Part VI

NONPARAMETRIC TESTS

In this part we cover *classical* nonparametric tests such as the sign test, the signed-rank sum test and the Wilcoxon test. Further nonparametric tests are found throughout the book and listed with their parametric alternatives such as the Spearman rank correlation coefficient and the Fisher exact test. This arrangement is arbitrary but we think that it serves the intention of this book best. Also, nonparametric tests is a broad term and we refrain from tediously classifying each test according to such criteria.

In SAS you can perform most of the tests presented here with the procedure `PROC NPAR1WAY`. The procedure also allows to calculate the exact distribution and therefore exact p-values. However, this is very cumbersome and time consuming (even taking days to calculate). Furthermore exact p-values via Monte Carlo estimation can be calculated. We do not present this here and refer the reader to the SAS documentation. In most cases the asymptotic p-values should suffice. In R the tests are covered by single functions, described here.

Statistical Hypothesis Testing with SAS and R, First Edition. Dirk Taeger and Sonja Kuhnt.

8

Tests on location

In this chapter we present nonparametric tests for the location parameter. The simplest one is the sign test, with the only assumption that the data are sampled from a continuous distribution. This test has its foundation in the early eighteenth century (Arbuthnot 1710). Some of the tests presented here find their parametric analogs in the one- and two-sample t-tests. So, these tests are good alternatives if the Gaussian distribution assumption appears to be violated. We show how to perform one-, two-, and K-sample nonparametric tests on the location parameter in SAS and R. Tables of critical values can be found, for example, in Owen (1962) and in Hollander and Wolfe (1999) as well as in many other textbooks.

8.1 One-sample tests

In this section we deal with the question if the median of a population differs from a predefined value. The most straightforward test is the sign test. However, if a symmetric distribution can be assumed the Wilcoxon signed-rank test is a better alternative.

8.1.1 Sign test

Description: Tests if the location (median m) of a population differs from a specific value m_0.

Assumptions:
- Data are measured at least on an ordinal scale.
- The random variable X follows a continuous distribution with median m.

Hypotheses:
(A) $H_0 : m = m_0$ vs $H_1 : m \neq m_0$
(B) $H_0 : m = m_0$ vs $H_1 : m > m_0$
(C) $H_0 : m = m_0$ vs $H_1 : m < m_0$

Test statistic:

$$V = \sum_{i=1}^{n} D_i \quad \text{with} \quad D_i = \begin{cases} 1, & \text{if} \quad X_i - m_0 > 0 \\ 0, & \text{if} \quad X_i - m_0 < 0 \end{cases}$$

Statistical Hypothesis Testing with SAS and R, First Edition. Dirk Taeger and Sonja Kuhnt.

Test decision: Reject H_0 if for the observed value v of V
(A) $v < b_{n;\alpha/2}$ or $v > n - b_{n;\alpha/2}$
(B) $v > n - b_{n;\alpha}$
(C) $v < b_{n;\alpha}$
where $b_{n;\alpha}$ is the α-quantile of the binomial distribution with parameters n and $p = 0.5$.

p-value: (A) $p = 2\min(P(V > v), 1 - P(V > v))$
(B) $p = P(V > v)$
(C) $p = 1 - P(V > v)$
where $P(V < v)$ is the cumulative distribution function of a binomial distribution with parameters n and $p = 0.5$.

Annotations:
- The test statistic $V \sim B(n, 0.5)$ is a binomial distribution.
- Observations equal to the value m_0 are not considered in the calculation of the distribution of the test statistic as due to the assumption of a continuous distribution such observations appear with probability zero.
- If the case $X_i - \mu_0 = 0$ happens due to rounding errors, ect., one way to deal with it is to ignore them and reduce the number n of observations accordingly (Gibbons 1988).

Example: To test the hypothesis that the median systolic blood pressure of a specific population equals 120 mmHg. The dataset contains observations of 55 patients (dataset in Table A.1).

SAS code

```
*** Variant 1 ***;
* Only for hypothesis (A);
proc univariate data=blood_pressure mu0=120 loccount;
 var mmhg;
run;

*** Variant 2;

* Calculate the median;
proc means data=blood_pressure median;
 var mmhg;
 output out=sign2 median=m;
run;

* Calculate test statistic;
data sign1;
 set blood_pressure;
 mu0=120;                      * Set mu0=120 to test for;
 if mmhg-mu0=0 then delete;    * Delete values equal to mu0;
 s=(mmhg-mu0>0);               * Set to 1 if values
                               * greater mu0;
run;
```

```
proc summary n;             * Count the number of
  var s;                    * 'successes' = test statistic v;
 output out=sign3 n=n       * and sample size;
                  sum=v;
run;

* Put median(=m) and sample size (=n)
  and test stattsic (=v) in one dataset;
data sign4;
 merge sign2 sign3;
 keep m n v;
run;

* Calculation of p-values;
data sign5;
 set sign4;
 format pvalue_A
        pvalue_B
        pvalue_C pvalue.;
 f=n-v;                       * Number of 'failures';

 * Decide which tail must be used for one-tailed tests;
 diff=m-120;
 if diff>=0 then
  do;
   pvalue_B=probbnml(0.5,n,f);
   pvalue_C=probbnml(0.5,n,v);
  end;
 if diff<0 then
  do;
   pvalue_B=probbnml(0.5,n,v);
   pvalue_C=probbnml(0.5,n,f);
  end;
 pvalue_A=min(2*min(pvalue_B,pvalue_C),1);
run;

* Output results;
proc print;
 var n v m pvalue_A pvalue_B pvalue_C;
run;
```

SAS output

```
Variant 1

         Tests for Location: Mu0=120

Test           -Statistic-     -----p Value------
Sign           M         6     Pr >= |M|    0.1337

      Location Counts: Mu0=120.00
```

```
      Count                 Value

      Num Obs >  Mu0          33
      Num Obs ^= Mu0          54
      Num Obs < Mu0           21

Variant 2

n     v    m    pvalue_A    pvalue_B    pvalue_C
54   33   134   0.1337      0.0668      0.9620
```

Remarks:

- Variant 1 calculates the p-value only for hypothesis (A). `PROC UNIVARIATE` does not provide p-values for the one-sided hypothesis.
- `mu0=`*value* is optional and indicates the m_0 to test against. Default is 0.
- `loccount` is optional and prints out the number of observations greater than, equal to, or less than m_0.
- Variant 2 calculates p-values for all three hypotheses.

R code

```
# Set value mu0 to test against
mu0<-120

# Calculate differences between values and mu0
d<-blood_pressure$mmhg-mu0

# Calculate number of differences not equal to zero
n<-length(d[d!=0])

# Calculate test statistic
v<-length(d[d>0])

# Calculation of p-values
pvalue_A<-binom.test(v,n,0.5,alternative="two.sided")

# Decide which tail must be used for one-tailed tests
diff<-median(blood_pressure$mmhg)-mu0

if (diff >=0){
    pvalue_B<-binom.test(v,n,0.5,alternative="greater")
    pvalue_C<-binom.test(v,n,0.5,alternative="less")
}

if (diff < 0) {
    pvalue_B<-binom.test(v,n,0.5,alternative="less")
    pvalue_C<-binom.test(v,n,0.5,alternative="greater")
}
```

```
# Output results
pvalue_A$p.value
pvalue_B$p.value
pvalue_C$p.value
```

R output

```
> pvalue_A$p.value
[1] 0.1336742
> pvalue_B$p.value
[1] 0.06683712
> pvalue_C$p.value
[1] 0.9620476
```

Remarks:

- There is no core R function that can be used directly to calculate the sign test.
- However, as the test statistic follows a binomial distribution, the p-value can be calculated easily with the function `binom.test`.

8.1.2 Wilcoxon signed-rank test

Description: Tests if the location (median m) differs from a specific value m_0.

Assumptions:
- Data are measured at least on an ordinal scale.
- The random variables $X_i, i = 1, \ldots, n$, follow continuous distributions, which might differ, but are all symmetric about the same median m.

