Lecture Notes in Applied Statistics, Spring 2011

Rui Manuel Castro
Eindhoven University of Technology.
## Contents

### 1 Introduction: Nonparametric statistics

#### 2 Nonparametric estimation of distribution functions and quantiles

2.1 A confidence band for $\hat{F}_n$ ........................................... 3
2.2 Confidence intervals for the distribution function at a fixed point .......... 4
2.3 Confidence intervals for quantiles using order statistics .................. 7

### 3 Goodness-of-Fit Tests

3.1 Simple GoF using the Empirical CDF ................................... 10
3.2 Consistency under the alternative ....................................... 12
3.3 Composite GoF tests ..................................................... 13
3.3.1 A Bootstrap approach for composite GoF tests .................. 15
3.4 Chi-Square-type GoF tests ............................................. 16
3.4.1 Composite $\chi^2$-GoF tests ........................................ 17
3.5 Probability Plotting and Quantile-Quantile Plots ...................... 18

### 4 Permutation tests

4.1 Final Remarks ............................................................ 24

### 5 Tests based on Signs, Runs and Ranks

5.1 Introduction ............................................................... 26
5.2 Sign test ................................................................. 26
5.2.1 Paired sample procedures .......................................... 27
5.3 Ranks ..................................................................... 28
5.3.1 The Wilcoxon Signed-Rank Test .................................. 29
5.4 The General Two-Sample Problem .................................... 30
5.4.1 The Wald-Wolfowitz Runs Test ................................... 31
5.4.2 The Wilcoxon Rank-Sum Test ..................................... 31
5.5 Further Reading .......................................................... 33

### 6 Measures of Association

6.1 Towards General Measures of Association .............................. 34
6.2 Kendall’s Tau ............................................................. 35
6.2.1 Estimating $\tau(X,Y)$ from Data ................................... 36
6.3 Spearman correlation ..................................................... 37

### 7 Useful R-commands

7.1 Empirical distribution function and related ............................ 38
7.2 Goodness of fit .......................................................... 38
7.3 Permutations ............................................................. 38
7.4 Rank tests and measures of association ................................ 38
Abstract

These notes are an edited version of notes for the course “Applied Statistics 2010” by Frank van der Meulen. The author of the current version is the sole responsible for the errors it might (and certainly does) contain.
1 Introduction: Nonparametric statistics

The term non-parametric statistics often takes a different meaning for different authors. For example, Wolfowitz (1942):

We shall refer to this situation (where a distribution is completely determined by the knowledge of its finite parameter set) as the parametric case, and denote the opposite case, where the functional forms of the distributions are unknown, as the non-parametric case.


Nonparametric statistics can and should be broadly defined to include all methodology that does not use a model based on a single parametric family.

Wasserman (2005)

The basic idea of nonparametric inference is to use data to infer an unknown quantity while making as few assumptions as possible.

Bradley (1968)

The terms nonparametric and distribution-free are not synonymous... Popular usage, however, has equated the terms... Roughly speaking, a nonparametric test is one which makes no hypothesis about the value of a parameter in a statistical density function, whereas a distribution-free test is one which makes no assumptions about the precise form of the sampled population.

The term distribution-free is used quite often in the statistical learning theory community, to refer to an analysis that makes no assumptions on the distribution of training examples (and only assumes these are independent and identically distributed samples from an unknown distribution).

For our purposes we will say a model is nonparametric if it is not a parametric model. In the simplest form we speak of a parametric model if

\[ X \text{ (observation vector)} \sim P, \]

and \( P = P_\theta \) where \( \theta \in \Theta \subseteq \mathbb{R}^d \). That is, the model is completely specified by \( d < \infty \) parameters. Of course parametric models can be quite complicated (e.g., graphical models, Hidden Markov Models, etc.) but the key feature is that they can be described by a finite set of parameters, and the dimension of the model (number of parameters) is fixed a priori regardless of the data.

Fifty years ago, most nonparametric methods were not feasible in practice, due to limited computing power. Nowadays, this has changed due to rapid developments in computing science. Nonparametric models can be advantageous since they offer much more flexibility to model the data. Advantages of parametric methods include:

- Convenience: parametric models are generally easier to work with.
- Efficiency: If parametric model is correct, then parametric methods are more efficient than their nonparametric counterpart. However, the loss in efficiency of nonparametric methods is often small.
• Interpretation: Sometimes parametric models are easier to interpret.

Disadvantages include:

• Sometimes it is hard to find a suitable parametric model.

• Parametric methods are often only suitable for interval-scaled data, nonparametric methods based on order statistics work for ordinal data as well.

• Nonparametric methods are often less sensitive to outliers.

• Parametric methods have a high risk of *Mis-specification*.

Within non-parametric models, as described above, one can also consider some finer characterizations. *Semiparametric models* are somewhere in between parametric and nonparametric models. It is hard to give a precise definition. The typical setting is the case where the statistical model has a natural parametrization \((\theta, \eta) \mapsto P_{\theta, \eta}\), where \(\theta\) is a Euclidean parameter and \(\eta\) lies in some infinite-dimensional set. Often the main parameter of interest is \(\theta\) and \(\eta\) is a nuisance quantity.

1.1 Example 1. *Regression*: Observe \((X_1, Y_1), \ldots, (X_n, Y_n)\) and suppose

\[
Y_i = r(X_i) + \epsilon_i ,
\]

where \(r\) is a convex function and \(\{\epsilon_i\}\) is a sequence of independent \(N(0, \sigma^2)\)-distributed random variables.

2. *Logistic regression*: Observe a binary \(Y\) and covariate-vector \(Z\). Suppose

\[
\Pr(Y = 1) = \frac{1}{1 + e^{-r(Z)}} .
\]

Suppose \(Z = (Z_1, Z_2)\) and

\[
r(z_1, z_2) = \eta(z_1) + \theta' z_2 .
\]

3. *Interval Censoring*: Let the time of “death” be \(T\). Let \(C\) be some “check-up time”. Observe couples

\[
(C, 1\{T \leq C\}) .
\]

Assume \(T\) and \(C\) are independent, Model \(T\) by a Weibull distribution and leave \(C\) unspecified.

2 Nonparametric estimation of distribution functions and quantiles

In this section we consider what is undoubtedly one of the simplest non-parametric estimators, namely the empirical Cumulative Distribution Function (CDF). Suppose \(X_1, \ldots, X_n\) is a random sample from an unknown distribution function \(F(x) = \Pr(X \leq x)\). The empirical (cumulative) distribution function is defined as

\[
\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\} ,
\]
where  

\[ 1\{X_i \leq x\} = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{otherwise} \end{cases}. \]

This is a natural estimator of \( F \), and it is essentially the CDF of a distribution that puts mass \( 1/n \) on each data point. The following are some properties of the empirical CDF.

- Note that, for a fixed point \( x \) the quantity \( n\hat{F}_n(x) \) has a binomial distribution with parameters \( n \) and success probability \( F(x) \). Therefore

\[
\mathbb{E} [\hat{F}_n(x)] = F(x) \quad \text{and} \quad \text{Var}(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}.
\]

Furthermore, using Chebyshev’s inequality we have

\[
\Pr(|\hat{F}_n(x) - F(x)| \geq \epsilon) \leq \frac{F(x)(1-F(x))}{n\epsilon^2},
\]

which implies that \( \hat{F}_n(x) \) converges in probability to \( F(x) \) as \( n \to \infty \). In particular the strong law of large numbers applies here and implies a stronger result, namely \( \hat{F}_n(x) \) converges to \( F(x) \) almost surely, for every fixed \( x \in \mathbb{R} \).

- Chebyshev’s inequality above is rather loose, and although it can be used to get pointwise confidence bounds these will be extremely conservative. In particular in light of the central limit theorem \( \Pr(|\hat{F}_n(x) - F(x)| \geq \epsilon) \) should scale roughly like \( e^{-\epsilon^2} \) instead of \( 1/\epsilon^2 \). A stronger concentration of measure inequality that can be applied in this setting is Hoeffding’s inequality, which implies that

\[
\Pr(|\hat{F}_n(x) - F(x)| \geq \epsilon) \leq 2e^{-2n\epsilon^2}.
\]

- The Glivenko-Cantelly theorem implies a much stronger convergence result, namely the convergence holds uniformly in \( x \):

\[
\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0.
\]

- Finally a uniform concentration inequality, the Dvoretzky-Kiefer-Wolfowith (DKW) inequality, holds in this case as well. For any \( \epsilon > 0 \) and any \( n > 0 \)

\[
\Pr(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \geq \epsilon) \leq 2e^{-2n\epsilon^2}.
\]

### 2.1 A confidence band for \( \hat{F}_n \)

Using the DKW inequality we can get a confidence band for \( \hat{F}_n \). Namely rewriting DKW we get

\[
\Pr(\forall x \in \mathbb{R} \quad |\hat{F}_n(x) - F(x)| \leq \epsilon) > 1 - 2e^{-2n\epsilon^2}.
\]

Equating \( \alpha = 2e^{-2n\epsilon^2} \), implying that \( \epsilon = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} \) we get

\[
\Pr \left( \forall x \in \mathbb{R} \quad |\hat{F}_n(x) - F(x)| \leq \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} \right) > 1 - \alpha.
\]
Taking into consideration that \( F(x) \in [0,1] \) we can get a slightly more refined result. Let

\[
L(x) = \max \left\{ \hat{F}_n(x) - \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}, 0 \right\},
\]

\[
U(x) = \min \left\{ \hat{F}_n(x) + \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}, 1 \right\},
\]

Then, for any CDF \( F \) and all \( n \)

\[
\Pr(L(x) \leq F(x) \leq U(x) \text{ for all } x) > 1 - \alpha.
\]

### 2.2 Confidence intervals for the distribution function at a fixed point

In this section we look at various ways to construct a confidence interval (CI) for \( p = F(x) \). Since \( Y = nF_n(x) \) has a Bin\((n,p)\) distribution, such an interval can be obtained from a confidence interval for a Binomial success probability, given observation \( Y \). There are many ways to do it, and below we list four possibilities.

1. **Exact (Clopper-Pearson):** Here we use the connection between the critical region of a point-null hypothesis and confidence intervals. The null hypothesis \( H_0 : p = p_0 \) is rejected for an observed \( y \) if

\[
\Pr(Y \geq y) \leq \alpha/2 \quad \text{or} \quad \Pr(Y \leq y) \leq \alpha/2,
\]

where \( Y \sim \text{Bin}(n,p_0) \). For the observed value \( y \), a CI for \( p \) is then given by the complement of the critical region, which is

\[
\{ p : P_p(Y \geq y) > \alpha/2 \text{ and } P_p(Y \leq y) > \alpha/2 \}.
\]

This interval is in general conservative, with coverage probability always greater or equal to \( 1 - \alpha \). However, due to the discreteness of \( Y \) exact coverage is not possible.

2. **Asymptotic (Wald):** Let \( \hat{p}_n = Y/n \). By the central limit theorem

\[
\sqrt{n} \frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \xrightarrow{D} N(0,1),
\]

where we use the symbol \( \xrightarrow{D} \) for “convergence in distribution”. We also know that \( \hat{p}_n \) converges almost surely (and hence in probability) to \( p \), therefore we can use Slutsky’s theorem to replace \( p \) in the denominator by \( \hat{p}_n \),

\[
\sqrt{n} \frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \xrightarrow{D} N(0,1).
\]

Isolating \( p \), we obtain the following CI for \( p \)

\[
\left[ \hat{p}_n - z_{\alpha/2} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + z_{\alpha/2} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \right],
\]

where \( z_{\alpha/2} \) denotes the upper \( \alpha/2 \)-quantile of the standard normal distribution.
3. **Asymptotic, using a variance stabilizing transformation**: Instead of estimating \( p \) in the denominator by \( \hat{p}_n \) one can also use a variance stabilizing transformation. Let \( \phi \) be a function that is differentiable at point \( p \) and has non-zero derivative (we will see what is a good choice for this function in a bit). By the \( \delta \)-method \(^1\)

\[
\sqrt{n}(\phi(\hat{p}_n) - \phi(p)) \xrightarrow{D} \mathcal{N}(0, p(1 - p)(\phi'(p))^2).
\]

Now take \( \phi \) so that \( \phi'(p) = 1/(\sqrt{p(1 - p)}) \). After some clever manipulation it is clear the function we want is \( \phi(x) = 2 \arcsin{\sqrt{x}} \). Therefore

\[
2\sqrt{n}(\arcsin{\sqrt{\hat{p}_n}} - \arcsin{\sqrt{p}}) \xrightarrow{D} \mathcal{N}(0,1).
\]

Since \( \phi \) is monotone, we obtain that the following approximate \( 1 - \alpha \) confidence interval

\[
\left[\sin^2\left(\arcsin{\sqrt{\hat{p}_n}} - \frac{z_{\alpha/2}}{2\sqrt{n}}\right), \sin^2\left(\arcsin{\sqrt{\hat{p}_n}} + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)\right].
\]

with probability approximately \( 1 - \alpha \).

