Part V

CORRELATION

In this part we deal with tests on coefficients measuring types of association between two variables. The correlation coefficient of two random variables is a measure of the linear relationship between them and takes values between -1 and $+1$. If the correlation coefficient is -1 there is a perfect negative linear relation, if it is $+1$ there is a perfect positive linear relation, both with probability one. The sample correlation coefficient, also called *Pearson's product moment correlation coefficient*, aims at measuring the strength of the linear dependence based on a sample from the two random variables. It can be applied to data coming from a joint continuous bivariate distribution. The *Spearman rank correlation coefficient* more generally measures a monotonic relationship. We also discuss the *partial correlation coefficient* and a test on the difference between two correlation coefficients.

7

Tests on association

We first present one-sample tests for the Pearson product moment correlation coefficient and the Spearman rank correlation coefficient. Next we cover a test on the partial correlation coefficient and a test for two correlation coefficients.

7.1 One-sample tests

7.1.1 Pearson's product moment correlation coefficient

Description: Tests if the Pearson's product moment correlation coefficient ρ differs from a specific value ρ_0.

Assumptions:
- Data are measured on an interval or ratio scale.
- The relationship between X and Y is linear.
- Data pairs (x_i, y_i), $i = 1, \ldots, n$, are randomly sampled from a random vector (X, Y), which follows a bivariate normal distribution.

Hypotheses: (A) $H_0 : \rho = \rho_0$ vs $H_1 : \rho \neq \rho_0$
(B) $H_0 : \rho \leq \rho_0$ vs $H_1 : \rho > \rho_0$
(C) $H_0 : \rho \geq \rho_0$ vs $H_1 : \rho < \rho_0$

Test statistic:

$$(a) \quad \rho_0 = 0 : \quad T = \rho \frac{\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

$$(b) \quad \rho_0 \neq 0 : \quad Z = 0.5 \left[\ln \left(\frac{1+\rho}{1-\rho} \right) - \ln \left(\frac{1+\rho_0}{1-\rho_0} \right) \right] \bigg/ \frac{1}{\sqrt{n-3}}$$

$$\text{with} \quad \rho = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Test decision: (a) Reject H_0 if for the observed value t of T
                    (A) $t < t_{\alpha/2, n-2}$ or $t > t_{1-\alpha/2, n-2}$
            (b) Reject H_0 if for the observed value z of Z
                    (A) $z < z_{\alpha/2}$ or $z > z_{1-\alpha/2}$
                    (B) $z > z_{1-\alpha}$
                    (C) $z < z_{\alpha}$

p-value: (a) (A) $p = 2\,P(T \leq (-|t|))$
            (b) (A) $p = 2\Phi(-|z|)$
                (B) $p = 1 - \Phi(z)$
                (C) $p = \Phi(z)$

Annotations:
- The test statistic T is only used to test if $\rho_0 = 0$ and it follows a t-distribution with $n - 2$ degrees of freedom (Zar 1984, p. 309).
- $t_{\alpha, n-2}$ is the α-quantile of the t-distribution with $n - 2$ degrees of freedom.
- If $\rho_0 \neq 0$ the test statistic Z is used, which is based on the so-called Fisher's variance-stabilizing transformation $\frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right)$. Fisher (1921) has shown that this transformation is approximately a standard normal distribution.
- z_{α} is the α-quantile of the standard normal distribution.
- To ensure a better approximation to the normal distribution of the test statistic Z the term $\rho_0/2(n-1)$ can be subtracted from the numerator (Anderson 2003, p. 134).

Example: Of interest is the correlation between height and weight of a population of students. For the sake of the example the tests for the two cases (a) $\rho_0 = 0$ and (b) $\rho_0 = 0.5$ are to be conducted based on values for 20 students (dataset in Table A.6).

SAS code

```
* a) Test the hypothesis H0: Rho=0;
proc corr data=students pearson;
 var height weight;
run;

* b) Test the hypothesis H0: Rho=0.5;
proc corr data=students fisher(rho0=0.5 biasadj=yes
                                  type=twosided);
 var height weight;
run;
```

SAS output

```
a)
 Pearson Correlation Coefficients, N = 20
     Prob > |r| under H0: Rho=0
```

```
            height          weight
height      1.00000         0.61262
                            0.0041


weight      0.61262         1.00000
            0.0041
```

b)
 Pearson Correlation Statistics (Fisher's z Transformation)

```
            With            -----H0:Rho=Rho0-----
Variable    Variable        Rho0        p Value


height      weight          0.50000     0.5345
```

Remarks:

- To invoke test (a) use the keyword *pearson*.