Hypotheses:
(A) $H_0 : m = m_0$ vs $H_1 : m \neq m_0$
(B) $H_0 : m = m_0$ vs $H_1 : m > m_0$
(C) $H_0 : m = m_0$ vs $H_1 : m < m_0$

Test statistic:

$$W^+ = \sum_{i=1}^{n} R_i \mathbb{1}_{]0,\infty[}\left\{X_i - m_0\right\}, \text{ with } R_i = rank(X_i), \text{ for } i = 1, \ldots, n$$

Test decision: Reject H_0 if for the observed value w^+ of W^+
(A) $w^+ \geq w_{\alpha/2}$ or $w^+ \leq \frac{n(n+1)}{2} - w_{\alpha/2}$
(B) $w^+ \geq w_\alpha$
(C) $w^+ \leq \frac{n(n+1)}{2} - w_\alpha$

p-value:
(A) $p = 2\min(1 - P(W^+ < w^+), P(W^+ \leq w^+))$
(B) $p = 1 - P(W^+ < w^+)$
(C) $p = P(W^+ \leq w^+)$

Annotations:

- For the calculation of the test statistic, first the absolute differences are ranked from the lowest to the highest values. W^+ is the sum of the ranks of the differences with positive sign and W^- is the corresponding sum of ranks of the differences with negative sign. The test statistic W^+ or W^- can be used, but usually W^+ is used for the Wilcoxon signed-rank test (Wilcoxon 1945, 1949).
- w_α denotes the upper-tail probabilities for the distribution of W^+ under the null hypothesis, for example, given in table A.4 of Hollander and Wolfe (1999).
- Observations equal to the value m_0 are not considered in the calculation of the test statistic as observations are sampled from a continuous distribution and this case should happen with zero probability. If $X_i - m_0 = 0$ occurs due to rounding errors, etc., the number n of observations must be reduced accordingly for the calculation of the test statistic. In the case of ties, that is, the absolute differences have the same values, mid ranks are assigned (Hollander and Wolfe 1999, p. 38).
- For higher sample sizes the calculation of the distribution of the test statistic W^+ is tedious. For $n \geq 20$ it holds that W^+ is approximately Gaussian distributed with mean $E(W^+) = n(n+1)/4$ and variance $Var(W^+) = n(n+1)(2n+1)/24$. Hence $Z = \frac{W^+ - E(W^+)}{\sqrt{Var(W^+)}}$ is used as test statistic and compared with quantiles of the standard normal distribution. In the case of ties the variance is given by $Var(W) = \left[n(n+1)(2n+1) - \frac{1}{2}\sum_{k=1}^{n_t}(t_k^3 - t_k)\right]/24$, where n_t denotes the number of groups with ties and t_k the number of ties in group k (Hollander and Wolfe 1999, p. 38).

Example: To test the hypothesis that the median systolic blood pressure of a specific population equals 120 mmHg. The dataset contains observations of 55 patients (dataset in Table A.1).

SAS code

```
*** Variant 1 ***;
* Only for hypothesis (A);
proc univariate data=blood_pressure mu0=120 loccount;
 var mmhg;
run;

*** Variant 2;
* Hypothesis (A), (B), and (C) via Gaussian approximation;

* Calculate signs of the differences to mu0=120;
data wilcox1;
set blood_pressure;
```

```
 d=abs(mmhg-120);
 if mmhg-120>0 then sign="+";
 if mmhg-120>0 then sign="-";
 if mmhg-120=0 then delete; *delete observations
                                        equal to mu0;
run;

* Calculate ranks of the absolute differences;
proc rank data=wilcox1 out=wilcox2;
 var d;
 ranks r;
run;

* Sort by signs;
proc sort;
 by sign;
run;

* Calculate W+;
proc summary data=wilcox2;
 var r;
 by sign;
 output out=wilcox4 sum=W;
run;

* Calculate used observation size,
  taking zero differences into account;
proc summary data=wilcox2;
 var r;
 output out=wilcox5  n=n;
run;

* Keep only W+ and merge sample size to it;
data wilcox6;
 merge wilcox4 wilcox5;
 if _N_=1;
run;

* Now compute correction factor for
  the variance because of ties;
proc sort data=wilcox2;
 by d;
run;

proc summary data=wilcox2;
 var r;
 by d;
 output out=ties1 sum=sum_ranks;
run;

data ties2;
 set ties1;
 g=_FREQ_*(_FREQ_**3-_FREQ_);
run;
```

```
proc summary;
 var g;
 output out=ties3 sum=g_ranks;
run;

* g_ranks is the correction factor for the variance;
data ties4;
 set ties3;
 keep g_ranks;
  g_ranks=g_ranks/48;
run;

* Merge test statistic W+, used observations n,
  and variance correction factor g_ranks together;
data wilcox7;
 merge wilcox6 ties4;
run;

* Calculate test statistic z which
  is Gaussian distributed;
data wilcox8;
 set wilcox7;
 format pvalue_A pvalue_B pvalue_C pvalue.;

 mean=n*(n+1)/4;
 var=n*(n+1)*(2*n+1)/24-g_ranks;

 * Test statistic;
 z=(W-mean)/sqrt(var);

 * Decide which tail must be used for one-tailed tests;
 diff=n*(n+1)/2-W;   * Calculate the difference
                       between W+ and n*(n+1)/2;
 if diff>=0 then
  do;                * Case n*(n+1)/2 >  W+;
   pvalue_B=probnorm(-abs(z));
   pvalue_C=1-probnorm(-abs(z));
  end;
 if diff<0 then
  do;                * Case n*(n+1)/2 < W+;
   pvalue_B=1-probnorm(-abs(z));
   pvalue_C=probnorm(-abs(z));
  end;
  pvalue_A=2*min(probnorm(-abs(z)),1-probnorm(-abs(z)));
run;

* Output results;
proc print label;
 var n w z pvalue_A pvalue_B pvalue_C;
  label n="Used observations"
        w="W+"
        z="Z-statistic";
run;
```

SAS output

```
Variant 1
          Tests for Location: Mu0=120

Test            -Statistic-     -----p Value------
Signed Rank     S       402     Pr >= |S|    0.0003

         Location Counts: Mu0=120.00

          Count                 Value
         Num Obs >  Mu0           33
         Num Obs ^= Mu0           54
         Num Obs < Mu0            21

Variant 2

   Used
observations    W+      Z-statistic
    54        1144.5       3.46619

pvalue_A    pvalue_B    pvalue_C
 0.0005      0.0003      0.9997
```

Remarks:

- Variant 1 calculates the p-value only for hypothesis (A). `PROC UNIVARIATE` does not provide p-values for the one-sided hypotheses.
- `mu0=`*value* is optional and indicates the value m_0 to test against. Default is 0.
- `loccount` is optional and prints out the number of observations greater than, equal to, or less than m_0.
- SAS uses a different test statistic: $S = W^+ - E(W^+)$. This yields a value 402 instead of 1144.5. If $n \leq 20$ the exact distribution of S is used for the calculation of the p-value. Otherwise an approximation to the t-distribution is applied (Iman 1974). Hence the p-value from PROC UNIVARIATE differs from the one calculated by using the Gaussian approximation.
- Variant 2 calculates p-values for all three hypotheses and uses the common Gaussian approximation but should only be employed for sample sizes $\geq$20.

R code

```
wilcox.test(blood_pressure$mmhg,mu=120,exact=FALSE,
                     correct=TRUE,alternative="two.sided")
```

R output

```
Wilcoxon signed rank test with continuity correction

data:  blood_pressure$mmhg
V = 1144.5, p-value = 0.0005441
alternative hypothesis: true location is not equal to 120
```

Remarks:

- `mu=`*value* is optional and indicates the value m_0 to test against. Default is 0.
- `exact=`*value* is optional. If *value* is TRUE an exact p-value is computed, if it is FALSE an approximative p-value is computed. If `exact` is not specified or NULL (default value) an exact p-value will be computed if the sample size is less than 50 and no ties are present. Otherwise, the Gaussian distribution is used.
- `correct=`*value* is optional. If the *value* is TRUE (default value) a continuity correction to the Gaussian approximation is used, that is, a value of 0.5 is subtracted or added to the numerator of the Z-statistic.
- `alternative=`*"value"* is optional and defines the type of alternative hypothesis: "two.sided"= true location is not equal to m_0 (A); "greater"=true location is greater than m_0 (B); "less"=true location is less than m_0 (C). Default is "two.sided".