4. **Wilson Method**: In the Wald method we made use of Slutsky’s theorem, but we don’t really need that extra step. Namely we can take the following approximate (asymptotic) equality

\[
\Pr\left(-z_{\alpha/2} \leq \sqrt{\frac{\hat{p}_n - p}{\hat{p}(1 - p)}} \leq z_{\alpha/2}\right) \approx 1 - \alpha.
\]

Solving for \( p \) in the above gives the desired confidence interval, which has endpoints

\[
\hat{p}_n + \frac{z_{\alpha/2}^2}{2n} \pm \frac{\hat{p}_n(1 - \hat{p}_n) + z_{\alpha/2}^2/(4n)}{1 + z_{\alpha/2}^2/n}.\]

5. **Hoeffding’s inequality**: One can also use Hoeffding’s inequality to construct a CI, in the same fashion as the DKW-based confidence band we computed above. Details are left to the reader.

To judge the actual performance of the various confidence intervals we should assess their coverage probability, that is, if the CI takes the form \([l(Y), u(Y)]\) we want to evaluate \( \Pr(p \in [l(Y), u(Y)]) \) for \( Y \sim \text{Bin}(n, p) \). This is in general not so easy, so we conduct a simulation study, where we instead look at the coverage fraction. For now, pick \( n = 10 \) and \( p = 0.6 \). We generated 1000 \( \text{Bin}(n, p) \) random variables. For each realization, we compute the Clopper-Pearson, Wald (asymptotic), Wilson and variance stabilizing 95\% confidence interval. Subsequently, we compute the fraction of times that the CIs contain the true parameter \( p = 0.6 \). This is the actual (observed) coverage fraction of the interval. Next, we repeat this procedure for values of \( p \) on a fine grid on \([0, 1]\) (keep \( n = 10 \)) and make a figure in which we plot the fraction of times that the observed coverage against \( p \) in \([0, 1]\).

We did the same for \( n = 25 \) and \( n = 100 \). The results are in figure 1. From these figures we

\(^1\)Let \( Y_n \) be a sequence of random variables and \( \phi : \mathbb{R} \to \mathbb{R} \) a map that is differentiable at \( \mu \) and \( \phi' \neq 0 \). Suppose \( \sqrt{n}(Y_n - \mu) \xrightarrow{D} \mathcal{N}(0, 1) \). Then \( \sqrt{n}(\phi(Y_n) - \phi(\mu)) \xrightarrow{D} \mathcal{N}(0, (\phi'(\mu))^2) \).
Figure 1: Coverage of confidence intervals for various parameters of a binomial distribution.
see that exact confidence intervals are conservative: the actual coverages are above 95%. Wilson confidence intervals tend to behave well, except near the boundaries \((p \approx 0 \text{ and } p \approx 1)\). For \(n\), large, variance stabilized asymptotic confidence intervals have good performance as well.

R code for producing these figures is in the file `confint_binom.r` available in the course website. Note that confidence intervals are available from the library `Hmisc` using the `binconf` command.

2.1 Exercise (Computer exercise) Data on the magnitudes of earthquakes near Fiji are available from R. Just type `quakes`. For help on this dataset type ?quakes. Estimate the distribution function. Find an approximate 95% confidence interval for \(F(4.9) - F(4.3)\) using a Wilson-type confidence interval. (Wasserman, exercise 11, chapter 2.)

2.2 Exercise Compare the coverage of the confidence intervals above with that of a Hoeffding type confidence interval (e.g., as described in chapter 1 of Wasserman). Make a picture similar to figure 1 that includes these CIs.

2.3 Confidence intervals for quantiles using order statistics

If \(X_1, \ldots, X_n\) is a sample from a continuous distribution \(F\), then the sample can be ordered uniquely. If \(X_{(i)}\) denotes the \(i\)-th order statistic, then

\[
X_{(1)} < X_{(2)} < \cdots < X_{(n)} .
\]

Some distributional properties of order statistics can be found for example in (Gibbons and Chakraborti (1992)). Here we follow section 2.8 of this book. A quantile of a continuous distribution of a random variable \(X\) is a real number which divides the area under the probability density function into two parts of specified amounts. Let the \(p\)-th quantile be denoted by \(q_p\) for all \(p \in (0, 1)\). In terms of the cumulative distribution function, then, \(q_p\) is defined in general to be any number which is the solution to the equation

\[
F(q_p) = p .
\]

If \(F\) is continuous and strictly increasing a unique solution exists. If there are flat pieces in the graph of \(F\), or jumps (i.e., \(X\) is not a continuous random variable), then it is customary to define

\[
F^{-1}(y) := \inf\{x : F(x) \geq y\} ,
\]

and define the \(p\)-th quantile as \(F^{-1}(p)\).

The quantile \(q_p\) can be estimated by \(\hat{F}_n^{-1}(p)\). If \(p \in (\frac{i-1}{n}, \frac{i}{n}]\), then \(\hat{F}_n^{-1}(p) = X_{(i)}\), the \(i\)-th order statistic \(^2\). Suppose that a confidence interval for \(q_p\) is desired for some specified value of \(p\). A logical choice for the confidence interval end points are two order statistics \(X_{(r)}\) and \(X_{(s)}\). For a given significance level \(\alpha\) we need \(r\) and \(s\) to be such that \((r < s)\)

\[
\Pr(X_{(r)} < q_p \leq X_{(s)}) \geq 1 - \alpha .
\]

\(^2\)Another common definition of a sample quantile \(\hat{q}_n(p)\) is based on requiring that the \(i\)th order statistic is the \(i/(n + 1)\) quantile. The computation of \(\hat{q}_n(p)\) runs as follows: \(\hat{q}_n(p) = x_{(k)} + \eta(x_{(k+1)} - x_{(k)})\) with \(k = \lfloor p(n + 1) \rfloor\) and \(\eta = p(n + 1) - k\) (\([x]\) denotes the integer part of \(x\)). Note that if \(p = 0.5\), this definition agrees with the usual definition of the sample median.
Now
\[ \{q_p > X(r)\} = \{X(r) < q_p \leq X(s)\} \cup \{q_p > X(s)\}, \]
and the two events on the right hand side are disjoint. Therefore,
\[ \Pr(X(r) < q_p \leq X(s)) = \Pr(X(r) < q_p) - \Pr(X(s) < q_p). \tag{1} \]
Therefore to construct a confidence interval we need to study the event \( \{X(r) \leq u\} \) for some \( u \in \mathbb{R} \). Let’s first begin by looking at the event \( \{X(r) \leq u\} \). This event holds if and only if at least \( r \) of the sample values \( X_1, \ldots, X_n \) are \( \leq u \). The number of samples values that are \( \leq u \),
\[ N = \sum_{i=1}^{n} 1\{X_i \leq u\}, \]
has a binomial distribution with success probability \( \Pr(X_i \leq u) = F(u) \). Thus
\[ \Pr(X(r) \leq u) = \Pr(N \geq r) = \sum_{i=r}^{n} \binom{n}{i} F(u)^i (1 - F(u))^{n-i}. \]
If \( F \) is continuous, then \( \lim_{\epsilon \downarrow 0} F(u - \epsilon) = F(u) \) implying that
\[ \Pr(X(r) < u) = \sum_{i=r}^{n} \binom{n}{i} F(u)^i (1 - F(u))^{n-i}. \]
Finally taking \( u = q_p = F^{-1}(p) \) and putting everything together in equation (1) we get
\[ \Pr(X(r) < q_p \leq X(s)) = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1 - p)^{n-i}. \]
Now \( r < s \) need to be chosen such that this probability is at least \( 1 - \alpha \). Dividing the mass \( \alpha/2 \) evenly over the right and left sides we see that we should choose \( r < s \) such that \( s - r \) is minimal and
\[ \Pr(B \geq s) \leq \alpha/2 \quad \text{and} \quad \Pr(B < r) \leq \alpha/2, \]
where \( B \) is a binomial random variable with parameters \( n \) and \( p \). Therefore
\[ r = \max\{k \in \{0, 1, \ldots, n - 1\} : \Pr(B \leq k) \leq \alpha/2\} + 1. \]
If such a value of \( k \) does not exist (since \( \Pr(B \leq 0) > \alpha/2 \)), then set \( X(r) = -\infty \) in the confidence interval. Similarly
\[ s = \min\{k \in \{0, 1, \ldots, n\} : \Pr(B \leq k) \geq 1 - \alpha/2\} + 1. \]
If \( s = n + 1 \) (since the minimum value for \( k \) equals \( n \)), then set \( X(s) = \infty \).

2.3 Exercise Write an R-function that implements the above confidence interval for a given dataset, confidence level and quantile \( p \). Simulate 10,000 times a sample of size 10 from a standard normal distribution and compute the coverage of the confidence interval with \( p = 0.5 \). Repeat for sample sizes 50, 100 and 1000.

2.4 Exercise A manufacturer wants to market a new brand of heat-resistant tiles which may be used on the space shuttle. A random sample of size \( m \) of these tiles is put on a test and the heat resistance capacities of the tiles are measured. Let \( X(1) \) denote the smallest of these measurements. The manufacturer is interested in finding the probability that in a future test
(performed by, say, an independent agency) of a random sample of \( n \) of these tiles at least \( k \) \((k = 1, 2, \ldots, n)\) will have a heat resistance capacity exceeding \( X_{(1)} \) units. Assume that the heat resistance capacities of these tiles follows a continuous distribution with cdf \( F \).

**a.** Show that the probability of interest is given by \( \sum_{r=k}^{n} a_r \), where

\[
a_r = \frac{mn!(r + m - 1)!}{r!(n + m)!}.
\]

**b.** Show that

\[
\frac{a_r}{a_{r-1}} = \frac{r + m - 1}{r}
\]

a relationship that is useful in calculating \( a_r \).

**c.** Show that the number of tiles, \( n \), to be put on a future test such that all of the \( n \) measurements exceed \( X_{(1)} \) with probability \( p \) is given by

\[
n = \frac{m(1 - p)}{p}.
\]

((Gibbons and Chakraborti (1992)), exercise 2.31)
3 Goodness-of-Fit Tests

Often times we have some data and want to test if a particular model (or model class) is a good fit. For instance, it is common to make normality assumptions for simplicity, but often it is necessary to check if these assumptions are reasonable. In Goodness-of-Fit (GoF) tests we strive to check the compatibility of the data with a fixed single model (simple GoF) or with a model in a given class (composite GoF).

Let \( X_1, \ldots, X_n \) be i.i.d. samples from an unknown distribution. If we wish to infer whether this sample comes from a certain hypothesized distribution \( F_0 \) this problem can be cast as the following hypothesis test:

\[
H_0 : F = F_0 \quad \text{vs.} \quad H_1 : F \neq F_0 .
\]

This is known as the simple Goodness-of-Fit (GoF)-problem. The composite GoF problem arises when we want to test whether the distribution of the sample belongs to a certain class \( \mathcal{F} \) of distribution functions. In this case we consider the testing problem

\[
H_0 : F \in \mathcal{F} \quad \text{vs.} \quad H_1 : F \notin \mathcal{F} .
\]

This null hypothesis is of composite type, which typically makes a formal analysis much more difficult. A typical application of such a test arises when we fit a linear model and want to check whether the normality assumption on the residuals is reasonable.

