Volume 19 Number 3
March 1995

ISSN 0167-9473
CSDA 19(3) (1995) 265–368

# COMPUTATIONAL STATISTICS & DATA ANALYSIS

Incorporating Statistical Software Newsletter

The official journal of

iasc

The International Association for Statistical Computing
A Section of The International Statistical Institute

North-Holland

# SSN

# The Statistical Software Newsletter

**This issue:**



SSNinCSDA 19 (1995) N° 3: 351-368

# Preference of equivalence tests with standardized mean difference demonstrated by an application

**Ernst Schuster**

Institut für Medizinische Informatik, Statistik und Epidemiologie,
Universität Leipzig, Liebigstr. 27, D-04103 Leipzig, Germany

**Summary**: The disadvantages of non-standardized and the advantages of standardized equivalence tests are demonstrated by an extreme example. This is the summary of a talk presented at the occasion of Statistical Computing '94 in Günzburg, Germany, on June 21, 1994. (SSNinCSDA 19, 353-356 (1994))

**Keywords:** equivalence test, standardized mean difference, bioequivalence

*Received: July 1994* *Revised: January 1995*

## I. Introduction

Equivalence tests are used for studies to prove, that there are no practical relevant differences between two treatments. Therefore the type I error of equivalence tests consists of a false decision in favour of the hypothesis, that the treatments have "essentially identical" effects. With the equivalence in a narrow sense (in contrast to the "one-sided" equivalence, which will not be described here) the alternative hypothesis consists of an region limited on both sides. Let $\theta_0$ be a distribution parameter, which sensibly specifies the interesting effect in the population. The equivalence hypothesis assures the equality of $\theta_0$ with an appropriate nominal value $\theta_0$ "except for practically irrelevant deviations". These are described by a mostly symmetric interval near $\theta_0$, i.e. $(\theta_0-\varepsilon, \theta_0+\varepsilon)$ with $\varepsilon>0$ small. Often $\theta_0$ is equal to 0.

With the equivalence test the *null hypothesis of non-equivalence*

$$H_0: \theta \le \theta_0-\varepsilon \text{ or } \theta \ge \theta_0+\varepsilon$$

contrasts the *alternative hypothesis of equivalence acceptance*

$$H_A: \theta_0-\varepsilon < \theta < \theta_0+\varepsilon .$$

The choice of the distribution parameters $\theta$ is decisive. For two independent samples with the same variance $\sigma^2$, the parameter $\theta$ is often equal to $(\mu_1 - \mu_2)/\sigma$ the standardized mean difference (Wellek, 1993, 1994).

Another possibility $\theta = (\mu_1 - \mu_2)$ (Westlake, 1976) can lead to contradictions, as is shown, in the following example.

## II. Application sample

The following is an example of a bioequivalence study with two preparations of Nifedin. In this two-periods-crossover-study the randomly chosen patients of the first group get the test-preparation at first and then the reference-preparation, the participants of the second group get the reference-preparation at first and then the test-preparation. Serum concentration of the substance is measured at fixed times. The area (AUC) under the polygon curve of measure of values is the measure for the bioavailability. The AUC-values $A_i$ and $B_i$ are shown in Table 1. The suggested parametric model is used in the analysis. The model proceeds from a normally distributed logarithm of the period quotient, i.e.

$$X_i = \ln(A_{i1} / A_{i2}) \sim N(\mu_1, \sigma^2)$$
$$\text{with } \mu_1 = \ln(\mu_T / \mu_R) + \ln(\pi_1 / \pi_2).$$

The suggestion for the second group is analogous, i.e. $Y_i = \ln(B_{i1} / B_{i2})$ is normally distributed $Y_i \sim N(\mu_2, \sigma^2)$ with $\mu_2 = \ln(\mu_R / \mu_T) + \ln(\pi_1 / \pi_2)$. $\mu_T$ is the effect of the test-preparation, $\mu_R$ the effect of reference and the $\pi_i$ are the period effects. Calculations of these values are shown in Table 1. For the illustration of the asserted contradictions two values have been changed appropriately (Table 1).

