Elements of Nonparametric Statistics

Nicholas Henderson

2020-01-15

Contents

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

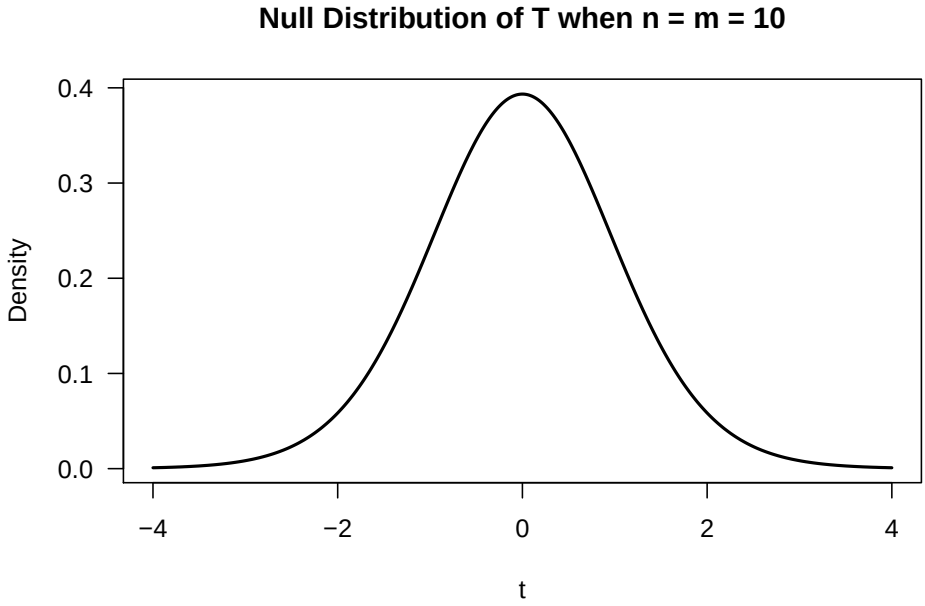
$$T = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}, \tag{1.4}$$

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.

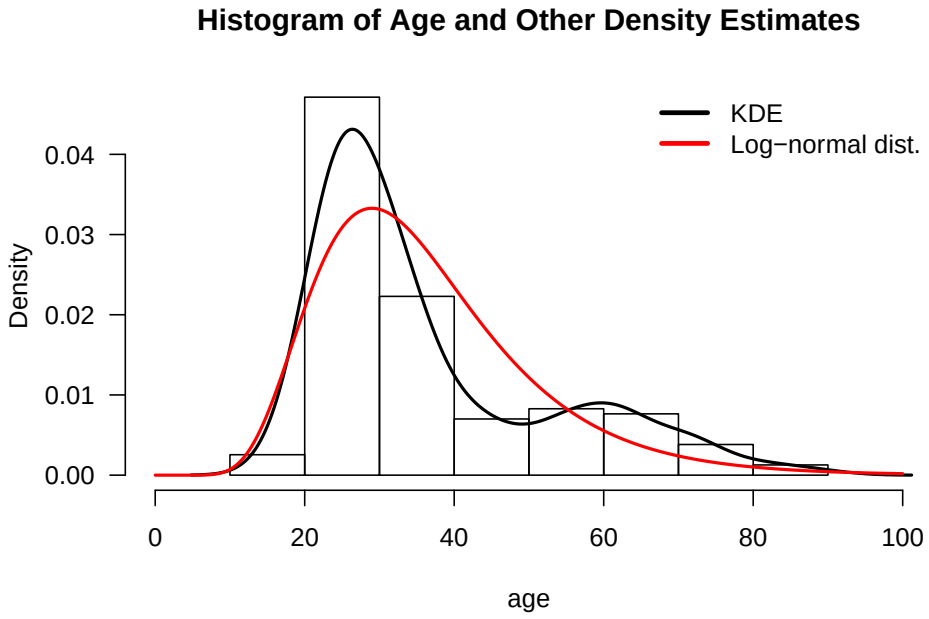
- The density function of X_i is often estimated by a kernel density estimator (KDE). This is defined as

