Elements of Nonparametric Statistics

Nicholas Henderson

2020-01-30

Contents

Preface **5**

1 Introduction **7**
1.1 What is Nonparametric Statistics? 7
1.2 Outline of Course 8
1.3 Example 1: Nonparametric vs. Parametric Two-Sample Testing . 8
1.4 Example 2: Nonparametric Estimation 11
1.5 Example 3: Confidence Intervals 13
1.6 Example 4: Nonparametric Regression with a Single Covariate . 16
1.7 Example 5: Classification and Regression Trees (CART) 18

2 Working with R **21**

I Nonparametric Testing **23**

3 Rank and Sign Statistics **25**
3.1 Ranks 25
3.2 The Wilcoxon Rank Sum (WRS) Test: A Two-Sample Test . . . 27
3.3 One Sample Tests 36
3.4 Power and Comparisons with Parametric Tests 43
3.5 Linear Rank Statistics in General 52
3.6 Additional Reading 57

4 Rank Tests for Multiple Groups **59**
4.1 The Kruskal-Wallis Test 60
4.2 Performing the Kruskal-Wallis Test in R 63
4.3 Comparison of Specific Groups 65
4.4 An Additional Example 66
4.5 Additional Reading 67

5 Permutation Tests **69**
5.1 Notation 69
5.2 Permutation Tests for the Two-Sample Problem 70

5.3 The Permutation Test as a Conditional Test 75
5.4 A Permutation Test for Correlation 78
5.5 A Permutation Test for Variable Importance in Regression and Machine Learning 80

Preface

This book will serve as the main source of course notes for Biostatistics 685/Statistics 560, Winter 2020.

Chapter 1

Introduction

1.1 What is Nonparametric Statistics?

What is Parametric Statistics?

- Parametric models refer to probability distributions that can be fully described by a fixed number of parameters that do not change with the sample size.
- Typical examples include
  - Gaussian
  - Poisson
  - Exponential
  - Beta
- Could also refer to a regression setting where the mean function is described by a fixed number of parameters.

What is Nonparametric Statistics?

- It is difficult to give a concise, all-encompassing definition, but nonparametric statistics generally refers to statistical methods where there is not a clear parametric component.
- A more practical definition is that nonparametric statistics refers to flexible statistical procedures where very few assumptions are made regarding the distribution of the data or the form of a regression model.

- The uses of nonparametric methods in several common statistical contexts are described in Sections 1.3 - 1.7.

1.2 Outline of Course

This course is roughly divided into the following 5 categories.

1. **Nonparametric Testing**

- Rank-based Tests
- Permutation Tests

1. **Estimation of Basic Nonparametric Quantities**

- The Empirical Distribution Function
- Density Estimation

1. **Nonparametric Confidence Intervals**

- Bootstrap
- Jacknife

1. **Nonparametric Regression Part I (Smoothing Methods)**

- Kernel Methods
- Splines
- Local Regression

1. **Nonparametric Regression Part II (Machine Learning Methods)**

- Decision Trees/CART
- Ensemble Methods

1.3 Example 1: Nonparametric vs. Parametric Two-Sample Testing

Suppose we have data from two groups. For example, outcomes from two different treatments.

- **Group 1 outcomes**: $X_1, \ldots, X_n$ an i.i.d (independent and identically distributed) sample from distribution function F_X. This means that

$$F_X(t) = P(X_i \leq t) \quad \text{for any } 1 \leq i \leq n$$

- **Group 2 outcomes**: $Y_1, \ldots, Y_m$ an i.i.d. sample from distribution function F_Y.

$$F_Y(t) = P(Y_i \leq t) \quad \text{for any } 1 \leq i \leq n$$

- To test the impact of a new treatment, we usually want to test whether or not F_X differs from F_Y in some way. This can be stated in hypothesis testing language as

$$\begin{aligned} H_0 &: F_X = F_Y \quad (\text{ populations are the same}) \\ H_A &: F_X \neq F_Y \quad (\text{ populations are different}) \end{aligned} \tag{1.1}$$

Parametric Tests

- Perhaps the most common parametric test for (1.1) is the **t-test**. The t-test assumes that

$$F_X = \text{Normal}(\mu_x, \sigma^2) \quad \text{and} \quad F_Y = \text{Normal}(\mu_y, \sigma^2) \tag{1.2}$$

- Under this parametric assumption, the hypothesis test (1.1) reduces to

$$H_0 : \mu_x = \mu_y \quad \text{vs.} \quad H_A : \mu_x \neq \mu_y \tag{1.3}$$

- The standard t-statistic (with a pooled estimate of σ^2) is the following

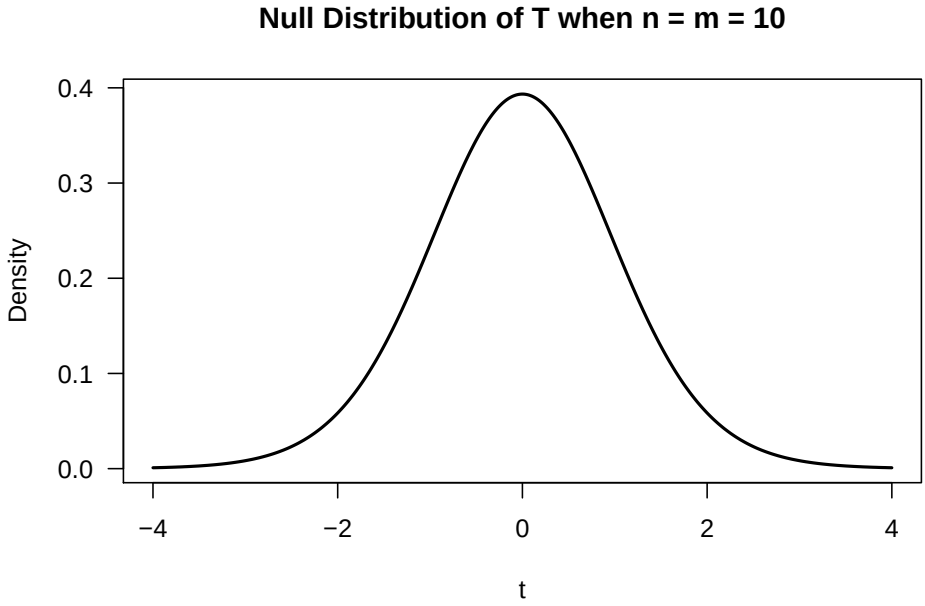
$$T = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}, \tag{1.4}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$ are the group-specific sample means and s_p^2 is the pooled estimate of σ^2

$$s_p^2 = \frac{1}{m+n-2}\Big\{ \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2 \Big\} \tag{1.5}$$

- The t-test is based on the **null distribution** of T - the distribution of T under the null hypothesis.

- Under the assumption of normality, the null distribution of T is a t distribution with $n + m - 2$ degrees of freedom.



- Notice that the null distribution of T depends on the parametric assumption that both $F_X = \text{Normal}(\mu_x, \sigma^2)$ and $F_Y = \text{Normal}(\mu_y, \sigma^2)$. Appealing to the Central Limit Theorem, one could argue that is a quite reasonable assumption.
- In addition to using the assumption that $F_X = \text{Normal}(\mu_x, \sigma^2)$ and $F_Y = \text{Normal}(\mu_y, \sigma^2)$, we used this parametric assumption (at least implicitly) in the formulation of the hypothesis test itself because we assumed that any difference between F_X and F_Y would be fully described by difference in μ_x and μ_y.
- So, in a sense, you are using the assumption of normality twice in the construction of the two-sample t-test.

Nonparametric Tests

- Two-sample nonparametric tests are meant to be "distribution-free". This means the null distribution of the test statistic does not depend on any parametric assumptions about the two populations F_X and F_Y.
- Many such tests are based on **ranks**. The distribution of the ranks under the assumption that $F_X = F_Y$ do not depend on the form of F_X (assuming F_X is continuous).
- Also, the statements of hypotheses tests for nonparametric tests should not rely on any parametric assumptions about F_X and F_Y.
- For example, $H_A : F_X \neq F_Y$ or $H_A : F_X \geq F_Y$.

- Nonparametric tests usually tradeoff power for greater robustness.
- In general, if the parametric assumptions are correct, a nonparametric test will have less power than its parametric counterpart.
- If the parametric assumptions are not correct, parametric tests might have inappropriate type-I error control or lose power.

1.4 Example 2: Nonparametric Estimation

- Suppose we have n observations $(X_1, \ldots, X_n)$ which are assumed to be i.i.d. (independent and identically distributed). The distribution function of X_i is F_X.
- Suppose we are interested in estimating the entire distribution function F_X rather than specific features of the distribution of X_i such as the mean or standard deviation.
- In a **parametric** approach to estimating F_X, we would assume the distribution of X_i belongs to some parametric family of distributions. For example,
  - $X_i \sim \text{Normal}(\mu, \sigma^2)$
  - $X_i \sim \text{Exponential}(\lambda)$
  - $X_i \sim \text{Beta}(\alpha, \beta)$

- If we assume that $X_i \sim \text{Normal}(\mu, \sigma^2)$, we only need to estimate 2 parameters to fully describe the distribution of X_i, and the number of parameters will not depend on the sample size.
- In a nonparametric approach to characterizing the distribution of X_i, we need to instead estimate the entire distribution function F_X or density function f_X.
- The distribution function F_X is usually estimated by the **empirical distribution function**

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq t), \tag{1.6}$$

  where $I()$ denotes the indicator function. That is, $I(X_i \leq t) = 1$ if $X_i \leq t$, and $I(X_i \leq t) = 0$ if $X_i > t$.
- The empirical distribution function is a discrete distribution function, and it can be thought of as an estimate having n "parameters.

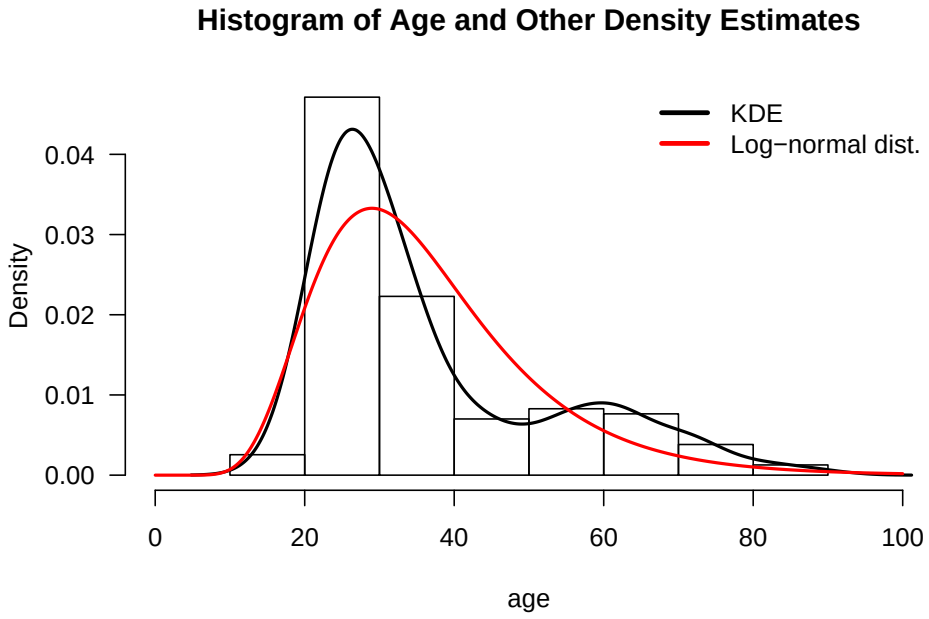
- The density function of X_i is often estimated by a kernel density estimator (KDE). This is defined as

$$\hat{f}_n(t) = \frac{1}{nh_n} \sum_{i=1}^{n} K\Big(\frac{t - X_i}{h_n}\Big). \tag{1.7}$$

- $K()$ - the kernel function

- h_n - the bandwidth

- The KDE is a type of smoothing procedure.



1.5 Example 3: Confidence Intervals

- Inference for a wide range of statistical procedures is based on the following argument

$$\hat{\theta}_n \text{ has an approximate Normal}\left(\theta, \widehat{\text{Var}}(\hat{\theta}_n)\right) \text{ distribution} \tag{1.8}$$

- Above, $\hat{\theta}_n$ is an estimate of a parameter θ, and $\widehat{\text{Var}}(\hat{\theta}_n)$ is an estimate of the variance of $\hat{\theta}_n$.
- $se_n = \sqrt{\widehat{\text{Var}}(\hat{\theta}_n)}$ is usually referred to as the **standard error**.
- 95% confidence intervals are reported using the following formula

$$[\hat{\theta}_n - 1.96se_n, \hat{\theta}_n + 1.96se_n] \tag{1.9}$$

- Common examples of this include:
  1. $\hat{\theta}_n = \bar{X}_n$.

  In this case, appeals to the Central Limit Theorem would justify approximation (1.8). The variance of $\hat{\theta}_n$ would be σ^2, and the standard error would typically be $se_n = \hat{\sigma}/\sqrt{n}$.

  2. $\hat{\theta}_n$ = Maximum Likelihood Estimate of θ.

In this case, asymptotics would justify the approximate distribution $\hat{\theta}_n \sim \text{Normal}(\theta, \frac{1}{nI(\theta)})$, where $I(\theta)$ denotes the Fisher information. The standard error in this context is often $se_n = \{nI(\hat{\theta}_n)\}^{-1/2}$.

- Confidence intervals using (1.8) rely on a parametric approximation to the sampling distribution of the statistic $\hat{\theta}_n$.
- Moreover, even if one wanted to use something like (1.8), working out standard error formulas can be a great challenge in more complicated situations.

- The **bootstrap** is a simulation-based approach for computing standard errors and confidence intervals.
- The bootstrap does not rely on any particular parametric assumptions and can be applied in almost any context (though bootstrap confidence intervals can fail to work as desired in some situations).
- Through resampling from the original dataset, the bootstrap uses many possible alternative datasets to assess the variability in $\hat{\theta}_n$.

	OriginalDat	Dat1	Dat2	Dat3	Dat4
Obs. 1	0.20	0.20	0.80	0.20	0.30
Obs. 2	0.50	0.20	0.80	0.20	0.70
Obs. 3	0.30	0.30	0.50	0.80	0.20
Obs. 4	0.80	0.30	0.70	0.50	0.50
Obs. 5	0.70	0.70	0.20	0.30	0.20
theta.hat	0.50	0.34	0.60	0.40	0.38

- In the above example, we have 4 **boostrap replications** for the statistic $\hat{\theta}$:

$$\hat{\theta}^{(1)} = 0.34 \tag{1.10}$$
$$\hat{\theta}^{(2)} = 0.60 \tag{1.11}$$
$$\hat{\theta}^{(3)} = 0.40 \tag{1.12}$$
$$\hat{\theta}^{(4)} = 0.38 \tag{1.13}$$

- In the above example, the bootstrap standard error for $\hat{\theta}_n$ would be the

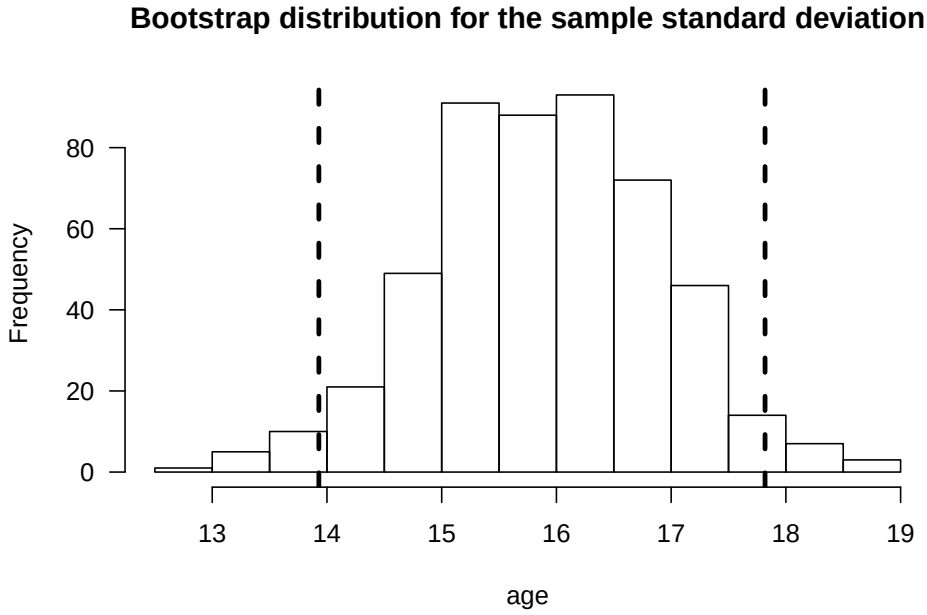


Figure 1.1: Bootstrap distribution of the sample standard deviation for the age variable from the kidney fitness data. Dasjed vertical lines are placed at the 2.5 and 97.5 percentiles of the bootstrap distribution.

standard deviation of the bootstrap replications

$$\begin{aligned} se_{boot} &= \Big(\frac{1}{3}\sum_{b=1}^{4}\{\hat{\theta}^{(b)} - \hat{\theta}^{(-)}\}^2\Big)^{1/2} \\ &= \Big((0.34-0.43)^2/3 + (0.60-0.43)^2/3 + (0.40-0.43)^2/3 + (0.38-0.43)^2/3\Big)^{1/2} \\ &= 0.116 \end{aligned} \tag{1.14}$$

where $\hat{\theta}^{(-)} = 0.43$ is the average of the bootstrap replications.

- One would then report the confidence interval $[\hat{\theta} - 1.96 \times 0.116, \hat{\theta} + 1.96 \times 0.116]$. In practice, the number of bootstrap replications is typically much larger than 4.

