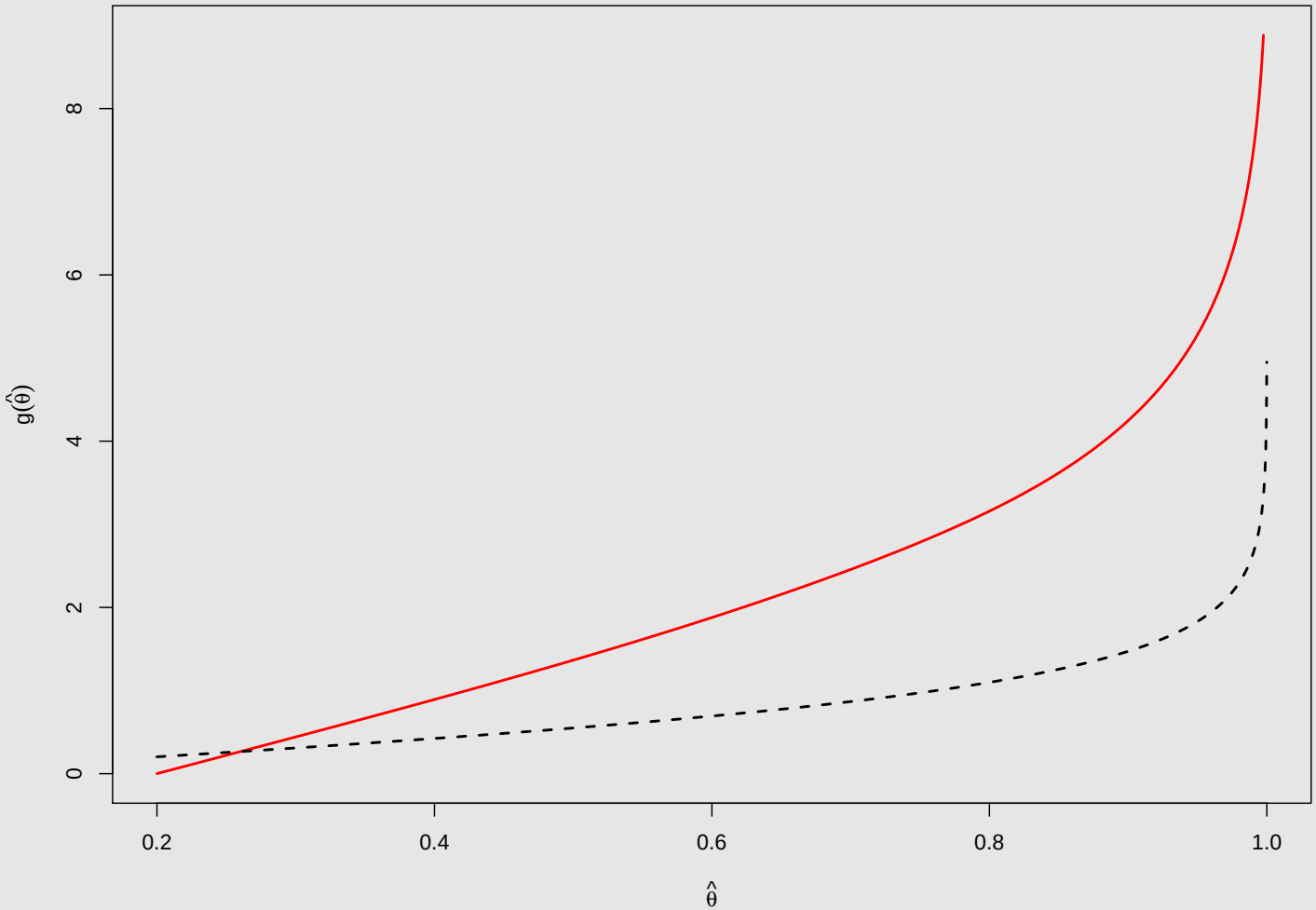
$B_1 = 10000$ bootstrap resamples with $B_2 = 500$ inner resamples.