- The output of test (a) is a matrix of height*weight. In the first row there is the estimated correlation coefficient ($\rho = 0.61262$). The second row contains the p-value of the test (p-value=0.0041).

- The above p-value is for hypothesis (A). The p-value for hypothesis (B) can be easily calculated. First the value t of the test statistic T must be calculated. Because here $\rho = 0.61262$ we get $t = 3.2885$ and $n - 2 = 18$ as degrees of freedom of the corresponding t-distribution. The estimated correlation coefficient is positive and therefore `p=probt(-abs(t),18)=0.0021` is the p-value of hypothesis (B). In the same way the p-value for hypothesis (C) is calculated by `p=1-prob(-abs(t),18)=0.9989`.

- Test (b) uses the Fisher transformation and can be requested by using the keyword *fisher*. Some optional parameters within the brackets are possible.

- `rho0=`*value* specifies the null hypothesis. The default is `rho0=0`.

- `biasadj=`*value* specifies if the bias adjustment is made (`biasadj=yes`) or not (`biasadj=no`). The default is `biasadj=yes`.

- However this option has no influence on the p-value. The p-value of the bias corrected test is always reported. This is strange –although mentioned in the SAS documentation– because using the non bias corrected test will yield an uncorrected confidence interval and a bias corrected p-value. To calculate the bias uncorrected p-values just calculate the Z-value. Here it is $z = 0.6753865$. So, the p-value of hypothesis (B) is calculated as `pB=probnorm(-abs(0.6753865))=0.2497` because the estimated correlation coefficient is above $\rho_0 = 0.5$, the p-value of hypothesis (C) is then `pC=1-probnorm(-abs(0.6753865))=0.7503`, and the p-value of hypothesis (A) is `pA=2*min(pB,pC)=0.4994`.

- `type=`*value*: for hypothesis (A) `type=twosided`; for hypothesis (B) `type=lower`; and for hypothesis (C) `type=upper`. Default is `type=twosided`.

R code

```
# a) Test the hypothesis H0: Rho=0
cor.test(students$height,students$weight,
               alternative="two.sided",method="pearson")

# b) Test the hypothesis H0: Rho=0.5;

# Define rho_0
rho_0=0.5;

# Calculate correlation coefficient
rho<-cor(students$height,students$weight)

# Calculate number of observations
n<-length(students$height)

# Calculate bias factor
b<-rho_0/(2*(n-1))

# Test statistic without bias factor
Z<-0.5*(log((1+rho)/(1-rho))-log((1+rho_0)/(1-rho_0)))
                                          *sqrt(n-3)

# p-values for hypothesis (A), (B), and (C)
pvalue_A=2*min(pnorm(-abs(Z)),1-pnorm(-abs(Z)))

if (rho >= 0){
  pvalue_B=pnorm(-abs(Z))
  pvalue_C=1-pnorm(-abs(Z))
}

if (rho < 0) {
 pvalue_B=1-pnorm(-abs(Z))
 pvalue_C=pnorm(-abs(Z))
}

# Output results
"p-values for tests without bias factor"
pvalue_A
pvalue_B
pvalue_C

# Test statistic with bias factor
Z_b<-(0.5*(log((1+rho)/(1-rho))-log((1+rho_0)/(1-rho_0)))-b)
                                          *sqrt(n-3)

# p-values for hypothesis (A), (B), and (C)
pvalue_A=2*min(pnorm(-abs(Z_b)),1-pnorm(-abs(Z_b)))

if (rho >= 0){
  pvalue_B=pnorm(-abs(Z_b))
  pvalue_C=1-pnorm(-abs(Z_b))
}
```

```
if (rho < 0) {
 pvalue_B=1-pnorm(-abs(Z_b))
 pvalue_C=pnorm(-abs(Z_b))
}

# Output results
"p-values for tests with bias factor"
pvalue_A
pvalue_B
pvalue_C
```