There are various approaches to GoF testing, and here we will focus on two of them: (i) procedures based on the empirical CDF; (ii) Chi-squared type tests. The Chi-square tests are an obvious choice when the hypothesized distributions (and data) are discrete, and the empirical CDF methods are very adequate for the continuous case.

3.1 Simple GoF using the Empirical CDF

The basic idea follows from the properties of the ECDF seen before. Under \( H_0 \), for \( n \) sufficiently large, \( \hat{F}_n \) is close to \( F_0 \). Recall that, under \( H_0 \), the Glivenko-Cantelli theorem tells us that

\[
\sup_t |\hat{F}_n(t) - F_0(t)| \xrightarrow{n \to \infty} 0 ,
\]

as \( n \to \infty \). Hence any discrepancy measure between \( \hat{F}_n \) and \( F_0 \) serves as a test statistic. A good statistic must be somewhat easy to compute and characterize. Here are some important examples.

- The Kolmogorov-Smirnov (KS) test statistic
  \[
  D_n := \sup_t |\hat{F}_n(t) - F_0(t)| .
  \]

- The Cramér-Von Mises (CvM) statistic
  \[
  C_n := \int (\hat{F}_n(t) - F_0(t))^2 dF_0(t) .
  \]

- The Anderson-Darling (AD) statistic
  \[
  A_n := \int \frac{(\hat{F}_n(t) - F_0(t))^2}{F_0(t)(1 - F_0(t))} dF_0(t) .
  \]
Although these expressions might look somewhat complicated they can be simplified significantly if $F_0$ is continuous. Note that $\hat{F}_n$ is piecewise constant and $F_0$ is a non-decreasing function, therefore the maximum deviation between $\hat{F}_n(t)$ and $F_0(t)$ must occur in a neighborhood of the points $X(i)$, and so

$$D_n = \max_{1 \leq i \leq n} \max \{ |\hat{F}_n(X(i)) - F_0(X(i))|, |\hat{F}_n(X(i)) - F_0(X(i))| \},$$

where $X(i) := X(i) - \epsilon$ for an arbitrarily small $\epsilon$ (there is a slight abuse of notation here). Now, taking into account that $F_0$ is continuous $F_0(X(i)) = F_0(X(i))$. Furthermore $\hat{F}_n(X(i)) = i/n$ and $\hat{F}_n(X(i)) = (i - 1)/n$ and therefore

$$D_n = \max_{1 \leq i \leq n} \max \{|i/n - F_0(X(i))|, |(i - 1)/n - F_0(X(i))| \}.$$

Finally define $U_i = F_0(X_i)$, and let $U(i)$ denote the corresponding order statistics (note that, since $F_0$ is monotone $U(i) = F_0(X(i))$). We can therefore write the above expression as

$$D_n = \max_{1 \leq i \leq n} \max \{|i/n - U(i)|, |(i - 1)/n - U(i)| \}.$$

In a similar fashion we get simplified expressions for the CvM and AD statistics:

$$nC_n = \frac{1}{12n} + \sum_{i=1}^{n} \left( U(i) - \frac{2i - 1}{2n} \right)^2, \quad (2)$$

and

$$nA_n = -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1) \left[ \log U(i) + \log(1 - U(n-i+1)) \right].$$

Now suppose that indeed we are working under $H_0$, which means the data $\{X_i\}$ came from the continuous distribution $F_0$. In that case $U_i$ are i.i.d. uniform random variables in $[0, 1]$ (see footnote\(^3\)) and therefore we conclude that, under $H_0$, the distribution of $D_n$, $C_n$ and $A_n$ does not depend on the underlying distribution $F_0$. In other words these statistics are distribution free under the null hypothesis. So, to devise simple GoF tests it suffices to study the case when the null hypothesis is the uniform distribution in $(0,1)$.

To use the above test statistics one needs to know the properties of their distribution. For small $n$ these have been tabulated, and for large $n$ we can use asymptotics. The analytical study requires some machinery of empirical processes theory, and is out of the scope of these notes. As mentioned before it suffices to study the case $F_0 = \text{Unif}(0,1)$. Let $U_i \sim \text{Unif}(0,1)$ and define $\hat{U}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{U_i \leq t\}$. Note that in this case

$$D_n \overset{D}{=} \sup_t \left| \hat{U}_n(t) - \Pr(U \leq t) \right| = \sup_{t \in [0,1]} \left| \hat{U}_n(t) - t \right|.$$

\(^3\)This result is known as the Probability Integral Transform: If $X$ has continuous distribution $F$ then $Y = F(X)$ has a uniform distribution supported in $(0,1)$. 

11
A well-known result from empirical processes theory states that the process $t \mapsto \sqrt{n}(\hat{U}_n(t) - t)$ converges in distribution to a process $B_0$, which is known as a standard Brownian Bridge on $[0,1]$. This is a Gaussian process defined for $t \in [0,1]$ with $\mathbb{E}[B_0(t)]$ and $\text{Cov}(B_0(s), B_0(t)) = \min\{s, t\} - st$. Now, with a bit of handwaving (a formal treatment requires the use of invariance principles) we have

$$\sqrt{n}D_n \xrightarrow{D} \sup_t |B_0(t)|,$$

as $n \to \infty$. Similarly

$$nC_n \xrightarrow{D} \int_0^1 B_0^2(t)dt \quad \text{and} \quad nA_n \xrightarrow{D} \int_0^1 \frac{B_0^2(t)}{t(1-t)}dt.$$

Fortunately the asymptotic distributions above can be studied analytically and we have

$$\lim_{n \to \infty} P_{F_0}(\sqrt{n}D_n \leq \lambda) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 \lambda^2},$$

$$\lim_{n \to \infty} P_{F_0}(nC_n > x) = \frac{1}{\pi} \sum_{j=1}^{\infty} (-1)^{j+1} \int_{(2j-1)^2 \pi^2}^{4j^2 \pi^2} \frac{-\sqrt{y}}{\sin(\sqrt{y})} e^{-xy/2} \frac{1}{y} dy.$$

Finally $nA_n \xrightarrow{D} A$, with

$$A = \sum_{j=1}^{\infty} \frac{Y_j}{j(j+1)},$$

where $Y_i \overset{i.i.d.}{\sim} \chi_1^2$.

### 3.2 Consistency under the alternative

A very important and pleasant property of the tests we’ve seen is that these are consistent under any alternative. This means that if the true distribution is not $F_0$ then eventually, as $n \to \infty$, we will reject the null hypothesis no matter what the true distribution is. Let’s see this in the case of the $KS$ statistic.

Let $G_n$ be the CDF of $\sqrt{n}D_n$ under $F_0$. That is

$$G_n(t) = P(\sqrt{n}D_n \leq t).$$

We reject the null hypothesis (with significance $\alpha$) if $\sqrt{n}D_n > G^{-1}_n(1 - \alpha)$, with $0 < \alpha < 1$. We will show the following result

#### 3.1 Lemma

If the data $\{X_i\}_{i=1}^{n}$ comes from a distribution $F \neq F_0$ then

$$P_F(\sqrt{n}D_n > G^{-1}_n(1 - \alpha)) \to 1,$$

as $n \to \infty$. 

12
Proof Since $F \neq F_0$ there is at least one point $a$ such that $F_0(a) \neq F(a)$. Now
\[
P_F(\sqrt{n}D_n > G_n^{-1}(1-\alpha)) = P_F(\sqrt{n} \sup_t |\hat{F}_n(t) - F_0(t)| > G_n^{-1}(1-\alpha))
\]
\[= P_F(\sqrt{n} \sup_t |\hat{F}_n(t) - F(t) + F(t) - F_0(t)| > G_n^{-1}(1-\alpha))
\]
\[\geq P_F(\sqrt{n} |\hat{F}_n(a) - F(a) + F(a) - F_0(a)| > G_n^{-1}(1-\alpha))
\]
\[\geq P_F(\sqrt{n} |F(a) - F_0(a)| - \sqrt{n} |\hat{F}_n(a) - F(a)| > G_n^{-1}(1-\alpha)) ,
\]
where the last step follows from $|x+y| \geq |x| - |y|$. Now note that the CLT implies $\sqrt{n}|\hat{F}_n(a) - F(a)| = O_P(1)$ (meaning $\forall \delta > 0 \ \exists \epsilon < \infty : P_F(\sqrt{n}|\hat{F}_n(a) - F(a)| \leq c) \geq 1-\delta$) and that $\sqrt{n}|F(a) - F_0(a)| \to \infty$. Therefore we conclude that $P_F(D_n > G_n^{-1}(1-\alpha))$ converges to one as $n \to \infty$. \hfill \Box

A similar argument applies also to $C_n$ and $A_n$.

3.2 Exercise Investigate the power of of the Cramér-Von Mises test for testing that a sample is Uniformly distributed over $(0,1)$. Use $\alpha = 0.01$. Consider the alternative $N(1/2,1/10)$. Consider $n = 10, 25, 100$. You may use Monte-Carlo simulation to find a critical value for the test statistic.

3.3 Exercise Verify the computational formula for $C_n$ in (2). To make your task easier you may make the extra assumption that $F_0$ is strictly increasing. (Note: The following fact might come in handy: $\sum_{i=1}^{n} (2i - 1)^2 = \frac{4n^3}{3} - \frac{n}{3}$).

3.3 Composite GoF tests

As alluded before, the composite GoF scenario is significantly more complicated. However, it is also more relevant from a practical standpoint, since we are hardly ever in the situation that we want to test e.g. if a sample is from an exponential distribution with parameter 2, or from a normal distribution with parameters 1.05 and 0.56. Often, the composite GoF-problem comes down to testing
\[
H_0 : F \in \{F_\theta : \theta \in \Theta\} \quad \text{vs.} \quad F \notin \{F_\theta : \theta \in \Theta\}.
\]

As an example $F_\theta$ may be the exponential distribution with mean $\theta$.

Perhaps the simplest idea to come to mind in this case is to compare the “best” distribution in the class with the empirical CDF. This can be done by estimating the parameter $\theta$ from the data (denote this estimator by $\hat{\theta}_n$) and comparing $\hat{F}_n$ with $F_{\hat{\theta}_n}$, where $F \in \mathcal{F}$. Therefore we end up with the following test statistics.
\[
D_n = \sup_t |\hat{F}_n(t) - F_{\hat{\theta}_n}(t)| \quad \text{or} \quad C_n = \int (\hat{F}_n(t) - F_{\hat{\theta}_n}(t))^2 dF_{\hat{\theta}_n}(t) ,
\]
and similarly for the AD statistic. Important remark: Plugging in a parameter estimate affects the distribution of these statistics, and the distribution under the null will be heavily influenced by the type of estimator you use. Therefore one can no longer use the distributions derived in the previous session. Practitioners often mistakenly plug in $\hat{\theta}_n$ and subsequently use an ordinary KS-test or CvM-test. This will result in inadequate testing procedures. Some tests are specifically designed for GoF with estimated parameters.
3.4 Example The Lilliefors test (1967) is an adaptation of the KS-test for which the null hypothesis equals $H_0 : X_1, \ldots, X_n$ is a sample from a normal distribution with unknown parameters. The unknown (population) mean and variance are estimated by the sample mean and sample variance. The distribution of this statistic under the null has been tabulated (by Monte-Carlo simulation).

3.5 Example The Jarque-Bera test (1980) is a test for normality that is especially popular in the econometrics literature. This test is based on the sample kurtosis and skewness.

$$
\text{skewness} \quad b_1 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^3
$$

$$
\text{kurtosis} \quad b_2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^4
$$

Under normality $\sqrt{n}b_1 \xrightarrow{D} N(0, 6)$ and $\sqrt{n}(b_2 - 3) \xrightarrow{D} N(0, 24)$. The Jarque-Bera statistic is defined by

$$
\text{JB} = n(b_1^2/6 + (b_2 - 3)^2/24)
$$

Its limiting distribution is Chi-squared with 2 degrees of freedom.