8.2 Two-sample tests

In this section we deal with the question if two populations of the same shape differ by their location. More formally, two populations with independent distributions F and G are assumed to have the same shape and the hypothesis $H_0 : F(x) = G(x)$ for all x vs $H_1 : F(x) = G(x - \Delta)$ for one x, with $\Delta \neq 0$, is considered. One-sided test problems can be formulated analogously. So we test on a shift in the distributions. We first treat the Wilcoxon rank-sum test for which the t-test is the parametric alternative if F and G are Gaussian distributions. The second test is the Wilcoxon matched-pairs signed-rank test which treats the case of paired samples.

8.2.1 Wilcoxon rank-sum test (Mann–Whitney U test)

Description: Tests if two independent populations differ by a shift in location.

Assumptions:

- Data are measured at least on an ordinal scale.
- Samples $X_i, i = 1, \ldots, n_1$ and $Y_j, j = 1, \ldots, n_2$ are randomly drawn from X and Y.
- The distributions of the random variables X and Y are continuous with distribution functions G and F, X and Y are independent.

Hypotheses:
(A) $H_0 : F(t) = G(t)$ vs $H_1 : F(t) = G(t - \Delta)$ with $\Delta \neq 0$
(B) $H_0 : F(t) = G(t)$ vs $H_1 : F(t) = G(t - \Delta)$ with $\Delta > 0$
(C) $H_0 : F(t) = G(t)$ vs $H_1 : F(t) = G(t - \Delta)$ with $\Delta < 0$

Test statistic:
For $n_1 \leq n_2$ the test statistic is given by:
$W =$ sum of ranks of $X_1, \ldots, X_{n_1}$ in the combined sample

Test decision:
Reject H_0 if for the observed value w of W
(A) $w \geq w_{\alpha/2}$ or $w \leq n_1(n_1 + n_2 + 1) - w_{\alpha/2}$
(B) $w \geq w_\alpha$
(C) $w \leq n_1(n_1 + n_2 + 1) - w_\alpha$

p-value:
(A) $p = 2\min(P(W \geq w), 1 - P(W \geq n_1(n_1 + n_2 + 1) - w))$
(B) $p = P(W \geq w)$
(C) $p = 1 - P(W \geq n_1(n_1 + n_2 + 1) - w)$

Annotations:

- For the calculation of the test statistic, first combine both samples and rank the combined sample from the lowest to the highest values. W is the sum of the ranks of sample X. It is also possible to use the sum of ranks of Y in the combined sample as test statistic. Usually the sum of ranks of the sample with the smallest sample size is used (Mann and Whitney 1947; Wilcoxon 1949).
- w_α denotes the upper-tail probabilities for the distribution of W under the null hypothesis, for example, given in table A.6 of Hollander and Wolfe (1999).
- In case of ties, that is, observations with the same values, mid ranks are assigned, resulting in an approximate test (Hollander and Wolfe 1999, p.108).
- For higher sample sizes the calculation of the distribution of the test statistic W is tedious. A Gaussian approximation can be used with $E(W) = n_1(n_1 + n_2 + 1)/2$ and variance $Var(W) = n_1 n_2(n_1 + n_2 + 1)/12$ and test statistic $Z = \frac{W - E(W)}{\sqrt{Var(W)}}$ (Hollander and Wolfe 1999, p.108).
- In the case of ties the variance needs to be modified for the Gaussian approximation. Let n_t be the number of groups with ties and t_k the number of ties in group k then $Var(W) = (n_1 n_2/12) \times \left[n_1 + n_2 + 1 - \sum_{k=1}^{n_t} (t_k^3 - t_k)/((n_1 + n_2)(n_1 + n_2 - 1)) \right]$.

Example: To test the hypothesis that the two populations of healthy subjects and subjects with hypertension are equal in location with respect to their mean systolic blood pressure. The dataset contains $n_1 = 25$ healthy subject (status=0) and $n_2 = 30$ subjects with hypertension (status=1) (dataset in Table A.1).

SAS code

```
proc npar1way data=blood_pressure wilcoxon correct=yes;
 class status;
 var mmhg;
 exact wilcoxon;
run;
```

SAS output

```
    Wilcoxon Scores (Rank Sums) for Variable mmhg
           Classified by Variable status

                 Sum of    Expected   Std Dev      Mean
status    N      Scores    Under H0   Under H0     Score
-----------------------------------------------------------
 0        25      343.0     700.0     59.129843    13.720
 1        30     1197.0     840.0     59.129843    39.900

             Average scores were used for ties.

                  Wilcoxon Two-Sample Test

             Statistic (S)                 343.0000

             Normal Approximation
             Z                              -6.0291
             One-Sided Pr <  Z               <.0001
             Two-Sided Pr > |Z|              <.0001

             t Approximation
             One-Sided Pr <  Z               <.0001
             Two-Sided Pr > |Z|              <.0001

             Exact Test
             One-Sided Pr <=  S           4.702E-13
             Two-Sided Pr >= |S - Mean|   9.414E-13

            Z includes a continuity correction of 0.5.
```

Remarks:

- The parameter `wilcoxon` enables the Wilcoxon rank-sum test of the procedure NPAR1WAY.
- `correct=`*value* is optional. If *value* is YES than a continuity correction for the normal approximation is used. The default is NO.
- `exact wilcoxon` is optional and applies an additional exact test. Note, the computation of an exact test can be very time consuming.
- SAS also invokes a t-distribution approximation in addition to the normal approximation.

- Besides the two-sided p-value SAS also reports a one-sided p-value. Which one is printed depends on the Z-statistic. If the value of the Z-statistic is greater than zero the p-value for the right-tailed test is printed, otherwise the p-value for the left-tailed test is printed.
- In this example the sum of scores for the healthy subjects is 343.0 compared with 1197.0 for the people with hypertension. So there is evidence of a locations shift in the sense that the median of healthy subjects is lower than the median of unhealthy subjects. The p-value for hypothesis (C) is $P(\Pr < z) < 0.0001$ and the p-value for hypothesis (B) is $1 - P(\Pr < z) = 1$.

R code

```
x<-blood_pressure$mmhg[blood_pressure$status==0]
y<-blood_pressure$mmhg[blood_pressure$status==1]

wilcox.test(x,y,exact=FALSE,correct=TRUE,
                                  alternative="two.sided")
```

R output

```
     Wilcoxon rank sum test with continuity correction

data:  x and y
W = 18, p-value = 1.649e-09
alternative hypothesis: true location shift is
                                          not equal to 0
```

Remarks:

- `exact=`*value* is optional. If *value* is TRUE an exact p-value is computed, if it is FALSE an approximative p-value is computed. If `exact` is not specified or NULL (default value) an exact p-value is only computed if the sample size is less than 50 and no ties are present. Otherwise, the Gaussian distribution is used. In the case of ties R cannot compute an exact test.
- `correct=`*value* is optional. If the *value* is TRUE (default value) a continuity correction to the Gaussian approximation is used, that is, a value of 0.5 is subtracted or added to the numerator of the Z-statistic.
- `alternative=`*"value"* is optional and defines the type of alternative hypothesis: "two.sided"= true location shift is not equal to 0 (A); "greater"=true location shift is greater than 0 (B); "less"=true location shift is less than 0 (C). Default is "two.sided".
- The reported test statistic W is in fact the Mann–Whitney U test statistic. It is calculated as $U = W - n_1(n_1 + 1)/2$. From the SAS output we know that $W = 343$ is the sum of scores with $n_1 = 25$. So, $U = 343 - 25 * 26/2 = 18$. The tests based on both statistics are equivalent.

8.2.2 Wilcoxon matched-pairs signed-rank test

Description: Tests if the location (median m) of the difference of populations is zero, in the case of paired samples.

Assumptions:
- Data are measured on an interval or ratio scale.
- The random variables X and Y are observed in pairs with observations $(x_i, y_i) \quad i = 1, \ldots, n$.
- The differences $D_i = X_i - Y_i$ are independent and identically distributed.
- The distribution of the D_i is continuous and symmetric around the median m.