R output

a)

```
Pearson's product-moment correlation

data:  students$height and students$weight
t = 3.2885, df = 18, p-value = 0.004084
alternative hypothesis: true correlation is not equal to 0
sample estimates:
      cor
0.6126242
```

b)

```
[1] "p-Values for tests without bias factor"
> pvalue_A
[1] 0.4994302
> pvalue_B
[1] 0.2497151
> pvalue_C
[1] 0.7502849

[1] "p-Values for tests with bias factor"
> pvalue_A
[1] 0.5345107
> pvalue_B
[1] 0.2672554
> pvalue_C
[1] 0.7327446
```

Remarks:

- The function cor.test() tests only the hypothesis where $\rho_0 = 0$.

- method="pearson" invokes this test. This method is the default of cor.test().

- alternative=*"value"* is optional and indicates the type of alternative hypothesis: "two.sided" (A); "greater" (B); "less" (C). Default is "two.sided".

- For the test with $\rho_0 \neq 0$ no standard R function is available.

7.1.2   Spearman's rank correlation coefficient

Description: Tests if the Spearman rank correlation coefficient ρ_r differs from a specific value ρ_0.

Assumptions:
- Data are measured at least on an ordinal scale.
- The relationship between X and Y is monotonic.
- The random variables X and Y follow continuous distributions.
- The realizations of both random variables are converted into ranks r_i and s_i, $i = 1, \ldots, n$, with corresponding random variables R_i and S_i.

Hypotheses:
(A) $H_0 : \rho_r = \rho_0$ vs $H_1 : \rho_r \neq \rho_0$
(B) $H_0 : \rho_r \leq \rho_0$ vs $H_1 : \rho_r > \rho_0$
(C) $H_0 : \rho_r \geq \rho_0$ vs $H_1 : \rho_r < \rho_0$

Test statistic:

(a) $\rho_0 = 0$: $T = \dfrac{\rho_r \sqrt{n-2}}{\sqrt{1 - \rho_r^2}}$

(b) $\rho_0 \neq 0$: $Z = 0.5 \left[\ln \left(\dfrac{1 + \rho_r}{1 - \rho_r} \right) - \ln \left(\dfrac{1 + \rho_0}{1 - \rho_0} \right) \right] \Big/ \dfrac{1}{\sqrt{n-3}}$

with $\rho_r = \dfrac{\sum\limits_{i=1}^{n} (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum\limits_{i=1}^{n} (R_i - \bar{R})^2 \sum\limits_{i=1}^{n} (S_i - \bar{S})^2}}$,

where $\bar{R} = \dfrac{1}{n} \sum\limits_{i=1}^{n} R_i$ and $\bar{S} = \dfrac{1}{n} \sum\limits_{i=1}^{n} S_i$

Test decision:
(a) Reject H_0 if for the observed value t of T
  (A) $t < t_{\alpha/2, n-2}$ or $t > t_{1-\alpha/2, n-2}$
(b) Reject H_0 if for the observed value z of Z
  (A) $z < z_{\alpha/2}$ or $z > z_{1-\alpha/2}$
  (B) $z > z_{1-\alpha}$
  (C) $z < z_{\alpha}$

p-value:
(a) (A) $p = 2\, P(T \leq (-|t|))$
(b) (A) $p = 2\Phi(-|z|)$
  (B) $p = 1 - \Phi(z)$
  (C) $p = \Phi(z)$

Annotations:
- The test statistic T is only used to test if $\rho_0 = 0$ and it is t-distributed with $n - 2$ degrees of freedom (Zar 1972).
- $t_{\alpha, n-2}$ is the α-quantile of the t-distribution with $n - 2$ degrees of freedom.
- If $\rho_0 \neq 0$ the test statistic Z is used. It is approximately a standard normal distribution (Fieller *et al.* 1957, 1961).

