

# 2

## Comparison of Two Samples

### 2.1 Introduction

Problems of comparing two samples arise frequently in medicine, sociology, agriculture, engineering, and marketing. The data may have been generated by observation or may be the outcome of a controlled experiment. In the latter case, randomization plays a crucial role in gaining information about possible differences in the samples which may be due to a specific factor. Full nonrestricted randomization means, for example, that in a controlled clinical trial there is a constant chance of every patient getting a specific treatment. The idea of a blind, double blind, or even triple blind set-up of the experiment is that neither patient, nor clinician, nor statistician, know what treatment has been given. This should exclude possible biases in the response variable, which would be induced by such knowledge. It becomes clear that careful planning is indispensable to achieve valid results.

Another problem in the framework of a clinical trial may consist of the fact of a systematic effect on a subgroup of patients, e.g., males and females. If such a situation is to be expected, one should stratify the sample into homogeneous subgroups. Such a strategy proves to be useful in planned experiments as well as in observational studies.

Another experimental set-up is given by a *matched-pair design*. Subgroups then contain only one individual and pairs of subgroups are compared with respect to different treatments. This procedure requires pairs to be homogeneous with respect to all the possible factors that may

exhibit an influence on the response variable and is thus limited to very special situations.

## 2.2 Paired $t$ -Test and Matched-Pair Design

In order to illustrate the basic reasoning of a matched-pair design, consider an experiment, the structure of which is given in Table 2.1.

Pair	Treatment		Difference
	1	2	
1	$y_{11}$	$y_{21}$	$y_{11} - y_{21} = d_1$
2	$y_{12}$	$y_{22}$	$y_{12} - y_{22} = d_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n$	$y_{1n}$	$y_{2n}$	$y_{1n} - y_{2n} = d_n$
			$\bar{d} = \sum d_i / n$

TABLE 2.1. Response in a matched-pair design.

We consider the linear model already given in (1.8). Assuming that

$$d_i \stackrel{i.i.d.}{\sim} N(\mu_d, \sigma_d^2), \quad (2.1)$$

the best linear unbiased estimator of  $\mu_d$ ,  $\bar{d}$ , is distributed as

$$\bar{d} \sim N\left(\mu_d, \frac{\sigma_d^2}{n}\right). \quad (2.2)$$

An unbiased estimator of  $\sigma_d^2$  is given by

$$s_d^2 = \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1} \sim \frac{\sigma_d^2}{n-1} \chi_{n-1}^2 \quad (2.3)$$

such that under  $H_0 : \mu_d = 0$  the ratio

$$t = \frac{\bar{d}}{s_d} \sqrt{n} \quad (2.4)$$

is distributed according to a (central)  $t$ -distribution.

A two-sided test for  $H_0 : \mu_d = 0$  versus  $H_1 : \mu_d \neq 0$  rejects  $H_0$ , if

$$|t| > t_{n-1; 1-\alpha(\text{two-sided})} = t_{n-1; 1-\alpha/2}. \quad (2.5)$$

A one-sided test  $H_0 : \mu_d = 0$  versus  $H_1 : \mu_d > 0$  ( $\mu_d < 0$ ) rejects  $H_0$  in favor of  $H_1 : \mu_d > 0$ , if

$$t > t_{n-1; 1-\alpha}. \quad (2.6)$$

$H_0$  is rejected in favor of  $H_1 : \mu_d < 0$ , if

$$t < -t_{n-1; 1-\alpha}. \quad (2.7)$$

### Necessary Sample Size and Power of the Test

We consider a test of  $H_0$  versus  $H_1$  for a distribution with an unknown parameter  $\theta$ . Obviously, there are four possible situations, two of which

Decision	Real situation	
	$H_0$ true	$H_0$ false
$H_0$ accepted	Correct decision	False decision
$H_0$ rejected	False decision	Correct decision

TABLE 2.2. Test decisions.

lead to a correct decision. The probability

$$P_{\theta}(\text{reject } H_0 \mid H_0 \text{ true}) = P_{\theta}(H_1 \mid H_0) \leq \alpha \quad \text{for all } \theta \in H_0 \quad (2.8)$$

is called the probability of a *type I error*.  $\alpha$  is to be fixed before the experiment. Usually,  $\alpha = 0.05$  is a reasonable choice. The probability

$$P_{\theta}(\text{accept } H_0 \mid H_0 \text{ false}) = P_{\theta}(H_0 \mid H_1) \geq \beta \quad \text{for all } \theta \in H_1 \quad (2.9)$$

is called the probability of a *type II error*. Obviously, this probability depends on the true value of  $\theta$  such that the function

$$G(\theta) = P_{\theta}(\text{reject } H_0) \quad (2.10)$$

is called the *power* of the test. Generally, a test on a given  $\alpha$  aims to fix the type II error at a defined level or beyond. Equivalently, we could say that the power should reach, or even exceed, a given value. Moreover, the following rules apply:

- (i) the power rises as the sample size  $n$  increases, keeping  $\alpha$  and the parameters under  $H_1$  fixed;
- (ii) the power rises and therefore  $\beta$  decreases as  $\alpha$  increases, keeping  $n$  and the parameters under  $H_1$  fixed; and
- (iii) the power rises as the difference  $\delta$  between the parameters under  $H_0$  and under  $H_1$  increases.

We bear in mind that the power of a test depends on the difference  $\delta$ , on the type I error, on the sample size  $n$ , and on the hypothesis being one-sided or two-sided. Changing from a one-sided to a two-sided problem reduces the power.