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

- $K()$ - 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

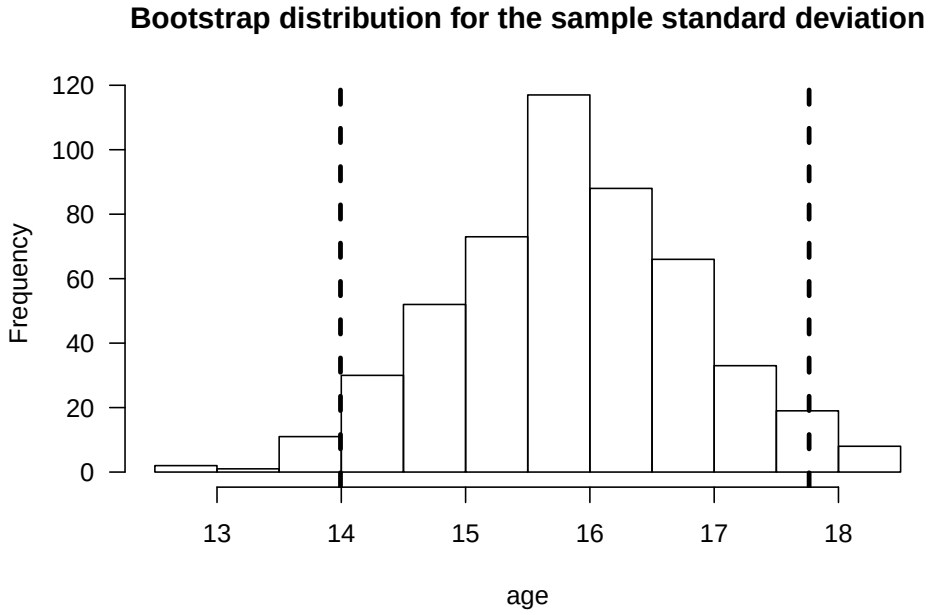


Figure 1.1: Bootstrap distribution of the sample standard deviation for the age variable from the kidney fitness data. Dasjed vertical lines are placed at the 2.5 and 97.5 percentiles of the bootstrap distribution.

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.

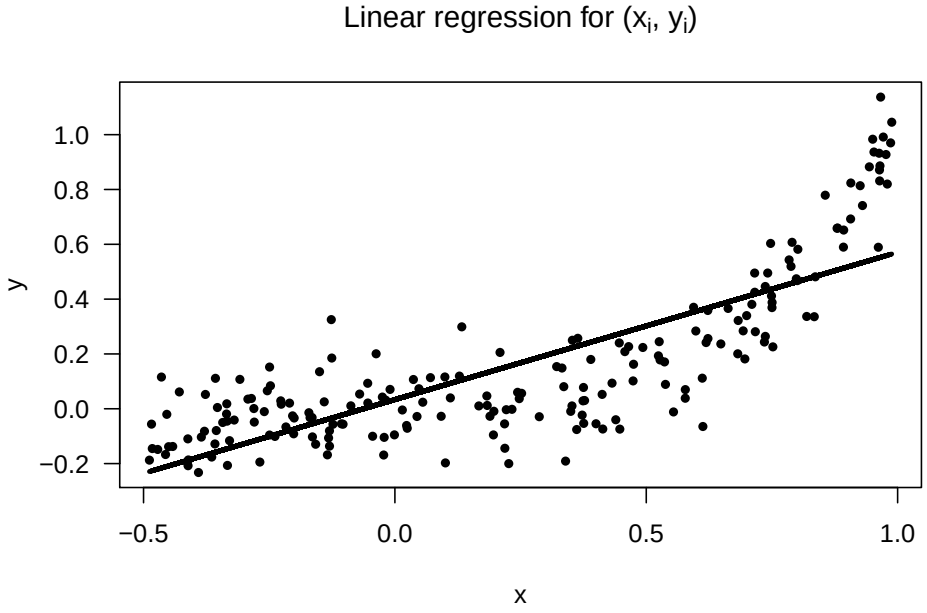
- It is often better to construct confidence intervals using the percentiles from the bootstrap distribution of $\hat{\theta}$ rather than use a confidence interval of the form: $\hat{\theta} \pm 1.96 \times se_{boot}$.