- z_α is the α-quantile of the standard normal distribution.
- The transformation $\frac{1}{2} \ln \left(\frac{1+\rho_r}{1-\rho_r} \right)$ is called a Fisher transformation.
- Instead of using the factor $1/\sqrt{n-3}$, Fieller *et al.* (1957) proposed using the variance factor $\sqrt{1.060/(n-3)}$ to ensure a better approximation to the normal curve.
- A bias adjustment can be conducted by subtracting the term $\rho_0/2(n-1)$ from the numerator of the test statistic Z (Anderson 2003, p. 134).
- The Spearman rank order coefficient can also be written in terms of the rank differences $D_i = R_i - S_i, i = 1, \ldots, n$:

$$\rho_r = 1 - \frac{6D}{n(n^2-1)} \quad \text{with} \quad D = \sum_{i=1}^{n} D_i^2.$$

- In case of ties usually mid ranges are used to calculate the correlation coefficient (Sprent 1993, p. 175).

Example: Of interest is the association between the height and weight in a population of students. For the sake of the example the two hypotheses (a) $\rho_r = 0$ and (b) $\rho_r = 0.5$ are to be tested based on values for 20 students (dataset in Table A.6).

SAS code

```
* a) Test the hypothesis H0: Rho=0;
proc corr data=students spearman;
 var height weight;
run;

* b) Test the hypothesis H0: Rho=0.5;
proc corr data=students spearman fisher(rho0=0.5 biasadj=no
                                        type=twosided);
 var height weight;
run;
```

SAS output

```
a)
 Spearman Correlation Coefficients, N = 20
          Prob > |r| under H0: Rho=0

              height          weight
height       1.00000         0.70686
                             0.0005

weight       0.70686         1.00000
             0.0005
```

```
b)
 Spearman Correlation Statistics (Fisher's z Transformation)


            With          ------H0:Rho=Rho0-----

 Variable    Variable            Rho0      p Value

 height      weight          0.50000      0.1892
```

Remarks:

- To invoke test (a) use the keyword *spearman*.

- The output of test (a) is a matrix of height*weight. In the first row there is the estimated correlation coefficient ($\rho = 0.70686$). The second row contains the p-value of the test (p-value=0.0005).

- The above p-value is for hypothesis (A). The p-value for hypothesis (B) can be easily calculated. First the value t of the test statistic T must be calculated. Because here $\rho = 0.70686$ we get $t = 4.23968$ and $n - 2 = 18$ as degrees of freedom of the corresponding t-distribution. Now p=probt(-abs(t),18)=0.00025 is the p-value of hypothesis (B), because the estimated correlation coefficient is positive. Therefore the p-value for hypothesis (C) is calculated by p=1-prob(-abs(t),18)=0.9998.

- Test (b) uses the Fisher transformation and can be requested by using the keyword *fisher*. Some optional parameters within the brackets are possible. SAS does not use the Fieller *et al.* (1957) and Fieller *et al.* (1961) recommendation for the variance factor.

- rho0=*value* specifies the null hypothesis. The default is rho0=0.

- biasadj=*value* specifies if the bias adjustment is made (biasadj=yes) or not (biasadj=no). The default is biasadj=yes. For the hypothesis $H_0 : \rho_r = \rho_0$ SAS always uses the bias adjustment.

- The option biasadj=*value* has no influence on the p-value. The p-value of the bias corrected test is always reported. This is strange–although mentioned in the SAS documentation–because using the non bias corrected test will yield an uncorrected confidence interval and a bias corrected p-value. To calculate the bias uncorrected p-values just calculate the Z-value. Here it is $z = 1.367115$. So, the p-value of hypothesis (B) is calculated as pB=probnorm(-abs(1.367115))=0.0858, because the estimated correlation coefficient is above $\rho_0 = 0.5$. The p-value of hypothesis (C) is pC=1-probnorm(-abs(1.367115))=0.9142, and the p-value of hypothesis (A) is pA=2*min(pB,pC)=0.1716.

- type=*value*: for hypothesis (A) type=twosided; for hypothesis (B) type=lower and for hypothesis (C) type=upper. Default is type=twosided.