The comparison of means in a matched-pair design yields the following relationship. Consider a one-sided test ( $H_0 : \mu_d = \mu_0$  versus  $H_1 : \mu_d = \mu_0 + \delta$ ,  $\delta > 0$ ) and a given  $\alpha$ . To start with, we assume  $\sigma_d^2$  to be known. We now try to derive the sample size  $n$  that is required to achieve a fixed power of  $1 - \beta$  for a given  $\alpha$  and known  $\sigma_d^2$ . This means that we have to settle  $n$

in a way that  $H_0 : \mu_d = \mu_0$ , with fixed  $\alpha$ , is accepted with probability  $\beta$ , although the true parameter is  $\mu_d = \mu_0 + \delta$ . We define

$$u := \frac{\bar{d} - \mu_0}{\sigma_d / \sqrt{n}}.$$

Then, under  $H_1 : \mu_d = \mu_0 + \delta$ , we have

$$\tilde{u} = \frac{\bar{d} - (\mu_0 + \delta)}{\sigma_d / \sqrt{n}} \sim N(0, 1). \quad (2.11)$$

$\tilde{u}$  and  $u$  are related as follows:

$$u = \tilde{u} + \frac{\delta}{\sigma_d} \sqrt{n} \sim N\left(\frac{\delta}{\sigma_d} \sqrt{n}, 1\right). \quad (2.12)$$

The null hypothesis  $H_0 : \mu_d = \mu_0$  is accepted erroneously if the test statistic  $u$  has a value of  $u \leq u_{1-\alpha}$ . The probability for this case should be  $\beta = P(H_0 \mid H_1)$ . So we get

$$\begin{aligned} \beta &= P(u \leq u_{1-\alpha}) \\ &= P\left(\tilde{u} \leq u_{1-\alpha} - \frac{\delta}{\sigma_d} \sqrt{n}\right) \end{aligned}$$

and, therefore,

$$u_\beta = u_{1-\alpha} - \frac{\delta}{\sigma_d} \sqrt{n},$$

which yields

$$n \geq \frac{(u_{1-\alpha} - u_\beta)^2 \sigma_d^2}{\delta^2} \quad (2.13)$$

$$= \frac{(u_{1-\alpha} + u_{1-\beta})^2 \sigma_d^2}{\delta^2}. \quad (2.14)$$

For application in practice, we have to estimate  $\sigma_d^2$  in (2.13). If we estimate  $\sigma_d^2$  using the sample variance, we also have to replace  $u_{1-\alpha}$  and  $u_{1-\beta}$  by  $t_{n-1;1-\alpha}$  and  $t_{n-1;1-\beta}$ , respectively. The value of  $\delta$  is the difference of expectations of the two parameter ranges, which is either known or estimated using the sample.

## 2.3 Comparison of Means in Independent Groups

### 2.3.1 Two-Sample $t$ -Test

We have already discussed the two-sample problem in Section 1.8. Now we consider the two independent samples

$$\begin{aligned} A &: x_1, \dots, x_{n_1}, \quad x_i \sim N(\mu_A, \sigma_A^2), \\ B &: y_1, \dots, y_{n_2}, \quad y_i \sim N(\mu_B, \sigma_B^2). \end{aligned}$$

Assuming  $\sigma_A^2 = \sigma_B^2 = \sigma^2$ , we may apply the linear model. To compare the two groups  $A$  and  $B$  we test the hypothesis  $H_0 : \mu_A = \mu_B$  using the statistic, *i.e.*,

$$t_{n_1+n_2-2} = (\bar{x} - \bar{y}) / s \sqrt{(n_1 n_2) / (n_1 + n_2)}.$$

In practical applications, we have to check the assumption that  $\sigma_A^2 = \sigma_B^2$ .

### 2.3.2 Testing $H_0 : \sigma_A^2 = \sigma_B^2 = \sigma^2$

Under  $H_0$ , the two independent sample variances

$$s_x^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2$$

and

$$s_y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y})^2$$

follow a  $\chi^2$ -distribution with  $n_1 - 1$  and  $n_2 - 1$  degrees of freedom, respectively, and their ratio follows an  $F$ -distribution

$$F = \frac{s_x^2}{s_y^2} \sim F_{n_1-1, n_2-1}. \quad (2.15)$$

Decision

Two-sided:

$$H_0 : \sigma_A^2 = \sigma_B^2 \text{ versus } H_1 : \sigma_A^2 \neq \sigma_B^2.$$

$H_0$  is rejected if

$$F > F_{n_1-1, n_2-1; 1-\alpha/2}$$

or

$$F < F_{n_1-1, n_2-1; \alpha/2} \quad (2.16)$$

with

$$F_{n_1-1, n_2-1; \alpha/2} = \frac{1}{F_{n_1-1, n_2-1; 1-\alpha/2}}. \quad (2.17)$$

One-sided:

$$H_0 : \sigma_A^2 = \sigma_B^2 \text{ versus } H_1 : \sigma_A^2 > \sigma_B^2. \quad (2.18)$$

If

$$F > F_{n_1-1, n_2-1; 1-\alpha}, \quad (2.19)$$

then  $H_0$  is rejected.

*Example 2.1.* Using the data set of Table 1.8, we want to test  $H_0 : \sigma_A^2 = \sigma_B^2$ . In Table 1.8 we find the values  $n_1 = n_2 = 10$ ,  $s_A^2 = \frac{26}{9}$ , and  $s_B^2 = \frac{18}{9}$ . This yields

$$F = \frac{26}{18} = 1.44 < 3.18 = F_{9,9;0.95}$$

so that we cannot reject the null hypothesis  $H_0 : \sigma_A^2 = \sigma_B^2$  versus  $H_1 : \sigma_A^2 > \sigma_B^2$  according to (2.19). Therefore, our analysis in Section 1.8 was correct.

### 2.3.3 Comparison of Means in the Case of Unequal Variances

If  $H_0 : \sigma_A^2 = \sigma_B^2$  is not valid, we are up against the so-called Behrens Fisher problem, which has no exact solution. For practical use, the following correction of the test statistic according to *Welch* gives sufficiently good results

$$t = \frac{|\bar{x} - \bar{y}|}{\sqrt{(s_x^2/n_1) + (s_y^2/n_2)}} \sim t_v \quad (2.20)$$

with degrees of freedom approximated by

$$v = \frac{(s_x^2/n_1 + s_y^2/n_2)^2}{(s_x^2/n_1)^2/(n_1 + 1) + (s_y^2/n_2)^2/(n_2 + 1)} - 2 \quad (2.21)$$

( $v$  is rounded). We have  $\min(n_1 - 1, n_2 - 1) < v < n_1 + n_2 - 2$ .