3.6 Example The Shapiro-Wilk test is another powerful test for normality. The test statistic is

$$
W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2} \quad (\in (0, 1]),
$$

where the weights $a_1, \ldots, a_n$ are specified by an adequate formula. Under $H_0$, the numerator is an estimator for $(n-1)\sigma^2$, whereas the denominator is also an estimator for $(n-1)\sigma^2$. Hence, under $H_0$, $W \approx 1$. Under $H_1$, the numerator is tends to be smaller. Therefore, we reject the null hypothesis for small values of $W$.

3.7 Example A simulation study to assess the performance of tests for normality. We compute the fraction of times that the null hypothesis of normality is rejected for a number of distributions (in total we simulated 1000 times).

<table>
<thead>
<tr>
<th>Results for n= 20</th>
<th>norm</th>
<th>cauchy</th>
<th>exp</th>
<th>t5</th>
<th>t10</th>
<th>t15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro</td>
<td>0.054</td>
<td>0.852</td>
<td>0.852</td>
<td>0.182</td>
<td>0.095</td>
<td>0.080</td>
</tr>
<tr>
<td>KS</td>
<td>0.038</td>
<td>0.206</td>
<td>1.000</td>
<td>0.067</td>
<td>0.050</td>
<td>0.046</td>
</tr>
<tr>
<td>AD</td>
<td>0.043</td>
<td>0.863</td>
<td>0.799</td>
<td>0.166</td>
<td>0.092</td>
<td>0.074</td>
</tr>
<tr>
<td>CvM</td>
<td>0.050</td>
<td>0.864</td>
<td>0.751</td>
<td>0.157</td>
<td>0.081</td>
<td>0.070</td>
</tr>
<tr>
<td>JB</td>
<td>0.025</td>
<td>0.807</td>
<td>0.516</td>
<td>0.162</td>
<td>0.067</td>
<td>0.060</td>
</tr>
</tbody>
</table>
Results for n= 50
norm  cauchy  exp  t5  t10  t15
Shapiro  0.065  0.994  1.000  0.360  0.152  0.100
KS       0.062  0.472  1.000  0.066  0.045  0.054
AD       0.055  0.994  0.998  0.289  0.123  0.073
CvM      0.055  0.738  0.989  0.249  0.113  0.070
JB       0.043  0.993  0.953  0.396  0.172  0.106

Results for n= 200
norm  cauchy  exp  t5  t10  t15
Shapiro  0.054  1.000  1.000  0.825  0.362  0.223
KS       0.044  0.999  1.000  0.084  0.058  0.047
AD       0.052  NA     NA     NA     0.258  0.136
CvM      0.044  0.003  0.981  0.689  0.213  0.107
JB       0.049  1.000  1.000  0.869  0.436  0.291

Results for n= 5000
norm  cauchy  exp  t5  t10  t15
Shapiro  0.056  1     1     1.000  1.000  0.997
KS       0.047  1     1     1.000  0.693  0.205
AD       0.058  NA    NA    NA    NA    0.989
CvM      0.061  0     0     0.067  1.000  0.962
JB       0.057  1     1     1.000  1.000  1.000

The R-code for obtaining this results is in the file “compare_gof_tests_normality.r”. The AD test implementation appears to have some problems for large sample sizes. Although most textbooks only treat the KS-test/Lilliefors test, from this simulation study it appears that this is a rather poor testing procedure in practice. The JB and Shapiro-Wilk seem to work significantly better when testing normality. (D’Agostino and Stephens (1986)) warn

... for testing for normality, the Kolmogorov-Smirnov test is only a historical curiosity. It should never be used. It has poor power in comparison to specialized tests such as Shapiro-Wilk, D’Agostino-Pearson, Bowman-Shenton, and Anderson-Darling tests.

As can be seen from this quote, there are many more specialized Gof-tests.

3.8 Exercise It seems that the Crámer - Von Mises test fails to reject data from a standard exponential distribution (there appears a zero in the table, where one would expect a number close to one). There is an error in the function cvm.test within the “nortest” library. Compute the correct value. You should use the function cvm.test.asymp which is posted on the webpage for this course.

3.3.1 A Bootstrap approach for composite Gof tests

When tables and asymptotic distributional results are not available one has to resort to simulation-based techniques to approach the testing problem. The following method, studied by (Stute et al. (1993)), is an application of the (parametric) bootstrap and under some weak conditions it results in accurate asymptotic approximations (in n) of the p-values.
X_{1}, \ldots, X_{n} \text{ i.i.d. samples from some unknown distribution.}

(i) Estimate \( \hat{\theta}_{n} \) from \( X_{1}, \ldots, X_{n} \) and construct the CDF \( \hat{F}_{\hat{\theta}_{n}} \).

(ii) Evaluate \( \bar{D}_{n}, \bar{C}_{n} \) or \( \bar{A}_{n} \). For the purpose of illustration let’s use the KS statistic in what follows,

\[
\bar{D}_{n} = \sup_{t} |\hat{F}_{n}(t) - F_{\hat{\theta}_{n}}(t)|.
\]

(iii) Generate \( B \in \mathbb{N} \) bootstrap samples of size \( n \) from \( F_{\hat{\theta}_{n}} \). Denote these \( B \) samples by \( \{X_{1,j}^{*}, \ldots, X_{n,j}^{*}\} \). The number of bootstrap samples \( B \) should be large to ensure a good approximation.

(iv) Compute \( \bar{D}_{j}^{*}, \bar{C}_{j}^{*} \) or \( \bar{A}_{j}^{*} \) as follows (example for the KS statistic)

\[
\bar{D}_{j}^{*} = \sup_{t} |\hat{F}_{j}^{*}(t) - F_{\hat{\theta}_{n}^{*}}(t)|,
\]

where \( \hat{\theta}_{n}^{*} \) is the estimate of \( \theta \) obtained using \( \{X_{1,j}^{*}, \ldots, X_{n,j}^{*}\} \) and \( \hat{F}_{j}^{*} \) is the empirical CDF of \( \{X_{1,j}^{*}, \ldots, X_{n,j}^{*}\} \).

Now, if a test with significance \( \alpha \) is desired reject \( H_{0} \) if

\[
\bar{D}_{n} > \bar{D}_{(B(1 - \alpha) + 1)}^{*},
\]

where \( \bar{D}_{(B(1 - \alpha) + 1)}^{*} \) denoted the \( (B(1 - \alpha) + 1) \) order statistic of \( \{\bar{D}_{1}^{*}, \ldots, \bar{D}_{B}^{*}\} \). This is just an estimate of the \( t : P_{H_{0}}(\bar{D}_{n} > t) = \alpha \).

Alternatively you can compute an approximate \( p \)-value using

\[
p \approx \frac{\# \{i : \bar{D}_{i}^{*} \geq \bar{D}_{n} \}}{B},
\]

and reject \( H_{0} \) if \( p < \alpha \).

3.4 Chi-Square-type GoF tests

This is a simple approach to GoF for both discrete and continuous random variables. It has several advantages, namely

- Suitable for both continuous and discrete settings
- Easy to use, even in high dimensions (so far we have been discussing only the one-dimensional setting).

However, there is a drawback: for continuous random variables the procedure requires some arbitrary choices (that must be done before seeing any data). As a consequence some information is lost, and these tests no longer have the property of being consistent against any alternative.

First consider the simple GoF-problem. Suppose \( S = \text{supp}(F_{0}) \). Fix \( k \) and let

\[
S = \bigcup_{i=1}^{k} A_{i}.
\]
be a partition of $S$ (meaning the sets $A_i$ are disjoint). For discrete random variables there is a natural choice for such a partition. For continuous random variables a partition can be obtained by forming appropriate cells using some knowledge about $F_0$. Usually one chooses the number of cells to satisfy $5 \leq k \leq 15$. It is often hard to fully justify a certainly chosen partition or value for $k$.

Define $F_i$ to be the observed frequency in cell $A_i$

$$F_i = \# \{ j : X_j \in A_i \} .$$

Under $H_0$, $e_i := \mathbb{E}_0 F_i = n P_0(X \in A_i)$ (the subscript emphasizes that the expectation and probability have to be computed under the null hypothesis). Under $H_0$ we expect $e_i$ and $F_i$ to be close. Any discrepancy measure between these two quantities can serve as a basis for a test statistic. In particular we defined the chi-square statistic as

$$Q = \sum_{i=1}^{k} \frac{(F_i - e_i)^2}{e_i} .$$

It is not hard to show that $Q$ converges to a $\chi^2$-distribution with $k - 1$ degrees of freedom. As a rule of thumb, this approximation is reasonable if all the expected cell frequencies $e_i$ are at least 5.

Instead of a Chi-square test, a likelihood-ratio test can be used. Given the partition of the data in classes, the vector $(F_1, \ldots, F_k)$ has a multinomial distribution with parameters $\theta = (\theta_1, \ldots, \theta_k)$. Under $H_0$, $\theta_i = e_i/n$. The likelihood $L(\theta_1, \ldots, \theta_k)$ is therefore proportional to

$$\prod_{i=1}^{k} \theta_i^{F_i}, \quad F_i \in \{0, 1, \ldots, n\}, \quad \sum_{i=1}^{k} F_i = n, \quad \sum_{i=1}^{k} \theta_i = 1 .$$

The Likelihood-Ratio (LR) statistic is thus given by

$$LR = \frac{L(e_1, \ldots, e_k)}{\sup_{\theta} L(\theta)} = \frac{\prod_{i=1}^{k} (e_i/n)^{F_i}}{\prod_{i=1}^{k} (F_i/n)^{F_i}} = \prod_{i=1}^{k} \left( \frac{e_i}{F_i} \right)^{F_i} ,$$

since the maximum likelihood estimator for $\theta_i$ is given by $F_i/n$. It is not hard to show that, asymptotically, the Chi-square and LR-test are equivalent. This can be done by comparing $-2 \log LR$ with $Q$, and is left as an exercise to the reader.

### 3.4.1 Composite $\chi^2$-GoF tests

In the composite GoF case things get more complicated as you one might expect. In particular the expected cell frequencies have to be estimated from the data. By plugging-in these estimators the distribution of $Q$ under the null is going to change. The properties of that distribution depend on the properties of the estimator. If the estimators are chosen using a maximum likelihood principle then, under some mild assumptions, the resulting limiting distribution will be chi-square with $k - 1 - s$ degrees of freedom, where $s$ is the number of independent parameters that must be estimated for calculating the estimated expected cell frequencies. So in case we test for normality, $s = 2$ (mean and variance).
3.9 Exercise  A quality control engineer has taken 50 samples of size 13 from a production process. The number of defectives for these samples are recorded below.

<table>
<thead>
<tr>
<th>Number of defects</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6 or more</td>
<td>0</td>
</tr>
</tbody>
</table>

(i) Suppose we want to check if a Poisson distribution is a good model for the number of defects in each sample. If \( X_i \) denotes the number of defects in the \( i \)-th sample, then we henceforth assume that \( X_1, \ldots, X_n \) is a random sample from a Poisson distribution, say with parameter \( \mu \). Show that the maximum likelihood estimate for the \( \mu \) is given by \( \hat{\mu} = 1.6 \).

(ii) Under the assumption of a Poisson distribution, the expected number of samples that have \( k \) defects (\( k = 0, 1, \ldots \)) can be estimated by \( \hat{E}_k = 50e^{-\hat{\mu}}\hat{\mu}^k/(k!) \). Compute the value of the test statistic

\[
Q = \sum_{i \in \mathcal{I}} \frac{(F_i - \hat{E}_i)^2}{\hat{E}_i},
\]

Here \( \mathcal{I} \) denotes the 7 classes of number of defects, i.e. \{0, 1, 2, \ldots, 5, 6 or more\} defects. \( F_i \) denotes the observed count in class \( i \).

(iii) Asymptotically, under the null hypothesis, \( Q \) has a \( \chi^2 \)-distribution with \( 7 - 1 - 1 = 5 \) degrees of freedom (we lose one degree of freedom for estimating \( \mu \)). Check whether the Chi-square test rejects the null hypothesis that the data are a random sample from a Poisson distribution. Use significance level \( \alpha = 0.05 \).