R code

```
# a) Test the hypothesis H0: Rho=0
cor.test(students$height,students$weight,
         alternative="two.sided",method="spearman",
         exact=NULL,continuity=FALSE)

# b) Test the hypothesis H0: Rho=0.5;

# Define rho_0
rho_0=0.5

# Convert data into ranks
x<-rank(students$height)
y<-rank(students$weight)

# Calculate correlation coefficient
rho<-cor(x,y)

# Calculate number of observations
n<-length(students$height)

# Calculate bias factor
b<-rho_0/(2*(n-1))

# Test statistic without bias factor
Z<-0.5*(log((1+rho)/(1-rho))-log((1+rho_0)/(1-rho_0)))
                                            *sqrt(n-3)

# p-values for hypothesis (A), (B), and (C)
pvalue_A=2*min(pnorm(-abs(Z)),1-pnorm(-abs(Z)))

if (rho >= 0){
  pvalue_B=pnorm(-abs(Z))
  pvalue_C=1-pnorm(-abs(Z))
}

if (rho < 0) {
 pvalue_B=1-pnorm(-abs(Z))
 pvalue_C=pnorm(-abs(Z))
}

# Output results
"p-Values for tests without bias factor"
pvalue_A
pvalue_B
pvalue_C

# Test statistic with bias factor
Z_b<-(0.5*(log((1+rho)/(1-rho))-log((1+rho_0)/(1-rho_0)))-b)
                                            *sqrt(n-3)

# p-values for hypothesis A), B), and C)
pvalue_A=2*min(pnorm(-abs(Z_b)),1-pnorm(-abs(Z_b)))
```

```
if (rho >= 0){
  pvalue_B=pnorm(-abs(Z_b))
  pvalue_C=1-pnorm(-abs(Z_b))
}

if (rho < 0) {
 pvalue_B=1-pnorm(-abs(Z_b))
 pvalue_C=pnorm(-abs(Z_b))
}

# Output results
"p-values for tests with bias factor"
pvalue_A
pvalue_B
pvalue_C
```

R output

```
a)
        Spearman's rank correlation rho

data:  students$height and students$weight
S = 389.8792, p-value = 0.0004929
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7068578

b)
[1] "p-values for tests without bias factor"
> pvalue_A
[1] 0.1715951
> pvalue_B
[1] 0.08579753
> pvalue_C
[1] 0.9142025

[1] "p-values for tests with bias factor"
> pvalue_A
[1] 0.1892351
> pvalue_B
[1] 0.09461757
> pvalue_C
[1] 0.9053824
```

Remarks:

- The function `cor.test()` tests only the hypothesis where $\rho_0 = 0$.

- `method="spearman"` invokes this test.

- With optional parameter `exact=`*value* an exact test `exact=TRUE` can be performed or not `exact=NULL` [see Best and Roberts (1975) for details]. If ties are present no exact test can be performed. Default is no exact test.

- `continuity=`*value* is optional. If `continuity=TRUE` a continuity correction is applied (for the not exact test). Default is `continuity=FALSE`.

- `alternative=`*"value"* is optional and indicates the type of alternative hypothesis: "two.sided" (A); "greater" (B); "less" (C). Default is "two.sided".

- For the test with $\rho_0 \neq 0$ no standard R function is available.

7.1.3 Partial correlation

Description: Tests if the correlation coefficient $\rho_{XY.Z}$ of two random variables X and Y given a third random variable Z differs from zero.

Assumptions:
- (I) Data are measured at least on an ordinal scale.
- (II) Data are measured on an interval or ratio scale.
- For (II) the three random variables X, Y, and Z are assumed to follow a joint Gaussian distribution.
- A sample $((X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n))$ of size n is taken.

Hypotheses:
(A) $H_0 : \rho_{XY.Z} = 0$ vs $H_1 : \rho_{XY.Z} \neq 0$
(B) $H_0 : \rho_{XY.Z} \leq 0$ vs $H_1 : \rho_{XY.Z} > 0$
(C) $H_0 : \rho_{XY.Z} \geq 0$ vs $H_1 : \rho_{XY.Z} < 0$

Test statistic:

$$T = \frac{\rho_{XY.Z}\sqrt{n-3}}{\sqrt{(1-\rho_{XY.Z}^2)}}$$

with $\rho_{XY.Z} = \dfrac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{(1-\rho_{XZ}^2)(1-\rho_{YZ}^2)}}$

and $\rho_{XY}, \rho_{XZ}, \rho_{YZ}$ are the correlation coefficients

between these random variables, that is, (I) Spearman's correlation

coefficient (see Test 7.1.2) and (II) Pearson's correlation coefficient (see Test 7.1.1).