*Table 1: Raw data for a bioequivalence study of the Calcium antagonist Nifedin*

| | group 1 | | | | | group 2 | | |
|---|---|---|---|---|---|---|---|---|
| i | $A_{i1}$ (Test) | $A_{i2}$ (Ref.) | $X_i^{org}$ $=\ln(A_{i1}/A_{i2})$ | $X_i$ | | $B_{i1}$ (Ref.) | $B_{i2}$ (Test) | $Y_i$ $=\ln(B_{i1}/B_{i2})$ |
| 1 | 106.9 | 112.9 | -.05461 | -.35461 | | 217.3 | 195.2 | .10725 |
| 2 | 131.3 | 124.4 | .05398 | -.25398 | | 174.4 | 122.7 | .35161 |
| 3 | 81.4 | 89.5 | -.09486 | | | 155.8 | 188.2 | -.18893 |
| 4 | 154.7 | 134.9 | .13695 | | | 299.5 | 309.2 | -.03187 |
| 5 | 111.2 | 108.3 | .02643 | | | 157.6 | 153.5 | .02636 |
| 6 | 85.8 | 94.0 | -.09128 | | | 121.4 | 104.7 | .14799 |
| 7 | 295.2 | 418.6 | -.34926 | | | 143.9 | 119.3 | .18748 |
| 8 | 217.0 | 207.0 | .04718 | | | 157.0 | 146.8 | .06717 |
| 9 | 252.3 | 239.3 | .05290 | | | 114.5 | 138.2 | -.18813 |
| 10 | 157.9 | 207.3 | -.27221 | | | 71.0 | 70.3 | .00991 |
| | | | $x_{org}$ = -0.05448 | x = - 0.1153 | | | | y = 0.04888 |
| | | | $S_X$ = 0.1542 | $S_X$=0.181 | | | | $S_Y$=0.1652 |

*Note: Data (Wellek, 1994)*

### III. t-test

The results of a normal t-test in Table 2 demonstrate, that the two treatments differ with 5% significance level ( p = 0.049), i.e. the test-preparation is worse!

### IV. Equivalence test (EQ-test) for $\theta = \mu_1 - \mu_2$

For $\mu_T / \mu_R$ an equivalence region (EQ-region) of 80-120% is typical i.e. $0.8 < \mu_T / \mu_R < 1.2$ .

Since $\theta = (\mu_1 - \mu_2)$

$$= \ln(\mu_T / \mu_R) + \ln(\pi_1 / \pi_2) - [\ln(\mu_R / \mu_T) + \ln(\pi_1 / \pi_2)]$$

$$= 2 \ln(\mu_T / \mu_R),$$

the equivalence region for $\theta$ is $-0.4463 < \theta < 0.3646$.

According to the **interval inclusion test** (Westlake, 1976) two treatments are equivalent if the $(1-2\alpha)$%-confidence interval is included in the equivalence region.

The 90% confidence interval (-0,299; -0,030), which is derived in Table 2, is indeed included in the equivalence region. Thus we **have equivalence** if $\theta = (\mu_1 - \mu_2)$ is used, **although both treatments differ significantly.**

### V. Equivalence test (EQ-test) for $\theta = (\mu_1 - \mu_2) / \sigma$

For an EQ-test with a test-statistic T(x) critical region

$$\{ x \mid C_1 < T(x) < C_2 \}.$$

*Table 2: Results of the t-test for the example (SPSS, 1993)*

t-tests for independent samples

| Variable | Number of Cases | Mean | SD | SEM |
|---|---|---|---|---|
| X | | | | |
| GR 1. | 10 | -.1153 | .181 | .057 |
| GR 2. | 10 | .0489 | .165 | .052 |

Mean Difference = -.1642

Levene's Test for Equality of Variances: F= .546 P= .470

t-test for Equality of Means

| Variances | t-value | df | 2-Tail Value | SE of Diff | 90% CI for Diff |
|---|---|---|---|---|---|
| Equal | -2.12 | 18 | .048 | 078 | (-.299; -.030) |
| Unequal | -2.12 | 17.85 | .049 | 078 | (-.299; -.030) |

is determined by the condition, that $P_\theta[\, C_1 < T(X) < C_2] \le \alpha$ for all $\theta \in H_0$ with $\alpha$ being the significance level.

Here I only examine the special case, that the EQ-region is symmetrical near the null $(-\varepsilon, \varepsilon)$, as discussed in the introduction. Let T(X) be a continual, symmetrical near null with monotonously increasing density up to the expected value and

$$T(X)_{|\theta-\epsilon} \sim -T(X)_{|\theta-\epsilon} \text{ for all } \varepsilon>0.$$

Then according to Wellek (1994), the critical region is symmetrical near null:

$$\{ x \mid -C < T(x) < C \}$$

(C must be determined from the equation:

$$P_{|\theta-\epsilon} ( |T(X)| < C ) = \alpha).$$

Due to the suggested characteristics of T(X), the examined probability for other $\theta \in H_0$ is smaller at most. In the example T(X) is the t-statistics for two independent samples: where

$$T(X) = (m\, n / (n+m))^{1/2} ( \overline{X} - \overline{Y} ) / S$$

and C is the result of the corresponding quantile of the non-central t-distribution. For the calculation of C the non-central F-distribution with first degree of freedom 1 is more convenient (SAS, 1988):

$C_{m,n}(\alpha; \varepsilon) =$ { 100$\alpha$-percent point of F-distribution with 1, m+n-2 degrees of freedom and $NC = (mn / (m+n))\varepsilon^2\}^{1/2}$.