3.5 Probability Plotting and Quantile-Quantile Plots

Probability plots provide a visual way to do empirically do GoF tests. These are not formal tests, but provide a quick tool to check if a certain distributional assumption is somewhat reasonable.

Let \( \mathcal{F} \) be a location-scale distribution family, that is

\[
\mathcal{F} = \{ F_{a,b} : a \in \mathbb{R}, b > 0 \},
\]

for some distribution \( F \). In the above \( F_{a,b} \) is the CDF of \( a + bX \) when \( X \sim F \), that is

\[
F_{a,b}(x) = F \left( \frac{x-a}{b} \right).
\]

As an example, a \( \mathcal{N}(\mu, \sigma^2) \) random variable is obtained from a standard normal random variable \( Z \) by the linear transformation \( Z \rightarrow \mu + \sigma Z \).

Let \( X_1, \ldots, X_n \) be data from some distribution. Recall that \( \hat{F}_n(X_{(i)}) = i/n \). Therefore

\[
F_{a,b}^{-1} \left( \hat{F}_n(X_{(i)}) \right) = F_{a,b}^{-1} \left( \frac{i}{n} \right).
\]
Now, if the data comes from a distribution in $\mathcal{F}$ then $\hat{F}_n(X(i)) \approx F_{a,b}(X(i))$, and so

$$X(i) \approx F_{a,b}^{-1}\left(\frac{i}{n}\right) = a + bF^{-1}\left(\frac{i}{n}\right).$$

Said differently, if the data comes from a distribution in $\mathcal{F}$ we expect the points $\left(X(i), F^{-1}\left(\frac{i}{n+1}\right)\right)$ to lie approximately in a straight line. Note that we replaced $i/n$ by $i/(n+1)$: this is to ensure that we stay away from evaluating $F^{-1}(1)$, which can be infinite. The plot of the above points is commonly called the quantile-quantile plot.

The use of probability plots requires some training, but these are very commonly used and helpful. If we want to test for a distribution $F_{\theta}$ that is not in a location-scale family, then the preceding reasoning implies that the points $\left(F_{\theta}^{-1}\left(\frac{i}{n+1}\right), x(i)\right)$ should be on a straight line if $\theta$ is known. If $\theta$ is unknown, an estimator for it can be plugged in. Most software packages can generate such plots automatically. As pointed out in (Venables and Ripley (1997)) (page 165), a QQ-plot for e.g. a $t_{15}$-distribution can be generated by executing the following code in R:

```r
plot(qt(ppoints(x),9), sort(x))
```

(We assume the data are stored in a vector `x`). Adding the command `qqline(x)`, produces a straight line through the lower and upper quartiles. This helps assess whether the points are (approximately) on a straight line. You may also want to consider the function `qq.plot` in the library `car`, which gives a direct implementation of the QQ-plot.

3.10 Example Comparing QQ-normality plots and AD-test for normality. We simulate samples of size 10, 50, 100, 1000, 5000, 10000 from a standard normal and $t_{15}$ distribution. Figures 3.10 and 3 show QQ-normality plots for both cases respectively. Reading these figure from upper-left to lower-right, the corresponding AD $p$-values are:

<table>
<thead>
<tr>
<th>normal</th>
<th>$t_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.671</td>
<td>0.021</td>
</tr>
<tr>
<td>0.278</td>
<td>0.815</td>
</tr>
<tr>
<td>0.493</td>
<td>0.381</td>
</tr>
<tr>
<td>0.925</td>
<td>0.001</td>
</tr>
<tr>
<td>0.186</td>
<td>0.001</td>
</tr>
<tr>
<td>0.339</td>
<td>9.98e-08</td>
</tr>
</tbody>
</table>

For almost every purpose in practice, the difference between a $t_{15}$ and a Normal distribution is of no importance. However, as we obtain sufficiently many data, hypothesis tests will always detect any fixed deviation from the null hypothesis, as can be seen very clearly from the computed $p$-values. The R-code for producing these figures is in the file `compare_qq_testing.r`.

For large datasets there are difficulties with the interpretation of QQ-plots, as indicated by the following theorem.

3.11 Theorem Let $r_n$ be the correlation coefficient of the pairs $\left(X(i), \Phi^{-1}\left(\frac{i}{n+1}\right)\right)$, where $\Phi$ is the distribution function of the standard normal distribution. Let $F$ be the true CDF with
Figure 2: QQ-plots for samples of sizes 10, 50, 100, 1000, 5000, 10000 from a standard normal distribution. The upper-left figure is for sample size 10, the lower-right is for sample 10000.

Figure 3: QQ-plots for samples of sizes 10, 50, 100, 1000, 5000, 10000 from a $t_{15}$ distribution. The upper-left figure is for sample size 10, the lower-right is for sample 10000.
Table 1: Limiting values for $r_n$

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$\rho_F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>1</td>
</tr>
<tr>
<td>uniform</td>
<td>0.98</td>
</tr>
<tr>
<td>double exp</td>
<td>0.98</td>
</tr>
<tr>
<td>$t_3$</td>
<td>0.90</td>
</tr>
<tr>
<td>$t_5$</td>
<td>0.98</td>
</tr>
<tr>
<td>$\chi^2_3$</td>
<td>0.96</td>
</tr>
<tr>
<td>exponential</td>
<td>0.90</td>
</tr>
<tr>
<td>logistic</td>
<td>0.97</td>
</tr>
</tbody>
</table>

variance $\sigma^2$, then

$$\lim_{n \to \infty} r_n = \frac{1}{\sigma} \int_0^1 F^{-1}(x)\Phi^{-1}(x)dx =: \rho_F \quad a.s.$$  

See theorem 26.10 in (DasGupta (2008)). Table 1 provides values for the limiting value for various distribution. We conclude that asymptotically we get a perfect straight line in case of normality (as it should be). However, for many other distributions we obtain a correlation coefficient that is very close to one. It is hard to distinguish a set of points with correlation coefficient 0.97 from a set with correlation coefficient equal to 1 with the human eye. The difference is mainly in the tails. For small sample sizes, probability plots help to assess whether normality holds approximately, which is often all we need (for example in assessing approximate normality of residuals of linear models).
4 Permutation tests

The material for this topic is the article (Ernst (2004)), that is available in the course website. In this section we briefly describe how to implement simple permutation tests in R. We focus on making the main ideas clear. However, there are specialized packages to perform such tests in a very efficient way.

4.1 Example Byzantine coins. This is example 15.6 in (Kvam and Vidakovic (2007)). (Hendy and Charles (1970)) investigate the silver content (%Ag) of a number of Byzantine coins discovered in Cyprus. The coins are from the first and fourth coinage in the reign of King Manuel I, Commenus (1143-1180). First we visualize the data, using a stripplot.

```r
c1<-c(5.9, 6.8, 6.4, 7.0, 6.6, 7.7, 7.2, 6.9, 6.2)
c4<-c(5.3, 5.6, 5.5, 5.1, 6.2, 5.8, 5.8)
coindata<-data.frame(coinage=c(c1,c4),
                    type=c(rep("first",9),rep("fourth",7)))
stripchart(coinage~type,data=coindata,cex=1.5,pch=16,
          main='Stripchart of Byzantine Coins')
```

In this case let $X$’s denote the data relative to the first coinage and $Y$’s the data pertaining the fourth coinage, therefore $m = 9$, $n = 7$. The data are shown in figure 4 (left). We use a permutation test to see if the silver content of the coins was significantly different in the later coinage. The total number of permutation equals $\binom{16}{9} = 11440$. We choose the test statistic $T = \bar{Y} - \bar{X}$, and large or small values of $T$ indicate violation of $H_0$. For each of the 11440 permutations, we calculate the test statistic. An histogram showing the values of $T$ of all permutations (under $H_0$) is shown in figure 4 (right). We used the following code. This code merely serves as an easy to read example. This can be coded much more efficiently.

```r
twosample.perm.test<-function(x,y)
{
  z<-c(x,y)
  m<-length(x)
  n<-length(y)
  N<-n+m
  mat<-combn(N,m)

  poss<-ncol(mat)
  t<-teststat(x,y)
  T<-numeric(poss)
  for ( i in 1:poss)
  {
    x<-z[mat[,i]]
    y<-z[-mat[,i]]
    T[i]<-teststat(x,y)
  }
  return(list(T=T,t=t))
}
```
Figure 4: Left: stripplot of byzantine coin data. Right: histogram of test statistic $T$ for all permutations

teststat<-function(x,y) mean(y)-mean(x)

result<-twosample.perm.test(c1,c4)

hist(result$T,50, main='Histogram of T under H_0');abline(v=result$t, col='red', lwd=2);

p_value<-mean(abs(result$T-mean(result$T))>=abs(result$t-mean(result$T)))
cat('P-value for the test ',p_value,'
')

The observed statistic is 1.13. The p-value is computed as described in (Ernst (2004)), as we have a two-sided alternative. This results in a p-value that equals 0.000699, so we are quite confident we can reject the null hypothesis.

The conditional inference (coin) package can also do the calculations (and much faster) using the following code.

```
oneway_test(coinage~type,data=coindata,distribution="exact")
```

This package offers much more functionality for computing p-values for permutation tests. However, it takes some study to understand what’s going on and what user input is necessary.

Obviously, for large values of $n$ and $m$ exhaustively computing all permutations is unfeasible, even using clever tricks. Since all that is needed to compute the p-values is to evaluate probabilities under the null hypothesis, and it is easy to generate samples from the permutation distribution we can use Monte-Carlo methods. The idea is to generate $B$ samples from the permutation distribution and approximate the p-value by its sample counterpart.

$$
\hat{p} = 1 + \frac{1}{B+1} \sum_{i=1}^{B} 1\{|t_i - \frac{1}{B} \sum_{j=1}^{B} t_j| \geq |t - \frac{1}{B} \sum_{j=1}^{B} t_j|\}.
$$

We can even compute a CI for $\hat{p}$ since it is the proportion of successes in $B+1$ independent trials with success probability $p$. Putting all these together allows us to get a conservative p-value in the end.

A Monte-Carlo approximation is easily obtained using the boot-package.
library(boot)
T.fun<-function(data,i)
{
  d<-data[i]
  mean(d[10:16])-mean(d[1:9])
}
coins<-c(c1,c4)
B<-10000
perm.coin<-boot(coins,T.fun,R=B,sim="permutation")
# use argument permutation to sample without replacement
plot(perm.coin)
p.val<-(1+sum(abs(perm.coin$t-mean(perm.coin$t))>=
  abs(perm.coin$t0-mean(perm.coin$t))))/(B+1)
cat('Observed value of the test-statistic', perm.coin$t0,'
')
cat('approximate p-value equals', p.val,'

4.1 Final Remarks
It is important to notice that not all statistics can be used in permutation methods. For example, consider the studentized statistic

$$T_{m,n} = \frac{\bar{X}_m - \bar{Y}_n}{\sqrt{S^2_X/m + S^2_Y/n}},$$

which is the $t$-statistic for comparing the difference in means when sampling from independent normal distributions with variances that are unknown and not necessarily equal. If $X$ and $Y$ have a different variance under the null hypothesis, then the distribution of the combined sample of $X$ and $Y$ values is not invariant under permutations, and for this reason using $T_{m,n}$ violates a crucial assumption needed for permutation tests.

4.2 Exercise Seven patients each underwent three different methods of kidney dialysis. The following values were obtained for weight change in kilograms between dialysis sessions.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Treatment A</th>
<th>Treatment B</th>
<th>Treatment C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.90</td>
<td>2.97</td>
<td>2.67</td>
</tr>
<tr>
<td>2</td>
<td>2.56</td>
<td>2.45</td>
<td>2.62</td>
</tr>
<tr>
<td>3</td>
<td>2.88</td>
<td>2.76</td>
<td>1.84</td>
</tr>
<tr>
<td>4</td>
<td>2.73</td>
<td>2.20</td>
<td>2.33</td>
</tr>
<tr>
<td>5</td>
<td>2.50</td>
<td>2.16</td>
<td>1.27</td>
</tr>
<tr>
<td>6</td>
<td>3.18</td>
<td>2.89</td>
<td>2.39</td>
</tr>
<tr>
<td>7</td>
<td>2.83</td>
<td>2.87</td>
<td>2.39</td>
</tr>
</tbody>
</table>

We wish to test if there is a significant difference in the mean weight change among the three methods. Let $\bar{A}$, $\bar{B}$ and $\bar{C}$ denote the mean weight change in each group. Define $M = (\bar{A} + \bar{B} + \bar{C})/3$. A reasonable statistic to consider is

$$T = (\bar{A} - M)^2 + (\bar{B} - M)^2 + (\bar{C} - M)^2.$$
(a) Show that the test statistic $S = \bar{A}^2 + \bar{B}^2 + \bar{C}^2$ is equivalent to $T$, in the sense that it results in the same testing procedure.

