Permutation Test

Max Turgeon

STAT 3150—Statistical Computing

Lecture Objectives

- Explain the difference between bootstrap and permutation tests.
- Apply permutation tests to two-sample problems.

Motivation

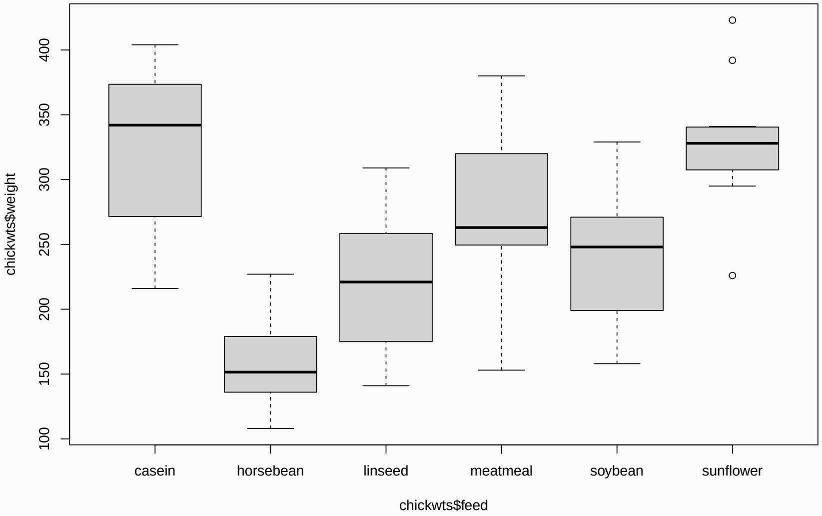
- When discussing bootstrap and jackknife, we constructed confidence intervals for our estimates.
- Confidence intervals can be used for *hypothesis testing*:
    - Check if value of population parameter under the null hypothesis is contained in the interval.
- **Permutation tests** are a whole family of resampling strategies that can be used specifically for hypothesis testing.
    - And there are generally more powerful than bootstrap.

Motivating example—T test i

- We will use the `chickwts` dataset (available in base `R`).
- Contains 71 observations: chick weight, and the type of feed use.

```
boxplot(chickwts$weight ~ chickwts$feed)
```

Motivating example—T test ii



Motivating example—T test iii

- We will focus on two types of feed: *soybean* and *linseed*

```
soy_vec <- chickwts$weight[chickwts$feed == ”soybean”]
lin_vec <- chickwts$weight[chickwts$feed == ”linseed”]

c(length(soy_vec), length(lin_vec))
```

```
## [1] 14 12
```

- We are interested in whether different feed leads to differences in weight.

Motivating example—T test iv

- One way to formalize this into a hypothesis test is to test whether the *mean weight* is the same for both groups:

$$H_0 : \mu_S = \mu_L.$$

- Our estimators are the sample means for each group.
- In STAT 1150, we saw that we can use the t statistic to perform a t-test for two means.

```r
# By default, it assumes unequal variance
(fit <- t.test(soy_vec, lin_vec, var.equal = TRUE))
```

Motivating example—T test v

```
##
## Two Sample t-test
##
## data: soy_vec and lin_vec
## t = 1.3208, df = 24, p-value = 0.199
## alternative hypothesis: true difference in
means is not equal to 0
## 95 percent confidence interval:
## -15.57282 70.92996
## sample estimates:
## mean of x mean of y
## 246.4286 218.7500
```

Motivating example—T test vi

- We get a lot of information out of this:
    - The sample means are 246.4 and 218.8, respectively.
    - A 95% confidence interval for the *mean difference* is $(-15.6, 70.9)$
    - The p-value is 0.199
- Overall, we don't have enough evidence to reject the null hypothesis.
- But what were the assumptions?

Motivating example—T test vii

- It's helpful to recall what's actually going on.
    - Compute the sample means $\hat{\mu}_S$ and $\hat{\mu}_L$.
    - Compute the pooled variance $\hat{\sigma}^2$.
    - Construct the t-statistic $t = \frac{\hat{\mu}_S - \hat{\mu}_L}{\hat{\sigma}\sqrt{n_S^{-1} + n_L^{-1}}}$.
- If the null hypothesis holds and if **the weights are normally distributed with the same variance**, then t follows a t distribution on $n_S + n_L - 2$ degrees of freedom.