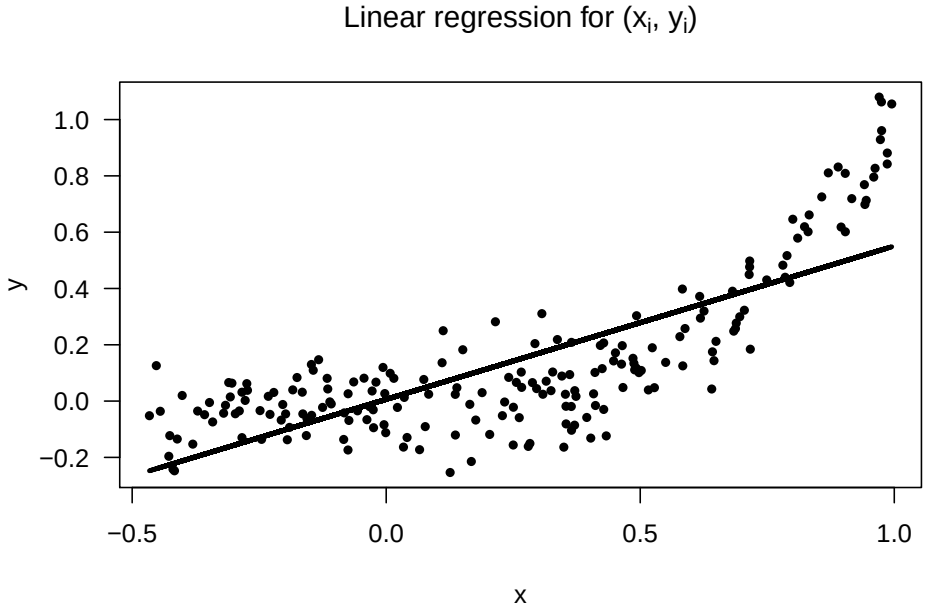
- It is often better to construct confidence intervals using the percentiles from the bootstrap distribution of $\hat{\theta}$ rather than use a confidence interval of the form: $\hat{\theta} \pm 1.96 \times se_{boot}$.

1.6 Example 4: Nonparametric Regression with a Single Covariate

- Regression is a common way of modeling the relationship between two different variables.
- Suppose we have n pairs of observations $(y_1, x_1), \ldots, (y_n, x_n)$ where y_i and x_i are suspected to have some association.
- Linear regression would assume that these y_i and x_i are related by the following

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1.15}$$

  with the assumption $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$ often made.
- In this model, there are only 3 parameters: $(\beta_0, \beta_1, \sigma^2)$, and the number of parameters stays fixed for all n.



- The nonparametric counterpart to linear regression is usually formulated in the following way

$$y_i = m(x_i) + \varepsilon_i \tag{1.16}$$

- Typically, one makes very few assumptions about the form of the mean function m, and it is not assumed m can be described by a finite number of parameters.
- There are a large number of nonparametric methods for estimating m.

- One popular method is the use of **smoothing splines**.

- With smoothing splines, one considers mean functions of the form

$$m(x) = \sum_{j=1}^{n} \beta_j g_j(x) \tag{1.17}$$

where $g_1, \ldots, g_n(x)$ are a collection of spline basis functions.

- Because of the large number of parameters in (1.17), one should estimate the basis function weights β_j through penalized regression

$$\text{minimize} \quad \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{n} \beta_j g_j(x_i) \Big)^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \Omega_{ij} \beta_i \beta_j \tag{1.18}$$

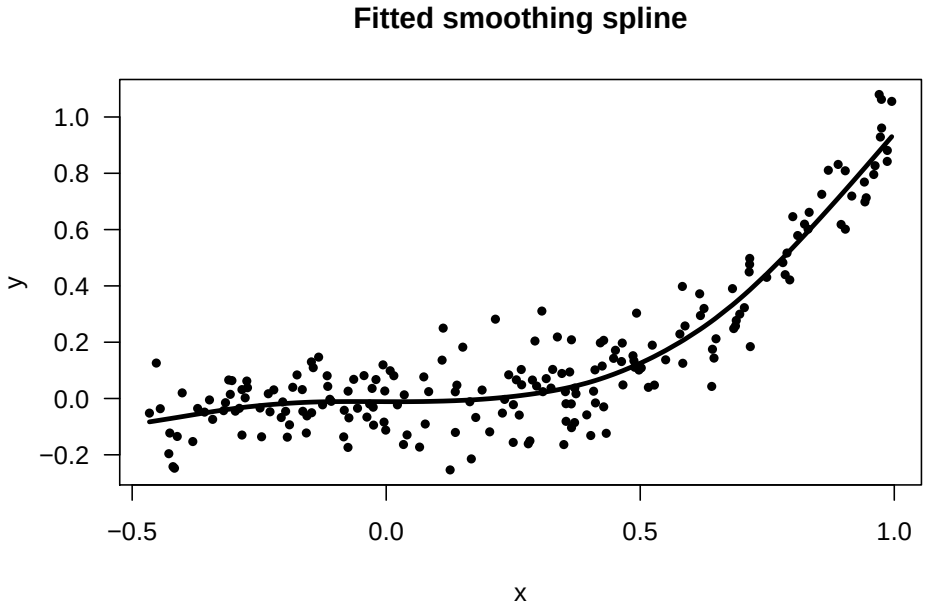
where $\Omega_{ij} = \int g_i''(t) g_j''(t) dt$.

- Using coefficient estimates $\hat{\beta}_1, \ldots, \hat{\beta}_n$ found from solving (1.17), the nonparametric estimate of the mean function is defined as

$$\hat{m}(x) = \sum_{j=1}^{n} \hat{\beta}_j g_j(x) \tag{1.19}$$

- While the estimation in (1.18) resembles parametric estimation for linear regression, notice that the number of parameters to be estimated will change with the sample size.

- Allowing the number of basis functions to grow with n is important. For a sufficiently large number of basis functions, one should be able to approximate the true mean function $m(x)$ arbitrarily closely.



1.7 Example 5: Classification and Regression Trees (CART)

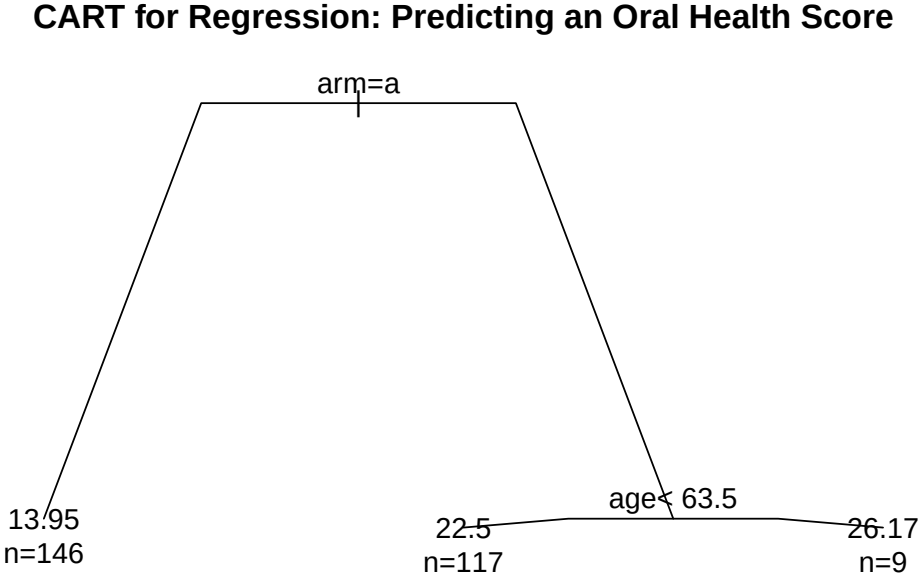
- Suppose we now have observations $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ where y_i is a continuous response and $\mathbf{x}_i$ is a p-dimensional vector of covariates.

- Regression trees are a nonparametric approach for predicting y_i from $\mathbf{x}_i$.

- Here, the regression function is a **decision tree** rather than some fitted curve.

- With a decision tree, a final prediction from a covariate vector $\mathbf{x}_i$ is obtained by answering a sequence of "yes or no" questions.

- When the responses y_i are binary, such trees are referred to as classification trees. Hence, the name: classification and regression trees (CART).



- Classification and regression trees are constructed through **recursive partitioning**.
- Recursive partitioning is the process of deciding if and how to split a given node into two child nodes.
- Tree splits are usually chosen to minimize the "within-node" sum of squares.
- The size of the final is determined by a process of "pruning" the tree with cross-validation determining the best place to stop pruning.

- Regression trees are an example of a more algorithmic approach to constructing predictions (as opposed to probability modeling in more traditional statistical methods) with a strong emphasis on predictive performance as measured through cross-validation.

- While single regression trees have the advantage of being directly interpretable, their prediction performance is often not that great.
- However, using collections of trees can be very effective for prediction and has been used in many popular learning methods. Examples include: random forests, boosting, and Bayesian additive regression trees (BART).
- Methods such as these can perform well on much larger datasets. We will discuss additional methods if time allows.

Chapter 2

Working with R

You can download R by visiting https://www.r-project.org/ and clicking on the **download R** link. Follow the instructions to complete installation. THe most recent version is version 3.6.2.

It is not necessary to use this, but I find **RStudio** to be a very useful integrated development environment (IDE) for computing with **R**. **RStudio** may be downloaded and installed by visiting https://rstudio.com/

Part I

Nonparametric Testing

Chapter 3

Rank and Sign Statistics

3.1 Ranks

3.1.1 Definition

- Suppose we have n observations $\mathbf{X} = (X_1, \ldots, X_n)$. The **rank** of the i^{th} observation R_i is defined as

$$R_i = R_i(\mathbf{X}) = \sum_{j=1}^{n} I(X_i \geq X_j) \tag{3.1}$$

  where

$$I(X_i \geq X_j) = \begin{cases} 1 & \text{if } X_i \geq X_j \\ 0 & \text{if } X_i < X_j \end{cases} \tag{3.2}$$

- The largest observation has a rank of n.
- The smallest observation has a rank of 1 (if there are no ties).
- I'm using the notation $R_i(\mathbf{X})$ to emphasize that the rank of the i^{th} observations depends on the entire vector of observations rather than only the value of X_i.
- You can compute ranks in **R** using the **rank** function:

```
x <- c(3, 7, 1, 12, 6)  ## 5 observations
rank(x)
```

```
## [1] 2 4 1 5 3
```

3.1.2 Handling Ties

- In the definition of ranks shown in (3.1), tied observations receive their maximum possible rank.
- For example, suppose that $(X_1, X_2, X_3, X_4) = (0, 1, 1, 2)$. In this case, one could argue whether both observations 2 and 3 should be ranked 2^{nd} or 3^{rd} while observations 1 and 4 should unambiguously receive ranks of 1 and 4 respectively.
- Under definition (3.1), both observations 2 and 3 receive a rank of 3.
- In **R**, handling ties that is consistent with definition (3.1) is done using the **ties.method = "max"** argument

```
x <- c(0, 1, 1, 2)
rank(x, ties.method="max")
```

```
## [1] 1 3 3 4
```

- The default in **R** is to replace the ranks of tied observations with their "average" rank

```
x <- c(0, 1, 1, 2)
rank(x)
```

```
## [1] 1.0 2.5 2.5 4.0
```

```
y <- c(2, 9, 7, 7, 3, 2, 1)
rank(y, ties.method="max")
```

```
## [1] 3 7 6 6 4 3 1
```

```
rank(y)
```

```
## [1] 2.5 7.0 5.5 5.5 4.0 2.5 1.0
```

- When defining ranks using the "average" or "midrank" approach to handling ties, replaces tied ranks with the average of the two "adjacent" ranks.
- For example, if we have a vector of ranks (R_1, R_2, R_3, R_4) where $R_2 = R_3 = 3$ and $R_1 = 4$ and $R_4 = 1$, then the vector of modified ranks using the "average" approach to handling ties would be

$$(R_1', R_2', R_3', R_4') = \left(4, \frac{4+1}{2}, \frac{4+1}{2}, 1\right) \tag{3.3}$$

- The "average" approach is the most common way of handling ties when computing the Wilcoxon rank sum statistic.

3.1.3 Properties of Ranks

Suppose $(X_1, \ldots, X_n)$ is random sample from a continuous distribution F (so that the probability of ties is zero). Then, the following properties hold for the associated ranks $R_1, \ldots, R_n$.

- Each R_i follows a discrete uniform distribution

$$P(R_i = j) = 1/n, \quad \text{for any } j = 1, \ldots, n. \tag{3.4}$$

- The expectation of R_i is

$$E(R_i) = \sum_{j=1}^{n} jP(R_i = j) = \frac{1}{n}\sum_{j=1}^{n} j = \frac{(n+1)}{2} \tag{3.5}$$

- The variance of R_i is

$$\text{Var}(R_i) = E(R_i^2) - E(R_i)^2 = \frac{1}{n}\sum_{j=1}^{n} j^2 - \Big(\frac{n+1}{2}\Big)^2 = \frac{n^2-1}{12} \tag{3.6}$$

- The random variables $R_1, \ldots, R_n$ are **not** independent (why?). However, the vector $\mathbf{R}_n = (R_1, \ldots, R_n)$ is uniformly distributed on the set of $n!$ permutations of $(1, 2, \ldots, n)$.

Exercise 3.1: Suppose X_1, X_2, X_3 are i.i.d. observations from a continuous distribution function F_X. Compute the covariance matrix of the vector of ranks $\big(R_1(\mathbf{X}), R_2(\mathbf{X}), R_3(\mathbf{X})\big)$.

Exercise 3.2: Again, suppose that X_1, X_2, X_3, X_4 are i.i.d. observations from a continuous distribution function F_X. Let $T = R_1(\mathbf{X}) + R_2(\mathbf{X})$. Compute $P(T = j)$ for $j = 3, 4, 5, 6, 7$.

3.2 The Wilcoxon Rank Sum (WRS) Test: A Two-Sample Test

3.2.1 Goal of the Test

- The Wilcoxon Rank Sum (WRS) test (sometimes referred to as the Wilcoxon-Mann-Whitney test) is a popular, rank-based two-sample test.
- The WRS test is used to test whether or not observations from one group tend to be larger (or smaller) than observations from the other group.

- Suppose we have observations from two groups: $X_1, \ldots, X_n \sim F_X$ and $Y_1, \ldots, Y_m \sim F_Y$.
- Roughly speaking, the WRS tests the following hypothesis

$$\begin{aligned} H_0: &\quad F_X = F_Y \quad \text{versus} \\ H_A: &\quad \text{Observations from } F_X \text{ tend to be larger than observations from } F_Y \end{aligned} \tag{3.7}$$

- What is meant by "tend to be larger" in the alternative hypothesis?
- Two common ways of stating the alternative hypothesis for the WRS include
  1. The stochastic dominance alternative

$$\begin{aligned} H_0: &\quad F_X = F_Y \quad \text{versus} \\ H_A: &\quad F_X \text{ is stochastically larger than } F_Y \end{aligned} \tag{3.8}$$

  2. The "shift" alternative

$$\begin{aligned} H_0: &\quad F_X = F_Y \quad \text{versus} \\ H_A: &\quad F_X(t) = F_Y(t - \Delta), \Delta > 0. \end{aligned} \tag{3.9}$$

- A distribution function F_X is said to be stochastically larger than F_Y if $F_X(t) \leq F_Y(t)$ for all t with $F_X(t) < F_Y(t)$ for at least one value of t.
- Note that the "shift alternative" implies stochastic dominance.
- Why do we need to specify an alternative?

- It is often stated that the WRS test is a test of equal medians.
- This is true under the assumption that the relevant alternative is of the form $F_X(t) = F_Y(t - \Delta)$.
- However, one could have a scenario where the two groups have equal medians, but the WRS test has a very high probability of rejecting H_0.
- In addition, in many applications, it is difficult to justify that the "shift alternative" is a reasonable model.
- An alternative is to view the WRS test as performing the following hypothesis test:

$$H_0: \quad P(X_i > Y_j) + \tfrac{1}{2}P(X_i = Y_j) = 1/2 \quad \text{versus} \tag{3.10}$$

$$H_A: \quad P(X_i > Y_j) + \tfrac{1}{2}P(X_i = Y_j) > 1/2 \tag{3.11}$$

  See Divine et al. (2018) for more discussion around this formulation of the WRS test.

- The hypothesis test (3.11) makes fewer assumptions about how F_X and F_Y are related and is, in many cases, more interpretable.
- For example, in medical applications, it is often more natural to answer the question: what is the probability that the outcome under treatment 1 is better than the outcome under treatment 2.
- The justification of hypothesis test (3.11) comes through the close connection between the WRS test statistic W and the Mann-Whitney statistic U_{MW}. Specifically, $W = U_{MW} + n(n+1)/2$. (Although, often U_{MW} is defined as $U_{MW} = mn + n(n+1)/2 - W$).
- The Mann-Whitney statistic divided by mn is an estimate of the probability:

$$P(X_i > Y_j) + \tfrac{1}{2}P(X_i = Y_j) = 1/2. \tag{3.12}$$

3.2.2 Definition of the WRS Test Statistic

- The WRS test statistic is based on computing the sum of ranks (ranks based on the pooled sample) in one group.
- If observations from group 1 tend to be larger than those from group 2, the average rank from group 1 should exceed the average rank from group 2.
- A sufficiently large value of the average rank from group 1 will allow us to reject H_0 in favor of H_A.

- We will define the pooled data vector $\mathbf{Z}$ as

$$\mathbf{Z} = (X_1, \ldots, X_n, Y_1, \ldots, Y_m) \tag{3.13}$$

  This is a vector with length $n + m$.
- The Wilcoxon rank-sum test statistic W for testing hypotheses of the form (3.8) is then defined as

$$W = \sum_{i=1}^{n} R_i(\mathbf{Z}) \tag{3.14}$$

- In other words, the WRS test statistic is the sum of the ranks for those observations coming from group 1 (i.e., the group with the X_i as observations).
- If the group 1 observations tend to, in fact, be larger than the group 2 observations, then we should expect the sum of the ranks in this group to be larger than the sum of the ranks from group 2.

- Under H_0, we can treat both X_i and Y_i as being observations coming from a common distribution function F.
- Hence, the expectation of $R_i(\mathbf{Z})$ under the null hypothesis is

$$E_{H_0}\{R_i(\mathbf{Z})\} = \frac{n+m+1}{2} \tag{3.15}$$

and thus the expectation of W under H_0

$$E_{H_0}(W) = \sum_{i=1}^{n} E_{H_0}\{R_i(\mathbf{Z})\} = \frac{n(n+m+1)}{2} \tag{3.16}$$

- It can be shown that the variance of W under the null hypothesis is

$$\text{Var}_{H_0}(W) = \frac{mn(m+n+1)}{12} \tag{3.17}$$

3.2.3 Computing p-values for the WRS Test

Exact Distribution

- The p-value is found by computing the probability

$$\text{p-value} = P_{H_0}(W \geq w_{obs}) \tag{3.18}$$

where w_{obs} is the observed WRS test statistic that we get from our data.