(b) Test the null hypothesis that there is no difference in mean weight change among treatments. Notice that now you can permute the 21 patients between 3 different groups, so there are $21!/(3 \times 7!) = 399072960$ possibilities. Use Monte-Carlo simulation to approximate the $p$-value of the permutation test.
5 Tests based on Signs, Runs and Ranks

5.1 Introduction

Statistical models are always, at best, an approximation. Methods that make few assumptions about the data are therefore desirable. However, there is no free lunch: gaining robustness by dropping assumptions always comes at the cost of a loss of efficiency. Notably, this loss can be quite modest.

5.1 Example The $t$-confidence interval for the mean is given by

$$C_n := [\bar{X}_n - t_{\alpha/2,n-1} \frac{S_n}{\sqrt{n}}, \bar{X}_n + t_{\alpha/2,n-1} \frac{S_n}{\sqrt{n}}].$$

Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, then $\Pr(\mu \in C_n) = 1 - \alpha$ for all $n, \mu$ and $\sigma$. That is, for normal data this interval is exact. By the central limit theorem, for each distribution $F$ for which $E_F X^2 < \infty$,

$$\lim_{n \to \infty} P_F(\mu_F \in C_n) = 1 - \alpha,$$

showing that this interval is asymptotically exact. However, it can be shown that if $\mathcal{F}$ denotes the set of all distribution functions with finite second moment, then

$$\lim_{n \to \infty} \inf_{F \in \mathcal{F}} P_F(\mu_F \in C_n) = 0.$$

This shows things can go really bad, and parametric procedures can fail dramatically under completely nonparametric models. Below we introduce robust non-parametric testing procedures that address these problems by making very few assumptions about the data.

5.2 Sign test

This is a simple alternative to the one-sample $t$-test, this time targeting the median of a distribution. Let $X_1, \ldots, X_n$ be an i.i.d. random sample from a distribution with median $M$. Assume this median is uniquely defined, in particular assume $\Pr(X_i = M) = 0$. We wish to test

$$H_0 : M = M_0 \text{ against } H_1 : M > M_0.$$

We will also consider the two-sided alternative $H_1 : M \neq M_0$. A suitable test statistic is just

$$T = \sum_{i=1}^n 1\{X_i > M_0\}.$$

Under $H_0$ $T \sim \text{Bin}(n, 1/2)$. If testing against the one-sided alternative $H_1 : M > M_0$ we should reject the null hypothesis is $T$ is large, therefore the rejection region is of the form $[c_\alpha, \infty)$ where $c_\alpha$ is the smallest integer such that

$$P_{H_0}(T \geq c_\alpha) = \sum_{k=c_\alpha}^n \binom{n}{k} 2^{-n} \leq \alpha.$$
The p-value for such a test is given simply by

\[ P_{H_0}(T \geq t) = 2^{-n} \sum_{k=t}^{n} \binom{n}{k} . \]

If the alternative is two-sided \( H_1 : M \neq M_0 \) then the rejection region is of the form \( (\infty, c_{\alpha/2}] \cup [c'_{\alpha/2}, \infty) \), where

\[ 2^{-n} \sum_{k=0}^{c_{\alpha/2}} \binom{n}{k} \quad \text{and} \quad 2^{-n} \sum_{k=c'_{\alpha/2}}^{n} \binom{n}{k} , \]

and \( c_{\alpha/2} \) and \( c'_{\alpha/2} \) are respectively the largest and smallest integers such that the above inequalities hold. Obviously \( n = c_{\alpha/2} + c'_{\alpha/2} \). The p-value for this test is simply given by

\[ 2 \min \left\{ \sum_{k=t}^{n} \binom{n}{k} 2^{-n}, \sum_{k=0}^{t} \binom{n}{k} 2^{-n} \right\} = 2^{1-n} \sum_{k=0}^{t \wedge (n-t)} \binom{n}{k} . \]

Finally, If \( n \) is large, binomial probabilities can be approximated by normal probabilities, due to the central limit theorem. More precisely, asymptotically \( T \sim \mathcal{N}(n/2, n/4) \).

Note that this median test also allows you to get confidence intervals for the median, yielding the same results as in the end of Section 2.

**Remark:** Zero differences occur if \( X_i = M_0 \) for at least one \( i \). It is common practice to ignore such observations and adjust \( n \) accordingly. An alternative is to count half of the zero differences as positive differences. Yet another possibility of handling the zeros is to assign to all zeros that sign which is least conductive to rejection of \( H_0 \); this is a strictly conservative approach. Is this notes however we will assume there are no zero-differences for simplicity.

### 5.2.1 Paired sample procedures

Suppose we observe the i.i.d. pairs

\[(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) . \]

In many applications one would like to test that \( Y \) is "larger" than \( X \) or vice versa. Similar as in the paired one sample t-test, we first form the differences

\[ D_i = X_i - Y_i . \]

Next assume that \( D_1, \ldots, D_n \) is a sample from a continuous distribution with (unique) median \( M \). We can use the sign-test to test the hypothesis that the median of the differences equals zero \( (H_0 : M = 0) \). This is not the same as testing that the difference of the medians of \( X \) and \( Y \) equals zero. In general \( M_D \neq M_X - M_Y \).

**5.2 Exercise** Give an example of two continuous distributions for \( X \) and \( Y \) such that \( M_D \neq M_X - M_Y \), where \( D = X - Y \).
5.3 Ranks

Sign tests provide a very simple, but crude approach to testing. Note that all the magnitude information is being ignored, and even the relative ordering of the samples (which provides quite a lot of information) is completely neglected. This is too dramatic, and we should hope to do better by retaining some of this information.

The rank of $X_i$ within the random sample $X_1, \ldots, X_n$ is given by

$$R(X_i) = \sum_{j=1}^{n} 1\{X_i - X_j \geq 0\} .$$

That is, the rank of $X_i$ is the number of elements in the sample that are smaller or equal to $X_i$. Therefore the rank assigns an integer number to the ordered sample. If there are no repetitions in the sample than the sample element $X_i$ corresponding to $X_{(1)}$ is assigned rank $R(X_i) = 1$ and so on.

**Important:** Note that the notation for the rank is slightly incomplete, as it doesn’t explicitly specify the complete sample used for the rank calculation. For this reason it is important to clearly state with respect to which sample is the rank computed upon, even if relatively clear from the context.

If the $X_j$ is a sample from a continuous distribution, then the random variable $R(X_i)$ follows a discrete uniform distribution on $\{1, 2, \ldots, n\}$ and so

$$\Pr(R(X_i) = j) = \frac{1}{n}, \quad \text{for} \quad j = 1, \ldots, n .$$

Allthough $X_1, \ldots, X_n$ are independent, $R(X_1), \ldots, R(X_n)$ are of course not. The first and second moments of $R(X_i)$ are given by

$$\mathbb{E}[R(X_i)] = \frac{1}{n} \sum_{j=1}^{n} j = \frac{n+1}{2}, \quad \mathbb{E}[R(X_j)^2] = \frac{1}{n} \sum_{j=1}^{n} j^2 = \frac{(n+1)(2n+1)}{6} ,$$

which implies

$$\text{Var} (R(X_j)) = \frac{n^2 - 1}{12} .$$

We will see later that $\text{Cov} (R(X_i), R(X_j)) = -(n+1)/12$ when $i \neq j$.

Rank based tests are permutation tests that are applied to the ranks of the observations, instead of the observations themselves. Rank tests do not depend on any distributional assumptions of the observations. Instead, the relative magnitude and ordering of the observations is utilized. This is particularly useful if observations are not on a numerical but on an ordinal scale.

Advantages of rank tests over classical tests include

- rank tests are less sensitive to outliers;
- rank tests often possess reasonable efficiency;
- rank tests are easy to explain;
- rank tests can be applied without a population model (a randomization model suffices);
- rank tests are applicable to ordinal data (whereas classical tests are strictly speaking not).
5.3.1 The Wilcoxon Signed-Rank Test

If we are willing to make the extra assumption that the data comes from a symmetric distribution under the null hypothesis we can take advantage of the extra information in the ranks effectively (more on this assumption later). Recall that the sign-test only utilizes the sign of the differences between each observation and the hypothesized median \( M_0 \). The magnitudes of these observations relative to \( M_0 \) are ignored. If these magnitudes are available and taken into account, better testing procedures can be expected. The Wilcoxon signed rank test is one of the best known examples of such a procedure.

Assume that the distribution of the observations is continuous and symmetric. We wish to test that the median equals a prescribed value \( M_0 \):

\[
H_0 : M = M_0.
\]

Define \( D_i = X_i - M_0 \), and let \( R(|D_i|) \) denote the rank of \(|D_i|\) among \( \{|D_1|, \ldots, |D_n|\} \), the WSiR-statistic is defined by

\[
T^+ = \sum_{i=1}^{n} R(|D_i|) 1\{D_i > 0\}.
\]

Large values of \( T^+ \) favor the alternative hypothesis \( H_1 : M > M_0 \). This is the positive rank sum of the absolute differences. Notice that if \( R(\cdot) = 1 \) this would be simply the sign test. In effect, we order the absolute differences \(|D_1|, |D_2|, \ldots, |D_n|\) from smallest to largest and assign them ranks 1, 2, \ldots, \( n \), while keeping track of the signs of the differences \( D_i \).

If the alternative hypothesis is \( H_1 : M < M_0 \), then we may use

\[
T^- = \sum_{i=1}^{n} R(|D_i|) 1\{D_i < 0\},
\]

instead. However, since \( T^+ + T^- = \sum_{i=1}^{n} i = \frac{n(n+1)}{2} \), tests based on \( T^+ \), \( T^- \) or \( T^+ - T^- \) are all equivalent. Now large or small values of

\[
T = T^+ - T^- = \sum_{i=1}^{n} R(|D_i|) \text{sgn}(D_i) = 2T^+ - \frac{n(n + 1)}{2}
\]

favor the alternative hypothesis \( H_1 : M \neq M_0 \). In the above

\[
\text{sgn}(x) = \begin{cases} 
1 & \text{if } x > 0 \\
0 & \text{if } x = 0 \\
-1 & \text{if } x < 0
\end{cases}
\]

A natural question arises to which is the distribution of \( T^+ \) under the null hypothesis. Under \( H_0 \), the random variables \( 1\{D_1 > 0\}, 1\{D_2 > 0\}, \ldots, 1\{D_n > 0\} \) are a random sample from a Bernoulli distribution with success probability 1/2. The following fact is going to be extremely useful.

5.3 Exercise Suppose that \( D_1, \ldots, D_n \) are mutually independent random variables, symmetric around 0 and with continuous distribution function. Show that the vectors \( (R(|D_1|), \ldots, R(|D_n|)) \) and \( (1\{D_1 > 0\}, \ldots, 1\{D_n > 0\}) \) are independent.
This exercise implies that
\[
\mathbb{E}[T^+] = \mathbb{E} \left[ \sum_{i=1}^{n} R(|D_i|)1\{D_i > 0\} \right]
= \sum_{i=1}^{n} \mathbb{E}[R(|D_i|)]\mathbb{E}[1\{D_i > 0\}]
= \frac{n(n+1)}{4}.
\]

In a similar way, under \(H_0\)
\[
\text{Var}(T^+) = \frac{n(n+1)(2n+1)}{24}.
\] (3)

An application of Lyapunov’s central limit theorem (omitting details here) yields
\[
\frac{T^+ - \mathbb{E}T^+}{\sqrt{\text{Var}(T^+)}} \xrightarrow{D} \mathcal{N}(0,1).
\]

For details, see (Hettmansperger (1984)).