Variance stabilizing transformations: Example

Example (LSAT vs GPA data)

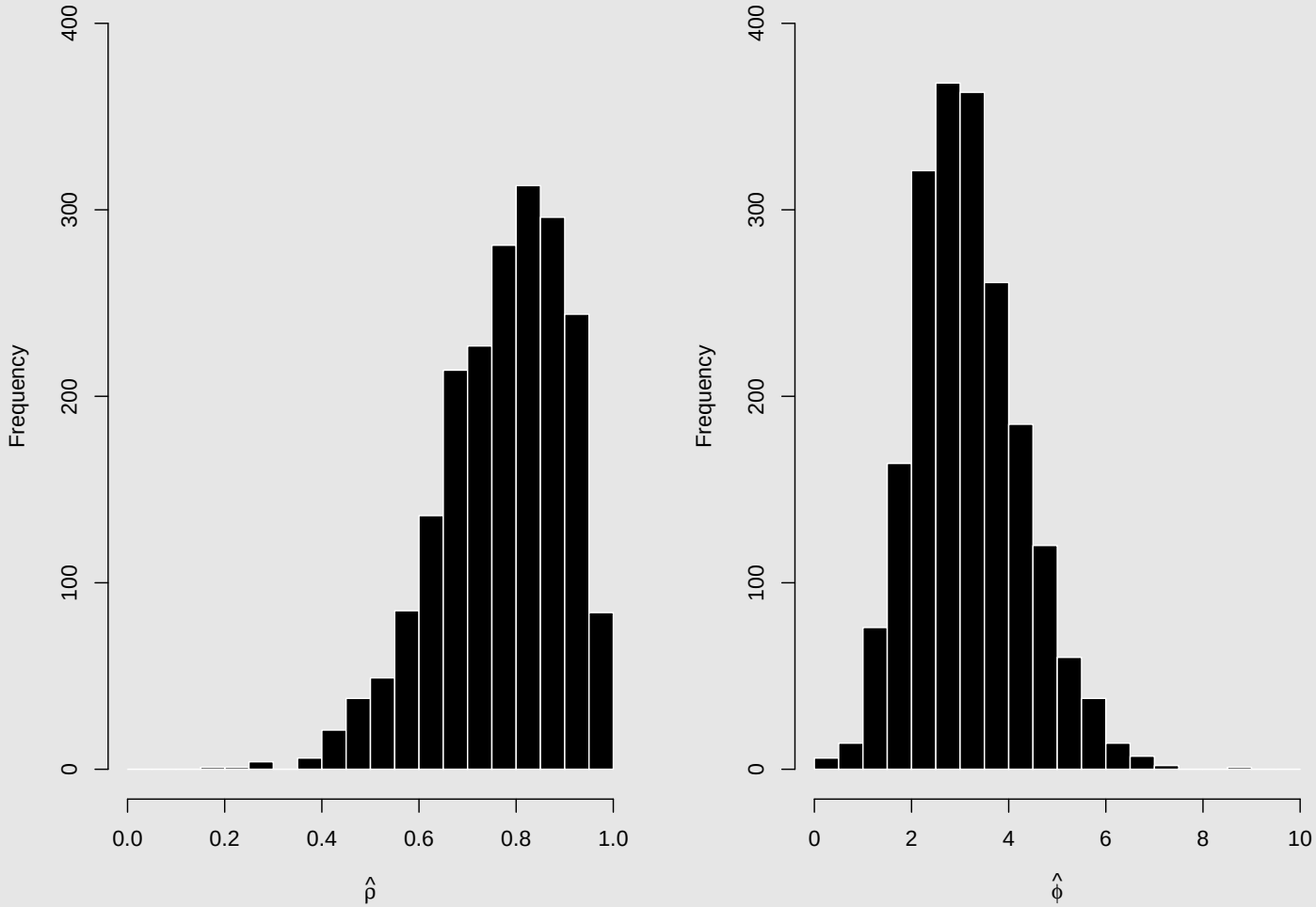
Estimated transform g (solid), and Fisher transform (dashed)



Variance stabilizing transformations: Example

Example (LSAT vs GPA data)

$B = 2000$ bootstrap samples of $\widehat{\rho}$ (left) and $\widehat{\phi}$ (right).