- Computing p-values for the WRS test requires us to work with the **null distribution** of W. That is, the distribution of W under the assumption that $F_X = F_Y$.
- The exact null distribution is found by using the fact that each possible ordering of the ranks has the same probability. That is,

$$P\{R_1(\mathbf{Z}) = r_1, \ldots, R_{n+m}(\mathbf{Z}) = r_{n+m}\} = \frac{1}{(n+m)!}, \tag{3.19}$$

where $(r_1, \ldots, r_{n+m})$ is any permutation of the set $\{1, 2, \ldots, n+m\}$. Note that the null distribution only depends on n and m.

- Also, there are $\binom{n+m}{n}$ possible ways to assign distinct ranks to group 1.
- Consider an example with $n = m = 2$. In this case, there are $\binom{4}{2} = 6$ distinct ways to assign 2 ranks to group 1. What is the null distribution of the WRS test statistic? Try to verify that

$$\begin{aligned} P_{H_0}(W = 7) &= 1/6 \\ P_{H_0}(W = 6) &= 1/6 \\ P_{H_0}(W = 5) &= 1/3 \\ P_{H_0}(W = 4) &= 1/6 \\ P_{H_0}(W = 3) &= 1/6. \end{aligned}$$

Large-Sample Approximate Distribution

- Looking at (3.14), we can see that the WRS test statistic is a sum of nearly independent random variables (at least nearly independent for large n and m).
- Thus, we can expect that an appropriately centered and scaled version of W should be approximately Normally distributed (recall the Central Limit Theorem).
- The standardized version $\tilde{W}$ of the WRS is defined as

$$\tilde{W} = \frac{W - E_{H_0}(W)}{\sqrt{\text{Var}_{H_0}(W)}} = \frac{W - n(n+m+1)/2}{\sqrt{mn(n+m+1)/12}} \tag{3.20}$$

- Under H_0, $\tilde{W}$ converges in distribution to a Normal$(0, 1)$ random variable.
- A p-value using this large-sample approximation would then be computed in the following way

$$\begin{aligned} \text{p-value} &= P_{H_0}(W \geq w_{obs}) = P\left(\frac{W - n(n+m+1)/2}{\sqrt{mn(n+m+1)/12}} \geq \frac{w_{obs} - n(n+m+1)/2}{\sqrt{mn(n+m+1)/12}}\right) \\ &= P_{H_0}\left(\tilde{W} \geq \frac{w_{obs} - n(n+m+1)/2}{\sqrt{mn(n+m+1)/12}}\right) = 1 - \Phi\left(\frac{w_{obs} - n(n+m+1)/2}{\sqrt{mn(n+m+1)/12}}\right), \end{aligned}$$

  where $\Phi(t)$ denotes the cumulative distribution function of a standard Normal random variable.
- Often, in practice, a continuity correction is applied when using this large-sample approximation. For example, we would compute the probability $P_{H_0}(W \geq w_{obs} - 0.5)$ with the Normal approximation rather than $P_{H_0}(W \geq w_{obs})$ directly.

- Many statistical software packages (including **R**) will not compute p-values using the exact distribution in the presence of ties.
- The **coin** package in **R** does allow you to perform a permutation test in the presence of ties.
- A "two-sided" Wilcoxon rank sum test can also be performed. The two-sided hypothesis tests could either be stated as

$$\begin{aligned} H_0: &\quad F_X = F_Y \quad \text{versus} \\ H_A: &\quad F_X \text{ is stochastically larger or smaller than } F_Y \end{aligned} \tag{3.21}$$

or

$$\begin{aligned} H_0: &\quad F_X = F_Y \quad \text{versus} \\ H_A: &\quad F_X(t) = F_Y(t-\Delta), \Delta \neq 0. \end{aligned} \tag{3.22}$$

or

$$H_0: \quad P(X_i > Y_i) + \tfrac{1}{2}P(X_i = Y_i) = 1/2 \quad \text{versus} \tag{3.23}$$

$$H_A: \quad P(X_i > Y_i) + \tfrac{1}{2}P(X_i = Y_i) \neq 1/2 \tag{3.24}$$

Exercise 3.3. Using the exact distribution, what is the smallest possible one-sided p-value associated with the WRS test for a fixed value of n and m (assuming the probability of ties is zero)?

3.2.4 Computing the WRS test in R

- To illustrate performing the WRS test in **R**, we can use the **wine** dataset from the **rattle.data** package. This dataset is also available from the UCI Machine Learning Repository.

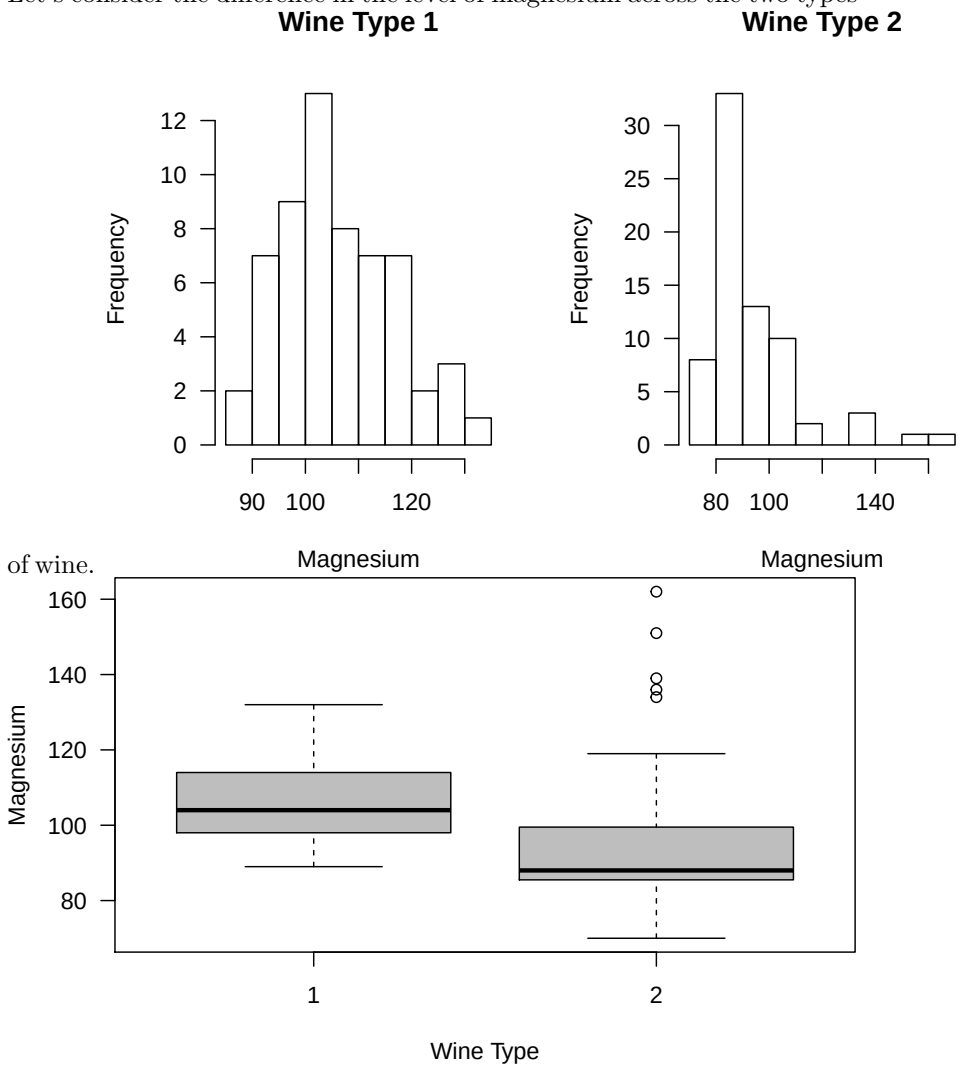
```
library(rattle.data)
head(wine)
```

```
##   Type Alcohol Malic  Ash Alcalinity Magnesium Phenols Flavanoids
## 1    1   14.23  1.71 2.43       15.6       127    2.80       3.06
## 2    1   13.20  1.78 2.14       11.2       100    2.65       2.76
## 3    1   13.16  2.36 2.67       18.6       101    2.80       3.24
## 4    1   14.37  1.95 2.50       16.8       113    3.85       3.49
## 5    1   13.24  2.59 2.87       21.0       118    2.80       2.69
## 6    1   14.20  1.76 2.45       15.2       112    3.27       3.39
##   Nonflavanoids Proanthocyanins Color  Hue Dilution Proline
## 1          0.28            2.29  5.64 1.04     3.92    1065
## 2          0.26            1.28  4.38 1.05     3.40    1050
## 3          0.30            2.81  5.68 1.03     3.17    1185
## 4          0.24            2.18  7.80 0.86     3.45    1480
## 5          0.39            1.82  4.32 1.04     2.93     735
## 6          0.34            1.97  6.75 1.05     2.85    1450
```

- This dataset contains three types of wine. We will only consider the first two.

```
wine2 <- subset(wine, Type==1 | Type==2)
wine2$Type <- factor(wine2$Type)
```

- Let's consider the difference in the level of magnesium across the two types of wine.



- Suppose we are interested in testing whether or not magnesium levels in Type 1 wine are generally larger than magnesium levels in Type 2 wine. This can be done with the following code

```
wilcox.test(x=wine2$Magnesium[wine2$Type==1], y=wine2$Magnesium[wine2$Type==2],
            alternative="greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  wine2$Magnesium[wine2$Type == 1] and wine2$Magnesium[wine2$Type == 2]
```

```
## W = 3381.5, p-value = 8.71e-10
## alternative hypothesis: true location shift is greater than 0
```

You could also use the following code (just be careful about the ordering of the levels of **Type**)

```
wilcox.test(Magnesium ~ Type, data=wine2, alternative="greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Magnesium by Type
## W = 3381.5, p-value = 8.71e-10
## alternative hypothesis: true location shift is greater than 0
```

- What is the value of the WRS test statistic? We can code this directly with the following steps:

```
W <- wilcox.test(x=wine2$Magnesium[wine2$Type==1], y=wine2$Magnesium[wine2$Type==2])

n <- sum(wine2$Type==1)
m <- sum(wine2$Type==2)
zz <- rank(wine2$Magnesium) ## vector of pooled ranks
sum(zz[wine2$Type==1])  ## The WRS test statistic
```

```
## [1] 5151.5
```

- The statistic returned by the **wilcox.test** function is actually equal to $W - n(n+1)/2$ not W

```
sum(zz[wine2$Type==1]) - n*(n + 1)/2
```

```
## [1] 3381.5
```

```
W$statistic
```

```
##      W
## 3381.5
```

- $\{W - n(n+1)/2\}$ is equal to the Mann-Whitney statistic. Thus, **W$statistic/(mn)** is an estimate of the probability $P(X_i > Y_j) + P(X_i = Y_j)/2$.

```
W$statistic/(m*n)
```

```
##         W
## 0.8072332
```

- Let's check how the Mann-Whitney statistic matches a simulation-based estimate of this probability

```
ind1 <- which(wine2$Type==1)
ind2 <- which(wine2$Type==2)
xgreater <- rep(0, 100)
for(k in 1:100) {
    xi <- sample(ind1, size=1)
    yi <- sample(ind2, size=1)
    xgreater[k] <- ifelse(wine2$Magnesium[xi] > wine2$Magnesium[yi], 1, 0) +
                   ifelse(wine2$Magnesium[xi] == wine2$Magnesium[yi], 1/2, 0)
}
mean(xgreater)  ## estimate of this probability
```

```
## [1] 0.825
```

3.2.5 Additional Notes for the WRS test

3.2.5.1 Comparing Ordinal Data

- The WRS test is often suggested when comparing categorical data which are **ordinal**.
- For example, we might have 4 categories:
  - Poor
  - Fair
  - Good
  - Excellent
- In this case, there is a natural ordering of the categories but any numerical values assigned to these categories would be arbitrary.
- In such cases, we might be interested in testing whether or not outcomes tend to be better in one group than the other rather than simply comparing whether or not the distribution is different between the two groups.
- A WRS test is useful here since we can still compute ranks without having to choose aribtrary numbers for each category.
- Thinking of the "probability greater than alternative (3.11)" or "stochastically larger than alternative (3.8)" interpretation of the WRS test is probably more reasonable than the "shift alternative (3.9)" interpretation.
- Note that there will probably be many ties when comparing ordinal data.

- The Hodges-Lehmann Estimator $\hat{\Delta}$ is an estimator of Δ in the location-shift model

$$F_X(t) = F_Y(t - \Delta)$$

- The Hodges-Lehmann is defined as the median difference among all possible (group 1, group 2) pairs. Specifically,

$$\hat{\Delta} = \text{median}\{(X_i - Y_j); i = 1, \ldots, n; j = 1, \ldots, m\}$$

- We won't discuss the Hodges-Lehmann estimator in detail in this course, but in many statistical software packages, the Hodges-Lehmann is often reported when computing the WRS test.
- In **R**, the Hodges-Lehmann estimator can be obtained by using the **conf.int=TRUE** argument in the **wilcox.test** function

```
WC <- wilcox.test(x=wine2$Magnesium[wine2$Type==1], y=wine2$Magnesium[wine2$Type==2],
                  conf.int=TRUE)
WC$estimate     ## The Hodges-Lehmann estimate
```

```
## difference in location
##               14.00005
```

3.3 One Sample Tests

3.3.1 The Sign Test

3.3.1.1 Motivation and Definition

- The **sign test** can be thought of as a test of whether or not the median of a distribution is greater than zero (or greater than some other fixed value θ_0).
- Frequently, the sign test is explained in the following context:
  - Suppose we have observations $D_1, \ldots, D_n$ which arise from the model

$$D_i = \theta + \varepsilon_i, \tag{3.25}$$

    where ε_i are iid random variables each with distribution function F_ϵ that is assumed to have a median of zero. Moreover, we will assume the density function $f_\varepsilon(t)$ is symmetric around zero.
- The distribution function of D_i is then

$$F_D(t) = P(D_i \leq t) = P(\varepsilon_i \leq t - \theta) = F_\epsilon(t - \theta) \tag{3.26}$$

- Likewise the density function $f_D(t)$ of D_i is given by

$$f_D(t) = f_\epsilon(t - \theta) \tag{3.27}$$

- In this context, θ is usually referred to as a **location parameter**.
- The goal here is to test $H_0 : \theta = \theta_0$ vs. $H_A : \theta > \theta_0$. (Often, $\theta_0 = 0$).

- This sort of test usually comes up in the context of **paired data**. Common examples include
  - patients compared "pre and post treatment"
  - students before and after the introduction of a new teaching method
  - comparison of "matched" individuals who are similar (e.g., same age, sex, education, etc.)
  - comparing consistency of measurements made on the same objects

	Baseline_Measure	Post_Treatment_Measure
Patient 1	Y1	X1
Patient 2	Y2	X2
Patient 3	Y3	X3
Patient 4	Y4	X4

- In such cases, we have observations X_i and Y_i for $i = 1, \ldots n$ where it is not necessarily reasonable to think of X_i and Y_i as independent.
- We can define $D_i = X_i - Y_i$ as the difference in the i^{th} pair.
- With this setup, a natural question is whether or not the differences D_i tend to be greater than zero or not.

- The **sign** statistic S_n is defined as

$$S_n = \sum_{i=1}^{n} I(D_i > 0) \tag{3.28}$$

- If the null hypothesis $H_0 : \theta = 0$ is true, then we should expect that roughly half of the observations will be positive.
- This suggests that we will reject H_0 if $S_n \geq c$ where c is a number that is greater than $n/2$.

3.3.1.2 Null Distribution and p-values

- Notice that the sign statistic defined in (3.28) is the sum of independent Bernoulli random variable.
- That is, we can think of $Z_i = I(D_i > 0)$ as a random variable with success probability $p(\theta)$ where the formula for $p(\theta)$ is

$$p(\theta) = P(Z_i = 1) = P(D_i > 0) = 1 - F_D(0) = 1 - F_\epsilon(-\theta) \tag{3.29}$$

- This implies that S_n is a binomial random variable with n trials and success probability $p(\theta)$. That is,

$$S_n \sim \text{Binomial}(n, p(\theta)) \tag{3.30}$$

- Because $p(0) = 1/2$, $S_n \sim \text{Binomial}(n, 1/2)$ under H_0.
- Notice that the "null distribution" of the sign statistic is "distribution free" in the sense that the distribution does not depend on the distribution of D_i.
- The p-value for the sign test can be computed by

$$\text{p-value} = P_{H_0}(S_n \geq s_{obs}) = \sum_{j=s_{obs}}^{n} P_{H_0}(S_n = j) = \sum_{j=s_{obs}}^{n} \binom{n}{j} \frac{1}{2^n}, \tag{3.31}$$

  where s_{obs} is the observed value of the sign statistic.

```
### How to compute the p-value for the sign test using R
xx <- rnorm(100)
sign.stat <- sum(xx > 0)
1 - pbinom(sign.stat - 1, size=100, prob=1/2) ## p-value for sign test
```

```
## [1] 0.7579408
```

- The reason that this is the right expression using **R** is that for any positive integer w

$$P_{H_0}(S_n \geq w) = 1 - P_{H_0}(S_n < w) = 1 - P_{H_0}(S_n \leq w - 1) \tag{3.32}$$

  and the **R** function **pbinom(t, n, prob)** computes $P(X \leq t)$ where X is a binomial random variable with n trials and success probability **prob**.
- You can also perform the one-sided sign test by using the **binom.test** function in **R**.

```
btest <- binom.test(sign.stat, n=100, p=0.5, alternative="greater")
btest$p.value
```

```
## [1] 0.7579408
```

3.3.1.3 Two-sided Sign Test

- Notice that the number of negative values of D_i can be expressed as

$$\sum_{i=1}^{n} I(D_i < 0) = n - S_n \tag{3.33}$$

  if there are no observations that equal zero exactly. Large value of $n - S_n$ would be used in favor of another possible one-sided alternative $H_A : \theta < 0$.