Test decision: Reject H_0 if for the observed value t of T
(A) $t < t_{\alpha/2, n-3}$ or $t > t_{1-\alpha/2, n-3}$
(B) $t > t_{1-\alpha, n-3}$
(C) $t < t_{\alpha, n-3}$

p-value:
(A) $p = 2\, P(T \leq (-|t|))$
(B) $p = 1 - P(T \leq t)$
(C) $p = P(T \leq t)$

Annotations:
- The test statistic T for the partial correlation coefficient, regardless of whether it is calculated with Pearson's correlation coefficient or Spearman's correlation coefficient, is t-distributed with $n - 3$ degrees of freedom (Sheskin 2007, p. 1459).

- $t_{\alpha,n-3}$ is the α-quantile of the t-distribution with $n-3$ degrees of freedom.
- The partial correlation can also be calculated as the correlation between the residuals of the linear regressions of X on a set of k variables and Y on a set of the same k variables. The degrees of freedom of the corresponding t-distribution of the test statistic is then $n-k-2$ (Kleinbaum *et al.* 1998, pp. 165-171).

Example: Of interest is the partial association between height and weight in a population of students given their sex. For the sake of the example all three hypotheses are tested based on values for 20 students (dataset in Table A.6).

SAS code

```
proc corr data=students pearson;
 var height weight;
 partial sex;
run;
```

SAS output

```
  Pearson Partial Correlation Coefficients, N = 20
       Prob > |r| under H0: Partial Rho=0

                          height          weight

height        1.00000         0.56914
                              0.0110

weight        0.56914         1.00000
                              0.0110
```

Remarks:

- With the keyword `pearson` the partial correlation based on Pearson's product moment correlation coefficient is performed. This is the default. Use the keyword *spearman* to calculate the rank based partial correlation coefficient.

- The keyword `partial` *variable* invokes the calculation of a partial correlation coefficient. The value *variable* stands for one or more variables on which the correlation is partialled. Note: In the case of more than one variable the degrees of freedom of the test statistic are changing.

- The output is a matrix of height*weight. In the first row there is the estimated correlation coefficient ($\rho = 0.56194$). The second row contains the p-value of the test (p-value=0.0110).

- The above p-value is for hypothesis (A). The p-value for hypotheses (B) and (C) can either be calculated via the Fisher transformation (see Test 7.1.1) or directly

by using the value of the test statistic T and comparing it to the corresponding t-distribution. Here $\rho = 0.56194$ and we get $t = 2.853939$. Furthermore $n - 3 = 17$ are the degrees of freedom of the corresponding t-distribution. Now p=probt(-abs(t),17)=0.0055 is the p-value of hypothesis (B) because the estimated correlation coefficient is greater than zero. Therefore the p-value for hypothesis (C) is calculated by p=1-prob(-abs(t),17)=0.9945.

R code

```
# Calculate correlation between variables
rho_wh<-cor(students$weight,students$height,
                                method="pearson")
rho_ws<-cor(students$weight,students$sex,method="pearson")
rho_hs<-cor(students$height,students$sex,method="pearson")

# Calculate number of observations
n<-length(students$height)

# Calculate partial correlation
rho_wh.s=(rho_wh-rho_ws*rho_hs)
                /sqrt((1-rho_ws^2)*(1-rho_hs^2))

# Calculate test statistic
t=(rho_wh.s*sqrt(n-3))/sqrt((1-rho_wh.s^2))

# Calculate p-values
pvalue_A=2*min(pt(-abs(t),n-3),1-pt(-abs(t),n-3))

if (rho_wh.s >= 0){
  pvalue_B=pt(-abs(t),n-3)
  pvalue_C=1-pt(-abs(t),n-3)
}

if (rho_wh.s < 0) {
 pvalue_B=1-pt(-abs(t),n-3)
 pvalue_C=pt(-abs(t),n-3)
}

# Output results
rho_wh.s
pvalue_A
pvalue_B
pvalue_C
```

R output

```
> rho_wh.s
[1] 0.5691401
> pvalue_A
[1] 0.01098247
```

```
> pvalue_B
[1] 0.005491237
> pvalue_C
[1] 0.9945088
```

Remarks:

- There is no core R function to calculate the test directly.