Hypotheses:
(A) $H_0 : m = 0$ vs $H_1 : m \neq 0$
(B) $H_0 : m = 0$ vs $H_1 : m > 0$
(C) $H_0 : m = 0$ vs $H_1 : m < 0$

Test statistic:

$$W^+ = \sum_{i=1}^{n} R_i \mathbb{1}_{]0,\infty[}(D_i), \text{ with } R_i = rank|D_i|, \quad \text{for} \quad i = 1, \ldots, n$$

Test decision: Reject H_0 if for the observed value w of W^+
(A) $w \geq w_{\alpha/2}$ or $w \leq w_{1-\alpha/2}$
(B) $w \geq w_{\alpha}$
(C) $w \leq w_{1-\alpha}$

p-value:
(A) $p = 2\min(P(W \geq w), 1 - P(W \geq w))$
(B) $p = P(W \geq w)$
(C) $p = 1 - P(W \geq w)$

Annotations:
- Critical values for the test can be found in McCornack 1965.
- The hypotheses can be extended to the case $H_0 : m = m_0$ vs $H_1 : m \neq m_0$ by using $D_i^* = X_i - Y_i - m_0$ instead of $D_i = X_i - Y_i$.
- Note, the hypothesis $m = 0$ does not equal the hypothesis $m_X = m_Y$ unless the random variables X and Y are symmetric distributed around their medians m_X or m_Y.
- This test is the nonparametric equivalent to the paired t-test (Test 2.2.5).

Example: To test that the difference of the median intelligence quotients before training (IQ1) and after training (IQ2) is zero. The dataset contains 20 subjects (dataset in Table A.2).

SAS code

```
data temp;
 set iq;
 diff=iq1-iq2;
run;

proc univariate data=temp mu0=0 loccount;
 var diff;
run;
```

SAS output

```
           Tests for Location: Mu0=0

Test            -Statistic-     -----p Value------

Signed Rank     S       -105    Pr >= |S|    <.0001

            Location Counts: Mu0=0.00

            Count                 Value

            Num Obs > Mu0             0
            Num Obs ^= Mu0           20
            Num Obs < Mu0            20
```

Remarks:

- `PROC UNIVARIATE` calculates only the p-value for hypothesis (A). To find the p-values of the one-sided hypotheses please refer to the example of Test 8.1.2.
- `mu0=`*value* is optional and indicates the value m_0 to test against. Default is 0.
- `loccount` is optional and prints out the number of observations greater than, equal to, or less than m_0.
- SAS uses a different test statistic: $S = W^+ - E(W^+)$. This yields a value of -105 instead of 0.
- If $n \leq 20$ the exact distribution of S is used for the calculation of the p-value. Otherwise an approximation to the t-distribution is applied (Iman 1974).

R code

```
wilcox.test(iq$IQ1,iq$IQ2,mu=0,paired=TRUE,exact=FALSE,
                      correct=FALSE,alternative="two.sided")
```

R output

```
        Wilcoxon signed rank test

data:  iq$IQ1 and iq$IQ2
V = 0, p-value = 1.711e-05
alternative hypothesis: true location shift is
                                          not equal to 0
```

Remarks:

- `mu=`*value* is optional an indicates the value m_0 to test against. Default is 0.
- `paired=`*TRUE* invokes this test. If *value* is FALSE or `paired` is missing Test 8.1.2 is instead performed.
- `exact=`*value* is optional. If *value* is TRUE an exact p-value is computed, if it is FALSE an approximative p-value is computed. If `exact` is not specified or NULL (default value) an exact p-value is computed if the sample size is less than 50 and no ties are present. Otherwise, the Gaussian distribution is used. In the case of ties this function cannot compute an exact test.
- `correct=`*value* is optional. If the *value* is TRUE (default value) a continuity correction to the Gaussian approximation is used, that is, a value of 0.5 is subtracted or added to the numerator of the Z-statistic.
- `alternative=`*"value"* is optional and defines the type of alternative hypothesis: "two.sided"= true location shift is not equal to 0 (A); "greater"=true location shift is greater than 0 (B); "less"=true location shift is less than 0 (C). Default is "two.sided".

8.3 *K*-sample tests

The Kruskal–Wallis test (Kruskal 1952; Kruskal and Wallis 1952) is the extension of the Wilcoxon rank-sum test (Test 8.2.1) for more than two independent samples.

8.3.1 Kruskal–Wallis test

Description: Tests if the location (median) of three or more populations is the same.

Assumptions:

- Data are measured at least on an ordinal scale.
- Samples $X_{j1}, \ldots, X_{jn_j}$ are independently taken from k populations, $j = 1, \ldots, k, N = n_1 + \ldots + n_k$.
- The k populations are described by independent random variables $X_1, \ldots X_k$ with continuous distribution and distribution functions $F_1, \ldots, F_k$.
- The distribution functions differ in their location, that is, they can be described by a distribution function $F(t)$ of a continuous distribution and constants τ_j with $F_j(t) = F(t - \tau_j), j = 1 \ldots k$.

Hypotheses: $H_0 : \tau_1 = \cdots = \tau_k$ vs $H_1 : \tau_l \neq \tau_m$ for at least one pair l, m with $l \neq m$

Test statistic:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{1}{n_j} \left(R_j - \frac{n_j(N+1)}{2} \right)^2$$

with R_j sum of ranks of $X_{j1}, \ldots, X_{jn_j}$ in the combined sample

Test decision: Reject H_0 if for the observed value h of H
$h \geq h_{k,(n_1,\ldots,n_k),\alpha}$

p-value: $p = P(H \geq h)$

Annotations:

- For the calculation of the test statistic, first combine all samples and rank the combined sample from the lowest to the highest values. R_j is the sum of the ranks of $X_{j1}, \ldots, X_{jn_j}$ in the combined sample.
- Critical values $h_{k,(n_1,\ldots,n_k);\alpha}$ for the test statistic H can be found in table A.12 of Hollander and Wolfe (1999).
- For an alternative large sample test it can be used that the test statistic H is asymptotically χ^2-distributed with $k-1$ degrees of freedom. Hence, the null hypothesis is rejected if $h > \chi^2_{k-1;1-\alpha}$.
- In the case of ties, that is, observations with the same values, mid ranks are used and H must be adjusted. Let n_t be the number of groups with ties and t_p the number of ties in group p then the test statistic $H' = H/B$ is used with $B = 1 - \frac{1}{N^3-N} \sum_{p=1}^{n_t} (t_p^3 - t_p)$. Now, the above test can be applied as an approximate test.

Example: To test the hypothesis that the diameters of workpieces produced by three different machines do not differ in location (median). A dataset is available with $n_1 = n_2 = n_3 = 10$ observations from each machine (dataset in Table A.3).

SAS code

```
proc npar1way data=workpieces wilcoxon;
 class machine;
 var diameter;
 exact wilcoxon;
run;
```

SAS output

```
The NPAR1WAY Procedure

   Wilcoxon Scores (Rank Sums) for Variable diameter
            Classified by Variable machine
```

```
                  Sum of     Expected    Std Dev      Mean
machine     N     Scores     Under H0    Under H0     Score
-----------------------------------------------------------
1           10    174.0      155.0       22.730303    17.40
2           10    147.0      155.0       22.730303    14.70
3           10    144.0      155.0       22.730303    14.40

                    Kruskal-Wallis Test

          Chi-Square                      0.7045
          DF                                   2
          Asymptotic Pr >  Chi-Square     0.7031
          Exact      Pr >= Chi-Square     0.7157
```

Remarks:

- The parameter `wilcoxon` yields the Kruskal–Wallis test of the procedure NPAR1WAY, if there are more than two levels in the classification variable.
- `exact wilcoxon` is optional and applies an additional exact test. Note that the computation of an exact test can be very time consuming.

R code

```
kruskal.test(workpieces$diameter ~ workpieces$machine)
```

R output

```
Kruskal-Wallis rank sum test

data:  workpieces$diameter by workpieces$machine
Kruskal-Wallis chi-squared = 0.7045, df=2, p-value = 0.7031
```

Remarks:

- Also the alternative code `kruskal.test(workpieces$diameter,workpieces$machine)` is possible.
- This function reports only the asymptotic p-value.

References

Arbuthnot J. 1710 An argument for divine providence, taken from the constant regularity observed in the birth of both sexes. *Philosophical Transactions of the Royal Society of London* **27**, 186–190.

Gibbons J.D. 1988 Sign tests. In *Encyclopedia of Statistical Sciences* (eds Kotz S., Johnson N.L. and Campbell B.), Vol. 8, pp. 471–475. John Wiley & Sons, Ltd.