5.4 Exercise  Prove expression (3).

5.5 Exercise  Show that under \(H_0\), \(T^+\) is distributed as \(\sum_{i=1}^{n} i\eta_i\), where \(\eta_1, \ldots, \eta_n\) is a i.i.d. random sample from a Bernoulli\(\left(\frac{1}{2}\right)\) distribution. This shows that the Wilcoxon signed rank statistic is distribution free under \(H_0\).

So, in applying a test using the statistic \(T^+\) we can either use the normal approximation to compute the rejection region (if \(n\) is large), use a Monte-Carlo approach, or compute the exact distribution by carefully enumerating all the possibilities (although you shouldn’t use a brute-force approach) - see (Gibbons and Chakraborti (1992)) for further details.

Remark:  The WSiR-test was actually devised as a pair-sampled data procedure. In particular under the consideration of a location model, where the data is \((X_1, Y_1), \ldots, (X_n, Y_n)\), where one wants to test
\[
H_0 : F_X = F_Y \quad \text{against} \quad H_1 : F_Y(x) = F_X(x - \Delta), \text{for some } \Delta \neq 0,
\]
and \(F_X\) and \(F_Y\) denote the marginal distributions of \(X_i\) and \(Y_i\), respectively. In this case the symmetry assumption of \(D_i = Y_i - X_i\) holds, and we can even use this test to get a confidence interval for \(\Delta\).

5.4 The General Two-Sample Problem

As we already considered in the population model for permutation tests let
\[
X_1, \ldots, X_m \sim F \quad \text{and} \quad Y_1, \ldots, Y_n \sim G,
\]
be independent samples from two populations with continuous distributions. We are generally interested in testing \(H_0 : F = G\) against a suitable alternative. If some strong assumptions are made we can devise very powerful tests, such as the two-sample \(t\)-test. However, we want to make as few assumptions as possible.
5.4.1 The Wald-Wolfowitz Runs Test

Take \(X_1, \ldots, X_m\) and \(Y_1, \ldots, Y_n\) and combine them in a single sequence \((Z_1, \ldots, Z_{n+m}) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)\). Now order the data \(Z_i\) from smallest to largest (assume no ties) and assign a label “X” or “Y” depending on the population of origin of that data point.

\[
\begin{array}{cccccc}
Z_{(1)} & Z_{(2)} & Z_{(3)} & Z_{(4)} & \cdots & Z_{(n+m)} \\
\downarrow & \downarrow & \downarrow & \downarrow & \cdots & \\
X & Y & Y & X & \cdots & Y
\end{array}
\]

Under \(H_0\) we expect all Xs and Ys to be well mixed. Define the number of runs as the number of sequences of identical symbols preceded and followed by either a different symbol or no symbol at all). In the sequence above we begin with a run of length 1 “X” followed by a run of length 2 “YY” and so on. A large number of runs indicates thorough mixing, and supports the null hypothesis.

Let \(R\) be the total number of runs. We can test \(H_0: F = G\) against \(H_1: F(x) \neq G(x)\) for some \(x\).

Such a test has a rejection region \(R \leq c_\alpha\) where \(c_\alpha\) is chosen such that \(P_{H_0}(R \leq c_\alpha) \leq \alpha\).

Note that this is also a permutation test, so you can obtain the distribution of \(R\) under the null by enumerating all possibilities. This is a bit of overkill, as remarkably in this case we can write the exact distribution. The following theorem is stated without proof (Gibbons and Chakraborti (1992))

5.6 Theorem Under the null hypothesis \(H_0\) and the assumption that \(F\) is a continuous distribution

\[
\Pr(R = r) = \begin{cases} 
\frac{2(m^{-1})(n^{-1})}{(m+n)^{-1}} & \text{if } r \text{ is even} \\
\frac{(r^{-1/2})(n^{-1})(r^{-1/2})(n^{-1})(r^{-1/2})(n^{-1})}{(m+n)^{-1}} & \text{if } r \text{ is odd}
\end{cases},
\]

where \(r = 2, 3, \ldots, n + m\).

This result is exact, but we can also consider asymptotics and show that when \(m/(n+m) \to \lambda\) as \(n \to \infty\) the standardized statistic

\[
Z = \frac{R - 2(n+m)\lambda(1-\lambda)}{2\sqrt{n+m\lambda(1-\lambda)}},
\]

converges to a standard normal distribution.

The Wald-Wolfowitz test is extremely general, and it is consistent under the alternative \(F \neq G\). This great generality also weakens its performance. If we are willing to do more assumptions however we can construct a very powerful rank-based test.

5.4.2 The Wilcoxon Rank-Sum Test

In this scenario we are going to consider the location model, already introduced. Suppose we want to test

\[
H_0: F = G \quad \text{against} \quad H_1: G(x) = F(x - \Delta), \text{for some } \Delta \neq 0.
\]
Let \( N = n + m \) and define \((Z_1, \ldots, Z_N) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)\). Let \( R(Z_i) \) be the rank of \( Z_i \) within \((Z_1, \ldots, Z_N)\) and define the Wilcoxon Rank-Sum statistic as

\[
W = \sum_{i=m+1}^{N} R(Z_i) .
\]

Large or small values of \( W \) favor the alternative hypothesis \( H_1 : \Delta \neq 0 \). Note that, again, the test obtained is just a permutation test applied to the ranks of the observations instead of the observations themselves. The distribution of \( W \) under the null-hypothesis does not depend on the unknown distribution \( F \). That is, the test is distribution-free.

The distribution of \( W \) under the \( H_0 \) can be obtained by exhaustive enumeration, or the quantiles can be approximated by Monte-Carlo simulation (just as in the permutation test scenario). However, we can also study the asymptotic properties of \( W \).

Under \( H_0 \)

\[
\mathbb{E}[W] = \mathbb{E} \left[ \sum_{i=m+1}^{N} R(Z_i) \right] = \sum_{i=m+1}^{N} \mathbb{E}[R(Z_i)] = n \frac{N+1}{2} .
\]

Similarly

\[
\text{Var}(W) = \text{Var} \left( \sum_{i=m+1}^{N} R(Z_i) \right) = \sum_{i=m+1}^{N} \text{Var}(R(Z_i)) + \sum_{i=m+1}^{N} \sum_{j=m+1, j \neq i}^{N} \text{Cov}(R(Z_i), R(Z_j)) .
\]

Now, symmetry comes to the rescue. Because the joint distribution of \( R(Z_i) \)'s must be symmetric under permutations the covariance term is not a function of \( i \neq j \), and depends only on \( N \). Therefore \( \text{Cov}(R(Z_i), R(Z_j)) = h(N) \), for some function \( h \). We can re-write

\[
\text{Var}(W) = n \frac{N^2 - 1}{12} + n(n-1)h(N) .
\]

To determine \( h(N) \) we can consider the particular case \( m = 0, N = n \). In this case

\[
\text{Var}(W) = \text{Var} \left( \sum_{i=1}^{N} R(Z_i) \right) = \text{Var} \left( \sum_{i=1}^{N} i \right) = 0 ,
\]

therefore

\[
0 = N \frac{N^2 - 1}{12} + N(N-1)h(N) .
\]

This implies that \( h(N) = -(n+1)/12 \), and so

\[
\text{Var}(W) = \frac{nm(N+1)}{12} .
\]

Whence, as \( \min(n, m) \rightarrow \infty \), using the Lyapunov’s central limit theorem (we omit details),

\[
\frac{W - n(N+1)}{\sqrt{nm(N+1)}} \overset{D}{\rightarrow} \mathcal{N}(0, 1) .
\]
5.7 Exercise The Mann-Whitney statistic is given by

\[ U = \sum_{i=1}^{m} \sum_{j=1}^{n} 1\{X_i < Y_j\} . \]

Assume no ties in the combined sample. Show that \( U \) is a linear function of \( W \) and therefore tests based on this statistic are equivalent to tests based on the Wilcoxon Sum Rank statistic.

5.5 Further Reading

An extensive collection of rank-tests can be found in the book by (Hollander and Wolfe (1999)). It is a good place to quickly see what types of tests are available in a specific setting. Asymptotic theory on rank statistics can be found in chapter 13 in (Van der Vaart (1998)). A short and accessible article on robust fitting of linear models using the Wilcoxon weighted norm is (McKean (2004)).

The tests we have discussed are distribution free, and in that sense nonparametric. Therefore, they possess good robustness properties. Robust statistics in general is primarily concerned with methods that are relatively insensitive to small deviations from parametric models. As such, robust statistics is part of parametric statistics. Good sources on robust methods are (Huber and Ronchetti (2009)) and (Hampel et al. (2005)).
6 Measures of Association

This section is based on sections 12.1 and 12.2 in (Gibbons and Chakraborti (1992)). In previous section we focus on the problem of testing differences between two distributions. Here we attempt to quantify, in some sense, the relationship between distributions.

If $X$ and $Y$ are two random variables with a bivariate probability distribution, then we know that

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X,Y).$$

Therefore the covariance between $X$ and $Y$ reflects the amount of association or correspondence between the two random variables. If the covariance has a large positive value, then large (small) values of $X$ are associated with large (small) values of $Y$. Conversely, if the covariance has a large negative value, then large (small) values of $X$ are associated with small (large) values of $Y$. To assist the interpretation of the covariance, we can scale it to $[-1,1]$. This results in the Pearson product-moment correlation coefficient (generally referred to simply as “correlation coefficient”), which is defined by

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

The range $[-1,1]$ is a simple consequence of the Cauchy-Schwarz inequality. The correlation coefficient quantifies the amount of linear relationship between $X$ and $Y$: for $a,b,c,d \in \mathbb{R}$

$$\rho(aX + b, cY + d) = \text{sgn}(ac)\rho(X,Y).$$

As you know well, in general zero correlation does not imply independence; take for example $X$ a standard normal random variable and $Y = X^2$. However, for multivariate normal distributions independence is equivalent to zero correlation.

In general we don’t have access to the joint distribution, but rather to a finite sample from this distribution $(X_1,Y_1),\ldots,(X_n,Y_n)$. The Pearson correlation coefficient can be estimated by its sample analogue

$$\hat{\rho}_n = \frac{\sum_{i=1}^n(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n(X_i - \bar{X}_n)^2}\sum_{j=1}^n(Y_j - \bar{Y}_n)^2}. \quad (4)$$

6.1 Towards General Measures of Association

If the joint distribution of $X$ and $Y$ is multivariate normal, then zero correlation does imply independence, making correlation perfectly suitable for modeling dependence in this setting. However, in nonparametric statistics, this reasoning has little significance and other measures of association should be equally acceptable. (Gibbons and Chakraborti (1992)) (chapter 12) gives a list of properties that a “good” measure $m(X,Y)$ of association should have:

1. Let $(X_1,Y_1)$ and $(X_2,Y_2)$ be two independent pairs of random variables with the same distribution as $(X,Y)$. If it is the case that we have with probability 1

$$(X_1 - X_2)(Y_1 - Y_2) > 0,$$

then $X$ and $Y$ are said to be perfectly concordant. In that case the association measure should take precisely the value $m(X,Y) = 1$. 

34
2. Conversely, in the same setting as above, if we have with probability 1 that

\[(X_1 - X_2)(Y_1 - Y_2) < 0 ,\]

then \(X\) and \(Y\) are said to be \textit{perfectly discordant}. In that case we require that \(m(X, Y) = -1\).

3. In all other cases the measure will take values in \((-1, 1)\).

4. The measure will be \(m(X, Y) = 0\) if \(X\) and \(Y\) are independent (the converse will in general not be true).

5. \(m(X, Y) = m(Y, X) = m(-X, -Y)\).

6. \(m(-X, Y) = m(X, -Y) = -m(X, Y)\).

7. The measure should be invariant under all transformations of \(X\) and \(Y\) for which order of magnitude is preserved.