*Example 2.2.* In material testing, two normal variables,  $A$  and  $B$ , were examined. The sample parameters are summarized as follows:

$$\begin{array}{llll} \bar{x} = & 27.99, & s_x^2 = & 5.98^2, & n_1 = & 9 \\ \bar{y} = & 1.92, & s_y^2 = & 1.07^2, & n_2 = & 10 \end{array}.$$

The sample variances are not equal

$$F = \frac{5.98^2}{1.07^2} = 31.23 > 3.23 = F_{8,9;0.95}.$$

Therefore, we have to use *Welch's test* to compare the means

$$t_v = \frac{|27.99 - 1.92|}{\sqrt{5.98^2/9 + 1.07^2/10}} = 12.91$$

with  $v \approx 9$  degrees of freedom. The critical value of  $t_{9;0.975} = 2.26$  is exceeded and we reject  $H_0 : \mu_A = \mu_B$ .

### 2.3.4 Transformations of Data to Assure Homogeneity of Variances

We know from experience that the two-sample  $t$ -test is more sensitive to discrepancies in the homogeneity of variances than to deviations from the assumption of normal distribution. The two-sample  $t$ -test usually reaches the level of significance if the assumption of normal distributions is not fully justified, but sample sizes are large enough ( $n_1, n_2 > 20$ ) and the homogeneity of variances is valid. This result is based on the central limit theorem. Analogously, deviations from variance homogeneity can have severe effects on the level of significance.

The following transformations may be used to avoid the inhomogeneity of variances:

- logarithmic transformation  $\ln(x_i)$ ,  $\ln(y_i)$ ; and
- logarithmic transformation  $\ln(x_i + 1)$ ,  $\ln(y_i + 1)$ , especially if  $x_i$  and  $y_i$  have zero values or if  $0 \leq x_i, y_i \leq 10$  (Woolson, 1987, p. 171).

### 2.3.5 Necessary Sample Size and Power of the Test

The necessary sample size, to achieve the desired power of the two-sample  $t$ -test, is derived as in the paired  $t$ -test problem. Let  $\delta = \mu_A - \mu_B > 0$  be the one-sided alternative to be tested against  $H_0 : \mu_A = \mu_B$  with  $\sigma_A^2 = \sigma_B^2 = \sigma^2$ . Then, with  $n_2 = a \cdot n_1$  (if  $a = 1$ , then  $n_1 = n_2$ ), the minimum sample size to preserve a power of  $1 - \beta$  (cf. (2.14)) is given by

$$n_1 = \sigma^2(1 + 1/a)(u_{1-\alpha} + u_{1-\beta})^2/\delta^2 \quad (2.22)$$

and

$$n_2 = a \cdot n_1 \quad \text{with } n_1 \text{ from (2.22).}$$

### 2.3.6 Comparison of Means without Prior Testing

$H_0 : \sigma_A^2 = \sigma_B^2$ ; Cochran-Cox Test for Independent Groups

There are several alternative methods to be used instead of the two-sample  $t$ -test in the case of unequal variances. The test of Cochran and Cox (1957) uses a statistic which approximately follows a  $t$ -distribution. The Cochran-Cox test is conservative compared to the usually used  $t$ -test. Substantially, this fact is due to the special number of degrees of freedom that have to be used. The degrees of freedom of this test are a weighted average of  $n_1 - 1$

and  $n_2 - 1$ . In the balanced case ( $n_1 = n_2 = n$ ) the Cochran–Cox test has  $n - 1$  degrees of freedom compared to  $2(n - 1)$  degrees of freedom used in the two-sample  $t$ -test. The test statistic

$$t_{c-c} = \frac{\bar{x} - \bar{y}}{s_{(\bar{x}-\bar{y})}} \quad (2.23)$$

with

$$s_{(\bar{x}-\bar{y})}^2 = \frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}$$

has critical values at:

$$\text{two-sided:} \quad (2.24)$$

$$t_{c-c(1-\alpha/2)} = \frac{s_x^2/n_1 \ t_{n_1-1;1-\alpha/2} + s_y^2/n_2 \ t_{n_2-1;1-\alpha/2}}{s_{(\bar{x}-\bar{y})}^2}, \quad (2.25)$$

$$\text{one-sided:} \quad (2.26)$$

$$t_{c-c(1-\alpha)} = \frac{s_x^2/n_1 \ t_{n_1-1;1-\alpha} + s_y^2/n_2 \ t_{n_2-1;1-\alpha}}{s_{(\bar{x}-\bar{y})}^2}. \quad (2.27)$$

The null hypothesis is rejected if  $|t_{c-c}| > t_{c-c}(1 - \alpha/2)$  (two-sided) (resp.,  $t_{c-c} > t_{c-c}(1 - \alpha)$  (one-sided,  $H_1: \mu_A > \mu_B$ )).

*Example 2.3.* (Example 2.2 continued).

We test  $H_0: \mu_A = \mu_B$  using the two-sided Cochran–Cox test. With

$$\begin{aligned} s_{(\bar{x}-\bar{y})}^2 &= \frac{5.98^2}{9} + \frac{1.07^2}{10} \\ &= 3.97 + 0.11 = 4.08 = 2.02^2 \end{aligned}$$

and

$$\begin{aligned} t_{c-c(1-\alpha/2)} &= \frac{3.97 \cdot 2.31 + 0.11 \cdot 2.26}{4.08} \\ &= 1.86, \end{aligned}$$

we get  $t_{c-c} = |27.99 - 1.92|/2.02 = 12.91 > 2.31$ , so that  $H_0$  has to be rejected.

## 2.4 Wilcoxon's Sign-Rank Test in the Matched-Pair Design

Wilcoxon's test for the differences of pairs is the nonparametric analog to the paired  $t$ -test. This test can be applied to a continuous (not necessarily normal distributed) response. The test allows us to check whether the differences  $y_{1i} - y_{2i}$  of paired observations  $(y_{1i}, y_{2i})$  are symmetrically distributed with median  $M = 0$ .