Variance stabilizing transformations: Example

Example (LSAT vs GPA data)

Comparison of 90 % intervals:

Method	Lower	Upper
Fisher transform	0.5090	0.9071
$\widehat{g}$	0.5299	0.9448
Bootstrap	0.5262	0.9475

The bootstrap interval is computed direct from the percentiles of the initial B_1 bootstrap resampled values.

Bootstrap Percentile Interval

The *bootstrap percentile interval* is computed direct from the percentiles of the bootstrap resampled values.

$$\widehat{l}_{\alpha/2} \quad : \quad \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\{\widehat{\theta}^{(b)} \leq \widehat{l}_{\alpha/2}\}} \doteq \alpha/2$$

$$\widehat{u}_{\alpha/2} \quad : \quad \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\{\widehat{\theta}^{(b)} \geq \widehat{u}_{\alpha/2}\}} \doteq \alpha/2$$

This method is invariant to transformation of the statistic of interest; it is a non-parametric version of the previous variance-stabilizing transform method which focusses on the quantiles of interest.

Typically, B needs to be large to get reliable percentile intervals. They can be more variable when the sampling distribution of the statistic is highly skewed.

Bias correction/Acceleration

An alternative approach uses the interval based on

$$\widehat{l}_{\alpha_1} \quad : \quad \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\{\widehat{\theta}^{(b)} \leq \widehat{l}_{\alpha_1}\}} \doteq \alpha_1$$

$$\widehat{u}_{\alpha_2} \quad : \quad \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\{\widehat{\theta}^{(b)} \leq \widehat{u}_{\alpha_2}\}} \doteq \alpha_2$$

that is, the empirical α_1 and α_2 quantiles, where

$$\begin{aligned} \alpha_1 &= \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_{\alpha/2}}{1 - \widehat{a}(\widehat{z}_0 + z_{\alpha/2})}\right) \\ \alpha_2 &= \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_{1-\alpha/2}}{1 - \widehat{a}(\widehat{z}_0 + z_{1-\alpha/2})}\right) \end{aligned}$$

where $\widehat{z}_0$ and $\widehat{a}$ are to be defined, and $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the tail quantiles of the standard normal.

Bias correction/Acceleration

For $\alpha = 0.1$, $z_{\alpha/2} = z_{0.05} = -1.645$, $z_{1-\alpha/2} = z_{0.95} = 1.645$.

- $\widehat{z}_0$ is a *bias-correction* constant defined by

$$\widehat{z}_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\{\widehat{\theta}^{(b)} \leq \widehat{\theta}\}}\right)$$

  that is, it measures the median bias of $\widehat{\theta}^{(b)}$
- $\widehat{a}$ is an *acceleration* factor which defined via *jackknife* quantities

$$\widehat{a} = \frac{\sum_{i=1}^{n}(\widehat{\theta}_{(.)} - \widehat{\theta}_{(i)})^3}{6\left\{\sum_{i=1}^{n}(\widehat{\theta}_{(.)} - \widehat{\theta}_{(i)})^2\right\}^{3/2}}$$

Bias corrected/accelerated percentile intervals have good theoretical properties.

Permutation tests

Permutation procedures utilize resampling to construct tests of hypothesis without using parametric assumptions.

The sampling distribution of a (test) statistic is constructed using resampling from the original data, but assuming the structure implied by the **null hypothesis**.

Thus, unlike the bootstrap which retains the original data structure during resampling, permutation tests impose a hypothesis driven structure.