- If we now want to test the two-sided alternative

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_A : \theta \neq 0$$

you would need to compute the probability under the null hypothesis of observing a "more extreme" observation than the one that was actually observed.

- Extreme is defined by thinking about the fact that we would have rejected H_0 if either S_n or $n - S_n$ were very large.
- For example, if $n = 12$, then the expected value of the sign statistic would be 6. If $s_{obs} = 10$, then the collection of "more extreme" events then this would be ≤ 2 and ≥ 10.
- The two-sided p-value is determined by looking at the tail probabilities on both sides

$$\text{p-value} = \begin{cases} P_{H_0}(S_n \geq s_{obs}) + P_{H_0}(S_n \leq n - s_{obs}) & \text{if } s_{obs} \geq n/2 \\ P_{H_0}(S_n \leq s_{obs}) + P_{H_0}(S_n \geq n - s_{obs}) & \text{if } s_{obs} < n/2 \end{cases} \tag{3.34}$$

- It actually works out that

$$\text{p-value} = \begin{cases} 2P_{H_0}(S_n \geq s_{obs}) & \text{if } s_{obs} \geq n/2 \\ 2P_{H_0}(S_n \leq s_{obs}) & \text{if } s_{obs} < n/2 \end{cases} \tag{3.35}$$

- Also, you can note that this p-value would be the same that you would get from performing the test $H_0 : p = 1/2$ vs. $H_A : p \neq 1/2$ when it is assumed that $S_n \sim \text{Binomial}(n, p)$.
- Another note: It is often suggested that one should drop observations which are exactly zero when performing the sign test.

3.3.2 The Wilcoxon Signed Rank Test

- The Wilcoxon signed rank test can be applied under the same scenario that we used the sign test.
- One criticism of the sign test is that it ignores the magnitude of the observations.
- For example, the sign test statistic S treats observations $D_i = 0.2$ and $D_i = 3$ the same.
- The **Wilcoxon signed rank statistic** T_n weights the signs of D_i by the rank of its absolute value.

- Specifically, the Wilcoxon signed rank statistic is defined as

$$T_n = \sum_{i=1}^{n} \text{sign}(D_i) R_i(|\mathbf{D}|) \tag{3.36}$$

where the sign function is defined as

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \tag{3.37}$$

- Here, $R_i(|\mathbf{D}|)$ is the rank of the i^{th} element from the vector $|\mathbf{D}| = (|D_1|, |D_2|, \ldots, |D_n|)$.
- Intuitively, the Wilcoxon signed rank statistic is measuring whether or not large values of $|D_i|$ tend to be associated with positive vs. negative values of D_i.

Discuss some of these in class

Exercise 3.4. Suppose we had data $(-2, 1, -1/2, 3/2, 3)$. What would be the value of the Wilcoxon signed rank statistic?

Exercise 3.5. Under the assumptions of model (3.25), what is the density function of $|D_i|$ and $-|D_i|$?

Exercise 3.6. Under the assumptions of model (3.25) and assuming that $\theta = 0$, show that the expectation of the Wilcoxon signed-rank statistic is 0.

3.3.2.1 Asymptotic Distribution

- As mentioned in the above exercise, the expectation of T_n under H_0 is zero.
- It can be shown that the variance under the null hypothesis is

$$\text{Var}_{H_0}(T_n) = \frac{n(2n+1)(n+1)}{6}$$

- Similar, to the large-sample approximation we used for the WRS test, we have the following asymptotic result for the Wilcoxon signed-rank test

$$\frac{T_n}{\sqrt{\text{Var}_{H_0}(T_n)}} \longrightarrow \text{Normal}(0,1) \quad \text{as } n \longrightarrow \infty \tag{3.38}$$

- Because the variance of T is dominated by the term $n^3/3$ for very large n, we could also say that under H_0 that

$$\frac{T_n}{\sqrt{n^3/3}} \longrightarrow \text{Normal}(0,1) \quad \text{as } n \longrightarrow \infty \tag{3.39}$$

  In other words, we can say that T_n has an approximately Normal$(0, n^3/3)$ for large n.

3.3.2.2 Exact Distribution

- The exact distribution of the Wilcoxon signed rank statistic T_n is somewhat more complicated than the exact distribution of the WRS test statistic. Nevertheless, there exists functions in **R** for working with this exact distribution.

3.3.3 Using R to Perform the Sign and Wilcoxon Tests

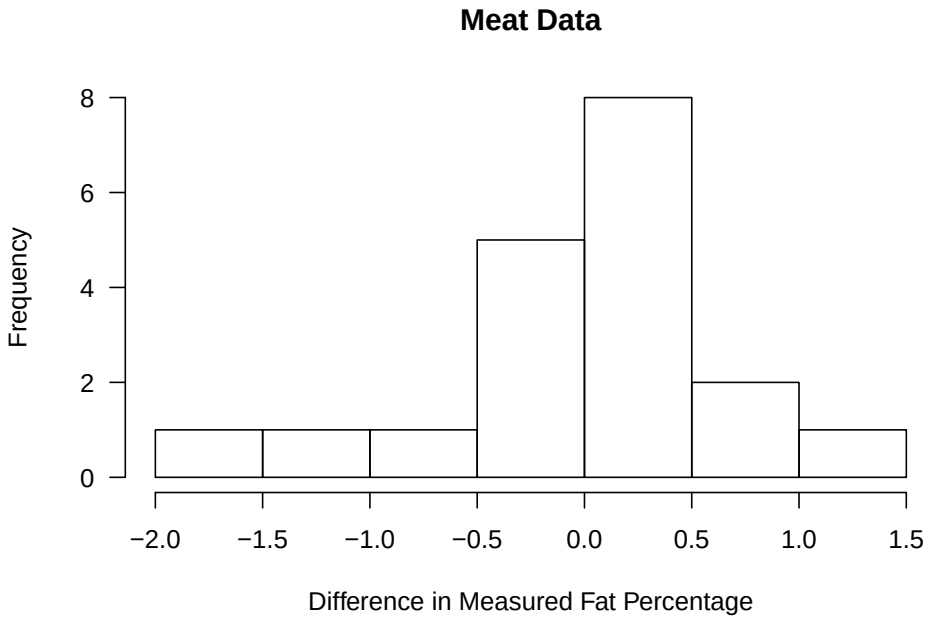
- Let's first look at the **Meat** data from the **PairedData R** package.
- This data set contains 20 observations with measures of fat percentage using different measuring techniques.

```
library(PairedData, quietly=TRUE, warn.conflicts=FALSE) ## loading PairedData package
data(Meat)  ## loading Meat data
head(Meat)
```

```
##   AOAC Babcock     MeatType
## 1 22.0    22.3       Wiener
## 2 22.1    21.8       Wiener
## 3 22.1    22.4       Wiener
## 4 22.2    22.5       Wiener
## 5 24.6    24.9   ChoppedHam
## 6 25.3    25.6 ChooppedPork
```

- Define the differences D_i as the **Babcock** measurements minus the **AOAC** measures. We will drop the single observation that equals zero.

```
DD <- Meat[,2] - Meat[,1]
DD <- DD[DD!=0]
hist(DD, main="Meat Data", xlab="Difference in Measured Fat Percentage", las=1)
```




```
summary(DD)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.60000 -0.25000  0.30000  0.04211  0.40000  1.10000
```

The Sign Test in R

- Let's first test the hypothesis $H_0 : \theta = 0$ vs. $H_A : \theta \neq 0$ using the two-sided sign test. This can be done using the **binom.test** function

```
binom.test(sum(DD > 0), n = length(DD), p=0.5)$p.value
```

```
## [1] 0.6476059
```

Wilcoxon Signed Rank Test in R

- You can actually use the function **wilcox.test** to perform the Wilcoxon signed rank test in addition to the Wilcoxon rank sum test. To perform the Wilcoxon signed rank test in **R**, you just need to enter data for the **x** argument and leave the **y** argument empty.

```
wilcox.test(x=DD)
```

```
## Warning in wilcox.test.default(x = DD): cannot compute exact p-value with
## ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  DD
```

```
## V = 118.5, p-value = 0.3534
## alternative hypothesis: true location is not equal to 0
```

- You will note that the p-value for the Wilcoxon signed rank test is lower than that of the sign test. In general, the Wilcoxon signed rank test is somewhat more "sensitive" than the sign test meaning that it will have a greater tendency to reject H_0 for small deviations from H_0.
- We can explore this sensitivity comparison with a small simulation study. We will consider a scenario where $D_i = 0.4 + \varepsilon_i$ with ε_i having a t distribution with 3 degrees of freedom.

```
set.seed(1327)
n.reps <- 500  ## number of simulation replications
samp.size <- 50  ## the sample size
wilcox.reject <- rep(0, n.reps)
sign.reject <- rep(0, n.reps)
for(k in 1:n.reps) {
    dsim <- .4 + rt(samp.size, df=3)
    wilcox.reject[k] <- ifelse(wilcox.test(x=dsim)$p.value < 0.05, 1, 0)
    sign.reject[k] <- ifelse(binom.test(sum(dsim > 0),
                                        n=samp.size, p=0.5)$p.value < 0.05, 1, 0)
}
mean(wilcox.reject)  ## proportion of times Wilcoxon signed rank rejected H0
```

```
## [1] 0.614
```

```
mean(sign.reject)  ## proportion of times Wilcoxon signed rank rejected H0
```

```
## [1] 0.488
```

3.4 Power and Comparisons with Parametric Tests

3.4.1 The Power Function of a Test

- The **power** of a test is the probability that a test rejects the null hypothesis when the alternative hypothesis is true.
- The alternative hypothesis H_A is usually characterized by a large range of values of the parameter of interest. For example, $H_A : \theta > 0$ or $H_A : \theta \neq 0$.
- For this reason, it is better to think of power as a function that varies across the range of the alternative hypothesis.
- To be more precise, we will define the power function as a function of some parameter θ where the null hypothesis corresponds to $\theta = \theta_0$ and

the alternative hypothesis represents a range of alternative values of θ.

- The power function $\gamma_n(\cdot)$ of a testing procedure is defined as

$$\gamma_n(\delta) = P_{\theta=\delta}\{\text{reject } H_0\} \qquad \text{for } \delta \in H_A.$$

- The notation $P_{\theta=\delta}\{\text{reject } H_0\}$ means that we are computing this probability under the assumption that the parameter of interest θ equals δ.

The Approximate Power Function of the Sign Test

- Let us consider the sign test for testing $H_0 : \theta = 0$ vs. $\theta > 0$.
- The sign test is based on the value of the sign statistic S_n.
- Recalling (3.30), we know that $S_n \sim \text{Binomial}(n, p(\theta))$. Hence,

$$\sqrt{n}(\tfrac{S_n}{n} - p(\theta)) \longrightarrow \text{Normal}\Big(0, p(\theta)(1-p(\theta))\Big) \quad \text{as } n \longrightarrow \infty \tag{3.40}$$

- The sign test will reject H_0 when $S_n \geq c_{\alpha,n}$ where the constant $c_{\alpha,n}$ is chosen so that $P_{H_0}(S_n \geq c_{\alpha,n}) = \alpha$. Using the large-sample approximation (3.40), you can show that

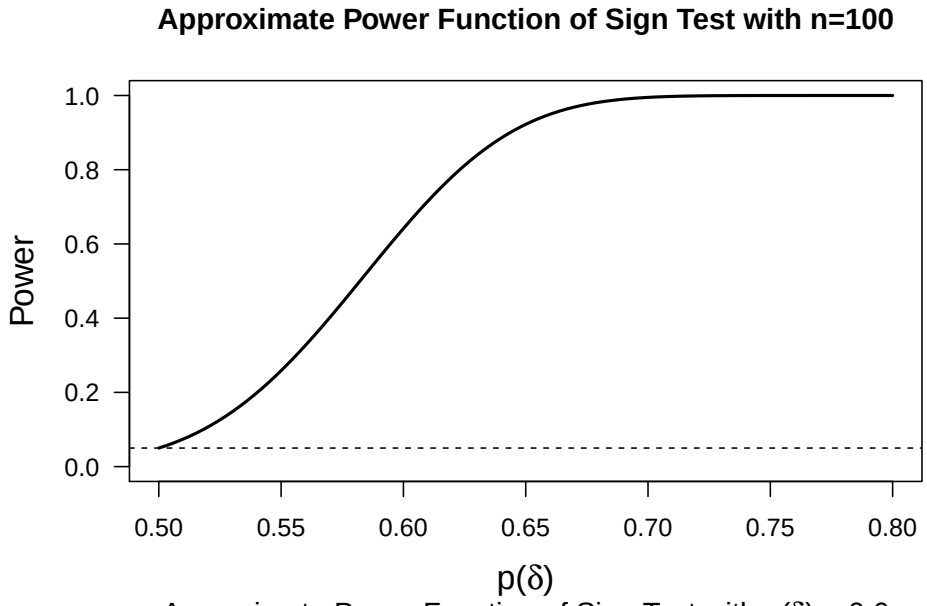
$$c_{\alpha,n} = \frac{n + \sqrt{n} z_{1-\alpha}}{2}, \tag{3.41}$$

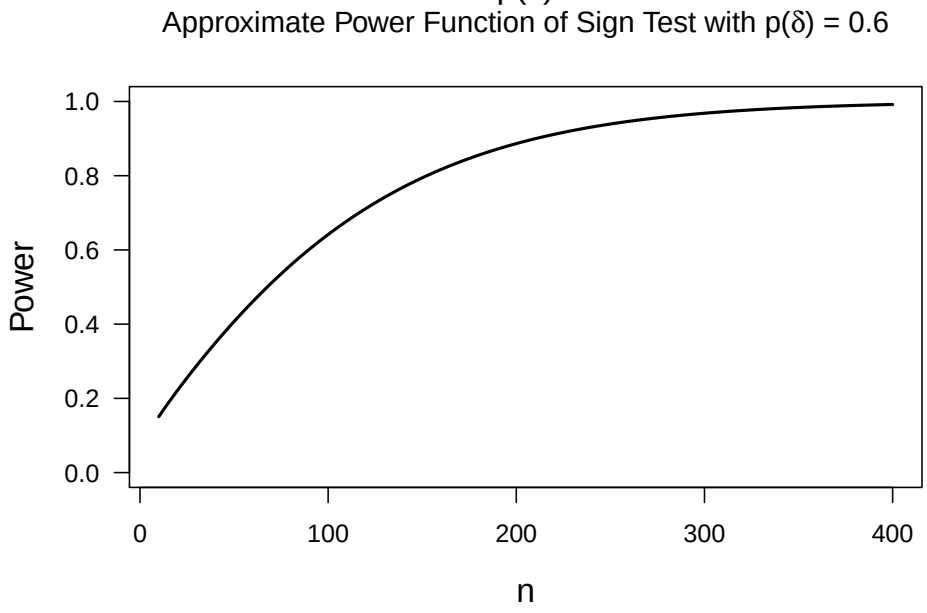
where $z_{1-\alpha}$ denotes the upper $1-\alpha$ quantile of the standard normal distribution. In other words, $\Phi(z_{1-\alpha}) = 1 - \alpha$.

- Also, when using large-sample approximation (3.40), the power of this test to detect a value of $\theta = \delta$ is given by

$$\begin{aligned}
\gamma_n(\delta) &= P_{\theta=\delta}\{S_n \geq c_{\alpha,n}\} = P_{\theta=\delta}\left\{ \frac{\sqrt{n}(S_n/n - p(\delta))}{\sqrt{p(\delta)(1-p(\delta))}} \geq \frac{\sqrt{n}(c_{\alpha,n}/n - p(\delta))}{\sqrt{p(\delta)(1-p(\delta))}} \right\} \\
&= 1 - \Phi\left(\frac{\sqrt{n}(c_{\alpha,n}/n - p(\delta))}{\sqrt{p(\delta)(1-p(\delta))}} \right) \\
&= 1 - \Phi\left(\frac{z_{1-\alpha}}{2\sqrt{p(\delta)(1-p(\delta))}} - \frac{\sqrt{n}(p(\delta) - 1/2)}{\sqrt{p(\delta)(1-p(\delta))}} \right)
\end{aligned} \tag{3.42}$$

- Notice that the power of the test depends more directly on the term $p(\delta) = P_{\theta=\delta}(D_i > 0)$. Recall from Section 3.3.1 that $p(\delta) = 1 - F_\epsilon(\delta)$, where F_ϵ is the distribution function of ε_i in the model $D_i = \theta + \varepsilon_i$.
- So, in any power or sample size calculation, it would be more sensible to think about plausible values for $p(\delta)$ rather than δ itself. Plus, $p(\delta)$ has the direct interpretation $p(\delta) = P_{\theta=\delta}(D_i > 0)$.





Exercise 3.7: Derive the formula for $c_{\alpha,n}$ shown in (3.41).

3.4.2 Power Comparisons and Asymptotic Relative Efficiency

- Notice that for the sign statistic power function shown in (3.42), we have that

$$\lim_{n\longrightarrow\infty} \gamma_n(\delta) = \begin{cases} \alpha & \text{if } \delta = 0 \\ 1 & \text{if } \delta > 0 \end{cases} \tag{3.43}$$

- The above type of limit for the power function is will be true for most "reasonable" tests.
- Indeed, a test whose power function satisfies (3.43) is typically called a **consistent** tests.

- If nearly all reasonable tests are consistent, then how can we compare tests with respect to their power?
- One approach is to use simulations to compare power for several plausible alternatives. While this can be useful for a specific application, it limits our ability to make more general statements about power comparisons.
- Another approach might be to determine for which values of (δ, n) one test has greater power than another. However, this could be tough to interpret (no test will be uniformly more powerful for all distributions) or even difficult to compute.
- One way to think about power is to think about the **relative efficiency** of two testing procedures. The efficiency of a test in this context is the sample size required to achieve a certain level of power.