Now the rule of decision is:

$$EQ \Leftrightarrow |T| < C_{m,n}(\alpha; \varepsilon) .$$

In the example, the test value T = -2.12 as shown in Table 2. The degrees of freedom of the F-distribution are 1 and 2*10 - 2=18. With $NC = (10^2/20)\ 1^2 = 5$ for $\varepsilon$=1, one obtains

$$C_{10,10}(0{,}05;\ 1) = 0.61365 < 2.12,$$

by writing C=sqrt(finv(.05, 1, 18, 5)) in a SAS (1988) DATA step and for $\varepsilon$=1.5:

$$C_{10,10}(0.05;\ 1.5) = 1.66969 < 2.12.$$

Thus according to this procedure the treatments **are not equivalent** even for $\varepsilon$ =1.5.

Finally the choice of $\varepsilon$ shall be examined. Because $X_i \sim N(\mu_1, \sigma^2)$ and $Y_i \sim N(\mu_2, \sigma^2)$ are normally distributed, the difference is normally distributed too with double variance $(X_i - Y_i) \sim N(\mu_1-\mu_2, 2\sigma^2)$. It follows $P(X_i > Y_i) = \Phi((\mu_1-\mu_2) / (\sigma\sqrt{2}))$. If one divides the EQ-region $-\varepsilon < (\mu_1-\mu_2)/\sigma < \varepsilon$ by $\sqrt{2}$ and goes over the inequalities with the monotonously in-

creasing standard normal distribution $\Phi$, a new EQ-

with $\tilde{\varepsilon} = \Phi(\varepsilon/\sqrt{2}) - 1/2$ for the above probability. The probability 1/2 corresponds to no difference, therefore $\tilde{\varepsilon}$ is so standardized, that it shows the deviation of 1/2. From Table 3 you can see, that $\varepsilon$ should not be greater than 1. Therefore, the probability is between 1/4 and 3/4 instead of 1/2.

*Table 3*

| $\varepsilon$ | $\tilde{\varepsilon}$ |
|---|---|
| 0.5 | 0.13 |
| 1 | 0.26 |
| 1.5 | 0.35 |

## VI. Conclusions

Wellek (1993, 1994) pointed to the problems of EQ-tests for $\theta = (\mu_1 - \mu_2)$. The example demonstrated shows that non-standardized EQ-tests should be avoided.

An exception only would be acceptable, if the equivalence region for location parameters is defined in a form, that small differences are unimportant concerning the special branch of science. Then the test of interval inclusion tests the conformity of the local parameters but not the conformity of the underlying distributions.

Hilgers (1994) proposed the combination of a one-

by him.

In my opinion only two simultaneous equivalence tests for location- and dispersion-parameters might be an alternative for standardization (Bauer et al., 1994).

## References

Bauer, P. & Bauer, M.M. (1994): Testing equivalence simultaneously for location and dispersion of two normally distributed populations. Biom. J. 36, 643-660

Hauschke, D., Steinijans, V.W. & Diletti, E. (1990): A distribution-free procedure for the statistical analysis of bioequivalence studies. Int. J. of Clinical Pharmacology, Therapy and Toxicology 28, 72-78

Hilgers, R.A. (1994): Tests auf "therapeutische Äquivalenz". (In: Pöppl, S.J. u.a. (Hrsg): Medizinische Informatik. München: MMV Medizin Verlag)

Munk, A. (1994): An improvement on commonly used tests in bioequivalence assessment. Biometrics 49, 1225-1230

SAS Language Guide for Personal Computers (1988): Release 6.03 Edition (Cary: NC: SAS Institute Inc.)

SPSS for Windows Base System User's Guide (1993): Release 6.0, (SPSS Inc.)

Wellek, S. (1993): Basing the analysis of comparative bioavailability trials on an individualized statistical definition of equivalence. Biom. J. 35, 47-55

Wellek, S. (1994): Statistische Methoden zum Nachweis von Äquivalenz. Stuttgart: Gustav Fischer Verlag