Permutation tests: Example

Example (Mouse data)

Recall the mouse survival time data:

Control C (y)	52	104	146	10	50	31	40	27	46
Treatment T (z)	94	197	16	38	99	141	23		

Suppose the null hypothesis is that there is no difference between the survival time distributions in the two groups.

Permutation tests: Example

Example (Mouse data)

We select a test statistic, $T = T(\boldsymbol{Y}, \boldsymbol{Z})$, suitable for testing the hypothesis; could use the two-sample t-statistic assuming *unequal* variances:

$$T = \frac{\overline{Y} - \overline{Z}}{\sqrt{\frac{s_Y}{n_Y} + \frac{s_Z}{n_Z}}}$$

where s_Y and s_Z are the unbiased sample variances. This statistic is derived from the perspective of normal random sampling, but the test itself will not rely on this assumption.

The sampling distribution of T *under the null hypothesis*, H_0, is computed by noting that, under H_0, the labelling of a mouse as being in the Control or Treatment group makes no difference to their random survival time.

Permutation tests: Example

Example (Mouse data)

Therefore, we compute the statistic for each of the possible relabellings of the $n = n_Y + n_Z$ measurements consistent with the original group sizes; that is, for relabelling j, we choose **at random without replacement** n_X measurements and denote them $\boldsymbol{y}_{(j)}$, denote the remaining n_Y measurements $\boldsymbol{z}_{(j)}$, and compute the resulting relabelled statistic $T_{(j)}$.

When all N possible relabellings have been examined, we have computed the statistics

$$t_{(1)}, \ldots, t_{(N)}$$

which define the permutation distribution.

Permutation tests: Example

Example (Mouse data)

The *critical values* for the test are defined via the empirical distribution of

$$t_{(1)}, \ldots, t_{(N)}$$

If $t = T(\boldsymbol{y}, \boldsymbol{z})$ is the observed test statistic for the original data, then empirical probabilities

$$\mathbb{P}_0[T_{(*)} \leq t] = \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{[t_{(j)}, \infty)}(t)$$

can be used to define the p-value (*achieved significance level*) in the two-tailed test as

$$\mathbb{P}_0[T_{(*)} \leq -|t|] + \mathbb{P}_0[T_{(*)} \geq |t|]$$

Permutation tests: Computation

Evaluating **all** possible relabellings can be a computationally demanding step; for the two-sample problem, the number of relabellings is

$$N = \binom{n}{n_Y}$$

For the mouse data, $n_Y = 9, n_Z = 7$, so

$$N = \binom{16}{9} = 11440$$

which is not too large.

If N gets too large, a fixed number of *randomly chosen* relabellings can be generated in place of the exhaustive calculation.

Permutation tests: Computation

If the permutations are randomly chosen, the procedure is termed a *Monte Carlo permutation* (or *Monte Carlo randomization*) procedure.