Hollander M. and Wolfe D.A. 1999 *Nonparametric Statistical Methods*, 2nd edn. John Wiley & Sons, Ltd.

Iman R.L. 1974 Use of a t-statistic as an approximation to the exact distribution of the Wilcoxon signed rank statistic. *Communications in Statistics* **3**, 795–806.

Kruskal W.H. 1952 A nonparametric test for the several sample problem. *Annals of Mathematical Statistics* **23**, 525–540.

Kruskal W.H. and Wallis W.A. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* **47**, 583–621.

Mann H. and Whitney D. 1947 On a test of whether one or two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* **18**, 50–60.

McCornack R.L. 1965 Extended tables of the Wilcoxon matched pairs signed rank statistics. *Journal of the American Statistical Association* **60**, 864–871.

Owen D.B. 1962 *Handbook of Statistical Tables*. Addison Wesley.

Wilcoxon F. 1945 Individual comparisons by ranking methods. *Biometrics* **1**, 80–83.

Wilcoxon F. 1949 *Some Rapid Approximate Statistical Procedures*. Stanford Research Laboratories, American Cyanamid Corporation.

9

Tests on scale difference

In this chapter we present nonparametric tests for the scale parameter. Actually, it is tested if two samples come from the same population where alternatives are characterized by differences in dispersion. These tests are called tests on the scale, spread or dispersion. The most famous one is the Siegel–Tukey test (Test 9.1.1). The introduced tests can be employed if the samples are not normally distributed, but the equality of median assumption is crucial.

9.1 Two-sample tests

9.1.1 Siegel–Tukey test

Description: Tests if the scale (variance) of two independent populations is the same.

Assumptions:
- Data are measured at least on an ordinal scale.
- Samples $X_i, i = 1, \ldots, n_1$ and $Y_j, j = 1, \ldots, n_2$ are independently drawn from the two populations, $n = n_1 + n_2$.
- The random variables X and Y are independent with continuous distribution functions F and G, scale parameters σ_X^2, σ_Y^2 and median m_X, m_Y. It holds that $m_X = m_Y$.
- F and G belong to the same distribution function with possibly differences in scale and location. Under the assumption of equal median, the hypothesis $H_0 : F(t) = G(t)$ reduces to $H_0 : \sigma_X = \sigma_Y$.

Hypotheses:
(A) $H_0 : \sigma_X = \sigma_Y$ vs $H_1 : \sigma_X \neq \sigma_Y$
(B) $H_0 : \sigma_X = \sigma_Y$ vs $H_1 : \sigma_X > \sigma_Y$
(C) $H_0 : \sigma_X = \sigma_Y$ vs $H_1 : \sigma_X < \sigma_Y$

Test statistic: For $n_1 < n_2$ the test statistic is given by:

$S =$ sum of ranks of $X_1, \ldots, X_{n_1}$ in the combined sample

Here ranks are assigned to the ordered combined sample as follows for n even

Statistical Hypothesis Testing with SAS and R, First Edition. Dirk Taeger and Sonja Kuhnt.

$$R_i = \begin{cases} 2i, & i \text{ even and } 1 < i < n/2 \\ 2(n-i)+2, & i \text{ even and } n/2 < i \leq n \\ 2i-1, & i \text{ odd and } 1 \leq i \leq n/2 \\ 2(n-i)+1, & i \text{ odd and } n/2 < i < n \end{cases}$$

If n is uneven, the above ranking is applied after the middle observation of the combined and ordered sample is discarded and the sample size is reduced to $n-1$.

Test decision: Reject H_0 if for the observed value s of S
(A) $s \geq s_{\alpha/2}$ or $s \leq n_1(n_1+n_2+1) - s_{\alpha/2}$
(B) $s \geq s_\alpha$
(C) $s \leq n_1(n_1+n_2+1) - s_\alpha$

p-value: (A) $p = 2\min(P(S \geq s), 1 - P(S \geq n_1(n_1+n_2+1) - s))$
(B) $p = P(S \geq s)$
(C) $p = 1 - P(S \geq n_1(n_1+n_2+1) - s)$

Annotations:

- Tables with critical values s_α can be found in Siegel and Tukey (1980). Due to the used ranking procedure the same tables for critical values can be used as for the Wilcoxon rank sum test for location.
- For the calculation of the test statistic, first combine both samples and rank the combined sample from the lowest to the highest values according to the above ranking scheme. Hence, the lowest value gets the rank 1, the highest value the rank 2, the second highest value the rank 3, the second lowest value the rank 4, the third lowest value the rank 5, and so forth. The above test statistic S is the sum of the ranks of the sample of X based on the assumption $n_1 \leq n_2$. The test can also be based on the ranks of Y-observations in the combined sample. Usually the sum of ranks of the sample with the smaller sample size is used due to arithmetic convenience (Siegel and Tukey 1980).
- The distribution with the larger scale will have the lower sum of ranks, because the lower ranks are on both ends of the combined sample.
- It is not necessary to remove the middle observation if the combined sample size is odd. The advantage of this is, that the sum of ranks of adjacent observations is always the same and therefore the sum of ranks is a symmetric distribution under H_0.
- For large samples the test statistic $Z = \frac{2S - n_1(n_1+n_2+1) \pm 1}{\sqrt{n_1(n_1+n_3+1)(n_2/3)}}$ can be used, which is approximately a standard normal distribution. The sign has to be chosen such that $|z|$ is smaller (Siegel and Tukey 1980).

Example: To test the hypothesis that the dispersion of the systolic blood pressure in the two populations of healthy subjects (status=0) and subjects with hypertension (status=1) is the same. The dataset contains $n_1 = 25$ observations for status=0 and $n_2 = 30$ observations for status=1 (dataset in Table A.1).

SAS code

```
proc npar1way data=blood_pressure correct=no st;
 var mmhg;
 class status;
 exact st;
run;
```

SAS output

```
                    The NPAR1WAY Procedure

           Siegel-Tukey Scores for Variable mmhg
              Classified by Variable status

                  Sum of    Expected   Std Dev     Mean
status    N       Scores    Under H0   Under H0    Score
-----------------------------------------------------------
0        25       655.0      700.0     59.001584   26.20
1        30       885.0      840.0     59.001584   29.50

               Average scores were used for ties.

                  Siegel-Tukey Two-Sample Test

                  Statistic             655.0000
                  Z                      -0.7627
                  One-Sided Pr <  Z       0.2228
                  Two-Sided Pr > |Z|      0.4456
```

Remarks:

- The parameter `st` enables the Siegel–Tukey test of the procedure NPAR1WAY.
- `correct=`*value* is optional. If *value* is YES than a continuity correction for the normal approximation is used. The default is NO.
- `exact st` is optional and applies an additional exact test. Note, the computation of an exact test can be very time consuming. This is the reason why in this example no exact p-values are given in the output.
- Besides the two-sided p-value SAS also reports a one-sided p-value; which one is printed depends on the Z-statistic. If it is greater than zero the right-sided p-value is printed. If it is less than or equal to zero the left-sided p-value is printed.
- In this example the sum of scores for the healthy subjects is 655.0 compared with 885.0 for the people with hypertension. So there is evidence that the scale of healthy subjects is higher than the scale of unhealthy subjects. In fact the variance of the healthy subjects is 124.41 and the variance of the unhealthy subjects is 120.05. Therefore the p-value for hypothesis (C) is $P(\Pr < Z) = 0.2228$ and the p-value for hypothesis (B) is $1 - P(\Pr < Z) = 0.7772$.
- In the case of odd sample sizes SAS does not delete the middle observation.

