Randomization and Jackknifing

Goals of Resampling Methods

Goal

We have estimated some statistic of interest, S, for a set of observed data. We want to know whether the value of S we estimate, $\hat{s}$, is likely to have been generated by chance or under a model captured by an appropriate null hypothesis.

Approach

Randomization tests and related methods allow one to address these questions for many types of statistics that are not amenable to classical analysis.

Advantages of Resampling Methods

In general, resampling methods are:

- Computer intensive
- Require few assumptions about the distributional properties
- Robust

Overview of Randomization Tests

Basic Idea

Compare $\hat{s}$ to the distribution of estimates of S obtained by randomly reordering the data.

Randomization tests: Example

Consider the following measurements of mandible length (in mm) from skeletons of male and female golden jackals (*Canis aureus*).

- Male: 120, 107, 110, 116, 114, ~~11~~ **111**, 113, 117, 114, 112
- Female: 110, 111, 107, 108, 110, 105, 107, 106, 111, 111

Question

Are the mean values for the two sexes different?

Randomization tests: Example cont

The means for the two sexes are:

- $\bar{x}_{male} = 113.4$
- $\bar{x}_{female} = 108.6$

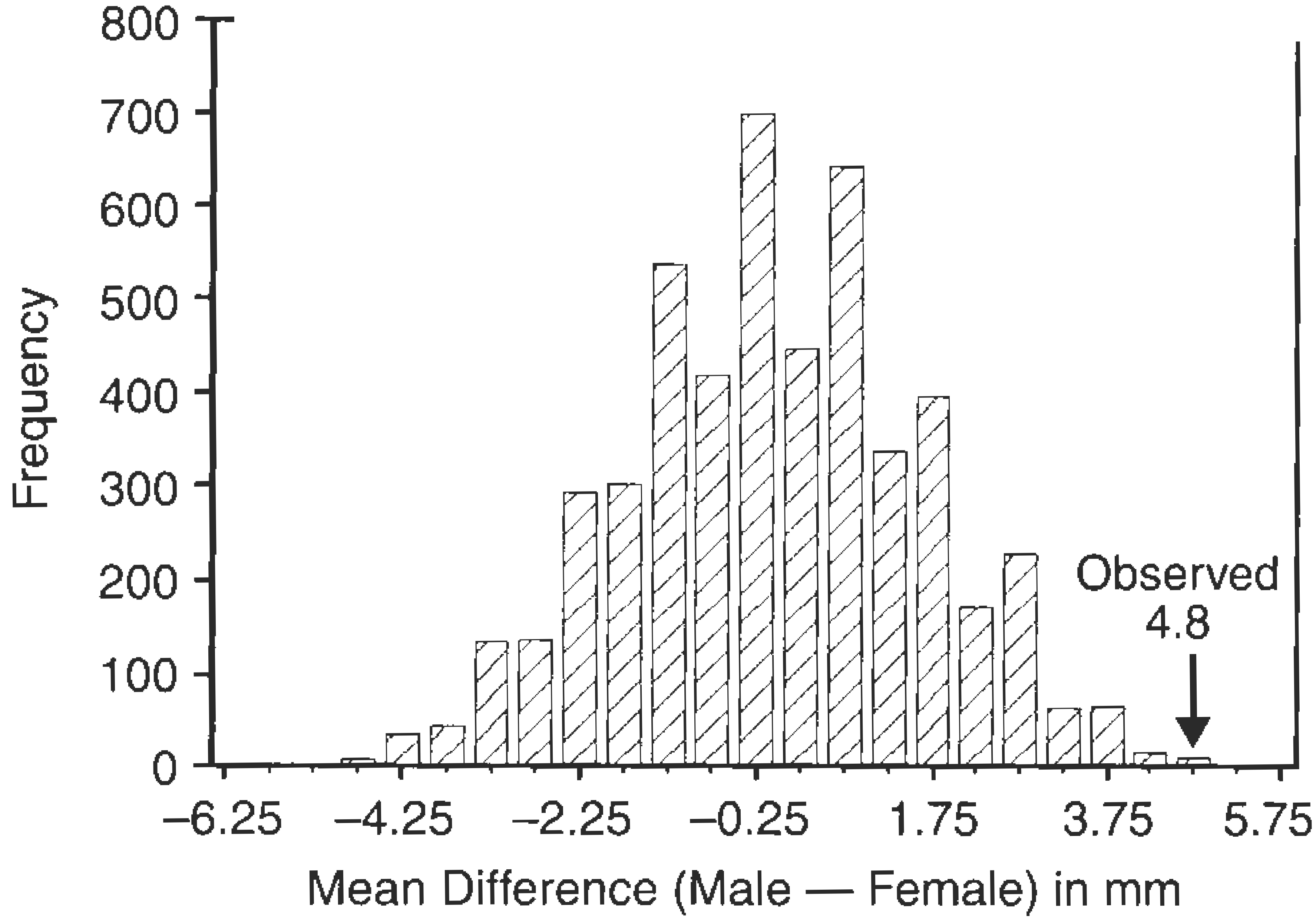
Standard deviations:

- $s_{male} = 3.72$
- $s_{female} = 2.27.$

To construct the randomization test:

- generate a large number of samples where we randomly reallocate 10 of the specimens to the male group and the remaining 10 we designate as female.
- for each randomized sample calculate the difference in means between the male and female groups
- Examine the distribution of the randomization distribution and ask whether the observed difference in means is atypical.

Randomization tests: Example cont



Advantages and Limitations of Randomization Tests

Advantages:

- Valid even without a random sample
- Can be designed to take particulars of a particular statistic into account
- When there is a classical, parametric equivalent there is often good agreement

Limitations:

- Not possible to generalize to a population of interest

Overiew of Jackknife Methodology

Basic Idea

Turn the problem of estimating any parameter of interest for n observations into a problem of estimating a sample mean.

Basic Mechanics

Calculate pseudovalues of S, s^*, leaving out a single observation at a time. Calculate mean and standard errors of pseudovalues to estimate confidence intervals for $\hat{s}$.

Jackknifing I

- Assume we're interested in some arbitrarily complex statistic that is a function of the n data points:

$$\hat{\Theta} = \phi(x_1, x_2, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots x_n)$$

- We define the ith **partial estimate** of Θ as

$$\hat{\Theta}_i = \phi(x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots x_n)$$

- We define the ith **pseudovalue** as:

$$\widehat{\Theta^*}_i = n\hat{\Theta} - (n-1)\hat{\Theta}_i$$

We will treat the pseudovalues as if they were independent random variables with mean $\hat{\Theta}$.

Jackknifing II

- From the pseudovalues we can calculate a **jackknife estimate** of Θ as follows

$$\widehat{\Theta*} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\Theta*}_i$$

- The variance of the pseudovalues, $\widehat{\Theta*}_i$, is given by:

$$Var(\widehat{\Theta*}) = \frac{(\sum \widehat{\Theta*}_i - \widehat{\Theta*})^2}{n-1}$$

- We can approximate the standard error of $\widehat{\Theta*}$ by calculating the standard error

$$SE_{jack} = \sqrt{\frac{Var(\widehat{\Theta*})}{n}}$$

- An approximate $(1-\alpha)$% confidence interval is given by

$$\widehat{\Theta*} \pm t_{\alpha/2,n-1}SE_{jack}$$

  where $t_{\alpha/2,n-1}$ is the valued that is exceeded with probability $\alpha/2$ for the t-distribution with $n-1$ degrees of freedom.

Jackknife Estimates: Example

Suppose we have a random sample of size $n = 20$ that consists of the following observations: 3.56, 0.69, 0.10, 1.84, 3.93, 1.25, 0.18, 1.13, 0.27, 0.50, 0.01, 0.61, 0.82, 1.70, 0.39, 0.11, 1.20, 1.21, 0.72

Let's use the jackknife to estimate confidence intervals for the standard deviation, σ of this sample.

- We calculate the jackknife pseudovalues, $\widehat{\sigma_1^*}, \ldots, \widehat{\sigma_{20}^*}$
- The mean of the jackknife pseudovalues, $\widehat{\sigma^*} = 1.096$.
- The variance of the pseudovalues, $Var(\widehat{\Theta^*}_i) = 1.488$
- The standard error of the pseudovalues, $\sqrt{\frac{Var(\widehat{\Theta^*}_i)}{n}} = 0.273$.
- The approximate 95% confidence interval is:
  $1.096 \pm 2.09 \times 0.273 = (0.53, 1.67)$

Advantages and Limitations of Jackknife Estimates

Advantages:

- Simple to calculate and not particularly computer intensive
- Jackknife estimates reduce bias

Limitations:

- Works best when observed sample is moderately large
- Jackknife confidence intervals can sometimes over or underestimate the true confidence interval so simulation studies to test the robustness of the jackknife for a statistic of interest are often warranted