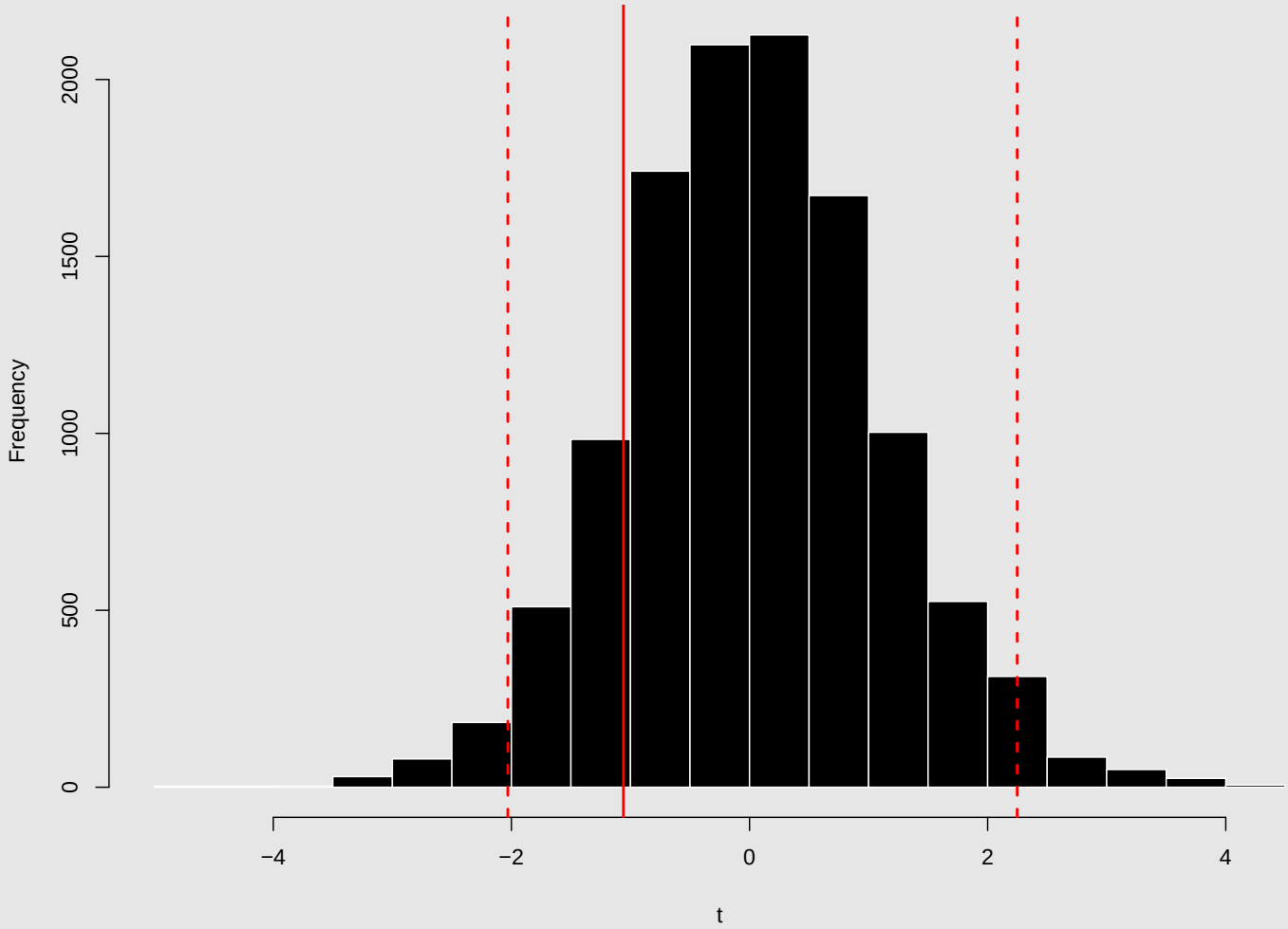
The choice of the Monte Carlo sample size depends on the quantity being estimated.

In any case, the procedure is *distribution-free*, but is reliant on being able to construct a suitable test statistic.

Permutation tests: Example

Example (Mouse data)

Permutation p-value = 0.3068.