R code

```
# Helper functions to find even or odd numbers
is.even <- function(x) x %% 2 == 0
is.odd  <- function(x) x %% 2 == 1

# Create a sorted matrix with first column the blood
# pressure and second column the status
data<-blood_pressure[order(blood_pressure$mmhg),]
x<-c(data$mmhg)
x<-cbind(x,data$status)

# If the sample size is odd then remove the observation
# in the middle
if (is.odd(nrow(x))) x<-x[-c(nrow(x)/2+0.5),]

# Calculate the (remaining) sample size
n<-nrow(x)

# y returns the Siegel-Tukey scores
y<-rep(0,times=n)

# Assigning the scores
for (i in seq(along=x)) {
 if (1<i & i <= n/2 & is.even(i))
 {
  y[i]<-2*i
 }
 else if (n/2<i & i<=n & is.even(i))
 {
  y[i]<-2*(n-i)+2
 }
 else if (1<=i & i <=n/2 & is.odd(i))
 {
  y[i]<-2*i-1
 }
 else if (n/2<i & i < n & is.odd(i))
 {
  y[i]<-2*(n-i)+1
 }
}

# Now mean scores must be created if necessary
t<-tapply(y,x[,1],mean) # Get mean scores for tied values
v<-strsplit(names(t), " ") # Get mmhg values

# r
r<-rep(0,times=n)

# Assign ranks and mean ranks to r
for (i in seq(along=r))
{
 for (j in seq(along=v))
```

```
 {
 if (x[i,1]==as.numeric(v[j])) r[i]=t[j]
 }
}

# Now calculate the test statistics S_0 (status 0)
# and S_1 (status 1) for both samples
S_0<-0
S_1<-0

for (i in seq(along=r)) {
 if(x[i,2]==0) S_0=S_0+r[i]
 if(x[i,2]==1) S_1=S_1+r[i]
}

# Calculate sample sizes for status=0 and status=1
n1<-sum(x[,2]==0)
n2<-sum(x[,2]==1)

# Choose the test statistic which belongs to the smallest
# sample size
if (n1<=n2) {
  # Choose the smaller |z| value
  z1<-(2*S_0-n1*(n+1)+1)/sqrt((n1*n2*(n+1)/3))
  z2<-(2*S_0-n1*(n+1)-1)/sqrt((n1*n2*(n+1)/3))
  if (abs(z1)<=abs(z2)) z=z1 else z=z2

  pvalue_B=1-pnorm(-abs(z))
  pvalue_C=pnorm(-abs(z))
}

if (n1>n2) {
  # Choose the smaller |z| value
  z1<-(2*S_1-n2*(n+1)+1)/sqrt((n1*n2*(n+1)/3))
  z2<-(2*S_1-n2*(n+1)-1)/sqrt((n1*n2*(n+1)/3))
  if (abs(z1)<=abs(z2)) z=z1 else z=z2

  pvalue_B=pnorm(-abs(z));
  pvalue_C=1-pnorm(-abs(z));
}

pvalue_A=2*min(pnorm(-abs(z)),1-pnorm(-abs(z)));

# Output results
print("Siegel-Tukey test")
n
S_0
S_1
z
pvalue_A
pvalue_B
pvalue_C
```

R output

```
[1] "Siegel-Tukey test"
> n
[1] 54
> S_0
[1] 600.5
> S_1
[1] 884.5
> z
[1] -1.027058
> pvalue_A
[1] 0.3043931
> pvalue_B
[1] 0.8478035
> pvalue_C
[1] 0.1521965
```

Remarks:

- There is no basic R function to calculate this test directly.
- In this implementation of the test, the observation in the middle of the sorted sample is removed. This is different to SAS and therefore the calculated values of the test statistic are not the same.
- In the case of ties – as in the above sample – the construction of ranks must be made in two passes. First the ranks are constructed in the ordered combined sample. Afterwards the mean of ranks of the tied observations are calculated.

9.1.2 Ansari–Bradley test

Description: Tests if the scale (variance) of two independent populations is the same.

Assumptions:
- Data are measured at least on an ordinal scale.
- Samples $X_i, i = 1, \ldots, n_1$ and $Y_j, j = 1, \ldots, n_2$ are independently drawn from the two populations, $n = n_1 + n_2$.
- The random variables X and Y are independent with continuous distribution functions F and G, scale parameters σ_X^2, σ_Y^2 and median m_X, m_Y. It holds that $m_X = m_Y$.
- F and G belong to the same distribution function with possibly differences in scale and location. Under the assumption of equal median, the hypothesis $H_0 : F(t) = G(t)$ reduces to $H_0 : \sigma_X = \sigma_Y$.

Hypotheses:
(A) $H_0 : \sigma_X = \sigma_Y$ vs $H_1 : \sigma_X \neq \sigma_Y$
(B) $H_0 : \sigma_X = \sigma_Y$ vs $H_1 : \sigma_X > \sigma_Y$
(C) $H_0 : \sigma_X = \sigma_Y$ vs $H_1 : \sigma_X < \sigma_Y$

Test statistic: For $n_1 < n_2$ the test statistic is given by:
$A =$ sum of ranks of $X_1, \ldots, X_{n_1}$ in the combined sample.

Here ranks are assigned to the ordered combined sample as follows for $n = n_1 + n_2$ even

$$R_i = \begin{cases} i, & 1 \le i \le n/2 \\ n - i + 1 & n/2 < i \le n \end{cases} \text{ and for odd } n\text{: } R_i = \begin{cases} i, & 1 \le i \le (n+1)/2 \\ n - i + 1 & (n+1)/2 < i \le n \end{cases}$$

Test decision: Reject H_0 if for the observed value a of A
(A) $a \ge c_{\alpha_1}$ or $a \le (c_{1-\alpha_2} - 1)$ with $\alpha_1 + \alpha_2 = \alpha$
(B) $a \ge c_\alpha$
(C) $a \le (c_{1-\alpha} - 1)$

p-value: (A) $p = 2\min(P(A \ge a), 1 - P(A \ge a))$
(B) $p = P(A \ge a)$
(C) $p = 1 - P(A \ge a)$

Annotations:

- For the calculation of the test statistic, first combine both samples and rank the combined sample from the lowest to the highest values according to the above ranking scheme. It means that for even sample size the series of ranks will be $1, 2, \ldots, n/2, \ldots, 2, 1$ and for odd sample size it will be $1, 2, \ldots, (n-1)/2, (n+1)/2, (n-1)/2, \ldots, 2, 1$. (Ansari and Bradley 1960). The distribution with the larger scale will have the lower sum of ranks because the lower ranks are on the both ends of the combined sample.
- Here, c_α denotes the upper-tail probability for the null distribution of the Ansari–Bradley statistic calculated for the sample with the smaller sample size; tables are given in Ansari and Bradley (1960) as well as in Hollander and Wolfe (1999, table A.8). In general, the test can alternatively be set up by using the sum of ranks of the sample with the larger sample size as the test statistic.
- In the case of tied observations mean ranks are used.
- For large sample sizes (n_1 and $n_2 \ge 20$) the test statistic A is asymptotically normally distributed. If no ties are present and $n = n_1 + n_2$ is even, then $E(A) = n_1(n+2)/4$ and $Var(A) = [n_1 n_2 (n+2)(n-2)]/[48(n+1)]$. If no ties are present and n is odd, then $E(A) = n_1(n+1)^2/[4n]$ and $Var(A) = [n_1 n_2 (n+1)(3+n^2)]/[48n^2]$. In the case of ties the expectation is the same, but the variance is somewhat different. Let g be the number of tied groups, t_j the number of tied observations in group j, and r_j the middle range in group j.

  If n is even, then $Var(A) = n_1 n_2 (16 \sum_{j=1}^{g} t_j r_j^2 - n(n+2)^2)/(16n(n-1))$.

  If n is odd, then $Var(A) = n_1 n_2 (16n \sum_{j=1}^{g} t_j r_j^2 - (n+1)^4)/(16n^2(n-1))$.

  (Hollander and Wolfe 1999, p. 145).

Example: To test the hypothesis that the dispersion of the systolic blood pressure in the two populations of healthy subjects (status=0) and subjects with hypertension (status=1) is the same. The dataset contains $n_1 = 25$ observations for status=0 and $n_2 = 30$ observations for status=1 (dataset in Table A.1).