- To use Pearson's product moment correlation coefficient use `method="pearson"` in the calculation of the pairwise correlation coefficients. To calculate the rank based partial correlation coefficient use `method="spearman"`.

- To use more than one partialled variable some different coding is necessary. Assume you want to calculate the partial correlation between X and Y with W and Z partialled out. Use the code:

```
x<-residuals(lm(x ~ w z))
y<-residuals(lm(y ~ w z))
rho_xy.wz<-cor(x,y)
t=(rho_xy.wz*sqrt(n-4))/sqrt((1-rho_xy.wz^2))
```

to calculate the partial correlation coefficient `rho_xy.wz` and the value of the test statistic T. The test statistic is t-distributed with $n - 4$ degrees of freedom.

7.2   Two-sample tests

7.2.1   z-test for two correlation coefficients (independent populations)

Description: Tests if two correlation coefficients ρ_1 and ρ_2 from independent populations differ from each other.

Assumptions:
- Data are measured on an interval or ratio scale.
- Data are randomly sampled from two independent bivariate Gaussian distributions with sample sizes n_1 and n_2.
- The parameters ρ_1 and ρ_2 are the correlation coefficients in the two populations.

Hypotheses: (A) $H_0 : \rho_1 = \rho_2$ vs $H_1 : \rho_1 \neq \rho_2$
(B) $H_0 : \rho_1 \leq \rho_2$ vs $H_1 : \rho_1 > \rho_2$
(C) $H_0 : \rho_1 \geq \rho_2$ vs $H_1 : \rho_1 < \rho_2$

Test statistic:

$$Z = 0.5 \left[\ln \left(\frac{1 + \rho_1}{1 - \rho_1} \right) - \ln \left(\frac{1 + \rho_2}{1 - \rho_2} \right) \right] \Bigg/ \sqrt{\frac{1}{(n_1 - 3)} + \frac{1}{(n_2 - 3)}}$$

$$\text{with } \rho_j = \frac{\sum\limits_{i=1}^{n}(X_{ij} - \overline{X_j})(Y_{ij} - \overline{Y_j})}{\sqrt{\sum\limits_{i=1}^{n}(X_{ij} - \overline{X_j})^2 \sum\limits_{i=1}^{n}(Y_{ij} - \overline{Y_j})^2}} \quad j = 1, 2.$$

Test decision: Reject H_0 if for the observed value z of Z
(A) $z < z_{\alpha/2}$ or $z > z_{1-\alpha/2}$
(B) $z > z_{1-\alpha}$
(C) $z < z_{\alpha}$

p-value: (A) $p = 2\Phi(-|z|)$
(B) $p = 1 - \Phi(z)$
(C) $p = \Phi(z)$

Annotations: • The test statistic Z is approximately a standard normal distribution (Sheskin 2007, pp. 1247–1248).
• The test statistic Z can be easily expanded to the case of k independent bivariate Gaussian distributions (Sheskin 2007, p. 1249).

Example: To test, if the correlation coefficients between height and weight in two populations of male and female students differ from each other. Observations from 10 male (sex=1) and 10 female (sex=2) students are given (dataset in Table A.6).

SAS code

```
* Sort data by sex;
proc sort data=students;
 by sex;
run;

* Calculate correlation coefficients of males and females;
proc corr data=students outp=corr_data;
 by sex;
 var height weight;
run;

* Make four datasets from the output: number of observations
* and correlation coefficients for male and female;
data n_male corr_male n_female corr_female;
 set corr_data;
 if _type_="N" and sex=1 then output n_male;
 if _type_="N" and sex=2 then output n_female;
 if _type_="CORR" and _name_="height" and sex=1
                                   then output corr_male;
 if _type_="CORR" and _name_="height" and sex=2
                                   then output corr_female;
run;

* Rename number of observations of males as n1;
data n_male;
 set n_male;
 rename height=n1;
 keep height;
run;
```

```
* Rename number of observations of females as n2;
data n_female;
 set n_female;
 rename height=n2;
 keep height;
run;

* Rename correlation coefficients of males as rho1;
data corr_male;
 set corr_male;
 rename weight=rho1;
 keep weight;
run;

* Rename correlation coefficients of females as rho2;
data corr_female;
 set corr_female;
 rename weight=rho2;
 keep weight;
run;

* Merge all data into a dataset with a single observation;
data corr;
 merge corr_male corr_female n_male n_female;
run;

* Calculate test statistic and p-values;
data corr_test;
 set corr;

 Z=0.5*(log((1+rho1)/(1-rho1))-log((1+rho2)/(1-rho2)))
                              /(sqrt(1/(n1-3)+1/(n2-3)));
 diff=rho1-rho2;

* p-values for hypothesis (A), (B), and (C);
 pvalue_A=2*min(probnorm(-abs(Z)),1-probnorm(-abs(Z)));

 if diff>=0 then
  do;
   pvalue_B=probnorm(-abs(Z));
   pvalue_C=1-probnorm(-abs(Z));
  end;

 if diff<0 then
  do;
   pvalue_B=1-probnorm(-abs(Z));
   pvalue_C=probnorm(-abs(Z));
  end;

run;

* Output results;
proc print;
 var rho1 rho2 pvalue_A  pvalue_B pvalue_C;
run;
```