The correlation coefficient fails to satisfy several of these properties (namely 1, 2 and 7). As an example of perfectly concordant random variables take \(X \sim \mathcal{N}(0, 1)\) and \(Y = X^3\). A simple calculation shows that \(\rho(X, Y) = 3/\sqrt{15}\). Note that the above properties are somewhat motivated in the same way as the ranks, as we seek association measures that are invariant under transformation preserving order of magnitude.

### 6.2 Kendall's Tau

A simple way to define a measure of association satisfying 1 to 7 is to make use of the concepts defined in properties 1 and 2. Let \((X_1, Y_1)\) and \((X_2, Y_2)\) be independent with the same distribution as \((X, Y)\). Define the probability of concordance by

\[p_c = \Pr((X_1 - X_2)(Y_1 - Y_2) > 0)\]

and the probability of discordance by

\[p_d = \Pr((X_1 - X_2)(Y_1 - Y_2) < 0)\]

Now simply define the \textit{Kendall coefficient} \(\tau\) as the difference

\[\tau(X, Y) = p_c - p_d \]

This measure of association satisfies all the listed properties. The only property that might not be obvious is the independence property 4: if \(X\) and \(Y\) are independent then,

\[
\begin{align*}
p_c &= \Pr((X_1 - X_2)(Y_1 - Y_2) > 0) \\
&= \Pr((X_1 - X_2) > 0 \quad ; \quad (Y_1 - Y_2) > 0) + \Pr((X_1 - X_2) < 0 \quad ; \quad (Y_1 - Y_2) < 0) \\
&= \Pr((X_1 - X_2) > 0) \Pr((Y_1 - Y_2) > 0) + \Pr((X_1 - X_2) < 0) \Pr((Y_1 - Y_2) < 0) \\
&= \Pr((X_1 - X_2) < 0) \Pr((Y_1 - Y_2) > 0) + \Pr((X_1 - X_2) > 0) \Pr((Y_1 - Y_2) < 0) \\
&= p_d ,
\end{align*}
\]
where the second-to-last step follows simply because $X_1 - X_2$ is a symmetric random variable. So we conclude that $\tau(X, Y) = 0$ if $X$ and $Y$ are independent. Of course the converse is in general not true.

If the marginal distributions of $X$ and $Y$ are continuous then it is easily seen that $p_c + p_d = 1$ and so $\tau = 2p_c - 1 = 1 - 2p_d$. Also in this case it is easy to interpret $\tau(X, Y)$: note that

$$\frac{1 + \tau(X, Y)}{1 - \tau(X, Y)} = \frac{1 + p_c - p_d}{1 - (p_c - p_d)} = \frac{2p_c}{2p_d} = \frac{p_c}{p_d}.$$ 

So the odds of a pair of observations being concordant is given by $(1 + \tau(X, Y))/(1 - \tau(X, Y))$. For example, if $\tau(X, Y) = 1/3$, then the probability of concordance is twice the probability of discordance.

Finally, for bivariate normal distributions it can be shown with a bit of manipulation that

$$\tau(X, Y) = \frac{2}{\pi} \arcsin \rho(X, Y).$$

### 6.2.1 Estimating $\tau(X, Y)$ from Data

A natural way to estimate $\tau(X, Y)$ from data is to plug-in estimates of $p_c$ and $p_d$. A natural estimator for the probability of concordance (discordance) is the relative frequency of the number of concordant (discordant) pairs. Suppose we have data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Define

$$A_{ij} = \text{sgn}(X_i - X_j)\text{sgn}(Y_i - Y_j),$$

where as usual

$$\text{sgn}(u) = \begin{cases} 
-1 & \text{if } u < 0 \\
0 & \text{if } u = 0 \\
1 & \text{if } u > 0 
\end{cases}.$$

Clearly $A_{ij}$ takes the value 1 if the pairs $i$ and $j$ are concordant, -1 if the pairs are discordant, and zero if the pairs are neither concordant nor discordant or because of a tie in either component. Obviously then $\Pr(A_{ij} = 1) = p_c$ and $\Pr(A_{ij} = -1) = p_d$ and so $\mathbb{E}[A_{ij}] = \tau(X, Y)$. An unbiased estimator for $\tau$ is therefore provided by

$$\hat{\tau}_n(X, Y) = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} A_{ij}.$$ 

This measure is known as Kendall’s sample tau coefficient. Curiously we can write $\hat{\tau}_n(X, Y)$ as follows: define $U_{ij} = \text{sgn}(X_i - X_j)$, $V_{ij} = \text{sgn}(Y_j - Y_i)$. If there are no ties (i.e. $U_{ij} \neq 0$ and $V_{ij} \neq 0$ if $i \neq j$) then

$$\hat{\tau}_n(X, Y) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} U_{ij} V_{ij}}{\sqrt{\left( \sum_{i=1}^{n} \sum_{j=1}^{n} U_{ij}^2 \right) \left( \sum_{i=1}^{n} \sum_{j=1}^{n} V_{ij}^2 \right)}}.$$

This expression resembles (4), and essentially replaces the data values by signed differences.

Suppose we wish to test the null hypothesis that the $X$ and $Y$ random variables are independent. Since $\tau(X, Y) = 0$ for independent random variables, we should reject the null hypothesis
for large values of $|\hat{\tau}_n(X,Y)|$. Clearly, for continuous distributions and under the null $\hat{\tau}_n(X,Y)$ is distribution free, and so we can just evaluate this distribution. Under $H_0$ there are $n!$ equally likely randomizations, and in principle the exact distribution of $\hat{\tau}_n(X,Y)$ can be calculated. For large values of $n$, we can use the asymptotic result
\[
\frac{3\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \hat{\tau}_n(X,Y) \xrightarrow{D} \mathcal{N}(0,1),
\]
as $n \to \infty$. This allows us to obtain to obtain critical for such a test.

6.3 Spearman correlation

As the main motivation behind the properties 1 through 7 is invariance under order preserving transformations a very simple idea is to use ranks, and apply the Pearson’s correlation coefficient to ranks instead of the original data. For computation, replace $(X_i, Y_i)$ by $(R_i, S_i)$, where $R_i = R(X_i)$ is the rank of $X_i$ within the sample $X_1, \ldots, X_n$ and $S_i = R(Y_i)$ is the rank of $Y_i$ within the sample $Y_1, \ldots, Y_n$. In what follows we will assume no ties in the data. The Spearman’s rank correlation coefficient is defined as
\[
\hat{\rho}_n^{(S)} = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{j=1}^n (S_j - \bar{S})^2}}.
\]
Under the assumption of no ties this expression can be heavily simplified as
\[
\bar{R} = \bar{S} = \frac{1}{n} \sum_{i=1}^n i = \frac{1}{2(n+1)} \quad \text{and} \quad \sum_{i=1}^n (R_i - \bar{R})^2 = \sum_{i=1}^n (S_i - \bar{S})^2 = \sum_{i=1}^n (i-(n+1)/2)^2 = \frac{n^2-1}{12}.
\]
If we define $D_i = R_i - S_i$, then
\[
\sum_{i=1}^n D_i^2 = 2 \frac{n(n^2-1)}{12} - 2 \sum_{i=1}^n (R_i - \bar{R}_n)(S_i - \bar{S}_n).
\]
Therefore,
\[
\hat{\rho}_n^{(S)} = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n^2-1)}.
\]
If there are ties, one can use midranks to define the Spearman rank correlation (see (Gibbons and Chakraborti (1992))). Unfortunately it is difficult to express the corresponding measure of association for the Spearman rank correlation coefficient, but one can check that indeed $\hat{\rho}_n^{(S)}$ satisfies the nice properties the outlined.

Suppose we wish to test that the correlation equals zero: $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$. A logical choice for a test statistic is given by $\hat{\rho}_n^{(S)}$. Instead of asymptotic approximations, we can use a permutation test. Fix $X_1, \ldots, X_n$. Under $H_0$ the law of $\hat{\rho}_n$ is invariant under permutations of the set $Y_1, \ldots, Y_n$. To obtain the randomization distribution of $\hat{\rho}_n^{(S)}$ we henceforth need to compute its value under $n!$ permutations. If $n$ is large we can also use the following asymptotic approximation
\[
\sqrt{n-1} \hat{\rho}_n^{(S)} \xrightarrow{D} \mathcal{N}(0,1),
\]
as $n \to \infty$. 37
7 Useful R-commands

7.1 Empirical distribution function and related

<table>
<thead>
<tr>
<th>Method</th>
<th>R-function</th>
<th>within package</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical distr. function</td>
<td>ecdf</td>
<td>stats</td>
</tr>
<tr>
<td>Binomial confidence interval</td>
<td>binconf</td>
<td>Hmisc</td>
</tr>
</tbody>
</table>

7.2 Goodness of fit

<table>
<thead>
<tr>
<th>Test</th>
<th>R-function</th>
<th>within package</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square GOF</td>
<td>chisq.test</td>
<td>stats</td>
</tr>
<tr>
<td>KS GOF</td>
<td>ks.test</td>
<td>stats</td>
</tr>
<tr>
<td>KS-distribution</td>
<td>ksdist</td>
<td>PASWR</td>
</tr>
<tr>
<td>QQ-plot</td>
<td>qq.plot</td>
<td>car</td>
</tr>
</tbody>
</table>

Specialized test for normality.

<table>
<thead>
<tr>
<th>Test</th>
<th>R-function</th>
<th>within package</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>shapiro.test</td>
<td>stats</td>
</tr>
<tr>
<td>Anderson Darling</td>
<td>ad.test</td>
<td>nortest</td>
</tr>
<tr>
<td>Cramer-Von Mises</td>
<td>cvm.test</td>
<td>nortest</td>
</tr>
<tr>
<td>(for composite normality)</td>
<td>cvm.test</td>
<td>nortest</td>
</tr>
<tr>
<td>Jarque-Bera</td>
<td>jarque.bera.test</td>
<td>tseries</td>
</tr>
</tbody>
</table>

7.3 Permutations

<table>
<thead>
<tr>
<th>Command</th>
<th>Result</th>
<th>within package</th>
</tr>
</thead>
<tbody>
<tr>
<td>choose(10,5)</td>
<td>compute choose(10,5)</td>
<td>base</td>
</tr>
<tr>
<td>factorial(5)</td>
<td>compute numbers of permutations of 5 elements</td>
<td>base</td>
</tr>
<tr>
<td>combinations(10,5)</td>
<td>list all combinations of choose(10,5)</td>
<td>base</td>
</tr>
<tr>
<td>permutations(5,5)</td>
<td>list all permutations of 5 elements</td>
<td>base</td>
</tr>
<tr>
<td>permutations(5,3)</td>
<td>list all permutations of the elements from choose(5,3)</td>
<td>base</td>
</tr>
<tr>
<td></td>
<td>i.e. total C(5,3) * 3! possible outcomes</td>
<td></td>
</tr>
</tbody>
</table>

7.4 Rank tests and measures of association

<table>
<thead>
<tr>
<th>Test</th>
<th>R-function</th>
<th>within package</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sign-test</td>
<td>SIGN.test</td>
<td>PASWR</td>
</tr>
<tr>
<td>Wilc. rank sum distr.</td>
<td>dwilcox</td>
<td>stats</td>
</tr>
<tr>
<td>Wilc. signed rank distr.</td>
<td>dsignrank</td>
<td>stats</td>
</tr>
<tr>
<td>Wilc. sign / rank sum</td>
<td>wilcox.test</td>
<td>stats</td>
</tr>
<tr>
<td>Kruskal-Wallis</td>
<td>kruskal.test</td>
<td>stats</td>
</tr>
<tr>
<td>Friedman</td>
<td>friedman.test</td>
<td>stats</td>
</tr>
<tr>
<td>correlation</td>
<td>cor.test</td>
<td>stats</td>
</tr>
<tr>
<td>spearman-correlation</td>
<td>spearman.test</td>
<td>Hmisc</td>
</tr>
<tr>
<td>Kendall’s tau</td>
<td>Kendall</td>
<td>Kendall</td>
</tr>
<tr>
<td>McNemar test</td>
<td>mcnemar.test</td>
<td>stats</td>
</tr>
<tr>
<td>Mantel-Haenszel test</td>
<td>mantelhaen.test</td>
<td>stats</td>
</tr>
</tbody>
</table>
References