Motivating example—T test viii

```
# Let's bootstrap
B <- 5000
data <- chickwts[chickwts$feed %in% c(”soybean”,
                                      ”linseed”), ]
n <- nrow(data)
```

Motivating example—T test ix

```r
results <- replicate(B, {
  indices <- sample(n, n, replace = TRUE)
  data_b <- data[indices,]

  soy_b <- data_b$weight[data_b$feed == ”soybean”]
  lin_b <- data_b$weight[data_b$feed == ”linseed”]
  mean(soy_b) - mean(lin_b)
})
```

Motivating example—T test x

```r
# 95% confidence interval
mu_diff <- mean(soy_vec) - mean(lin_vec)
se_boot <- sd(results)
c(mu_diff - 1.96*se_boot, mu_diff + 1.96*se_boot)
```

```
## [1] -12.38240  67.73955
```

- The bootstrap confidence interval is a bit narrower, but it still leads to the same conclusion.
- On the other hand, how can we compute a p-value?

Permutation tests i

- **Permutation tests** are a large family of resampling methods that can be used to test hypotheses of the form

$$H_0 : F = G,$$

  where F, G are the distribution functions of two different samples.
- You can see this as a *generalization* of the t-test in two ways:
  - We replace equality of means by equality of distributions.
  - We don't assume the data follows a normal distribution.

Permutation tests ii

- It can also be used to test for **independence**:
    - If we have two variables X, Y, with F_X, F_Y the marginal distributions and F_{XY} the joint distribution, independence is equivalent to $F_{XY} = F_X F_Y$.
- The main idea is as follows:
    - Let $X_1, \ldots, X_n \sim F$ and $Y_1, \ldots, Y_m \sim G$.
    - If $H_0 : F = G$ holds, then $X_1, \ldots, X_n, Y_1, \ldots, Y_m \sim F$.
    - Furthermore, **any** permutation of these $n + m$ random variables is also a sample from F!
    - This gives us a way to “generate” data under the null hypothesis.

Permutation tests iii

Algorithm

Let $N = n + m$, and let $\hat{\theta}$ be the estimate for the original sample.

1. Permute the observations to get a sample $Z_1, \dots, Z_N$.
2. Compute the estimate $\hat{\theta}^{(k)} = \hat{\theta}(Z_1, \dots, Z_N)$.
3. Repeat these two steps K times.
4. The permutation p-value is given by

$$\hat{p} = \frac{1 + \sum_{k=1}^{K} I(\hat{\theta}^{(k)} \geq \hat{\theta})}{K + 1}.$$

A few observations i

- The procedure is usually considered *approximate*, because we are not using all possible permutations.
    - In practice, 1000 permutations will give a good approximation for $\alpha = 0.05$.
- Assume that our estimator is the difference of means, like in the motivating example. To compute the permuted estimate $\hat{\theta}^{(k)}$, we compute the sample mean of the first n observations, the sample mean of the remaining m observations, and take the difference.
    - Remember: under the null hypothesis, group membership is meaningless!

A few observations ii

- In bootstrap, it was important to *preserve* the correlation structure between different variables. With permutation tests, the goal is to *break* the association in order to mimic the null hypothesis.
- Permutations = Sampling **without** replacement.

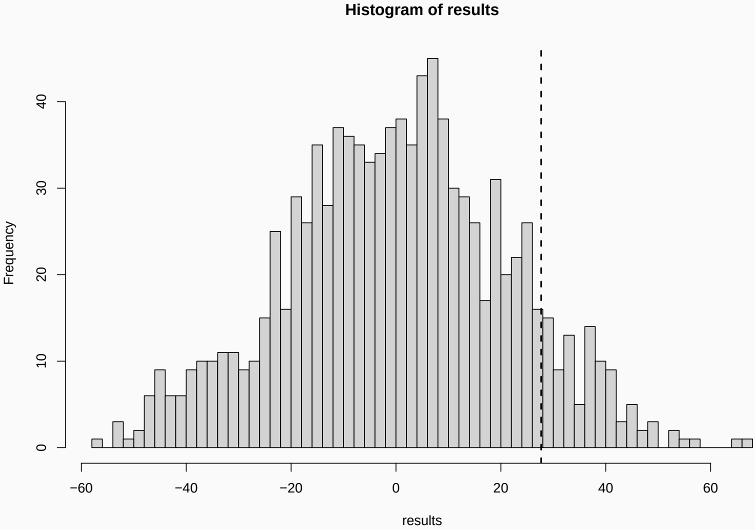
Example (cont'd) i

```r
K <- 1000 # Number of permutations
combined_data <- c(soy_vec, lin_vec) # Combine data
N <- length(combined_data)
results <- replicate(K, {
  perm_data <- combined_data[sample(N)] # Permute
  soy_perm <- perm_data[1:length(soy_vec)] # Allocate
  lin_perm <- perm_data[(length(soy_vec) + 1):N]
  mean(soy_perm) - mean(lin_perm)
})
```