- To find the asymptotic relative efficiency, we first need to derive the asymptotic power function.
- For our hypothesis $H_0 : \theta = \theta_0$ vs. $H_A : \theta > \theta_0$, this is defined as

$$\tilde{\gamma}(\delta) = \lim_{n\longrightarrow\infty} \gamma_n(\theta_0 + \delta/\sqrt{n})$$

- Considering the sequence of "local alternatives" $\theta_n = \theta_0 + \delta/\sqrt{n}$, we avoid the problem of the power always converging to 1.
- It can be shown that

$$\tilde{\gamma}(\delta) = 1 - \Phi\left(z_{1-\alpha} - \delta\frac{\mu'(\theta_0)}{\sigma(\theta_0)}\right) \tag{3.44}$$

as long as we can find functions $\mu(\cdot)$ and $\sigma(\cdot)$ such that

$$\frac{\sqrt{n}(V_n - \mu(\theta_n))}{\sigma(\theta_n)} \longrightarrow \text{Normal}(0,1) \tag{3.45}$$

where the test of $H_0 : \theta = \theta_0$ vs. $H_A : \theta > \theta_0$ is based on the test statistic V_n with rejection of H_0 occurring whenever $V_n \geq c_{\alpha,n}$. Statement (3.45) asssumes that the distribution of V_n is governed by θ_n for each n.

- The ratio $e(\theta_0) = \mu'(\theta_0)/\sigma(\theta_0)$ is the **asymptotic efficiency** of the test.
- When comparing two tests with efficiency $e_1(\theta_0)$ and $e_2(\theta_0)$, the asymptotic relative efficiency of test 1 vs. test 2 is defined as

$$ARE_{12}(\theta_0) = \Big(\frac{e_1(\theta_0)}{e_2(\theta_0)}\Big)^2 \tag{3.46}$$

Interpretation of Asymptotic Efficiency of Tests

- Roughly speaking, the asymptotic relative efficiency $ARE_{12}(\theta_0)$ approximately equals n_2/n_1 where n_1 is the sample size needed for test 1 to achieve power β and n_2 is the sample size needed for test 2 to achieve power β. This is true for an arbitrary β.
- To further justify this interpretation notice that, for large n, we should have

$$c_{\alpha,n} \approx \mu(\theta_0) + \frac{\sigma(\theta_0) z_{1-\alpha}}{\sqrt{n}} \tag{3.47}$$

(This approximation for $c_{\alpha,n}$ comes from the asymptotic statement in (3.45))

- Now, consider the power for detecting $H_A : \theta = \theta_A$ (where we will assume that θ_A is "close" to θ_0). Using (3.45), the approximate power in this setting is

$$\begin{aligned} P_{\theta_A}\Big(V_n \geq c_{\alpha,n}\Big) &= P_{\theta_A}\left(\frac{\sqrt{n}(V_n - \mu(\theta_A))}{\sigma(\theta_A)} \geq \frac{\sqrt{n}(c_{\alpha,n} - \mu(\theta_A))}{\sigma(\theta_A)}\right) \approx 1 - \Phi\left(\frac{\sqrt{n}(c_{\alpha,n} - \mu(\theta_A))}{\sigma(\theta_A)}\right) \\ &= 1 - \Phi\left(\frac{\sqrt{n}(\mu(\theta_0) - \mu(\theta_A))}{\sigma(\theta_A)} + \frac{z_{1-\alpha}\sigma(\theta_0)}{\sigma(\theta_A)}\right) \end{aligned} \tag{3.48}$$

- Hence, if we want to achieve a power level of β for the alternative $H_A : \theta = \theta_A$, we need the corresponding sample size $n_\beta(\theta_A)$ to satisfy

$$\frac{\sqrt{n_\beta(\theta_A)}(\mu(\theta_0) - \mu(\theta_A))}{\sigma(\theta_A)} + \frac{z_{1-\alpha}\sigma(\theta_0)}{\sigma(\theta_A)} = z_{1-\beta} \tag{3.49}$$

which reduces to

$$n_\beta(\theta_A) = \left(\frac{z_{1-\beta}\sigma(\theta_A) - z_{1-\alpha}\sigma(\theta_0)}{\mu(\theta_0) - \mu(\theta_A)} \right)^2 \approx \left(\frac{[z_{1-\beta} - z_{1-\alpha}]\sigma(\theta_0)}{(\theta_A - \theta_0)\mu'(\theta_0)} \right)^2 \tag{3.50}$$

- So, if we were comparing two testing procedures and we computed the approximate sample sizes $n^1_\beta(\theta_A)$ and $n^2_\beta(\theta_A)$ needed to reach β power for the alternative $H_A : \theta = \theta_A$, the sample size ratio (using approximation (3.50)) would be

$$\frac{n^2_\beta(\theta_A)}{n^1_\beta(\theta_A)} = \left(\frac{\mu_1'(\theta_0)\sigma_2(\theta_0)}{\mu_2'(\theta_0)\sigma_1(\theta_0)} \right)^2 = \text{ARE}_{12}(\theta_0) \tag{3.51}$$

- Notice that $\text{ARE}_{12}(\theta_0) > 1$ indicates that the test 1 is better than test 2 because the sample size required for test 1 would be less than the sample size required for test 2.
- It is also worth noting that our justification for the interpretation of $\text{ARE}_{12}(\theta_0)$ was not very rigorous or precise, but it is possible to make a more rigorous statement. See, for example, Chapter 13 of Lehmann and Romano (2006) for a more rigorous treatment of relative efficiency.
- In Lehmann and Romano (2006), they have a result that states (under appropriate assumptions) that

$$\lim_{\theta \downarrow \theta_0} \frac{N_2(\theta)}{N_1(\theta)} = ARE_{12}(\theta_0) \tag{3.52}$$

where $N_1(\theta)$ and $N_2(\theta)$ are the sample sizes required to have power β against alternative θ.

3.4.3 Efficiency Examples

The Sign Test

- Let us return to the example of the sign statistic S_n and its use in testing the hypothesis $H_0 : \theta = 0$ vs. $H_A : \theta > 0$.
- Notice that the sign test rejects $H_0 : \theta = 0$ for $V_n > c_{\alpha,n}$ where $V_n = S_n/n$ and S_n is the sign statistic.
- When V_n is defined this way (3.45) is satisfied when $\mu(\theta) = p(\theta)$ and $\sigma(\theta) = \sqrt{p(\theta)(1 - p(\theta))}$ where $p(\theta) = 1 - F_\epsilon(-\theta)$.

- Thus, the efficiency of the sign test for testing $H_0 : \theta = 0$ vs. $H_A : \theta > 0$ is
$$\frac{\mu'(0)}{\sigma(0)} = \frac{p'(0)}{\sqrt{p(0)(1-p(0))}} = 2f_\epsilon(0) \tag{3.53}$$
where $f_\epsilon(t) = F'_\epsilon(t)$.

The One-Sample t-test

- Assume that we have data $D_1, \ldots, D_n$ generated under the same assumption as in our discussion of the sign test and the Wilcoxon signed-rank test. That is,
$$D_i = \theta + \varepsilon_i, \tag{3.54}$$
where ε_i are assumed to have median 0 with ε_i having p.d.f. f_ε
- The one-sample t-test will reject $H_0 : \theta = 0$ whenever $V_n > c_{\alpha,n}$, where V_n is defined to be
$$V_n = \frac{\bar{D}}{\hat{\sigma}} \tag{3.55}$$
- Note that (3.45) will apply if we choose
$$\begin{aligned} \mu(\theta) &= E_\theta(D_i) = \theta \\ \sigma(\theta) &= \sqrt{\mathrm{Var}_\theta(D_i)} = \sqrt{\mathrm{Var}(\varepsilon_i)} = \sigma_\epsilon \end{aligned} \tag{3.56}$$
- These choices of $\mu(\theta)$ and $\sigma(\theta)$ work because
$$\begin{aligned} \frac{\sqrt{n}(V_n - \mu(\theta_n))}{\sigma(\theta_n)} &= \frac{\sqrt{n}(\bar{D} - \theta_n)}{\sigma_e} + \sqrt{n}\theta_n\Big(\frac{1}{\hat{\sigma}} - \frac{1}{\sigma_e}\Big) \\ &= \frac{\sqrt{n}(\bar{D} - \theta_n)}{\sigma_e} + \delta\Big(\frac{1}{\hat{\sigma}} - \frac{1}{\sigma_e}\Big) \\ &\longrightarrow \mathrm{Normal}(0,1) \end{aligned} \tag{3.57}$$
- So, the efficiency of the one-sample t-test is given by
$$\frac{\mu'(0)}{\sigma(0)} = \frac{1}{\sigma_e}$$

The Wilcoxon Rank Sum Test

- Using the close relation between the WRS test statistic and the Mann-Whitney statistic, the WRS test can be represented as rejecting H_0 when $V_N \geq c_{\alpha,N}$ where V_N is
$$V_N = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} I(X_i \geq Y_j) \tag{3.58}$$
and $N = n + m$.

- The power of the WRS test is usually analyzed in the context of the "shift alternative". Namely, we are assuming that $F_X(t) = F_Y(t-\theta)$ and test $H_0 : \theta = 0$ vs. $H_A : \theta > 0$.
- The natural choice for $\mu(\theta)$ is the expectation of V_N when θ is the true shift parameter.
- So, let $\mu(\theta) = P_\theta(X_i \geq Y_j)$. This can be written in terms of F_Y and f_Y:

$$\begin{aligned} \mu(\theta) &= \int_{-\infty}^{\infty} P_\theta(X_i \geq Y_j | Y_j = t) f_Y(t) dt = \int_{-\infty}^{\infty} P_\theta(X_i \geq t) f_Y(t) dt \\ &= \int_{-\infty}^{\infty} \{1 - F_X(t)\} f_Y(t) dt = 1 - \int_{-\infty}^{\infty} F_Y(t-\theta) f_Y(t) dt \end{aligned} \quad (3.59)$$

- You can show that (3.45) holds (see e.g, Chapter 14 of Van der Vaart (2000)) if you choose $\sigma^2(\theta)$ to be

$$\sigma^2(\theta) = \frac{1}{1-\lambda} \text{Var}\{F_Y(X_i)\} + \frac{1}{\lambda} \text{Var}\{F_Y(Y_i - \theta)\} \quad (3.60)$$

  Here, $n/(m+n) \longrightarrow \lambda$.
- Thus, the efficiency of testing $H_0 : \theta = 0$ for the WRS test is

$$e(0) = \frac{\mu'(0)}{\sigma(0)} = \frac{\int_{-\infty}^{\infty} f^2(t) dt}{\sigma(0)} \quad (3.61)$$

3.4.4 Efficiency Comparisons for Several Distributions

Sign Test vs. One-Sample t-test

- Comparisons of the Efficiency of the sign and one-sample t-test only require us to find $f_\epsilon(0)$ and σ_ϵ^2 for different assumptions about the residual density f_ϵ.
- For the Logistic(0,1) distribution, $f_\epsilon(0) = 1/4$ and the standard deviation is $\pi/\sqrt{3}$. Hence, the asymptotic relative efficiency of the sign test vs. the one-sample t-test would be $(\pi/2\sqrt{3})^2$.
- The relative efficiencies for the sign vs. t-test for other distributions are shown below

Distribution	Efficiency	
		(3.62)
Normal$(0,1)$	$2/\pi$	(3.63)
Logistic$(0,1)$	$\pi^2/12$	(3.64)
Laplace$(0,1)$	2	(3.65)
Uniform$(-1,1)$	$1/3$	(3.66)
t-dist$_\nu$	$[4(\nu/(\nu-2))\Gamma^2\{(\nu+1)/2\}]/[\Gamma^2(\nu/2)\nu\pi]$	(3.67)

WRS Test vs. Two-Sample t-test

- The relative efficiencies for the WRS test vs. the two-sample t-test for several distributions are shown below.

Distribution	Efficiency	
Normal$(0, 1)$	$3/\pi$	(3.69)
Logistic$(0, 1)$	$\pi^2/9$	(3.70)
Laplace$(0, 1)$	$3/2$	(3.71)
Uniform$(-1, 1)$	1	(3.72)
t-dist$_3$	1.24	(3.73)
t-dist$_5$	1.90	(3.74)
		(3.75)

3.4.5 A Power "Contest"

- To compare power across for specific sample sizes, effect sizes, and distributional assumptions, a simulation study can be more helpful than statements about asymptotic relative efficiency.
- Below shows the results of a simulation study in **R** which compares power for the one-sample testing problem.
- This simulation study compares the sign test, Wilcoxon signed rank test, and the one-sample t-test.
- It is assumed that $n = 200$ and that responses D_i are generated from the following model:

$$D_i = 0.2 + \varepsilon_i \tag{3.76}$$

- Three choices for the distribution of ε_i were considered:
  - $\varepsilon_i \sim$ Logistic$(0, 1)$
  - $\varepsilon_i \sim$ Normal$(0, 1)$
  - $\varepsilon_i \sim$ Uniform$(-3/2, 3/2)$
- The **R** code and simulation results are shown below.

```
set.seed(148930)
theta <- 0.2
n <- 200
nreps <- 500
RejectSign <- RejectWilcoxonSign <- RejectT <- matrix(NA, nrow=nreps, ncol=4)
for(k in 1:nreps) {
```

```
  xx <- theta + rlogis(n)
  yy <- theta + rnorm(n)
  zz <- theta + runif(n, min=-3/2, max=3/2)
  ww <- theta + (rexp(n, rate=1) - rexp(n, rate=1))/sqrt(2)

  RejectSign[k,1] <- ifelse(binom.test(x=sum(xx > 0), n=n, p=0.5)$p.value < 0.05, 1, 0)
  RejectWilcoxonSign[k,1] <- ifelse(wilcox.test(xx)$p.value < 0.05, 1, 0)
  RejectT[k,1] <- ifelse(t.test(xx)$p.value < 0.05, 1, 0)

  RejectSign[k,2] <- ifelse(binom.test(x=sum(yy > 0), n=n, p=0.5)$p.value < 0.05, 1, 0)
  RejectWilcoxonSign[k,2] <- ifelse(wilcox.test(yy)$p.value < 0.05, 1, 0)
  RejectT[k,2] <- ifelse(t.test(yy)$p.value < 0.05, 1,0)

  RejectSign[k,3] <- ifelse(binom.test(x=sum(zz > 0), n=n, p=0.5)$p.value < 0.05, 1, 0)
  RejectWilcoxonSign[k,3] <- ifelse(wilcox.test(zz)$p.value < 0.05, 1, 0)
  RejectT[k,3] <- ifelse(t.test(zz)$p.value < 0.05,1,0)

  RejectSign[k,4] <- ifelse(binom.test(x=sum(ww > 0), n=n, p=0.5)$p.value < 0.05, 1, 0)
  RejectWilcoxonSign[k,4] <- ifelse(wilcox.test(ww)$p.value < 0.05, 1, 0)
  RejectT[k,4] <- ifelse(t.test(ww)$p.value < 0.05, 1, 0)
}

power.results <- data.frame(Distribution=c("Logistic", "Normal", "Uniform", "Laplace")
                 SignTest=colMeans(RejectSign), WilcoxonSign=colMeans(RejectWilcoxonSig
                 tTest=colMeans(RejectT))
```

Distribution	SignTest	WilcoxonSign	tTest
Logistic	0.25	0.37	0.34
Normal	0.59	0.77	0.81
Uniform	0.44	0.87	0.90
Laplace	0.93	0.92	0.81

Table 3.1: Estimated power for three one-sample tests and three distributions. 500 simulation replications were used.

3.5 Linear Rank Statistics in General

3.5.1 Definition

- The Wilcoxon rank sum statistic is an example of a statistic from a more general class of rank statistics.
- This is the class of **linear rank statistics**.

- Suppose we have observations $\mathbf{Z} = (Z_1, \ldots, Z_N)$. A linear rank statistic is a statistic T_N that can be expressed as

$$T_N = \sum_{i=1}^{N} c_{iN} a_N\big(R_i(\mathbf{Z})\big) \tag{3.77}$$

- The terms $c_{1N}, \ldots, c_{NN}$ are usually called **coefficients**. These are fixed numbers and are not random variables.
- The terms $a_N(R_i(\mathbf{Z}))$ are commonly referred to as **scores**.
- Typically, the scores are generated from a given function ψ in the following way

$$a_N(i) = \psi\Big(\frac{i}{N+1}\Big) \tag{3.78}$$

Example: WRS statistic

- For the Wilcoxon rank sum test, we separated the data $\mathbf{Z} = (Z_1, \ldots, Z_N)$ into two groups.
- The first n observations were from group 1 while the last m observations were from group 2.
- The WRS statistic was then defined as

$$W = \sum_{i=1}^{n} R_i(\mathbf{Z}) \tag{3.79}$$

- In this case, the WRS statistic can be expressed in the form (3.77) if we choose the coefficients to be the following

$$c_{iN} = \begin{cases} 1 & \text{if } i \leq n \\ 0 & \text{if } i > n \end{cases} \tag{3.80}$$

and we choose the scores to be

$$a_N(i) = i \tag{3.81}$$

3.5.2 Properties of Linear Rank Statistics

- The expected value of the linear rank statistic (if the distribution of the Z_i is continuous) is

$$E(T_N) = N\bar{c}_N\bar{a}_N, \tag{3.82}$$

where $\bar{c}_N = \frac{1}{N}\sum_{j=1}^{N} c_{jN}$ and $\bar{a}_N = \frac{1}{N}\sum_{j=1}^{N} a_N(j)$

- The formula (3.82) for the expectation only uses the fact that $R_i(\mathbf{Z})$ has a discrete uniform distribution. So,

$$E\{a_N(R_i(\mathbf{Z})\} = \sum_{j=1}^{N} a_N(j) P\{R_i(\mathbf{Z}) = j\} = \sum_{j=1}^{N} \frac{a_N(j)}{N} = \bar{a}_N \tag{3.83}$$

  Using this, we can then see that

$$E(T_N) = \sum_{j=1}^{N} c_{jN} E\{a_N(R_i(\mathbf{Z}))\} = \sum_{j=1}^{N} c_{jN} \bar{a}_N = N \bar{c}_N \bar{a}_N \tag{3.84}$$

- A similar argument can show that the variance of T_N is

$$\mathrm{Var}(T_N) = \frac{N^2}{n-1} \sigma_a^2 \sigma_c^2, \tag{3.85}$$

  where $\sigma_c^2 = \frac{1}{N} \sum_{j=1}^{N} (c_{jN} - \bar{c}_N)^2$ and $\sigma_a^2 = \frac{1}{N} \sum_{j=1}^{N} (a_N(j) - \bar{a}_N)^2$

- To perform hypothesis testing when using a general linear rank statistics, working with the exact distribution or performing permutation tests can often be computationally demanding.
- Using a large-sample approximation is often easier.
- As long as a few conditions for the coefficients and scores are satisfied, one can state the following

$$\frac{T_N - E(T_N)}{\sqrt{\mathrm{Var}(T_N)}} \longrightarrow \mathrm{Normal}(0, 1), \tag{3.86}$$

  where, as we showed, both $E(T_N)$ and $\mathrm{Var}(T_N)$ both have closed-form expressions for an arbitrary linear rank statistic.