Permutation tests: Other statistics

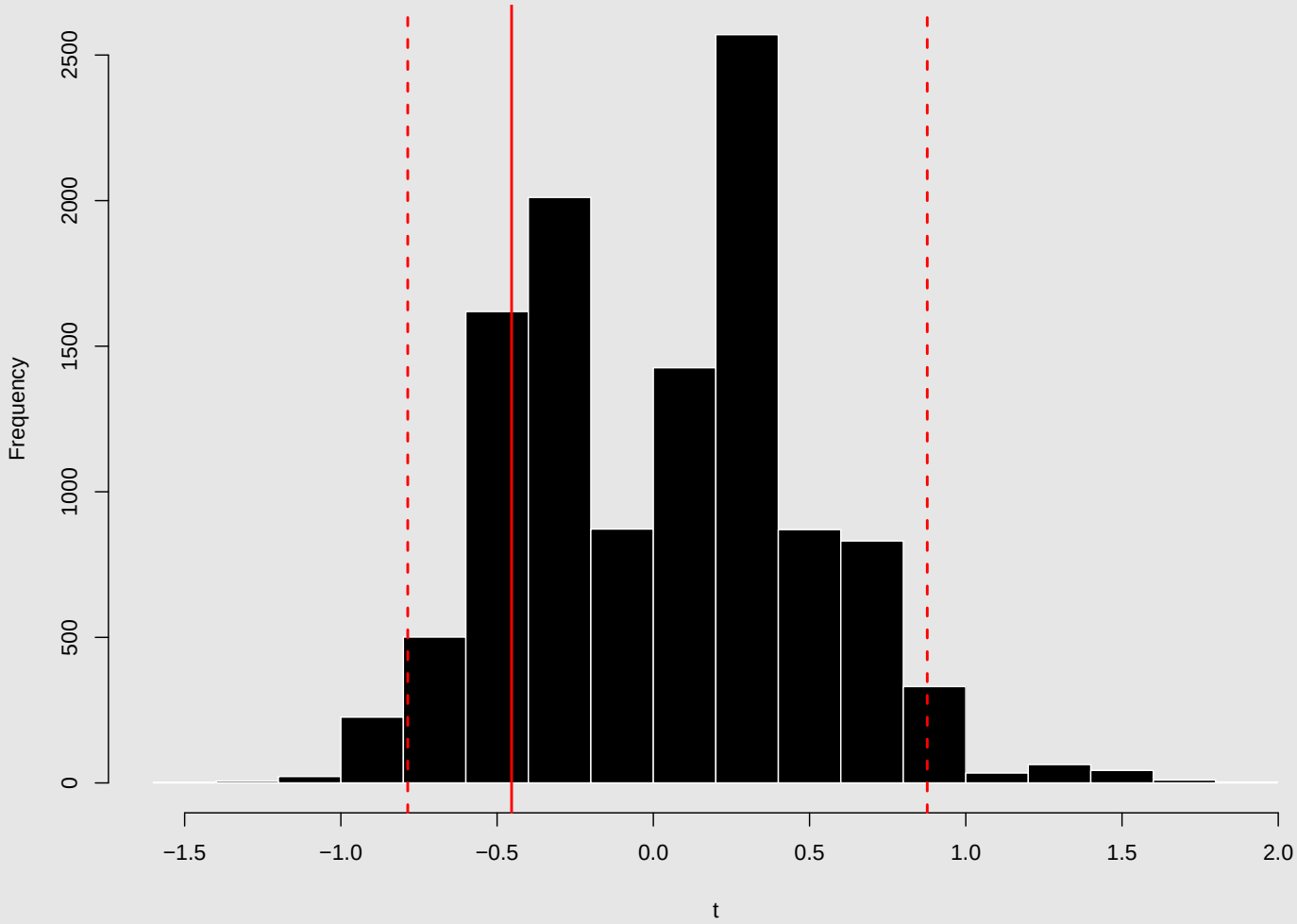
The permutation procedure can be used

- with other test statistics
  - tests for equality of variance
  - Kolmogorov-Smirnov
- in more general data collection situations
  - inference in regression
  - ANOVA
  - discrete data (chi-squared test, Fisher's Exact test)

Permutation tests: Other statistics

Example (Mouse data)