In the two-sided test problem, the hypothesis is given by

$$H_0 : M = 0 \quad \text{or, equivalently,} \quad H_0 : P(Y_1 < Y_2) = 0.5, \quad (2.28)$$

versus

$$H_1 : M \neq 0 \quad (2.29)$$

and in the one-sided test problem

$$H_0 : M \leq 0 \quad \text{versus} \quad H_1 : M > 0. \quad (2.30)$$

Assuming  $Y_1 - Y_2$  being distributed symmetrically, the relation  $f(-d) = f(d)$  holds for each value of the difference  $D = Y_1 - Y_2$ , with  $f(\cdot)$  denoting the density function of the difference variable. Therefore, we can expect, under  $H_0$ , that the ranks of absolute differences  $|d|$  are equally distributed amongst negative and positive differences. We put the absolute differences in ascending order and note the sign of each difference  $d_i = y_{1i} - y_{2i}$ . Then we sum over the ranks of absolute differences with positive sign (or, analogously, with negative sign) and get the following statistic (cf. Büning and Trenkler, 1978, p. 187):

$$W^+ = \sum_{i=1}^n Z_i R(|d_i|) \quad (2.31)$$

with

$$\begin{aligned} d_i &= y_{1i} - y_{2i}, \\ R(|d_i|) &: \text{rank of } |d_i|, \\ Z_i &= \begin{cases} 1, & d_i > 0, \\ 0, & d_i < 0. \end{cases} \end{aligned} \quad (2.32)$$

We also could sum over the ranks of negative differences ( $W^-$ ) and get the relationship  $W^+ + W^- = n(n+1)/2$ .

**Exact Distribution of  $W^+$  under  $H_0$**

The term  $W^+$  can also be expressed as

$$W^+ = \sum_{i=1}^n i Z_{(i)} \quad \text{with} \quad Z_{(i)} = \begin{cases} 1, & D_j > 0, \\ 0, & D_j < 0. \end{cases} \quad (2.33)$$

In this case  $D_j$  denotes the difference for which  $r(|D_j|) = i$  for given  $i$ . Under  $H_0 : M = 0$  the variable  $W^+$  is symmetrically distributed with center

$$E(W^+) = E\left(\sum_{i=1}^n i Z_{(i)}\right) = \frac{n(n+1)}{4}.$$

The sample space may be regarded as a set  $L$  of all  $n$ -tuples built of 1 or 0.  $L$  itself consists of  $2^n$  elements and each of these has probability  $1/2^n$

under  $H_0$ . Hence, we get

$$P(W^+ = w) = \frac{a(w)}{2^n} \quad (2.34)$$

with  $a(w)$  : number of possibilities to assign + signs to the numbers from 1 to  $n$  in a manner that leads to the sum  $w$ .

*Example:* Let  $n = 4$ . The exact distribution of  $W^+$  under  $H_0$  can be found in the last column of the following table:

$w$	Tuple of ranks	$a(w)$	$P(W^+ = w)$
10	(1 2 3 4)	1	1/16
9	(2 3 4)	1	1/16
8	(1 3 4)	1	1/16
7	(1 2 4), (3 4)	2	2/16
6	(1 2 3), (2 4)	2	2/16
5	(1 4), (2 3)	2	2/16
4	(1 3), (4)	2	2/16
3	(1 2), (3)	2	2/16
2	(2)	1	1/16
1	(1)	1	1/16
0		1	1/16
$\Sigma: 16/16 = 1$			

For example,  $P(W^+ \geq 8) = 3/16$ .

### Testing

Test A:

$H_0 : M = 0$  is rejected versus  $H_1 : M \neq 0$ , if  $W^+ \leq w_{\alpha/2}$  or  $W^+ \geq w_{1-\alpha/2}$ .

Test B:

$H_0 : M \leq 0$  is rejected versus  $H_1 : M > 0$ , if  $W^+ \geq w_{1-\alpha}$ .

The exact critical values can be found in tables (e.g., Table H, p. 373 in Büning and Trenkler, 1978). For large sample sizes ( $n > 20$ ) we can use the following approximation

$$Z = \frac{W^+ - E(W^+)}{\sqrt{\text{Var}(W^+)}} \stackrel{H_0}{\approx} N(0, 1),$$

i.e.,

$$Z = \frac{W^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}. \quad (2.35)$$

For both tests,  $H_0$  is rejected if  $|Z| > u_{1-\alpha/2}$  (resp.,  $Z > u_{1-\alpha}$ ).

### Ties

Ties may occur as *zero-differences* ( $d_i = y_{1i} - y_{2i} = 0$ ) and/or as *compound-differences* ( $d_i = d_j$  for  $i \neq j$ ). Depending on the type of ties, we use one of the following test:

- zero-differences test;
- compound-differences test; and
- zero-differences plus compound-differences test.

The following methods are comprehensively described in Lienert (1986, pp. 327–332).

#### 1. Zero-Differences Test

- (a) Sample reduction method of Wilcoxon and Hemelrijk (Hemelrijk, 1952):

This method is used if the sample size is large enough ( $n \geq 10$ ) and the percentage of ties is less than 10% ( $t_0/n \leq 1/10$ , with  $t_0$  denoting the number of zero-differences).

Zero-differences are excluded from the sample and the test is conducted using the remaining  $n_0 = n - t_0$  pairs.

- (b) Pratt's partial-rank randomization method (Pratt, 1959):

This method is used for small sample sizes with more than 10% of zero-differences.

The zero-differences are included during the association of ranks but are excluded from the test statistic. The exact distribution of  $W_0^+$  under  $H_0$  is calculated for the remaining  $n_0$  signed ranks. The probabilities of rejection are given by:

- Test A (two-sided):

$$P'_0 = \frac{2A'_0 + a'_0}{2^{n_0}}.$$

- Test B (one-sided):

$$P'_0 = \frac{A'_0 + a'_0}{2^{n_0}}.$$

Here  $A'_0$  denotes the number of orderings which give  $W_0^+ > w_0$  and  $a'_0$  denotes the number of orderings which give  $W_0^+ = w_0$ .

- (c) Cureton's asymptotic version of the partial-rank randomization test (Cureton, 1967):

This test is used for large sample sizes and many zero-differences ( $t_0/n > 0.1$ ). The test statistic is given by

$$Z_{W_0} = \frac{W_0^+ - E(W_0^+)}{\sqrt{\text{Var}(W_0^+)}}$$

with

$$\begin{aligned} E(W_0^+) &= \frac{n(n+1) - t_0(t_0+1)}{4}, \\ \text{Var}(W_0^+) &= \frac{n(n+1)(2n+1) - t_0(t_0+1)(2t_0+1)}{24}. \end{aligned}$$

Under  $H_0$ , the statistic  $Z_{W_0}$  follows asymptotically the standard normal distribution.

## 2. Compound-Differences Test

- (a) Shared-ranks randomization method.