3.5.3 Other Examples of Linear Rank Statistics

3.5.3.1 The van der Waerden statistic and the normal scores test

- Van der Waerden's rank statistic is used for two-sample problems where the first n observations come from group 1 while the last m observations come from group 2.
- Van der Waerden's rank statistic VW_N is defined as

$$VW_N = \sum_{j=1}^{n} \Phi^{-1}\left(\frac{\mathbf{R}_i(\mathbf{Z})}{N+1}\right) \tag{3.87}$$

- The function Φ^{-1} denotes the inverse of the cumulative distribution function of a standard Normal random variable.
- The statistic VW_N is a linear rank statistic with coefficients

$$c_{iN} = \begin{cases} 1 & \text{if } i \leq n \\ 0 & \text{if } i > n \end{cases} \tag{3.88}$$

  and scores determined by

$$a_N(i) = \Phi^{-1}\Big(\frac{i}{N+1}\Big) \tag{3.89}$$

- A test based on van der Waerden's statistic is often referred to as the **normal scores test**.
- The normal scores test is often suggested as an attractive test when the underlying data has an approximately normal distribution.
- If you plot a histogram of the van der Waerden scores $a_N(i)$ it should look roughly like a Gaussian distribution (if there are not too many ties).

3.5.3.2 The median test

- The median test is also a two-sample rank test.
- While the Wilcoxon rank sum test looks at the average rank within group 1, the median test instead looks at how many of the ranks from group 1 are less than the median rank (which should equal $(N+1)/2$).
- The test statistic M_N for the median test is defined as

$$M_N = \sum_{i=1}^{n} I\Big(R_i(\mathbf{Z}) \leq \frac{N+1}{2}\Big) \tag{3.90}$$

  because $(N+1)/2$ will be the median rank.
- This is a linear rank statistic with coefficients

$$c_{iN} = \begin{cases} 1 & \text{if } i \leq n \\ 0 & \text{if } i > n \end{cases} \tag{3.91}$$

  and scores

$$a_N(i) = \begin{cases} 1 & \text{if } i \leq (N+1)/2 \\ 0 & \text{if } i > (N+1)/2 \end{cases} \tag{3.92}$$

- The median test could be used to test whether or not observations from group 1 tend to be smaller than those from group 2.

3.5.4 Choosing the scores $a_N(i)$

- The rank tests we have discussed so far are nonparametric in the sense that their null distribution does not depend on any particular parametric assumptions about the distributions from which the observations arise.
- For power calculations, we often think of some parameter or "effect size" modifying the base distribution in some way.
- For example, we often think of the shift alternative $F_X(t) = F_Y(t - \theta)$ in the two-sample problem.

- In parametric statistics, when testing $H_0 : \theta = 0$ the most powerful test of $H_0 : \theta = \theta_0$ vs. $H_A : \theta = \theta_A$ is based on rejecting H_0 whenever the likelihood ratio is large enough:

$$\text{Reject } H_0 \text{ if:} \quad \frac{p_{\theta_A}(\mathbf{z})}{p_{\theta_0}(\mathbf{z})} \geq c_{\alpha,n} \tag{3.93}$$

  This is the Neyman-Pearson Lemma.

- The same property is true if we are considering tests based on ranks. The most powerful test for testing $H_0 : \theta = \theta_0$ vs. $H_A : \theta = \theta_A$ is based on

$$\text{Reject } H_0 \text{ if:} \quad \frac{P_{\theta_A}\Big(R_1(\mathbf{Z}), \ldots, R_N(\mathbf{Z})\Big)}{P_{\theta_0}\Big(R_1(\mathbf{Z}), \ldots, R_N(\mathbf{Z})\Big)} \geq c_{\alpha,n} \tag{3.94}$$

- The main difference between (3.93) and (3.93) is that the distribution $P_{\theta_A}\Big(R_1(\mathbf{Z}), \ldots, R_N(\mathbf{Z})\Big)$ is unknown unless we are willing to make certain distributional assumptions.
- Nevertheless, we can approximate this probability if θ_A is a location parameter "close" to θ_0

$$P_{\theta_A}\Big(R_1(\mathbf{Z}), \ldots, R_N(\mathbf{Z})\Big) \approx P_{\theta_0}\Big(R_1(\mathbf{Z}), \ldots, R_N(\mathbf{Z})\Big) + \frac{\theta_A}{N!} \sum_{i=1}^{N} c_{iN} E\left\{ \frac{\partial \log f(Z_{(i)})}{\partial Z} \right\} \tag{3.95}$$

where $Z_{(i)}$ denotes the i^{th} order statistic. See, for example, Chapter 13 of Van der Vaart (2000) for more details on the derivation of this approximation.

- So, large values of the linear rank statistic $T_N = \sum_{i=1}^{N} c_{iN} a_N(i)$ will approximately correspond to large values of $P_{\theta_A}\Big(R_1(\mathbf{Z}), \ldots, R_N(\mathbf{Z})\Big)$ if we

choose the scores to be

$$a_N(i) = E\left\{ \frac{\partial \log f(Z_{(i)})}{\partial Z} \right\} \tag{3.96}$$

- Linear rank statistics with scores generated this way are usually called **locally most powerful** rank test.

- The best choice of the scores will depend on what we assume about the density f.
- For example, if we assume that $f(z)$ is Normal$(0, 1)$, then

$$\frac{\partial \log f(z)}{\partial z} = -z \tag{3.97}$$

- The approximate expectation of the order statistics from a Normal$(0, 1)$ distribution are

$$E\{Z_{(i)}\} \approx \Phi^{-1}\left(\frac{i}{N+1} \right) \tag{3.98}$$

  This implies that the van der Waerden's scores are approximately optimal if we assume the distribution of the Z_i is Normal.
- This can also be worked out for other choices of $f(z)$.
- If $f(z)$ is a Logistic distribution, the optimal scores correspond to the Wilcoxon rank sum test statistic.
- If $f(z)$ is Laplace (meaning that $f(z) = \frac{1}{2}e^{-|z|}$), then the optimal scores correspond to the median test.

3.6 Additional Reading

- Additional reading which covers the material discussed in this chapter includes:
  - Chapters 3-4 from Hollander et al. (2013)

Chapter 4

Rank Tests for Multiple Groups

- We can roughly think of the tests discussed in Chapter 3 as being related to the parametric tests shown in the table below.

Parametric Test	**Nonparametric Tests**
One-Sample t-test	Wilcoxon Signed Rank/Sign Test
Two-Sample t-test	Wilcoxon Rank Sum/Normal Scores/Median Test

- The **Kruskal-Wallis** test can be though of as the nonparametric analogue of one-way analysis of variance (ANOVA).
- For $K \geq 3$ groups, one-way ANOVA considers the analysis of data arising from the following model

$$Y_{kj} = \mu_k + \varepsilon_{kj}, \qquad j = 1, \ldots, n_k; k = 1, \ldots, K \tag{4.1}$$

  where it is often assumed that $\varepsilon_{kj} \sim \text{Normal}(0, \sigma^2)$.
- Usually, the one-way ANOVA hypothesis of interest is something like

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_K \tag{4.2}$$

  which is sometimes referred to as the homogeneity hypothesis.
- A test of the hypothesis (4.2) is based on decomposing the observed vari-

ation in the responses Y_{kj}:

$$\begin{aligned}\underbrace{\sum_{k=1}^{K}\sum_{j=1}^{n_k}(Y_{kj}-\bar{Y}_{\cdot\cdot})^2}_{SST} &= \sum_{k=1}^{K}\sum_{j=1}^{n_k}(\bar{Y}_{k\cdot}-\bar{Y}_{\cdot\cdot})^2+\sum_{k=1}^{K}\sum_{j=1}^{n_k}(Y_{kj}-\bar{Y}_{k\cdot})^2 \\ &= \underbrace{\sum_{k=1}^{K} n_k(\bar{Y}_{k\cdot}-\bar{Y}_{\cdot\cdot})^2}_{SSA}+\underbrace{\sum_{k=1}^{K}\sum_{j=1}^{n_k}(Y_{kj}-\bar{Y}_{k\cdot})^2}_{SSE} \quad (4.3)\end{aligned}$$

where $\bar{Y}_{k\cdot} = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{kj}$ and $\bar{Y}_{\cdot\cdot} = \frac{1}{K}\sum_{k=1}^{K}\bar{Y}_{k\cdot}$.

- Large values of $SSA = \sum_{k=1}^{K} n_k(\bar{Y}_{k\cdot}-\bar{Y}_{\cdot\cdot})^2$ provide evidence against the null hypothesis (4.2). The alternative hypothesis here is that there is at least one pair of means μ_h, μ_l such that $\mu_h \neq \mu_l$.

4.1 The Kruskal-Wallis Test

4.1.1 Definition

- Instead of assuming (4.1) for the responses Y_{kj}, nonparametric way of thinking about this problem is to instead only assume that

$$Y_{kj} \sim F_k \quad (4.4)$$

  That is, $Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$ is an i.i.d. sample from F_k for each k.

- A nonparametric version of the one-way ANOVA homogeneity hypothesis is

$$H_0 : F_1 = F_2 = \ldots = F_K \quad (4.5)$$

- The "shift alternative" in this case can be stated as

$$H_A : F_k(t) = F(t-\Delta_k), \quad \text{for } k = 1, \ldots K \quad \text{and not all } \Delta_k \text{ equal} \quad (4.6)$$

- The Kruskal-Wallis test statistic is similar to the SSA term (defined in (4.3)) in the one-way ANOVA setting.

- Rather than comparing the group-specific means $\bar{Y}_{k\cdot}$ with the overall mean $\bar{Y}_{\cdot\cdot}$, the Kruskal-Wallis test statistic will be comparing the group-specific rank means $\bar{R}_{k\cdot}$ with their overall expectation under the null hypothesis.

- The Kruskal-Wallis test statistic is defined as

$$KW_N = \frac{12}{N(N-1)}\sum_{k=1}^{K} n_k\Big(\bar{R}_{k\cdot}-\frac{N+1}{2}\Big)^2, \quad \text{where } N = \sum_{k=1}^{K} n_k \quad (4.7)$$

- In (4.7), $\bar{R}_{k\cdot}$ is the average rank of those k^{th} group

$$\bar{R}_{k\cdot} = \frac{1}{n_k}\sum_{j=1}^{n_k} R_{kj}(\mathbf{Z}), \tag{4.8}$$

where $\mathbf{Z}$ denotes the pooled-data vector and $R_{kj}(\mathbf{Z})$ denotes the rank of Y_{kj} in the "pooled-data ranking".

- What is the expectation of $\bar{R}_{k\cdot}$ under the null hypothesis (4.5)?
- Again, if the null hypothesis is true, we can treat all of our responses Y_{kj} as just an i.i.d. sample of size N from a common distribution function F.
- Hence, as we showed in (3.5) from Chapter 3, $E\{R_{kj}(\mathbf{Z})\} = (N+1)/2$ under the assumption that the data are an i.i.d. sample from a common distribution function.
- So, the intuition behind the definition of KW_N is that the differences $\bar{R}_{k\cdot} - \frac{N+1}{2}$ should be small whenever the homogeneity hypothesis (4.5) is true.

- When $K = 2$, the following relationship between the Kruskal-Wallis statistic KW_N and the Wilcoxon rank sum test statistic W from Chapter 3 holds.

$$KW_N = \frac{12}{mn(N+1)}\Big(W - \frac{n(N+1)}{2}\Big)^2. \tag{4.9}$$

- Hence, the p-value from a Kruskal-Wallis test and a (two-sided) WRS test should be the same when $K = 2$.
- However, you cannot directly perform a one-sided test using the Kruskal-Wallis test.

- **Exercise 4.1** If $K = 2$, show that equation (4.9) holds.

An Example

- In this case, $N = 9$, $\bar{R}_{1\cdot} = 11/3$, $\bar{R}_{2\cdot} = 6$, and $\bar{R}_{3\cdot} = 16/3$. The Kruskall-Wallis statistic is

$$KW_N = \frac{1}{2}\Big\{3(11/3-5)^2 + 3(6-5)^2 + 3(16/3-5)^2\Big\} = 13/9 \tag{4.10}$$

Group	Y	Rank
Group 1	1.00	8
Group 1	-1.20	2
Group 1	-1.50	1
Group 2	0.00	5
Group 2	-0.10	4
Group 2	1.10	9
Group 3	0.90	7
Group 3	-0.40	3
Group 3	0.60	6

4.1.2 Asymptotic Distribution and Connection to One-Way ANOVA

- The Kruskal-Wallis statistic KW_N has an asymptotic chi-square distribution with $K-1$ degrees of freedom under the null hypothesis (4.5).
- This follows from the fact that $(\bar{R}_{k\cdot} - (N+1)/2)$ is approximately normally distributed for large n_k.
- **R** uses the large-sample approximation when computing the p-value for the Kruskal-Wallis test.

- The Kruskal-Wallis test can also be thought of as the test you would obtain if you applied the one-way ANOVA setup to the ranks of Y_{kj}.
- The one-way ANOVA test is based on the value of SSA where, as in (4.3), SSA is defined as

$$SSA = \sum_{k=1}^{K} n_k(\bar{Y}_{k\cdot} - \bar{Y}_{\cdot\cdot})^2 \tag{4.11}$$

- You then reject H_0, when $SSA/SSE = SSA/(SST - SSA)$ is sufficiently large.
- Notice that if we computed SSA using the ranks $R_{kj}(\mathbf{Z})$ rather than the observations Y_{kj}, we would get:

$$\begin{aligned} SSA_r &= \sum_{k=1}^{K} n_k(\bar{R}_{k\cdot} - \bar{R}_{\cdot\cdot})^2 \\ &= \sum_{k=1}^{K} n_k(\bar{R}_{k\cdot} - \frac{N+1}{2})^2 \\ &= \frac{N(N-1)}{12} KW_N \end{aligned} \tag{4.12}$$

- If you were applying ANOVA to the ranks of Y_{kj}, SST_r would be the following fixed constant:

$$\text{SST}_r = \frac{N(N+1)(N-1)}{12}$$

- So, any test of the homogeneity hypothesis would be based on just the value of SSA_r which as we showed in (4.12) is just a constant times the Kruskal-Wallis statistic.

4.2 Performing the Kruskal-Wallis Test in R

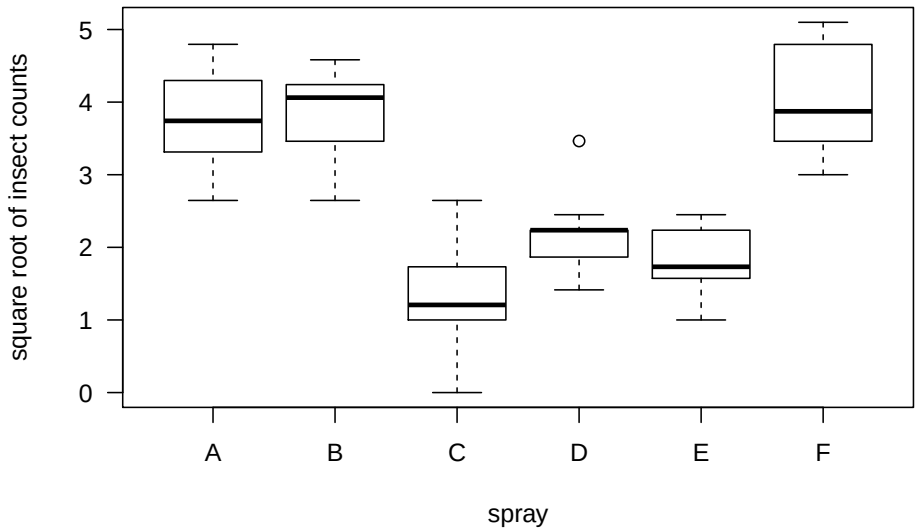
- We will look at performing Kruskal-Wallis tests in **R** by using the "InsectSprays" dataset.

```
head(InsectSprays)
```

```
##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

- This dataset has 72 observations.
- The variable **count** is the number of insects measured in some agricultural unit.
- The variable **spray** was the type of spray used on that unit.
- You could certainly argue that a standard ANOVA is not appropriate in this situation because the responses are counts, and for count data, the variance is usually a function of the mean.
- A generalized linear model with a log link function might be more appropriate.
- Applying a square-root transformation to count data is also a commonly suggested approach. (The square-root transformation is the "variance-stabilizing transformation" for Poisson-distributed data).

```
boxplot(sqrt(count) ~ spray, data=InsectSprays,las=1, ylab="square root of insect counts")
```



- Let us perform a test of homogeneity using both the one-way ANOVA approach and a Kruskal-Wallis test

```
anova(lm(sqrt(count) ~ spray, data=InsectSprays))
```

```
## Analysis of Variance Table
##
## Response: sqrt(count)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## spray      5 88.438 17.6876  44.799 < 2.2e-16 ***
## Residuals 66 26.058  0.3948
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
a  <- kruskal.test(sqrt(count) ~ spray, data=InsectSprays)
a$p.value
```

```
## [1] 1.510844e-10
```

- Notice that it applying the square root transformation or not does not affect the value of the Kruskal-Wallis statistic or the Kruskal-Wallis p-value.