Example (cont'd) ii

```r
theta_hat <- mean(soy_vec) - mean(lin_vec)
hist(results, 50)
abline(v = theta_hat, lty = 2, lwd = 2)
```

Example (cont'd) iii



Example (cont'd) iv

```r
# Is this the right p-value?
mean(c(theta_hat, results) >= theta_hat)
```

```
## [1] 0.0989011
```

```r
# What about this?
mean(abs(c(theta_hat, results)) >= abs(theta_hat))
```

```
## [1] 0.1928072
```

Example (cont'd) v

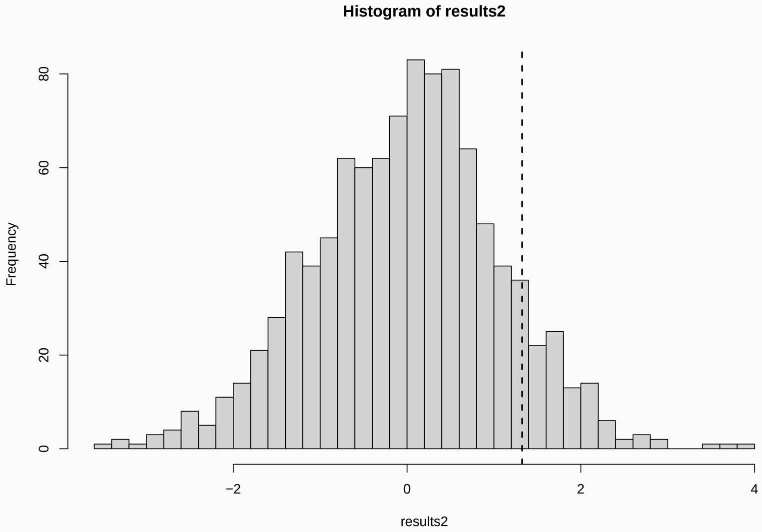
- We used the difference in sample means as our test statistic, but we can also use the t-statistic.

```r
results2 <- replicate(K, {
  perm_data <- combined_data[sample(N)] # Permute
  soy_perm <- perm_data[1:length(soy_vec)] # Allocate
  lin_perm <- perm_data[(length(soy_vec) + 1):N]
  t.test(soy_perm, lin_perm)$statistic
})
```

Example (cont'd) vi

```
t_hat <- t.test(soy_vec, lin_vec)$statistic
hist(results2, 50)
abline(v = t_hat, lty = 2, lwd = 2)
```

Example (cont'd) vii



Example (cont'd) viii

```r
# One-sided p-value
mean(c(t_hat, results2) >= t_hat)
```

```
## [1] 0.1058941
```

```r
# Two-sided p-value
mean(abs(c(t_hat, results2)) >= abs(t_hat))
```

```
## [1] 0.2247752
```

Other test statistics i

- We already saw above that we can use different test statistics for the same null hypothesis.
- On the other hand, you probably noticed that comparing means is probably not strict enough for $H_0 : F = G$.
    - Distributions can be different but have the same mean.
- One way to more directly compare the full distribution is the *Kolmogorov-Smirnov* test statistic:

$$D = \max_{1 \leq i \leq N} \left| F_n(Z_i) - G_m(Z_i) \right| ,$$

where F_n, G_m are the empirical CDFs of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, respectively.