SAS code

```
proc npar1way data=blood_pressure correct=no ab;
 var mmhg;
 class status;
 exact ab;
run;
```

SAS output

```
                  The NPAR1WAY Procedure

           Ansari-Bradley Scores for Variable mmhg
               Classified by Variable status

                     Sum of      Expected   Std Dev  Mean
status  N   Scores  Under H0    Under H0   Score    Score
-----------------------------------------------------------
0       25  334.0   356.363636  29.533137  13.360   13.360
1       30  450.0   427.636364  29.533137  15.000   15.000

             Average scores were used for ties.

              Ansari-Bradley Two-Sample Test

              Statistic             334.0000
              Z                      -0.7572
              One-Sided Pr < Z        0.2245
              Two-Sided Pr < |Z|      0.4489
```

Remarks:

- The parameter `ab` enables the Ansari–Bradley test of the procedure NPAR1WAY.
- `correct=`*value* is optional. If *value* is YES than a continuity correction for the normal approximation is used. The default is NO.
- `exact ab` is optional and applies an additional exact test. Note, the computation of an exact test can be very time consuming. This is the reason why in this example no exact p-values are given in the output.
- Besides the two-sided p-value SAS also reports a one-sided p-value; which one is printed depends on the Z-statistic. If the value of the Z-statistic is greater than zero the right-sided p-value is printed. If it is less than or equal to zero the left-sided p-value is printed.
- In this example the sum of scores for the healthy subjects is 334.0 compared with 450.0 for the people with hypertension. So there is evidence that the scale of healthy subjects is higher than the scale of unhealthy subjects. In fact the variance of the healthy subjects is 124.41 and the variance of the unhealthy subjects is 120.05. Therefore the p-value for hypothesis (C) is $P(\Pr < Z) = 0.2245$ and the p-value for hypothesis (B) is $1 - P(\Pr < Z) = 0.7775$.

R code

```
x<-blood_pressure$mmhg[blood_pressure$status==0]
y<-blood_pressure$mmhg[blood_pressure$status==1]

ansari.test(x,y,exact=NULL,alternative ="two.sided")
```

R output

```
        Ansari-Bradley test

data:  x and y
AB = 334, p-value = 0.4489
alternative hypothesis: true ratio of scales is not
                                                equal to 1
```

Remarks:

- `exact=`*value* is optional. If *value* is not specified or TRUE an exact p-value is computed if the combined sample size is less than 50. If it is NULL or FALSE the approximative p-value is computed. In the case of ties R cannot compute an exact test.
- R tests equivalent hypotheses of the type $H_0 : \sigma_x/\sigma_Y = 1$ vs $H_1 : \sigma_x/\sigma_Y \neq 1$ for hypothesis (A), and so on.
- `alternative=`*"value"* is optional and defines the type of alternative hypothesis: "two.sided"= true ratio of scales is not equal to 1 (A); "greater"=true ratio of scales is greater than 1 (C); "lower"=true ratio of scales is less than 1 (B). Default is "two.sided".

9.1.3 Mood test

Description: Tests if the scale (variance) of two independent populations is the same.

Assumptions:
- Data are measured at least on an ordinal scale.
- Samples $X_i, i = 1, \ldots, n_1$ and $Y_j, j = 1, \ldots, n_2$ are independently drawn from the two populations, $n = n_1 + n_2$.
- The random variables X and Y are independent with continuous distribution functions F and G, scale parameters σ_X^2, σ_Y^2 and median m_X, m_Y. It holds that $m_X = m_Y$.
- F and G belong to the same distribution function with possibly differences in scale and location. Under the assumption of equal median, the hypothesis $H_0 : F(t) = G(t)$ reduces to $H_0 : \sigma_X = \sigma_Y$.

Hypotheses:
(A) $H_0 : \sigma_X = \sigma_Y$ vs $H_1 : \sigma_X \neq \sigma_Y$
(B) $H_0 : \sigma_X = \sigma_Y$ vs $H_1 : \sigma_X > \sigma_Y$
(C) $H_0 : \sigma_X = \sigma_Y$ vs $H_1 : \sigma_X < \sigma_Y$

Test statistic: For $n_1 < n_2$ the test statistic is given by:
$M = \sum_{i=1}^{n_1} \left(R_i - \frac{n_1+n_2+1}{2} \right)^2$
where R_i is the rank of the ith X-observation in the combined sample

Test decision: Reject H_0 if for the observed value m of M
(A) $a \geq c_{\alpha_1}$ or $a \leq (c_{1-\alpha_2} - 1)$ with $\alpha_1 + \alpha_2 = \alpha$
(B) $a \geq c_\alpha$
(C) $a \leq (c_{1-\alpha} - 1)$

p-value: (A) $p = 2\min(P(A \geq a), 1 - P(A \geq a))$
(B) $p = P(A \geq a)$
(C) $p = 1 - P(A \geq a)$

Annotations:

- Tables with critical values c_α can be found in Laubscher *et al.* (1968).
- For the calculation of the test statistic, first combine both samples and rank the combined sample from the lowest to the highest values. Above test statistic M is the sum of the quadratic distance of the ranks of the X-observations from the median of all ranks based on the assumption $n_1 \leq n_2$. The test can also be based on the ranks of Y-observations in the combined sample. Usually the sum of ranks of the sample with the smaller sample size is used.
- In the case of tied observations mid ranks are used. However, tied observations only influence the test statistics if they are between the X- and Y-observations.
- For large sample sizes ($n_1 + n_2 \geq 20$) the test statistic is asymptotically normally distributed with $E(M) = n_1[(n_1 + n_2)^2 - 1]/12$ and $Var(M) = [n_1 n_2 (n_1 + n_2 + 1)((n_1 + n_2)^2 - 4)]/180$ (Mood 1954).

Example: To test the hypothesis that the dispersion of the systolic blood pressure in the two populations of healthy subjects (status=0) and subjects with hypertension (status=1) is the same. The dataset contains $n_1 = 25$ observations for status=0 and $n_2 = 30$ observations for status=1 (dataset in Table A.1).

SAS code

```
proc npar1way data=blood_pressure correct=no mood;
 var mmhg;
 class status;
 exact mood;
run;
```

SAS output

```
                The NPAR1WAY Procedure

            Mood Scores for Variable mmhg
            Classified by Variable status

               Sum of     Expected    Std Dev       Mean
status   N     Scores     Under H0    Under H0      Score
-----------------------------------------------------------
0        25   6864.0     6300.0      837.786511    274.560
1        30   6996.0     7560.0      837.786511    233.200

               Average scores were used for ties.

                     Mood Two-Sample Test

               Statistic             6864.0000
               Z                        0.6732
               One-Sided Pr >  Z        0.2504
               Two-Sided Pr > |Z|       0.5008
```

Remarks:

- The parameter `mood` enables the Mood test of the procedure NPAR1WAY.
- `correct=`*value* is optional. If *value* is YES than a continuity correction for the normal approximation is used. The default is NO.
- `exact mood` is optional and applies an additional exact test. Note, the computation of an exact test can be very time consuming. This is the reason why in this example no exact p-values are given in the output.
- Besides the two-sided p-value SAS also reports a one-sided p-value; which one is printed depends on the Z-statistic. If the observed value of the Z-statistic is greater than zero the right-sided p-value is printed. If it is less than or equal to zero the left-sided p-value is printed.
- In this example the sum of scores for the healthy subjects is 6864.0 compared with 6996.0 for the people with hypertension. So there is evidence that the scale of healthy subjects is higher than the scale of unhealthy subjects. In fact the variance of the healthy subjects is 124.41 and the variance of the unhealthy subjects is 120.05. Therefore the p-value for hypothesis (C) is $1 - P(\Pr > Z) = 0.7496$ and the p-value for hypothesis (B) is $P(\Pr > Z) = 0.2504$.

R code

```
x<-blood_pressure$mmhg[blood_pressure$status==0]
y<-blood_pressure$mmhg[blood_pressure$status==1]

mood.test(x,y,alternative ="two.sided")
```

R output

```
 Mood two-sample test of scale

data:  x and y
Z = 0.6765, p-value = 0.4987
alternative hypothesis: two.sided
```

Remarks:

- R handles ties differently to SAS. Instead of mid ranks a procedure by Mielke is used (Mielke 1967).
- `alternative=`*"value"* is optional and defines the type of alternative hypothesis: "two.sided"= true ratio of scales is not equal to 1 (A); "greater"=true ratio of scales is greater than 1 (C); "lower"=true ratio of scales is less than 1 (B). Default is "two.sided".

References

Ansari A.R. and Bradley R.A. 1960 Rank-sum tests for disperson. *Annals of Mathematical Statistics* **31**, 1174–1189.

Hollander M. and Wolfe D.A. 1999 *Nonparametric Statistical Methods*, 2nd edn. John Wiley & Sons, Ltd.