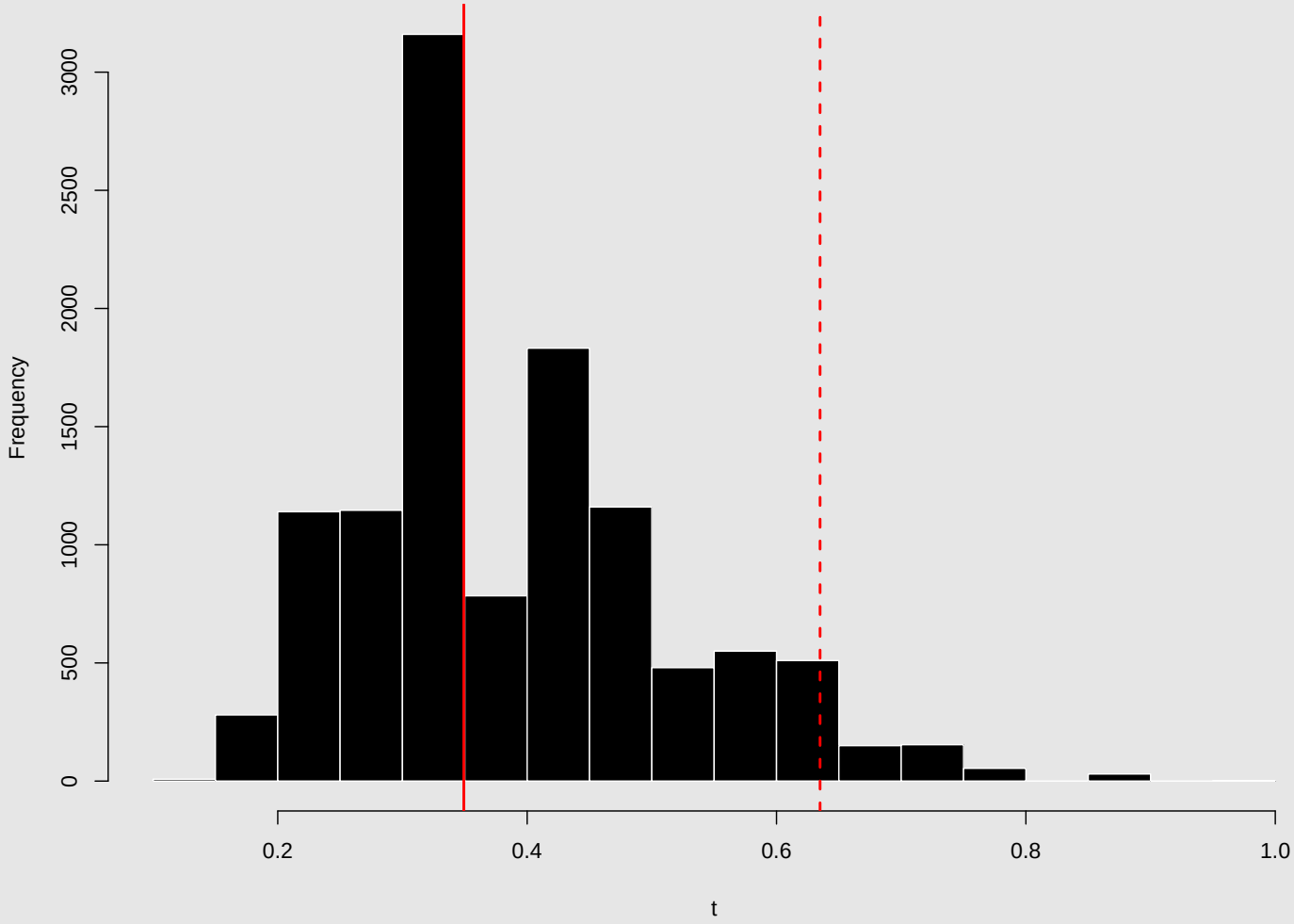
Test statistic $T = \log(s_Y/s_Z)$: two-sided p-value 0.3156



Permutation tests: Other statistics

Example (Mouse data)

Kolmogorov-Smirnov test one-sided p-value 0.5727 (32 possible values of $t_{(j)}$.



Hypothesis testing: summary

In general, hypothesis testing can be carried out using the bootstrap due to the connection between hypothesis testing and confidence intervals.

That is, a $(1-\alpha)\%$ confidence interval for θ that contains those θ_0 values that, in light of the observed data and estimate $\widehat{\theta}$, would NOT lead to the rejection of the null hypothesis

$$H_0 \ : \ \theta = \theta_0$$

at significance level α.