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

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

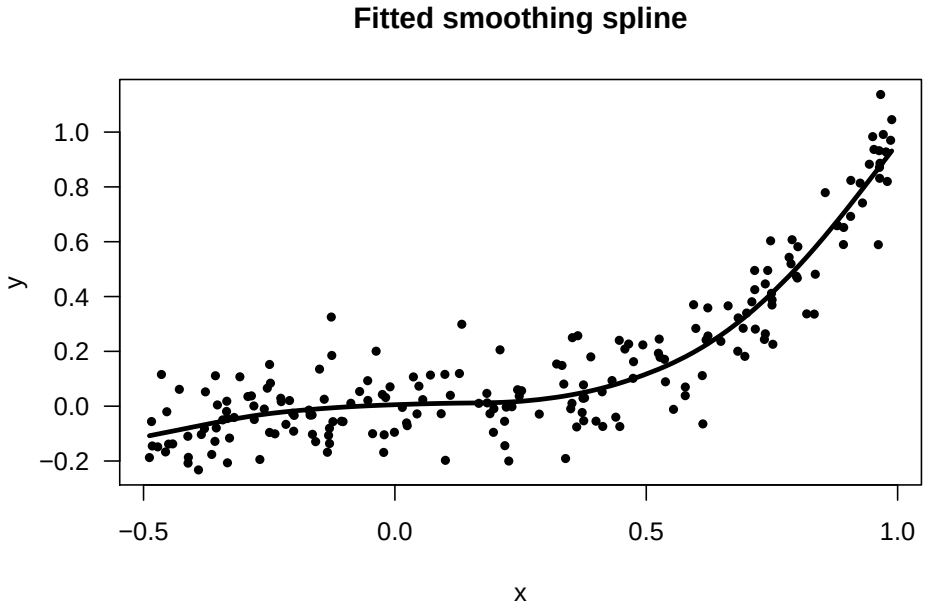
where $\Omega_{ij} = \int g_i''(t) g_j''(t) dt$.

- 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)

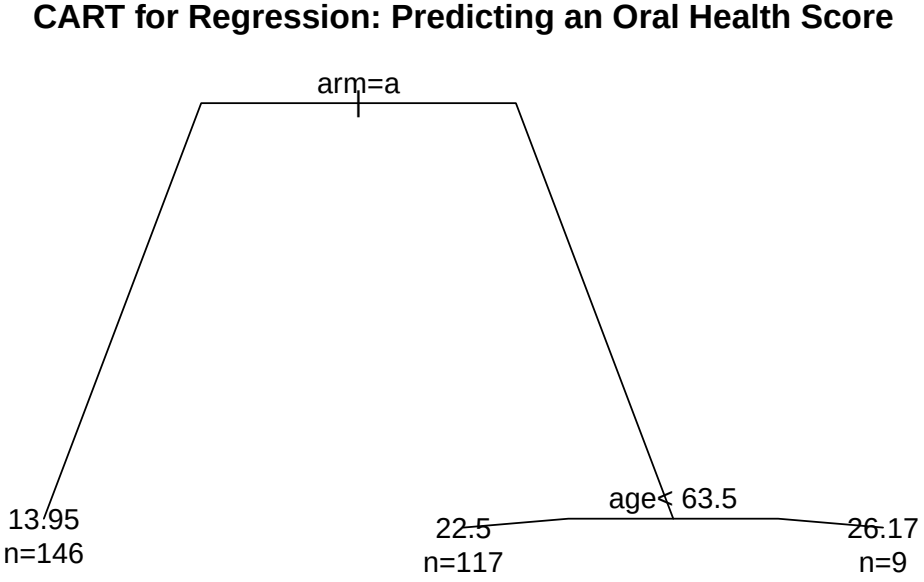
- Suppose we now have observations $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ where y_i is a continuous response and $\mathbf{x}_i$ is a p-dimensional vector of covariates.

- 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.