Laubscher N.F., Steffens F.E. and DeLange E.M. 1968 Exact critical values for Mood's distribution-free test statistic for dispersion and its normal approximation. *Technometrics* **10**, 497–508.

Mielke P.W. 1967 Note on some squared rank tests with existing ties. *Technometrics* **9**, 312–314.

Mood A.M. 1954 On the asymptotic efficiency of certain nonparametric two-sample tests. *Annals of Mathematical Statistics* **25**, 514–522.

Siegel S. and Tukey J.W. 1980 A nonparametric sum of ranks procedure for relative spread in unpaired samples. *Journal of the American Statistical Association* **55**, 429–445.

10

Other tests

In this chapter we present a well-known test for the problem if two independent samples are drawn from the same population or not. The test is based on very few assumptions, for example, it is not necessary to specify the distributions beyond the fact that they are continuous distributions.

10.1 Two-sample tests

10.1.1 Kolmogorov–Smirnov two-sample test (Smirnov test)

Description: Tests if two independent samples are sampled from the same distribution.

Assumptions:
- Data are at least measured on an ordinal scale.
- The random variables X_1 and X_2 are independent with continuous distribution functions $F_1(x)$ and $F_2(x)$.
- Samples $X_{1j}, j = 1, \ldots, n_1$ and $X_{2j}, j = 1, \ldots, n_2$ are independently drawn from the two populations.

Hypotheses:
(A) $H_0 : F_1(x) = F_2(x)$ vs $H_1 : F_1(x) \neq F_2(x)$ for at least one x
(B) $H_0 : F_1(x) = F_2(x)$ vs $H_1 : F_1(x) \geq F_2(x)$ with $F_1(x) \neq F_2(x)$ for at least one x
(C) $H_0 : F_1(x) = F_2(x)$ vs $H_1 : F_1(x) \leq F_2(x)$ with $F_1(x) \neq F_2(x)$ for at least one x

Test statistic:
(A) $D = \max_x |F_{n_1}(x) - F_{n_2}(x)|$
(B) $D^+ = \max_x (F_{n_1}(x) - F_{n_2}(x))$
(C) $D^- = \max_x (F_{n_2}(x) - F_{n_1}(x))$
where F_{n_1} and F_{n_2} denote the empirical distribution functions based on the two samples.

Statistical Hypothesis Testing with SAS and R, First Edition. Dirk Taeger and Sonja Kuhnt.

Test decision: Reject H_0 if for the observed value d of D

(A) $d \geq d_{1-\alpha,n_1,n_2}$

(B) $d^+ \geq d^+_{1-\alpha,n_1,n_2}$

(C) $d^- \geq d^-_{1-\alpha,n_1,n_2}$

The critical values $d_{1-\alpha,n_1,n_2}$, $d^+_{1-\alpha,n_1,n_2}$, $d^-_{1-\alpha,n_1,n_2}$ can be found for instance in Sheskin (2007, table A.23).

p-values:

(A) $p = P(D \geq d)$

(B) $p = P(D^+ \geq d^+)$

(C) $p = P(D^- \geq d^-)$

Annotations:

- The test statistics evaluate the maximum distances between the two empirical distribution functions.
- The Smirnov test can be presented as a rank test as the statistics can be written as supremum of linear rank statistics (Steck 1969).
- The test is known as the Kolmogorov–Smirnov test as well as the Smirnov test for two samples.

Example: To test the hypothesis that the two populations of healthy subjects (status=0) and subjects with hypertension (status=1) do not differ with respect to the distribution of their systolic blood pressure. The dataset contains $n_1 = 25$ observations for status=0 and $n_2 = 30$ observations for status=1 (dataset in Table A.1).

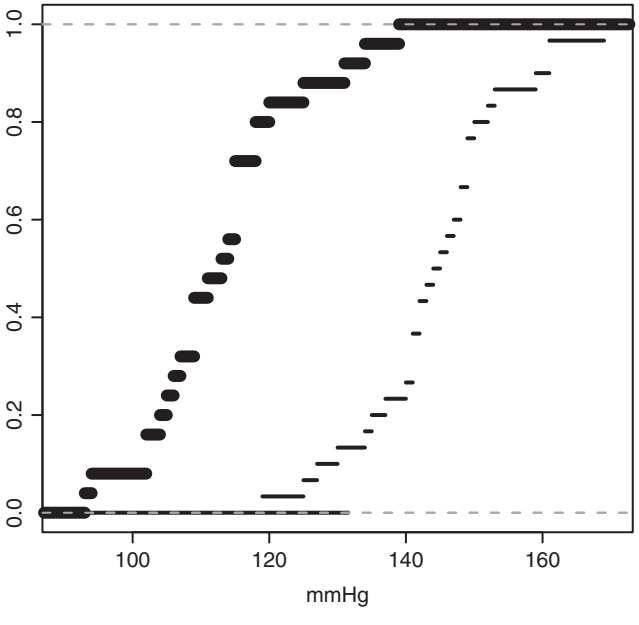

Figure 10.1 Cumulative empirical distribution functions of the blood pressure of healthy subjects (bold lines) and subjects with hypertension (non-bold lines).

SAS code

```
proc npar1way data=blood_pressure D;
 class status;
 var mmhg;
 exact edf;
run;
```

SAS output

```
              The NPAR1WAY Procedure

    Kolmogorov-Smirnov Test for Variable mmhg
          Classified by Variable status

                       EDF at    Deviation from Mean
status       N       Maximum        at Maximum
-----------------------------------------------------
0           25      0.880000          2.218182
1           30      0.066667         -2.024914
Total       55      0.436364

     Maximum Deviation Occurred at Observation 25
          Value of mmhg at Maximum = 125.0

              KS  0.4050    KSa  3.0034

    Kolmogorov-Smirnov Two-Sample Test (Asymptotic)

            D = max |F1 - F2|     0.8133
            Pr > D                <.0001

            D+ = max (F1 - F2)    0.8133
            Pr > D+               <.0001

            D- = max (F2 - F1)    0.0000
            Pr > D-               1.0000
```

Remarks:

- The option `D` enables the one-sided (B) and (C) tests in addition to the two-sided test (A). However, if only the two-sided test is desired, do not use any option or the option `EDF`.
- `exact edf` is optional and applies an additional exact test. Note, the computation of an exact test can be very time consuming. Although this option is given in the listing, the output is generated without this option because it would have taken too much time to calculate the exact p-values even for this tiny dataset.
- D^+ is the test statistic for hypothesis (B) and D^- is the test statistic for hypothesis (C). From Figure 10.1 it can be seen that the cumulative distribution function of the healthy subjects is above the cumulative distribution function of the subjects

with hypertension. Accordingly hypothesis (B) is rejected while hypothesis (C) is not.

R code

```
x<-blood_pressure$mmhg[blood_pressure$status==0]
y<-blood_pressure$mmhg[blood_pressure$status==1]

ks.test(x,y,alternative="two.sided",exact=FALSE)
```

R output

```
Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.8133, p-value = 2.923e-08
alternative hypothesis: two-sided
```

Remarks:

- `alternative=`*"value"* is optional and defines the type of alternative hypothesis: "two.sided"= the cumulative distribution functions of $F_1(x)$ and $F_2(x)$ do not differ (A); "greater"= the cumulative distribution function of $F_1(x)$ lies above $F_2(x)$ (C); "less"=the cumulative distribution function of $F_1(x)$ lies below $F_2(x)$ (B). Default is "two.sided".
- `exact=`*value* is optional. If *value* is not specified or TRUE an exact p-value is computed if the product of the sample sizes is less than 10 000, otherwise only the approximative p-value is computed. In the case of ties or a one-sided alternative no exact test is computed.
- D^+ is the test statistic for hypothesis (B) with option `alternative=`*"greater"* and D^- is the test statistic for hypothesis (C) with option `alternative=`*"less"*. From Figure 10.1 it can be seen that the cumulative distribution function of the healthy subjects is above the cumulative distribution function of the subjects with hypertension. Accordingly hypothesis (B) is rejected while hypothesis (C) is not.

References

Sheskin D. 2007 *Handbook of Parametric and Nonparametric Statistical Procedures*, 4nd edn. Chapman & Hall.

Steck G.P. 1969 The Smirnov two sample tests as rank tests. *The Annals of Mathematical Statistics* **40**, 1449–1466.