Other test statistics ii

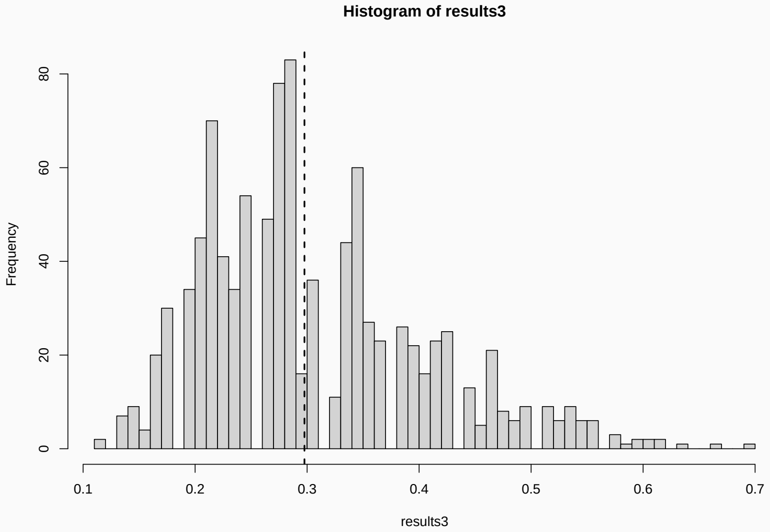
- The asymptotic distribution of D under the null hypothesis is known, but difficult to compute.
- Permutation tests are a simple alternative.

Example (cont'd) i

```
results3 <- replicate(K, {
  perm_data <- combined_data[sample(N)] # Permute
  soy_perm <- perm_data[1:length(soy_vec)] # Allocate
  lin_perm <- perm_data[(length(soy_vec) + 1):N]
  ks.test(soy_perm, lin_perm)$statistic
})
```

```
D_hat <- ks.test(soy_vec, lin_vec)$statistic
hist(results3, 50)
abline(v = D_hat, lty = 2, lwd = 2)
```

Example (cont'd) ii



Example (cont'd) iii

```
# Only one-sided p-value
mean(c(D_hat, results3) >= D_hat)
```

```
## [1] 0.4405594
```

Another example i

- We will use the same dataset, but compare sunflower and linseed feeds.

```
sun_vec <- chickwts$weight[chickwts$feed==”sunflower”]
c(length(sun_vec), length(lin_vec))
```

```
## [1] 12 12
```

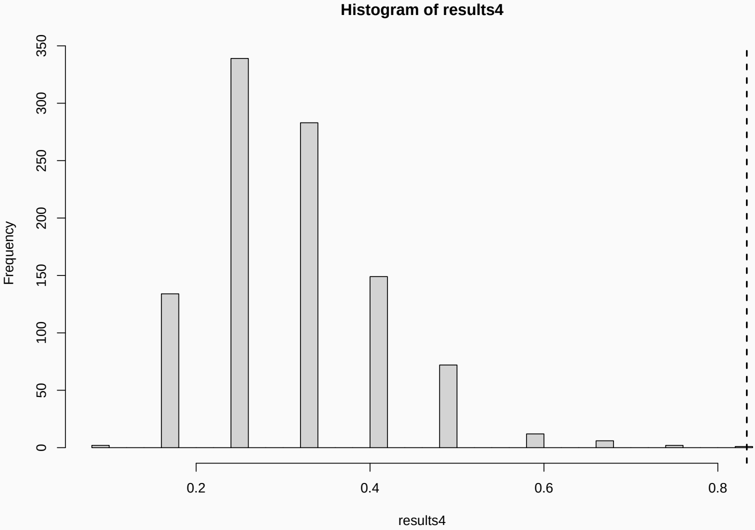
Another example ii

```
K <- 1000 # Number of permutations
combined_data <- c(sun_vec, lin_vec) # Combine data
N <- length(combined_data)
results4 <- replicate(K, {
  perm_data <- combined_data[sample(N)] # Permute
  sun_perm <- perm_data[1:length(sun_vec)] # Allocate
  lin_perm <- perm_data[(length(sun_vec) + 1):N]
  ks.test(sun_perm, lin_perm)$statistic
})
```

Another example iii

```r
D_hat <- ks.test(sun_vec, lin_vec)$statistic
hist(results4, 50)
abline(v = D_hat, lty = 2, lwd = 2)
```

Another example iv



Another example v

```
# Only one-sided p-value
mean(c(D_hat, results4) >= D_hat)
```

```
## [1] 0.001998002
```

Final remarks

- This is our last module on resampling methods.
- We discussed **jackknife**, **bootstrap** and **permutation tests**.
  - Bootstrap and jackknife have similar goals, but bootstrap is almost always better.
  - Permutation tests are *specifically* for hypothesis testing.
- Permutation tests are usually more powerful than looking at bootstrap confidence intervals.
  - Meaning, the probability of rejecting the null hypothesis when it **doesn't** hold is higher with permutation tests.
- Different test statistics will give different results.
  - Monte Carlo simulations is helpful in understanding when we should choose a given test statistic.