SAS output

```
 rho1        rho2        pvalue_A     pvalue_B     pvalue_C
0.49002     0.85390     0.16952      0.91524      0.084761
```

Remarks:

- There is no SAS procedure to calculate this test directly.

R code

```
# Calculate correlation coefficient for males
male.height<-students$height[students$sex==1]
male.weight<-students$weight[students$sex==1]
rho1<-cor(male.height,male.weight)

# Calculate number of observations for males
n1<-length(students$height[students$sex==1])

# Calculate correlation coefficient for females
female.height<-students$height[students$sex==2]
female.weight<-students$weight[students$sex==2]
rho2<-cor(female.height,female.weight)

# Calculate number of observations for females
n2<-length(students$height[students$sex==2])

# Test statistic
Z<-0.5*(log((1+rho1)/(1-rho1))-log((1+rho2)/(1-rho2)))/
                                (sqrt(1/(n1-3)+1/(n2-3)))
diff=rho1-rho2

# p-values for hypothesis A), B), and C)
pvalue_A=2*min(pnorm(-abs(Z)),1-pnorm(-abs(Z)))

if (diff >=0){
  pvalue_B=pnorm(-abs(Z))
  pvalue_C=1-pnorm(-abs(Z))
}

if (diff < 0) {
 pvalue_B=1-pnorm(-abs(Z))
 pvalue_C=pnorm(-abs(Z))
}

# Output results
"Correlation coefficient for males:"
rho1
"Correlation coefficient for females:"
rho2
"p-Values"
```

```
pvalue_A
pvalue_B
pvalue_C
```

R output

```
[1] "Correlation coefficient for males:"
> rho1
[1] 0.4900237

[1] "Correlation coefficient for females:"
> rho2
[1] 0.8539027

[1] "p-Values"
> pvalue_A
[1] 0.1695216
> pvalue_B
[1] 0.9152392
> pvalue_C
[1] 0.0847608
```

Remarks:

- There is no R function to calculate this test directly.

References

Anderson T.W. 2003 *An Introduction to Multivariate Analysis*, 3rd edn. John Wiley & Sons, Ltd.

Best D.J. and Roberts D.E. 1975 Algorithm AS 89: the upper tail probabilities of Spearman's rho. *Applied Statistics* **24**, 377–379.

Fieller E.C., Hartley H.O. and Pearson E.S. 1957 Tests for rank correlation coefficients I. *Biometrika* **44**, 470–481.

Fieller E.C., Hartley H.O. and Pearson E.S. 1961 Tests for rank correlation coefficients II. *Biometrika* **48**, 29–40.

Fisher R.A. 1921 On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron* **1**, 3–32.

Kleinbaum D.G., Kupper L.L., Muller K.E. and Nizam A. 1998 *Applied Regression Analysis and Other Multivariable Methods*. Duxbury Press.

Sprent P. 1993 *Applied Nonparametric Statistical Methods*, 2nd edn. Chapman & Hall.

Sheskin D.J. 2007 *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall.

Zar J.H. 1972 Significance testing of the Spearman rank correlation coefficient. *Journal of the Acoustical Society of America* **67**, 578–580.

Zar J.H. 1984 *Biostatistical Analysis*, 2nd edn. Prentice-Hall.