Practice problems–Permutation Tests

Problem 1

In the lecture notes on Permutation tests, I mentioned that bootstrap hypothesis tests tend to be less powerful than permutation tests. Design and conduct a simulation study to compare the power between a bootstrap and permutation test, using a difference of means as the test statistic.

You will need to specify the following quantities:

- The number of simulations N.
- The sample sizes n_X, n_Y of both samples.
- The distribution F_X of the sample $X_1, \ldots, X_{n_X}$.
- The distribution F_Y of the sample $Y_1, \ldots, Y_{n_Y}$.
- The difference $\Delta = \mu_X - \mu_Y$ between the population means.

Problem 2

Show that the following statistics lead to equivalent permutation tests for the equality of two population means:

a) the sum of the observations in the smallest sample;
b) the difference between the sample means;
c) the t-statistic.

In other words, given a set of permutations of the data and significance level α, one of these tests will reject the null hypothesis if and only if all tests reject the null hypothesis.

Hint: The sum of all the observations, the sum of the squares of all the observations, and the sample sizes are invariant under permutations.

Problem 3

Consider the `varespec` dataset from the `vegan` package:

```
library(vegan)
data(varespec)
dim(varespec)
```

```
## [1] 24 44
```

Each column represents cover values for 44 species of lichen, measured in 24 different sites. The first 16 rows correspond to grazed pastures; the last 8 rows correspond to ungrazed pastures. See `?vegan::varespec` for more details.

For this problem, we will focus on one species: *Cladonia coccifera*.

```r
# Cover values for Clad. Coccifera
comb_data <- varespec$Cladcocc
# Two groups: grazed vs ungrazed
groups <- factor(c(rep(1, 16), rep(2, 8)),
                 labels = c("grazed", "ungrazed"))
```

We are interested in comparing the distribution of cover values of *Cladonia coccifera* in grazed and ungrazed pastures.

a. Use bootstrap to construct a 95% confidence interval for the **difference in means**. You can choose any type of confidence interval.
b. Using a permutation test, compute a p-value using the **difference in sample means**, under the null hypothesis that both population means are equal.
c. Using a permutation test, compute a p-value using the **t statistic**, under the null hypothesis that both population means are equal. You can use the statistic from Welch's t-test, or from a two-sample t-test with equal variances.
d. Using a permutation test, compute a p-value using the **ratio of sample variances** as your statistic, under the null hypothesis that both population variances are equal. (*Hint*: Think very carefully about whether you want a one-sided or two-sided p-value, and what it means for a ratio of variances to be "more extreme" than a certain value. It may be helpful to transform the ratio first.)
e. **(Bonus)**: Using a permutation test, compute a p-value using the **Kolmogorov-Smirnov** test statistic, under the null hypothesis that both distributions are equal. Give an interpretation for the results of all five hypothesis tests.

Problem 4

We will use the `fastfood` dataset from the `openintro` package.

This dataset contains nutritional information about several meals offered in major fast-food chains. We will focus on two restaurants: Taco Bell and Burger King. To answer the questions below, we first need to **restrict our dataset to only these two restaurants**. by running the codebelow:

a. Using the difference in sample means as your statistic, perform a permutation test and compute a p-value for the null hypothesis that the distribution of calories is the same for both Taco Bell and Burger King. At significance level $\alpha = 0.05$, should we reject the null hypothesis?
b. Using bootstrap, construct a 95% confidence interval for the difference in means. You can choose any type of bootstrap confidence interval, except the percentile interval.
c. Consider the simple linear regression model where `calories` is the outcome and `sodium` is the covariate. Perform a residual analysis for this model.
d. Construct a 95% bootstrap confidence interval for the estimate of the slope (i.e. the coefficient corresponding to `sodium`). Be explicit about which type of resampling approach you chose, and explain why you chose it.

Problem 5

Consider the following 2×2 contingency table, describing gender distribution and whether a student's major is statistics, for a given course:

	Male students	Female students
Stats major	1	9

	Male students	Female students
Non-Stats major	11	3

The null hypothesis is that male and female students are equally likely to have statistics as their major.

Using Fisher's exact test, compute a p-value for this null hypothesis.

Problem 6

This topic is more advanced, and I've included it only for your personal interest.

As we briefly mentioned in class, permutation tests can be used to test for independence. Recall the general idea: suppose we have two variables X, Y (they could also be random vectors), with F_X, F_Y the ma rginal distributions and F_{XY} the joint distribution. Then independence is equivalent to $F_{XY} = F_X F_Y$. This means that independence between X and Y corresponds to the following null hypothesis:

$$H_0 : F_{XY} = F_X F_Y.$$

If we assume our data follows a normal distribution, then independence is equivalent to having no correlation. In that special case, we could test for independence by building a correlation test:

$$H_0 : \mathrm{Corr}(X, Y) = 0.$$

But in general, independence is stricter than no correlation. For that reason, we need a different test statistic.

Székely, Rizzo, and Bakirov (2007) introduced the notion of **distance correlation** for exactly that purpose: they were looking for a measure of dependence that would be equal to zero if and only if the variables X, Y are independent. Their definition is valid for random vectors $\mathbf{X}, \mathbf{Y}$ of different lengths. But for the purposes of this discussion, we will focus on X, Y random variables.

Here are the definitions. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of size n (so both variables are measured on the same experimental unit). Define two $n \times n$ matrices:

$$a_{k\ell} = |X_k - X_\ell|, \qquad b_{k\ell} = |Y_k - Y_\ell|.$$

Next, create two new $n \times n$ matrices from the previous ones by subtracting the row means, the column means, and adding back the overall mean:

$$A_{k\ell} = a_{k\ell} - \bar{a}_{k\cdot} - \bar{a}_{\cdot\ell} + \bar{a}_{\cdot\cdot},$$
$$B_{k\ell} = b_{k\ell} - \bar{b}_{k\cdot} - \bar{b}_{\cdot\ell} + \bar{b}_{\cdot\cdot}.$$

Next, define the **empirical distance covariance** as follows:

$$V(X, Y) = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} A_{k\ell} B_{k\ell},$$

$$V(X) = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} A_{k\ell}^2,$$

$$V(Y) = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} B_{k\ell}^2.$$

Finally, the **empirical distance correlation** is defined as:

$$R(X, Y) = \frac{V(X, Y)}{\sqrt{V(X) V(Y)}},$$

as long as the denominator is nonzero. Otherwise, we set $R(X, Y) = 0$.

How should we permute the data? If the data were independent, then pairing observations (X_i, Y_i) would be meaningless: the pair would be as likely as (X_i, Y_j) for any $j = 1, \ldots, n$. This gives us a h int for how the data should be permuted: permute the values of the vector $(X_1, \ldots, X_n)$ or $(Y_1, \ldots, Y_n)$ (no need to do both), and recompute the estimate.

a. Create a function `compute_distcor`, which takes two vectors `xvec` and `yvec` as input, and outputs the distance correlation.
b. Using one of the datasets we saw in class (e.g. `bootstrap::law` or `bootstrap::scor`), test the independence hypothesis using the distance correlation and a permutation test.