In small samples and for any percentage of compound-differences we assign averaged ranks to the compound-differences. The exact distributions as well as one- and two-sided critical values, are calculated as shown in Test 1(b).

- (b) Approximated compound-differences test.

If we have a larger sample ( $n > 10$ ) and a small percentage of compound-differences ( $t/n \leq 1/5$  with  $t$  = the number of compound-differences), then we assign averaged ranks to the compounded values. The test statistic is calculated and tested as usual.

- (c) Asymptotic sign-rank test corrected for ties.

This method is useful for large samples with  $t/n > 1/5$ .

In equation (2.36) we replace  $\text{Var}(W^+)$  by a corrected variance (due to the association of ranks)  $\text{Var}(W_{\text{corr.}}^+)$  given by

$$\text{Var}(W_{\text{corr.}}^+) = \frac{n(n+1)(2n+1)}{24} - \sum_{j=1}^r \frac{t_j^3 - t_j}{48},$$

with  $r$  denoting the number of groups of ties and  $t_j$  denoting the number of ties in the  $j$ th group ( $1 \leq j \leq r$ ). Unbounded observations are regarded as groups of size 1. If there are no ties, then  $r = n$  and  $t_j = 1$  for all  $j$ , e.g., the correction term becomes zero.

### 3. Zero-Differences Plus Compound-Differences Test

These tests are used if there are both zero-differences and compound-differences.

- (a) Pratt's randomization method.

For small samples which are cleared up for zeros ( $n_0 \leq 10$ ), we proceed as in Test 1(b) but additionally assign averaged ranks to the compound-differences.

- (b) Cureton's approximation method.

In larger zero-cleared samples the test statistic is calculated analogously to Test 3(a). The expectation  $E(W_0^+)$  equals that in Test 1(c) and is given by

$$E(W_0^+) = \frac{n(n+1) - t_0(t_0+1)}{4}.$$

The variance in Test 1(c) has to be corrected due to ties and is given by

$$\text{Var}_{\text{corr.}}(W_0^+) = \frac{n(n+1)(2n+1) - t_0(t_0+1)(2t_0+1)}{24} - \sum_{j=1}^r \frac{t_j^3 - t_j}{48}.$$

Finally, the test statistic is given by

$$Z_{W_0, \text{corr.}} = \frac{W_0^+ - E(W_0^+)}{\sqrt{\text{Var}_{\text{corr.}}(W_0^+)}}. \quad (2.36)$$

## 2.5 Rank Test for Homogeneity of Wilcoxon, Mann and Whitney

We consider two independent continuous random variables,  $X$  and  $Y$ , with unknown distribution or nonnormal distribution. We would like to test whether the samples of the two variables are samples of the same population (homogeneity). The so-called  $U$ -test of Wilcoxon, Mann, and Whitney is a rank test. As the Kruskal Wallis test (as the generalization of the Wilcoxon test) defines the null hypothesis that  $k$  populations are identical, *i.e.*, testing for the homogeneity of these  $k$  populations, the Mann Whitney Wilcoxon test could also be seen as a test for homogeneity for the case  $k = 2$  (cf. Gibbons, (1976), p. 173). This is the nonparametric analog of the  $t$ -test and is used if the assumptions for the use of the  $t$ -test are not justified or called into question. The relative efficiency of the  $U$ -test compared to the  $t$ -test is about 95% in the case of normally distributed variables. The  $U$ -test is often used as a quick test or as a control if the test statistic of the  $t$ -test gives values close to the critical values.

The hypothesis to be tested is  $H_0$  : the probability  $P$  to observe a value from the first population  $X$  that is greater than any given value of the population  $Y$  is equal to 0.5. The two-sided alternative is  $H_1$  :  $P \neq 0.5$ . The one-sided alternative  $H_1$  :  $P > 0.5$  means that  $X$  is *stochastically larger than*  $Y$ .

We combine the observations of the samples  $(x_1, \dots, x_m)$  and  $(y_1, \dots, y_n)$  in ascending order of ranks and note for each rank the sample it belongs to. Let  $R_1$  and  $R_2$  denote the sum of ranks of the  $X$ - and  $Y$ -samples, respectively. The test statistic  $U$  is the smaller of the values  $U_1$  and  $U_2$ :

$$U_1 = m \cdot n + \frac{m(m+1)}{2} - R_1, \quad (2.37)$$

$$U_2 = m \cdot n + \frac{n(n+1)}{2} - R_2, \quad (2.38)$$

with  $U_1 + U_2 = m \cdot n$  (control).

$H_0$  is rejected if  $U \leq U(m, n; \alpha)$  (Table 2.3 contains some values for  $\alpha = 0.05$  (one-sided) and  $\alpha = 0.10$  (two-sided)).

	$n$									
$m$	2	3	4	5	6	7	8	9	10	
4	—	0	1							
5	0	1	2	4						
6	0	2	3	5	7					
7	0	2	4	6	8	11				
8	1	3	5	8	10	13	15			
9	1	4	6	9	12	15	18	21		
10	1	4	7	11	14	17	20	24	27	

TABLE 2.3. Critical values of the  $U$ -test ( $\alpha = 0.05$  one-sided,  $\alpha = 0.10$  two-sided).

In the case of  $m$  and  $n \geq 8$ , the excellent approximation

$$u = \frac{U - m \cdot n/2}{\sqrt{m \cdot n(m+n+1)/12}} \sim N(0, 1) \quad (2.39)$$

is used. For  $|u| > u_{1-\alpha/2}$  the hypothesis  $H_0$  is rejected (type I error  $\alpha$  two-sided and  $\alpha/2$  one-sided).

*Example 2.4.* We test the equality of means of the two series of measurements given in Table 2.4 using the  $U$ -test. Let variable  $X$  be the flexibility of PMMA with silan and let variable  $Y$  be the flexibility of PMMA without silan. We put the  $(16 + 15)$  values of both series in ascending order, apply ranks and calculate the sums of ranks  $R_1 = 231$  and  $R_2 = 265$  (Table 2.5).