```
kruskal.test(count ~ spray, data=InsectSprays)
```

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  count by spray
## Kruskal-Wallis chi-squared = 54.691, df = 5, p-value = 1.511e-10
```

- This invariance to data transformation is not true for the standard one-way ANOVA.

```
anova(lm(count ~ spray, data=InsectSprays))
```

```
## Analysis of Variance Table
##
## Response: count
##           Df Sum Sq Mean Sq F value    Pr(>F)
## spray      5 2668.8  533.77  34.702 < 2.2e-16 ***
## Residuals 66 1015.2   15.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.3 Comparison of Specific Groups

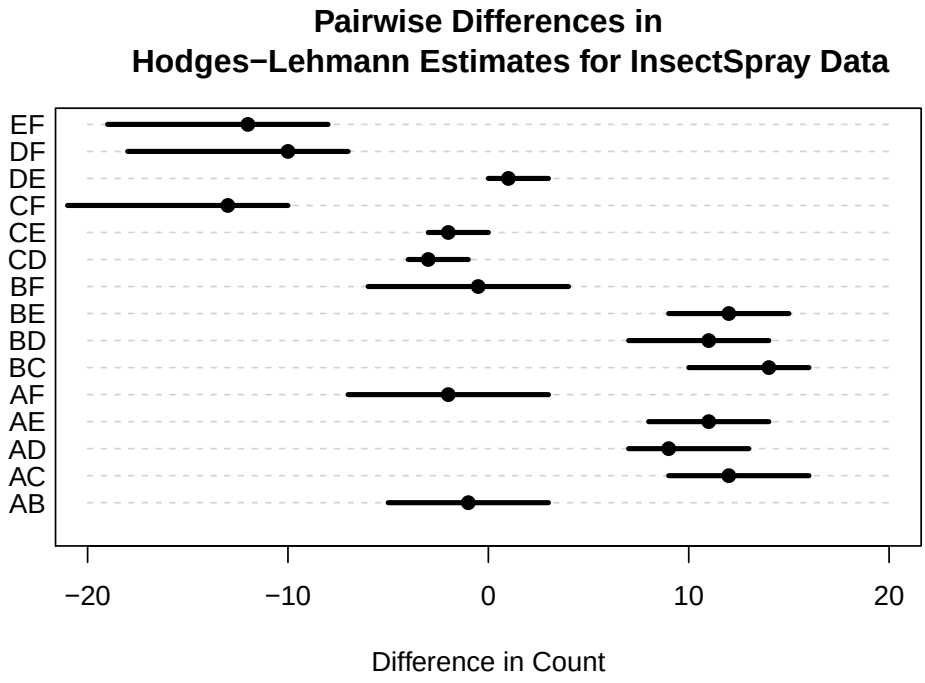
- A Kruskal-Wallis test performs a test of the overall homogeneity hypothesis

$$H_0 : F_1 = F_2 = \ldots = F_K \tag{4.13}$$

- However, a rejection of the homogeneity hypothesis does not indicate which group differences are primarily the source of this rejection nor does it provide any measure of the "magnitude" of the differences between each of the groups.
- Dunn's test is the suggested way to compute pairwise tests of stochatic dominance.
- Performing a series of pairwise Wilcoxon rank sum test can lead to violations of transitivity. For example, group A is "better" than B which is better than C, but group C is better than A.
- In **R**, Dunn's test can be performed using the **dunn.test** package.

- In traditional one-way ANOVA one often reports pairwise differences in the means and their associated confidence intervals.
- In the context of a Kruskal-Wallis test, one could report pairwise differences in the Hodges-Lehmann estimate though other comparisons may also be of interest.
- One nice approach is to use the proportional odds model interpretation of the Kruskal-Wallis test and then report the difference in the estimated

proportional odds coefficients. See Section 7.6 of http://hbiostat.org/doc/bbr.pdf for more details on the proportional odds model.



4.4 An Additional Example

- We will use the "cane" dataset from the **boot** package.

```
library(boot)
data(cane)
head(cane)
```

```
##     n  r  x var block
## 1  87 76 19   1     A
## 2 119  8 14   2     A
## 3  94 74  9   3     A
## 4  95 11 12   4     A
## 5 134  0 12   5     A
## 6  92  0  3   6     A
```

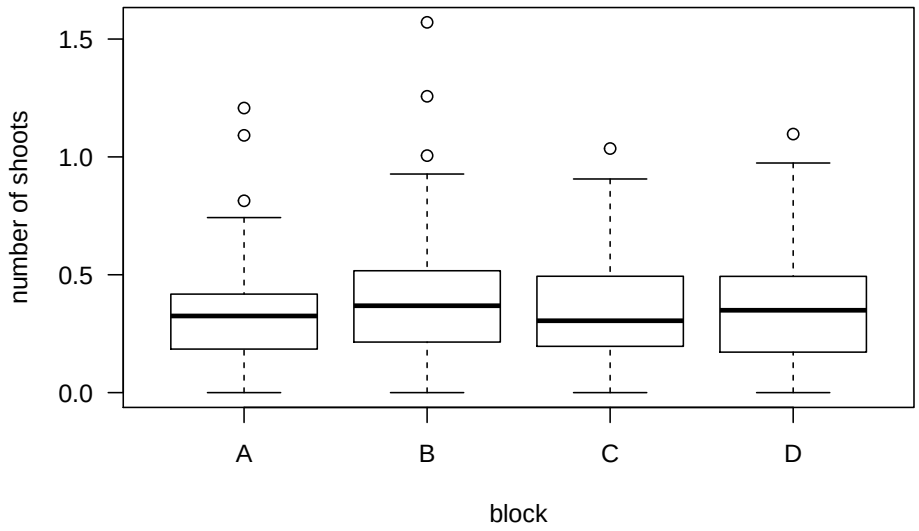
- These data come from a study trying to determine the susceptibility of different types of sugar cane to a particular type of disease.
- The variable **n** contains the total number of shoots in each plot.
- The variable **r** containes the total number of diseased shoots.

- We can create a new variable **prop** that measures the proportion of shoots that are diseased.

```
cane$prop <- cane$r/cane$n
```

- You could certainly argue that a logistic regression model is a better approach here, but we will analyze the transformed proportions using the arcsine square root transformation.

```
cane$prop.trans <- asin(sqrt(cane$prop))
boxplot(prop.trans ~ block, data=cane, las=1, ylab="number of shoots")
```




```
kruskal.test(prop ~ block, data=cane)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  prop by block
## Kruskal-Wallis chi-squared = 1.1355, df = 3, p-value = 0.7685
```

4.5 Additional Reading

- Additional reading which covers the material discussed in this chapter includes:
  - Chapters 6 from Hollander et al. (2013)

Chapter 5

Permutation Tests

- Permutation tests are a useful tool to avoid having to depend on specific parametric assumptions.
- Permutation tests are also useful in more complex modern applications where it can be difficult to work out the theoretical null distribution of a certain test statistic.

5.1 Notation

- A **permutation** π of a set S is a function $\pi : S \longrightarrow S$ is a function that is both one-to-one and onto.
- We will usually think of S as the set of observation indices in which case $S = \{1, \ldots, N\}$ for sample size N.
- Each permutation π of $S = \{1, \ldots, N\}$ defines a particular ordering of the elements of S. For this reason, a permutation is often expressed as the following ordered list

$$\pi = \big(\pi(1), \pi(2), \ldots, \pi(N)\big) \tag{5.1}$$

- In other words, we can think of a permutation of S as a particular ordering of the elements of S.
- For example, if $S = \{1, 2, 3\}$, and π_1 is a permutation of S defined as $\pi_1(1) = 3$, $\pi_1(2) = 1$, $\pi_1(3) = 2$, then this permutation expressed as an ordered list would be

$$\pi_1 = (3, 1, 2) \tag{5.2}$$

- There are 5 other possible permutations of S:

$$\begin{aligned}\pi_2 &= (1,2,3)\\ \pi_3 &= (2,1,3)\\ \pi_4 &= (1,3,2)\\ \pi_5 &= (3,2,1)\\ \pi_6 &= (2,3,1)\end{aligned}$$

- If S has N distinct elements, there are $N!$ possible permutations of S.
- We will let $\mathcal{S}_N$ denote the set of all permutations of the set $\{1,\ldots,N\}$.

5.2 Permutation Tests for the Two-Sample Problem

- A permutation test is motivated by the following reasoning.
  - If there is no real difference between the two groups, there is nothing "special" about the difference in means between the two groups.
  - The observed difference in the mean between the two groups should not be notably different than mean differences from randomly formed groups.
  - Forming "random" groups can be done by using many permutations of the original data.

5.2.1 Example 1

- Suppose we have observations from two groups $X_1,\ldots,X_n \sim F_X$ and $Y_1,\ldots,Y_m \sim F_Y$.
- Let $\mathbf{Z} = (Z_1,\ldots,Z_N)$ denote pooled data

$$(Z_1,\ldots,Z_N) = (X_1,\ldots,X_n,Y_1,\ldots,Y_m) \tag{5.3}$$

- For a permutation π of $\{1,\ldots,N\}$, we will let $\mathbf{Z}_\pi$ denote the corresponding permuted dataset

$$\mathbf{Z}_\pi = (Z_{\pi(1)}, Z_{\pi(2)},\ldots,Z_{\pi(N)}) \tag{5.4}$$

- The columns in the above table are just permutations of the original data $\mathbf{Z}$.

	OriginalData	Perm1	Perm2	Perm3	Perm4	Perm5
z1	0.60	-0.60	0.60	-0.90	0.70	0.60
z2	-0.80	-1.40	-0.60	0.70	-0.40	-0.60
z3	-0.60	0.70	0.20	0.60	-1.40	-0.80
z4	-0.90	0.20	-0.40	0.20	0.20	0.30
z5	0.30	-0.40	-1.30	-0.40	-0.90	-0.40
z6	-1.30	-1.30	-1.40	-0.60	-0.80	0.70
z7	0.20	0.30	0.70	-1.40	0.30	-0.90
z8	0.70	0.60	0.30	-1.30	0.60	0.20
z9	-1.40	-0.90	-0.80	0.30	-0.60	-1.40
z10	-0.40	-0.80	-0.90	-0.80	-1.30	-1.30
mean difference	0.16	0.12	0.12	0.80	0.00	0.36

Table 5.1: Example of Permuting a Vector of Responses. This example assumes n=m=5.

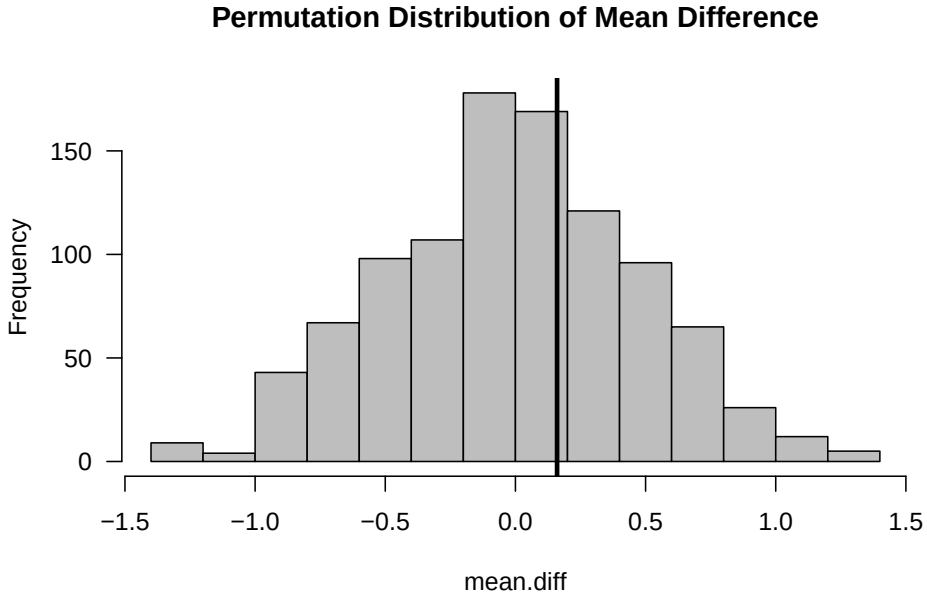
- Suppose we want to base a test on the difference in the means between the two groups

$$T_N(\mathbf{Z}) = \bar{X} - \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{m}\sum_{i=n+1}^{N} Z_i \tag{5.5}$$

- We will let t_{obs} denote the observed value of the mean difference. That is, $t_{obs} = T_N(\mathbf{Z}_{obs})$, where $\mathbf{Z}_{obs}$ is the vector of the observed data.

- Under the null hypothesis that $F_X = F_Y$, the observed mean difference should not be "abnormal" when compared with the mean differences from many other permutations of the data.

```
z <- c(0.6, -0.8, -0.6, -0.9, 0.3, -1.3, 0.2, 0.7, -1.4, -0.4) ## data
observed.diff <- mean(z[1:5]) - mean(z[6:10])  ## observed mean difference
nreps <- 1000
mean.diff <- rep(NA, nreps)
for(k in 1:nreps) {
    ss <- sample(1:10, size=10)  ## draw a random permutation
    z.perm <- z[ss]   ## form the permuted dataset
    mean.diff[k] <- mean(z.perm[1:5]) - mean(z.perm[6:10]) ## compute mean difference
                                                           ## for permuted dataset
}
hist(mean.diff, las=1, col="grey", main="Permutation Distribution of Mean Difference")
abline(v=observed.diff, lwd=3)
```



5.2.2 Permutation Test p-values

- The one-sided p-value for the permutation test is

$$\begin{aligned} \text{p-value} &= \frac{\text{number of permutations such that } T_N \geq t_{obs}}{N!} \\ &= \frac{1}{N!} \sum_{\pi \in \mathcal{S}_N} I\Big(T_N(\mathbf{Z}_\pi) \geq t_{obs}\Big) \end{aligned}$$

- The two-sided p-value for the two-sample problem would be

$$\text{p-value} = \frac{1}{N!} \sum_{\pi \in \mathcal{S}_N} I\Big(\Big|T_N(\mathbf{Z}_\pi)\Big| \geq |t_{obs}|\Big)$$

- As we did when producing the above histogram, the permutation-test p-value is often computed by using a large number of random permutations rather than computing the test statistic for every possible permutation.
- The Monte Carlo permutation-test p-value is defined as

$$\text{p-value}_{mc} = \frac{1}{S+1}\left[1 + \sum_{s=1}^{S} I\Big(T_N(\mathbf{Z}_{\pi_s}) \geq t_{obs}\Big)\right] \tag{5.6}$$

  where $\pi_1, \ldots, \pi_S$ are randomly drawn permutations
- The two-sided (Monte Carlo) p-value for the example shown in the above Table is

```
pval.mc <- (1 + sum(abs(mean.diff) >= abs(observed.diff)))/(nreps + 1)
round(pval.mc, 2)
```

```
## [1] 0.74
```

5.2.3 Example 2: Ratios of Means

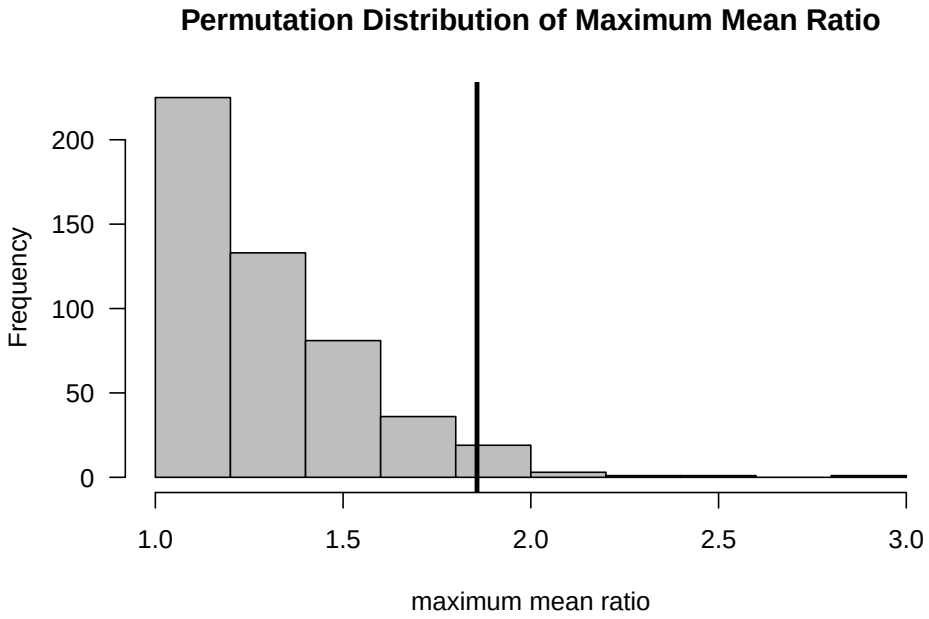
- With permutation tests, you are not limited to difference in means. You can choose the statistic $T_N(\mathbf{Z})$ to measure other contrasts of interest.

- For example, with nonnegative data you might be interested in the ratio of means between the two groups

$$T_N(\mathbf{Z}) = \max\left\{\bar{X}/\bar{Y}, \bar{Y}/\bar{X}\right\} \tag{5.7}$$

- Let us see how this works for a simulated example where we assume that $X_1, \ldots, X_n \sim$ Exponential(1) and $Y_1, \ldots, Y_m \sim$ Exponential(1/2)

```
set.seed(5127)
xx <- rexp(20, rate=1)
yy <- rexp(20, rate=0.5)
zz <- c(xx, yy) ## this is the original data
t.obs <- max(mean(zz[1:20])/mean(zz[21:40]), mean(zz[21:40])/mean(zz[1:20]))
nperms <- 500
mean.ratio <- rep(0, nperms)
for(k in 1:nperms) {
     ss <- sample(1:40, size=40)
     zz.perm <- zz[ss]
     mean.ratio[k] <- max(mean(zz.perm[1:20])/mean(zz.perm[21:40]),
                          mean(zz.perm[21:40])/mean(zz.perm[1:20]))
}
hist(mean.ratio, las=1, col="grey", main="Permutation Distribution of Maximum Mean Ratio",
     xlab="maximum mean ratio")
abline(v=t.obs, lwd=3)
```



- The two-side (Monte Carlo) permutation test p-value is:

```
pval.mc <- (1 + sum(mean.ratio >= t.obs))/(nperms + 1)
round(pval.mc, 2)
```

```
## [1] 0.04
```

5.2.4 Example 3: Differences in Quantiles

- Permutation tests are especially useful in problems where working out the null distribution is difficult, or when certain approximations of the null distributions are hard to justify.
- An example of this occurrs if you want to compare medians, or more generally, compare quantiles.
- The difference-in-quantiles statistic would be defined as

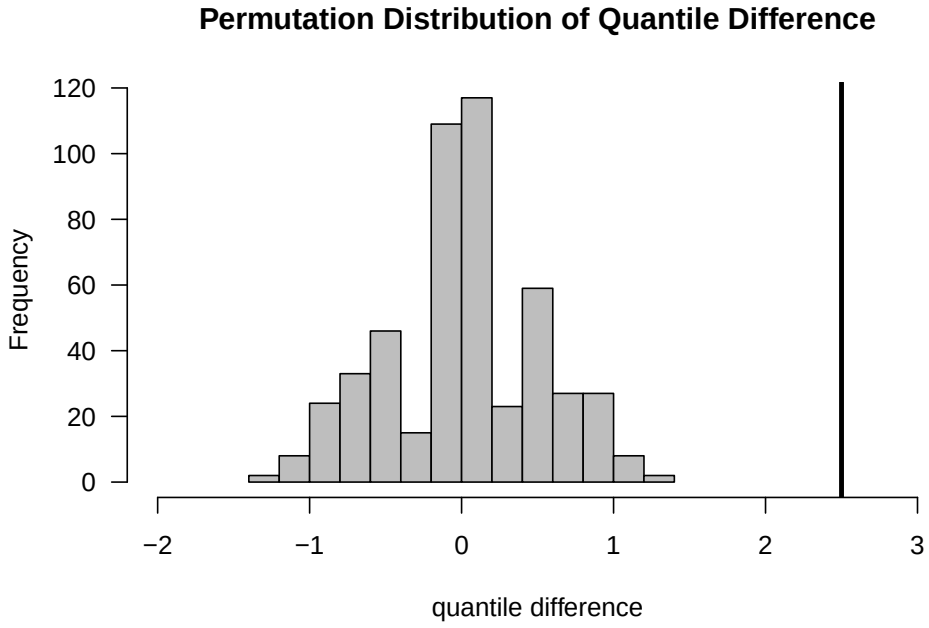
$$T_N(\mathbf{Z}) = Q_p(Z_1, \ldots, Z_n) - Q_p(Z_{n+1}, \ldots, Z_N) \tag{5.8}$$

  where $Q_p(X_1, \ldots, X_H)$ denotes the p^{th} quantile from the data $X_1, \ldots, X_H$.
- The difference in quantiles could be computed with the following **R** code:

```
z <- rnorm(10)
quantile(z[1:5], probs=.3) - quantile(z[6:10], probs=.3)
```

```
##       30%
## 0.2671133
```

- Note that setting **probs=.5** in the **quantile** function will return the median.



5.3 The Permutation Test as a Conditional Test

- A permutation test is an example of a **conditional test**.
- Typically, the p-value is defined as

$$\text{p-value} = P(T \geq t_{obs} | H_0) \tag{5.9}$$

  for some test statistic T.
- In other words, the p-value is the probability that a random variable following the null distribution exceeds t_{obs}.
- In many problems however, the null hypothesis H_0 is not determined by a single parameter but contains many parameters.
- For example, the null hypothesis in a t-test is really $H_0 : \mu_x = \mu_y$ and $\sigma > 0$. That is, the null hypothesis is true for many different values of σ.

- When H_0 contains many parameter values, one approach for computing a p-value is to choose the test statistic T so that its distribution is the same for every point in H_0.

- A more general approach is to instead compute a **conditional p-value**
- The conditional p-value is defined as

$$\text{p-value} = P(T \geq t_{obs} | S = s, H_0) \tag{5.10}$$

where S is a sufficient statistic for the unknown terms in H_0.

- A classic example of this is Fisher's exact test.

- A permutation test computes a conditional p-value where the sufficient statistic is the vector of order statistics $(Z_{(1)}, Z_{(2)}, \ldots, Z_{(N)})$.
- Recall that the order statistics are defined as

$$\begin{aligned} Z_{(1)} &= \text{smallest observation} && (5.11)\\ Z_{(2)} &= \text{second smallest observation} && (5.12)\\ &\vdots && (5.13)\\ Z_{(N)} &= \text{largest observation} && (5.14) \end{aligned}$$

- What is the conditional distribution of the observed data conditional on the observed order statistics?
- It is:

$$\begin{aligned} f_{Z_1,\ldots,Z_N | Z_{(1)},\ldots,Z_{(N)}}(z_{\pi(1)}, \ldots, z_{\pi(N)} | z_1, \ldots, z_N) &= \frac{f_{Z_1,\ldots,Z_N}(z_{\pi(1)}, \ldots, z_{\pi(N)})}{f_{Z_{(1)},\ldots,Z_{(N)}}(z_1, \ldots, z_N)} \\ &= \frac{f_{Z_i}(z_{\pi(1)}) \cdots f_{Z_i}(z_{\pi(N)})}{N! f_{Z_i}(z_1) \cdots f_{Z_i}(z_N)} \\ &= \frac{1}{N!} \end{aligned} \tag{5.15}$$

(See Chapter 5 of Casella and Berger (2002) for a detailed description of the distribution of order statistics)

- In other words, if we know the value of: $Z_{(1)} = z_1, \ldots, Z_{(N)} = z_N$, then any event of the form $\{Z_1 = z_{\pi(1)}, \ldots, Z_N = z_{\pi(N)}\}$ has an equal probability of occurring for any permutation chosen.
- This equal probability of $1/N!$ is only true under H_0 where we can regard the data as coming from a common distribution.

- If we are conditioning on $Z_{(1)} = z_1, \ldots, Z_{(N)} = z_N$, then the probability that $T_N(Z_1, \ldots, Z_N) \geq t$ is just the number of permutations of $(z_1, \ldots, z_N)$ such that the test statistic is greater than t divided by $N!$.

- In other words

$$P\Big\{T_N(Z_1,\ldots,Z_N) \geq t | Z_{(1)} = z_1, \ldots, Z_{(N)} = z_N\Big\} \quad (5.16)$$

$$= \frac{1}{N!}\sum_{\pi\in\mathcal{S}_N} I\Big(T_N(z_{\pi(1)},\ldots,z_{\pi(N)}) \geq t\Big) \quad (5.17)$$

- Let us consider a concrete example.

- Suppose we have a two-sample problem with four observations. The first two observations come from the first group while the last two observations come from the second group.

- The order statistics that we will condition on are:

$$Z_{(1)} = z_1 = -3 \quad (5.18)$$
$$Z_{(2)} = z_2 = -1 \quad (5.19)$$
$$Z_{(3)} = z_3 = 2 \quad (5.20)$$
$$Z_{(4)} = z_4 = 5 \quad (5.21)$$

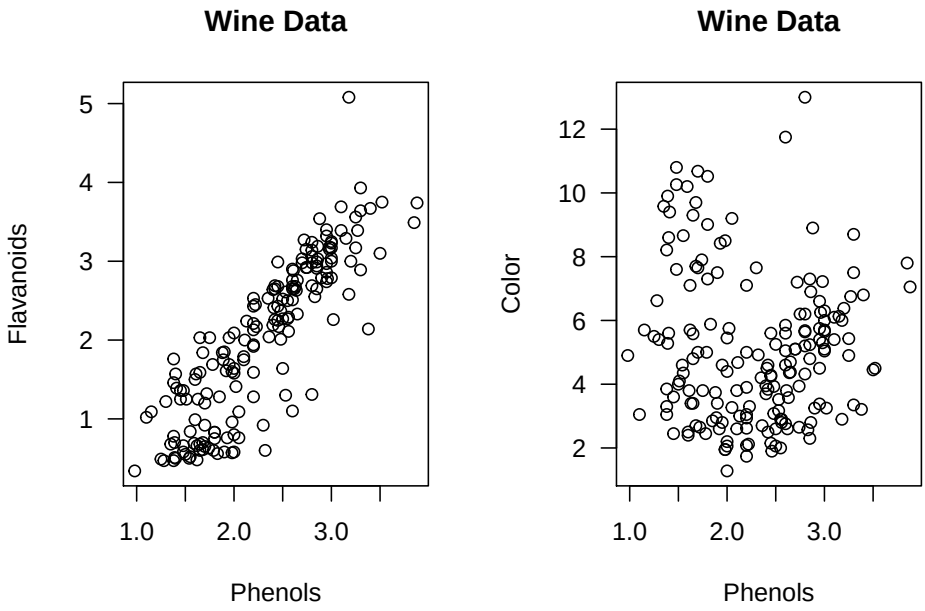
- If T_4 is the mean difference

$$T_4(Z_1,Z_2,Z_3,Z_4) = \frac{Z_1 + Z_2 - Z_3 - Z_4}{2} \quad (5.22)$$

what is the probability

$$P\Big\{T_4(Z_1,Z_2,Z_3,Z_4) \geq 2.5 | Z_{(1)} = z_1, Z_{(2)} = z_2, Z_{(3)} = z_3, Z_{(4)} = z_4\Big\} \quad (5.23)$$

5.4 A Permutation Test for Correlation



- Suppose we have N pairs of observations $(U_1, V_1), \ldots, (U_N, V_N)$
- The correlation between these pairs is defined as

$$\rho_{UV} = \frac{\text{Cov}(U_i, V_i)}{\sigma_U \sigma_V} \tag{5.24}$$

- The test statistic of interest here is the sample correlation

$$T_N(\mathbf{U}, \mathbf{V}) = \frac{\sum_{i=1}^N (U_i - \bar{U})(V_i - \bar{V})}{\sqrt{\sum_{i=1}^N (U_i - \bar{U})^2 \sum_{i=1}^N (V_i - \bar{V})^2}} \tag{5.25}$$

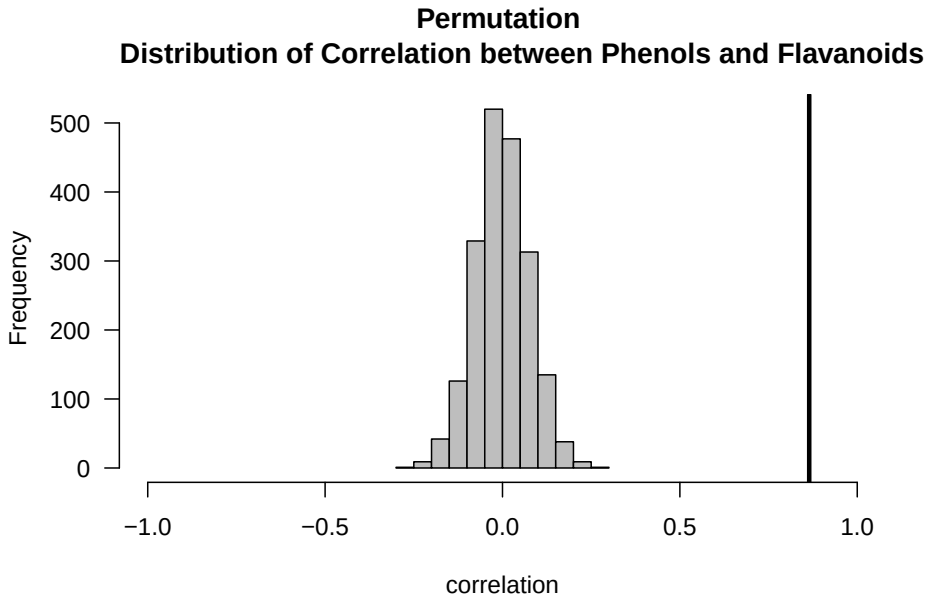
- To find the the permutation distribution, we only need to look at $T_N(\mathbf{U}_\pi, \mathbf{V})$ for different permutations π.
- In other words, we are computing correlation among pairs of the form $(U_{\pi(1)}, V_1), \ldots, (U_{\pi(N)}, V_N)$.
- We only need to look at $\mathbf{U}_\pi$ because this achieves the objective of randomly "switching observation pairs".

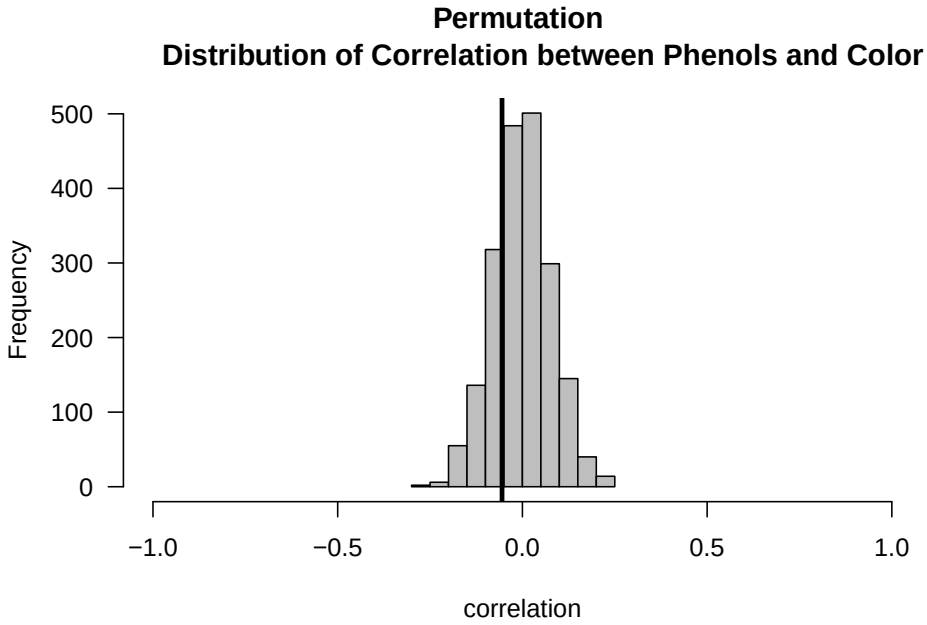
- The two-sided p-value for the permutation test of $H_0 : \rho_{UV} = 0$ vs. $H_A : \rho_{UV} \neq 0$ is

$$\text{p-value} = \frac{1}{N!} \sum_{\pi \in \mathcal{S}_N} I\Big(\Big|T_N(\mathbf{U}_\pi, \mathbf{V})\Big| \geq |t_{obs}|\Big)$$

```
library(rattle.data)
## Computing the permutation distribution for correlation
## between flavanoids and phenols

n.obs <- nrow(wine) ## number of observations
t.obs.pf <- cor(wine$Phenols, wine$Flavanoids) ## observed correlation
nperms <- 2000
cor.perm.pf <- rep(0, nperms)
for(k in 1:nperms) {
     ss <- sample(1:n.obs, size=n.obs)
     uu.perm <- wine$Phenols[ss]
     cor.perm.pf[k] <- cor(uu.perm, wine$Flavanoids)
}
hist(cor.perm.pf, xlim=c(-1, 1), las=1, col="grey", main="Permutation
     Distribution of Correlation between Phenols and Flavanoids",
     xlab="correlation")
abline(v=t.obs.pf, lwd=3)
```





- Now let us compute the p-values for both the Phenols/Flavanoids and Phenols/Color association tests.

```
pval.mc <- (1 + sum(abs(cor.perm.pf) >= abs(t.obs.pf)))/(nperms + 1)
round(pval.mc, 4)
```

```
## [1] 5e-04
```

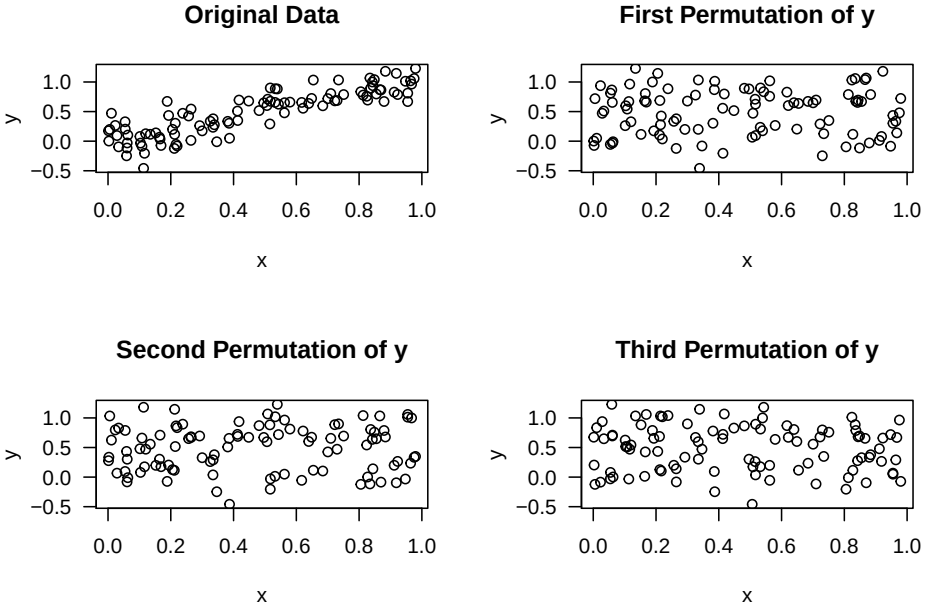
```
pval.mc <- (1 + sum(abs(cor.perm.pc) >= abs(t.obs.pc)))/(nperms + 1)
round(pval.mc, 4)
```

```
## [1] 0.4648
```

5.5 A Permutation Test for Variable Importance in Regression and Machine Learning

- The idea of permutation testing can also be applied in the context of regression.
- In regression, we have a series of responses $y_1, \ldots, y_N$, and we have a series of associated covariates vectors $\mathbf{x}_i$.
- For regression, we are going to perform permutations on the vector of responses $\mathbf{y} = (y_1, \ldots, y_N)$ and compute some measure for each permutation.

- For example, we might compute some measure of variable importance for each permutation.
- The idea here is that when permuting $\mathbf{y}$, the association between $\mathbf{y}$ and any "important covariates" should be lost.
- We want to see what the typical values of our variable importance measures will be when we break any association between $\mathbf{y}$ and a covariate.



Bibliography

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Divine, G. W., Norton, H. J., Barón, A. E., and Juarez-Colunga, E. (2018). The wilcoxon–mann–whitney procedure fails as a test of medians. *The American Statistician*, 72(3):278–286.

Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods*, volume 751. John Wiley & Sons.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.