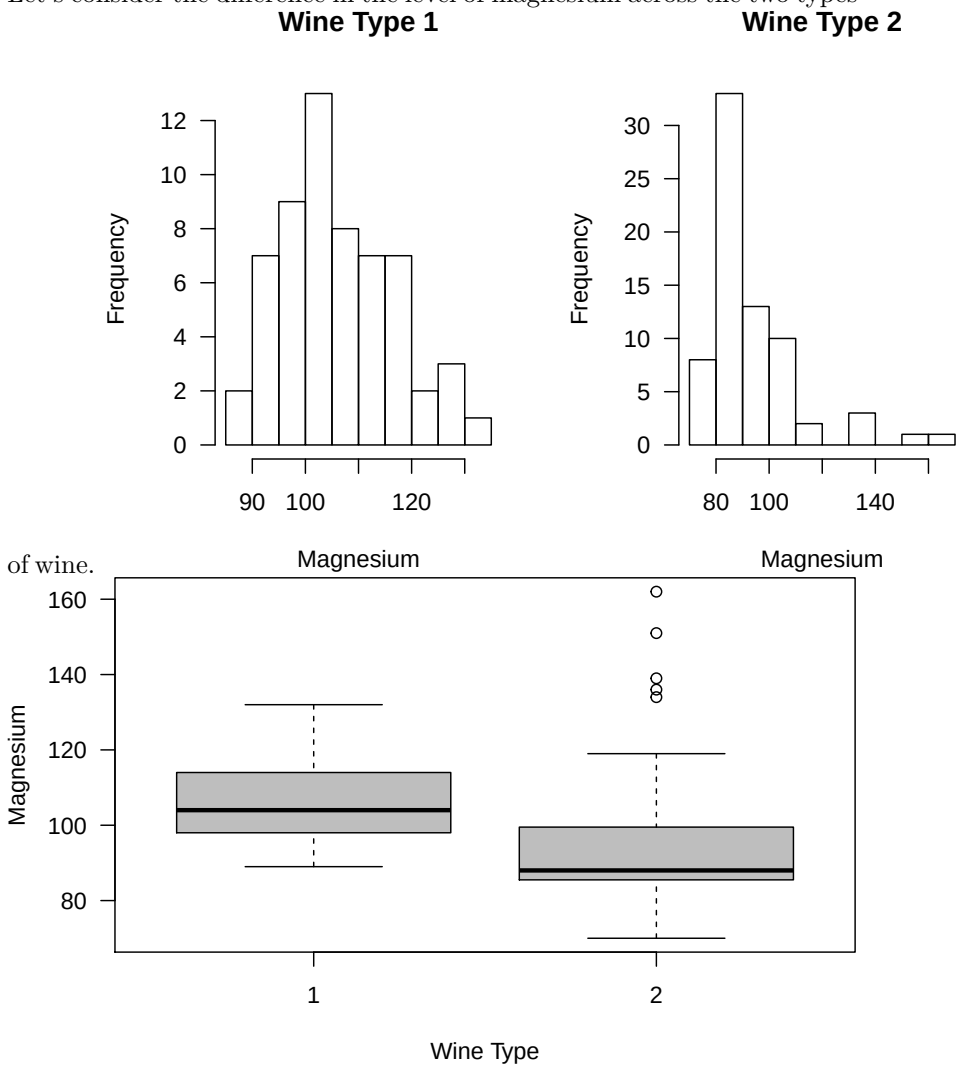
```
library(rattle.data)
head(wine)
```

```
##   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

```r
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)
}
mean(xgreater)  ## estimate of this probability
```

```
## [1] 0.76
```

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

- Suppose we have observations $D_1, \ldots, D_n$ which arise from the following model

$$D_i = \theta + \varepsilon_i,$$

  where ε_i are iid random variables each with distribution function F_ϵ that is assumed to have a median of zero.
- The distribution function of D_i is then

$$F_D(t) = P(D_i \leq t) = P(\varepsilon_i \leq t - \theta) = F_\epsilon(t - \theta) \tag{3.25}$$

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

$$f_D(t) = f_\epsilon(t - \theta) \tag{3.26}$$

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

Baseline_Measure

Post_Treatment_Measure

Patient 1

X1

Y1

Patient 2

X2

Y2

Patient 3

X3

Y3

Patient 4

X4

Y4

- 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

$$S = \sum_{i=1}^{n} I(D_i > 0) \tag{3.27}$$

- 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 $S \geq c$ 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 (3.27) 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

$$p(\theta) = P(Z_i > 0) = 1 - F_D(0) = 1 - F_\epsilon(-\theta) \tag{3.28}$$

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

$$S \sim \text{Binomial}(n, p(\theta)) \tag{3.29}$$

- 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

$$\text{p-value} = P_{H_0}(S \geq s_{obs}) = \sum_{j=s_{obs}}^{n} \binom{n}{j} \frac{1}{2^n}, \tag{3.30}$$

  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] 0.6913503
```

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

$$P_{H_0}(S \geq w) = 1 - P_{H_0}(S < w) = 1 - P_{H_0}(S \leq w - 1) \tag{3.31}$$

  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**.

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^+ weights the positive indicators $I(D_i > 0)$ by the rank of its absolute value.
- Specifically, the Wilcoxon signed rank statistic is defined as

$$T^+ = \sum_{i=1}^{n} I(D_i > 0) R_i(|\mathbf{D}|) \tag{3.32}$$

- Here, $R_i(\mathbf{D})$ is the rank of the i^{th} element from the vector $|\mathbf{D}| = (|D_1|, |D_2|, \ldots, |D_n|)$.

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?

- Expectation under the null hypothesis..

3.3.2.1 Exact Distribution

3.3.2.2 Asymptotic Distribution

3.4 Power and Comparisons with Parametric Tests

3.4.1 Power of Tests

- The **power** of a test is the probability that a test rejects the null hypothesis when the alternative hypothesis is true.

3.5 Thinking about Rank statistics more generally

3.6 Notes

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

Bibliography

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.