PMMA 2.2 Vol% quartz without silan	PMMA 2.2 Vol% quartz with silan
98.47	106.75
106.20	111.75
100.47	96.67
98.72	98.70
91.42	118.61
108.17	111.03
98.36	90.92
92.36	104.62
80.00	94.63
114.43	110.91
104.99	104.62
101.11	108.77
102.94	98.97
103.95	98.78
99.00	102.65
106.05	
$\bar{x} = 100.42$ $s_x^2 = 7.9^2$ $n = 16$	$\bar{y} = 103.91$ $s_y^2 = 7.6^2$ $m = 15$

TABLE 2.4. Flexibility of PMMA with and without silan (cf. Toutenburg, Toutenburg and Walther, 1991, p. 100).

Rank	1	2	3	4	5	6	7	8	9
Observation	80.00	90.92	91.42	92.36	94.63	96.67	98.36	98.47	98.70
Variable	X	Y	X	X	Y	Y	X	X	Y
Sum of ranks X	1		+3	+4			+7	+8	
Sum of ranks Y		2			+5	+6			+9
Rank	10	11	12	13	14	15	16	17	
Observation	98.72	98.78	98.97	99.00	100.47	101.11	102.65	102.94	
Variable	X	Y	Y	X	X	X	Y	X	
Sum of ranks X	+10	+11		+13	+14	+15		+17	
Sum of ranks Y			+12				+16		
Rank	18	19	20	21	22	23	24		
Observation	103.95	104.62	104.75	104.99	106.05	106.20	106.75		
Variable	X	Y	Y	X	X	X	Y		
Sum of ranks X	+18			+21	+22	+23			
Sum of ranks Y		+19	+20				+24		
Rank	25	26	27	28	29	30	31		
Observation	108.17	108.77	110.91	111.03	111.75	114.43	118.61		
Variable	X	Y	Y	Y	Y	X	Y		
Sum of ranks X	+25					+30			
Sum of ranks Y		+26	+27	+28	+29		+31		

TABLE 2.5. Computing the sums of ranks (Example 2.3, cf. Table 2.4).

Then we get

$$\begin{aligned}
 U_1 &= 16 \cdot 15 + \frac{16(16+1)}{2} - 231 = 145, \\
 U_2 &= 16 \cdot 15 + \frac{15(15+1)}{2} - 265 = 95, \\
 U_1 + U_2 &= 240 = 16 \cdot 15.
 \end{aligned}$$

Since  $m = 16$  and  $n = 15$  (both sample sizes  $\geq 8$ ), we calculate the test statistic according to (2.39) with  $U = U_2$  being the smaller of the two values of  $U$ :

$$u = \frac{95 - 120}{\sqrt{240(16 + 15 + 1)/12}} = -\frac{25}{\sqrt{640}} = -0.99,$$

and therefore  $|u| = 0.99 < 1.96 = u_{1-0.05F/2} = u_{0.975}$ .

The null hypothesis is not rejected (type I error 5% and 2.5% using two- and one-sided alternatives, respectively). The exact critical value of  $U$  is  $U(16, 15, 0.05_{\text{two-sided}}) = 70$  (Tables in Sachs, 1974, p. 232), *i.e.*, the decision is the same ( $H_0$  is not rejected).

### Correction of the $U$ -Statistic in the Case of Equal Ranks

If observations occur more than once in the combined and ordered samples  $(x_1, \dots, x_m)$  and  $(y_1, \dots, y_n)$ , we assign an averaged rank to each of them. The corrected  $U$ -test (with  $m + n = S$ ) is given by

$$u = \frac{U - m \cdot n/2}{\sqrt{[m \cdot n/S(S-1)][(S^3 - S)/12 - \sum_{i=1}^r (t_i^3 - t_i)/12]}}. \quad (2.40)$$

The number of groups of equal observations (ties) is  $r$ , and  $t_i$  denotes the number of equal observations in each group.

*Example 2.5.* We compare the time that two dentists B and C need to manufacture an inlay (Table 4.1). First, we combine the two samples in ascending order (Table 2.6).

Observation	19.5	31.5	31.5	33.5	37.0	40.0	43.5	50.5	53.0	54.0
Dentist	C	C	C	B	B	C	B	C	C	B
Rank	1	2.5	2.5	4	5	6	7	8	9	10
Observation	56.0	57.0	59.5	60.0	62.5	62.5	65.5	67.0	75.0	
Dentist	B	B	B	B	C	C	B	B	B	
Rank	11	12	13	14	15.5	15.5	17	18	19	

TABLE 2.6. Association of ranks (cf. Table 4.1) .

We have  $r = 2$  groups with equal data:

Group 1 : twice the value of 31.5;  $t_1 = 2$ ,

Group 2 : twice the value of 62.5;  $t_2 = 2$ .

The correction term then is

$$\sum_{i=1}^2 \frac{t_i^3 - t_i}{12} = \frac{2^3 - 2}{12} + \frac{2^3 - 2}{12} = 1.$$



The sums of ranks are given by

$$\begin{aligned} R_1 \text{ (dentist B)} &= 4 + 5 + \cdots + 19 = 130, \\ R_2 \text{ (dentist C)} &= 1 + 2.5 + \cdots + 15.5 = 60, \end{aligned}$$

and, according to (2.37), we get

$$U_1 = 11 \cdot 8 + \frac{11(11+1)}{2} - 130 = 24$$

and, according to (2.38),

$$\begin{aligned} U_2 &= 11 \cdot 8 + \frac{8(8+1)}{2} - 60 = 64, \\ U_1 + U_2 &= 88 = 11 \cdot 8 \text{ (control)}. \end{aligned}$$

With  $S = m + n = 11 + 8 = 19$  and with  $U = U_1$ , the test statistic (2.40) becomes

$$u = \frac{24 - 44}{\sqrt{\left[ \frac{88}{19 \cdot 18} \right] \left[ \frac{19^3 - 19}{12} - 1 \right]}} = -1.65,$$

and, therefore,  $|u| = 1.65 < 1.96 = u_{1-0.05/2}$ .

The null hypothesis  $H_0$  : *Both dentists need the same time to make an inlay* is not rejected. Both samples can be regarded as homogeneous and may be combined in a single sample for further evaluation.

We now assume the working time to be normally distributed. Hence, we can apply the  $t$ -test and get

$$\begin{aligned} \text{dentist B} : \bar{x} &= 55.27, s_x^2 = 12.74^2, n_1 = 11, \\ \text{dentist C} : \bar{y} &= 43.88, s_y^2 = 15.75^2, n_2 = 8, \end{aligned}$$

(see Table 4.1).

The test statistic (2.15) is given by

$$F_{10,7} = \frac{15.75^2}{12.74^2} = 1.53 < 3.15 = F_{10,7;0.95},$$

and the hypothesis of equal variance is not rejected. To test the hypothesis  $H_0 : \mu_x = \mu_y$  the test statistic (1.5) is used. The pooled sample variance is calculated according to (1.6) and gives  $s^2 = (10 \cdot 12.74^2 + 7 \cdot 15.75^2)/17 = 14.06^2$ . We now can evaluate the test statistic (1.5) and get

$$t_{17} = \frac{55.27 - 43.88}{14.06} \sqrt{\frac{11 \cdot 8}{11 + 8}} = 1.74 < 2.11 = t_{17;0.95(\text{two-sided})}.$$

As before, the null hypothesis is not rejected.

## 2.6 Comparison of Two Groups with Categorical Response

In the previous sections the comparisons in the matched-pair designs and in designs with two independent groups were based on the assumption of continuous response. Now we want to compare two groups with categorical response. The distributions (binomial, multinomial, and Poisson distributions) and the maximum-likelihood-estimation are discussed in detail in Chapter 8.

To start with, we first focus on binary response, e.g., to recover/not to recover from an illness, success/no success in a game, scoring more/less than a given level.

### 2.6.1 McNemar's Test and Matched-Pair Design

In the case of binary response we use the codings 0 and 1, so that the pairs in a matched design are one of the tuples of response (0, 0), (0, 1), (1, 0), or (1, 1). The observations are summarized in a  $2 \times 2$  table:

		Group 1		Sum
		0	1	
Group 2	0	$a$	$c$	$a + c$
	1	$b$	$d$	$b + d$
Sum		$a + b$	$c + d$	$a + b + c + d = n$

The null hypothesis is  $H_0 : p_1 = p_2$ , where  $p_i$  is the probability  $P(1 \mid \text{group } i)$  ( $i = 1, 2$ ). The test is based on the relative frequencies  $h_1 = (c + d)/n$  and  $h_2 = (b + d)/n$  for response 1, which differ in  $b$  and  $c$  (these are the frequencies for the discordant results (0, 1) and (1, 0)).

Under  $H_0$ , the values of  $b$  and  $c$  are expected to be equal or, analogously, the expression  $b - (b + c)/2$  is expected to be zero. For a given value of  $b + c$ , the number of discordant pairs follows a binomial distribution with the parameter  $p = 1/2$  (probability to observe a discordant pair (0, 1) or (1, 0)). As a result, we get  $E[(0, 1)\text{-response}] = (b + c)/2$  and  $\text{Var}[(0, 1)\text{-response}] = (b + c) \cdot \frac{1}{2} \cdot \frac{1}{2}$  (analogously, this holds symmetrically for  $[(1, 0)\text{-response}]$ ).

The following ratio then has expectation 0 and variance 1:

$$\frac{b - (b + c)/2}{\sqrt{(b + c) \cdot 1/2 \cdot 1/2}} = \frac{b - c}{\sqrt{b + c}} \stackrel{H_0}{\sim} (0, 1)$$

and follows the standard normal distribution for reasonably large  $(b + c)$  due to the central limit theorem. This approximation can be used for  $(b + c) \geq 20$ . For the continuity correction, the absolute value of  $|b - c|$  is decreased

by 1. Finally, we get the following test statistic:

$$Z = \frac{(b - c) - 1}{\sqrt{b + c}} \quad \text{if } b \geq c, \quad (2.41)$$

$$Z = \frac{(b - c) + 1}{\sqrt{b + c}} \quad \text{if } b < c. \quad (2.42)$$

Critical values are the quantiles of the cumulated binomial distribution  $B(b + c, \frac{1}{2})$  in the case of a small sample size. For larger samples (*i.e.*,  $b + c \geq 20$ ), we choose the quantiles of the standard normal distribution. The test statistic of McNemar is a certain combination of the two  $Z$ -statistics given above. This is used for a two-sided test problem in the case of  $b + c \geq 20$  and follows a  $\chi^2$ -distribution

$$Z^2 = \frac{(|b - c| - 1)^2}{b + c} \sim \chi_1^2. \quad (2.43)$$

*Example 2.6.* A clinical experiment is used to examine two different teeth-cleaning techniques and their effect on oral hygiene. The response is coded binary: reduction of tartar yes/no. The patients are stratified into matched pairs according to sex, actual teeth-cleaning technique, and age. We assume the following outcome of the trial:

		Group 1		Sum
		0	1	
Group 2	0	10	50	60
	1	70	80	150
Sum		80	130	210

We test  $H_0 : p_1 = p_2$  versus  $H_1 : p_1 \neq p_2$ . Since  $b + c = 70 + 50 > 20$ , we choose the McNemar statistic

$$Z^2 = \frac{(|70 - 50| - 1)^2}{70 + 50} = \frac{19^2}{120} = 3.01 < 3.84 = \chi_{1;0.95}^2$$

and do not reject  $H_0$ .

*Remark.* Modifications of the McNemar test can be constructed similarly to sign tests. Let  $n$  be the number of nonzero differences in the response of the pairs and let  $T_+$  and  $T_-$  be the number of positive and negative differences, respectively. Then the test statistic, analogously to the  $Z$ -statistics (2.41) and (2.42), is given by

$$Z = \frac{(T_+/n - 1/2) \pm n/2}{1/\sqrt{4n}}, \quad (2.44)$$

in which we use  $+n/2$  if  $T_+/n < 1/2$  and  $-n/2$  if  $T_+/n \geq 1/2$ . The null hypothesis is  $H_0 : \mu_d = 0$ . Depending on the sample size ( $n \geq 20$  or  $n < 20$ ) we use the quantiles of the normal or binomial distributions.

### 2.6.2 Fisher's Exact Test for Two Independent Groups

Regarding two independent groups of size  $n_1$  and  $n_2$  with binary response, we get the following  $2 \times 2$  table

	Group 1	Group 2	
1	$a$	$c$	$a + c$
0	$b$	$d$	$b + d$
	$n_1$	$n_2$	$n$

The relative frequencies of response 1 are  $\hat{p}_1 = a/n_1$  and  $\hat{p}_2 = c/n_2$ . The null hypothesis is  $H_0 : p_1 = p_2 = p$ . In this contingency table, we identify the cell with the smallest cell count and calculate the probability for this and all other tables with an even smaller cell count in the smallest cell. In doing so, we have to ensure that the marginal sums keep constant.

Assume  $(1, 1)$  to be the weakest cell. Under  $H_0$  we have, for response 1 in both groups (for given  $n$ ,  $n_1$ ,  $n_2$ , and  $p$ ):

$$P((a + c) | n, p) = \binom{n}{a + c} p^{a+c} (1 - p)^{n-(a+c)},$$

for Group 1 and response 1:

$$P(a | (a + b), p) = \binom{a + b}{a} p^a (1 - p)^b,$$

for Group 2 and response 1:

$$P(c | (c + d), p) = \binom{c + d}{c} p^c (1 - p)^d.$$

Since the two groups are independent, the joint probability is given by

$$P(\text{Group 1} = a \wedge \text{Group 2} = c) = \binom{a + b}{a} p^a (1 - p)^b \binom{c + d}{c} p^c (1 - p)^d$$

and the conditional probability of  $a$  and  $c$  (for the given marginal sum  $a + c$ ) is

$$\begin{aligned} P(a, c | a + c) &= \frac{\binom{a + b}{a} \binom{c + d}{c}}{\binom{n}{a + c}} \\ &= \frac{(a + b)! (c + d)! (a + c)! (b + d)!}{n!} \cdot \frac{1}{a! b! c! d!}. \end{aligned}$$

Hence, the probability to observe the given table or a table with an even smaller count in the weakest cell is

$$P = \frac{(a + b)! (c + d)! (a + c)! (b + d)!}{n!} \cdot \sum_i \frac{1}{a_i! b_i! c_i! d_i!},$$

with summation over all cases  $i$  with  $a_i \leq a$ . If  $P < 0.05$  (one-sided) or  $2P < 0.05$  (two-sided) hold, then hypothesis  $H_0 : p_1 = p_2$  is rejected.

*Example 2.7.* We compare two independent groups of subjects receiving either type *A* or type *B* of an implanted denture and observe whether it is lost during the healing process (8 weeks after implantation). The data are

	<i>A</i>	<i>B</i>	
Yes	2	8	10
No	10	4	14
	12	12	24

The two tables with a smaller count in the (yes | *A*) cell are

$$\begin{array}{|c|c|} \hline 1 & 9 \\ \hline 11 & 3 \\ \hline \end{array} \quad \text{and} \quad \begin{array}{|c|c|} \hline 0 & 10 \\ \hline 12 & 2 \\ \hline \end{array}$$

and, therefore, we get

$$P = \frac{10! \, 14! \, 12! \, 12!}{24!} \left( \frac{1}{2! \, 8! \, 10! \, 4!} + \frac{1}{1! \, 9! \, 11! \, 3!} + \frac{1}{0! \, 10! \, 12! \, 2!} \right) = 0.018,$$

$$\left. \begin{array}{ll} \text{one-sided test:} & P = 0.018 \\ \text{two-sided test:} & 2P = 0.036 \end{array} \right\} < 0.05.$$

Decision.  $H_0 : p_1 = p_2$  is rejected in both cases. The risk of loss is significantly higher for type *B* than for type *A*.

### Recurrence Relation

Instead of using tables, we can also use the following recurrence relation (cited by Sachs, 1974, p. 289):

$$P_{i+1} = \frac{a_i d_i}{b_{i+1} c_{i+1}} P_i.$$

In our example, we get

$$\begin{aligned} P &= P_1 + P_2 + P_3, \\ P_1 &= \frac{10! \, 14! \, 12! \, 12!}{24!} \frac{1}{2! \, 8! \, 10! \, 4!} \\ &= 0.0166, \\ P_2 &= \frac{2 \cdot 4}{11 \cdot 9} P_1 = 0.0013, \\ P_3 &= \frac{1 \cdot 3}{12 \cdot 10} P_2 = 0.0000, \end{aligned}$$

and, therefore,  $P = 0.0179 \approx 0.0180$ .

## 2.7 Exercises and Questions

- 2.7.1 What are the differences between the paired  $t$ -test and the two-sample  $t$ -test (degrees of freedom, power)?
- 2.7.2 Consider two samples with  $n_1 = n_2$ ,  $\alpha = 0.05$  and  $\beta = 0.05$  in a matched-pair design and in a design of two independent groups. What is the minimum sample size needed to achieve a power of 0.95, assuming  $\sigma^2 = 1$  and  $\delta^2 = 4$ .
- 2.7.3 Apply Wilcoxon's sign-rank test for a matched-pair design to the following table:

TABLE 2.7. Scorings of students who took a cup of coffee either before or after a lecture.

Student	Before	After
1	17	25
2	18	45
3	25	37
4	12	10
5	19	21
6	34	27
7	29	29

Does treatment  $B$  (coffee before) significantly influence the score?

- 2.7.4 For a comparison of two independent samples,  $X$  : leaf-length of strawberries with manuring  $A$ , and  $Y$  : manuring  $B$ , the normal distribution is put in question. Test  $H_0 : \mu_X = \mu_Y$  using the homogeneity test of Wilcoxon, Mann, and Whitney.

$A$	$B$
37	45
49	51
51	62
62	73
74	87
89	45
44	33
53	
17	

Note that there are ties.

- 2.7.5 Recode the response in Table 2.4 into binary response with: flexibility  $< 100 : 0$  ,

flexibility  $\geq 100 : 1$  ,

and apply Fisher's exact test for  $H_0 : p_1 = p_2$  ( $p_i = P(1 \mid \text{group } i)$ ).

- 2.7.6 Considering Exercise 2.7.3, we assume that the response has been binary recoded according to scoring higher/lower than average: 1/0. A sample of  $n = 100$  shows the following outcome:

		Before		
		0	1	
After	0	20	25	45
	1	15	40	55
		35	65	100

Test for  $H_0 : p_1 = p_2$  using McNemar's test.

Statistical Analysis of Designed Experiments, Third  
Edition

Toutenburg, H.; Shalabh

2009, XVIII, 615 p., Hardcover

ISBN: 978-1-4419-1147-6