Pak. J. Statist.
S. Ejaz Ahmed Special Issue
Vol. 18(2), pp.197-211 (2002)

A NEW PROPOSAL FOR PAIRWISE MULTIPLE COMPARISONS WITH REPEATED MEASUREMENTS

Ernst Schuster
Institute for Medical Informatics, Statistics and Epidemiology
University of Leipzig
Leipzig, Germany

Siegfried Kropf
Institute for Biometry and Medical Informatics
Otto von Guericke University
Magdeburg, Germany

Abstract

A new proposal is derived for multiple comparisons with repeated measurements. It is based on previous papers by Kropf (2000) and Kropf and Läuter (2002), where multiple comparison procedures are given for a set of dependent variables in the parametric one-sample problem. All hypotheses, i.e. all pairs of repeated measurements of interest, are ordered according to a data-dependent criterion and then tested sequentially in t tests for paired samples as long as significance occurs at the unadjusted error level. To obtain a suitable order of the hypotheses and hence a powerful procedure, a compound symmetry structure should approximately *be given. The application is demonstrated with data from a prospective study for patients suffering from rheumatoid arthritis. The properties of the procedure are demonstrated in simulation experiments showing good results especially for small samples.*

AMS Subject Classification: Primary 62J15, Secondary 62F03, 62F07, 62F35, 62H15, 62H10 and 62-07.

Key Words

Multiple comparisons, repeated measurements, ordered hypotheses, sequential procedure, parametric tests.

1. INTRODUCTION

Frequently a response variable is measured over several treatment conditions or periods of time. We consider the situation with a fixed set of measurements per individual, usually referred to as repeated measures situation. Particularly, we deal with the following multivariate Gaussian model:

$$x_j = \begin{pmatrix} x_{j1} \\ \vdots \\ x_{jk} \end{pmatrix} \sim N_k(\mu, \Sigma), \quad \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{pmatrix} \qquad (1)$$

for $j = 1,\ldots,n$ iid sample vectors consisting of the k generally dependent measurements x_{ji}, $i = 1,\ldots,k$ of the response variable. In many applications, a compound symmetry structure can approximately be assumed for the covariance matrix Σ, where all diagonal elements (variances) are equal and also all non-diagonal elements (covariances) are equal. In this paper, we do not need this assumption to derive the null distribution of the test statistics, but we will utilize this approximate structure to attain a high power in the tests.

In this situation, usually a multivariate analysis of variance is applied as a test for global effects over the different treatments or periods of time. Under the compound symmetry assumption, one can alternatively use a univariate ANOVA test (Timm,1980).

In addition to the test of the global hypothesis H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$, standard software packages typically consider orthogonal contrasts, such as Helmert contrasts or polynomial contrasts. If these contrasts are orthogonal, then the corresponding test statistics are independent from each other under the compound symmetry assumption. Nevertheless, it remains the problem of multiple testing because several tests are carried out to break down the original single question of differences among the k situations.

The so-called experimentwise error rate of a multiple comparison method is the supremum of the probability of making at least one type I error in all decisions of the procedure over all possible parameter configurations (cf., e.g., Hsu, 1996). The simplest way to ensure this experimentwise error rate is the Bonferroni adjustment, where each single test of the procedure uses the local level α/m for m simultaneous tests. Frequently, especially for nonlinear curves, the user wishes to carry out all pairwise comparisons and then α/m may be small.

Subsequently an alternative procedure for multiple comparisons is developed. As in the well known principle of testing a-priori ordered hypotheses (Maurer, Hothorn and Lehmacher, 1995), where the order is established on the basis of prior knowledge or experimental conditions, we want to test the hypotheses sequentially at the full α level as long as significance can be attained in this order. However, we want to derive the order of hypotheses from the data themselves without introducing α adjustment techniques. A data-driven order of hypotheses is also used in Holm's (Holm, 1979) procedure, where the variables would be ordered for increasing P values in the local tests. But then the P values have to be compared with the critical values α/m, $\alpha/(m-1),\ldots$, i.e., an α adjustment occurs.

Here, we utilize an alternative proposal by Kropf (2000) and Kropf and Läuter (2002) for ordering the hypotheses. It considers tests for the univariate hypotheses

$H_i : \mu_i = 0$ $(i = 1, \ldots, k)$ in the above Gauss model (1), i.e., the points in time are considered separately, not the differences among them:

- The k conditions are ordered for decreasing values of $\sum_{j=1}^{n} x_{ji}^2$ for $i = 1, \ldots, k$.
- In this order, the usual t tests for the hypothesis $\mu_i = 0$ are applied at the full level α as long as all tests are significant.

This procedure keeps the experimentwise error rate α even in the case of unequal variances at different points in time. The proof is given in the above papers utilizing results from multivariate analysis based on spherical distributions (Fang and Zhang, 1990; Läuter, Glimm and Kropf (1998). However, the assumption of equal variances is necessary in order to have an indication for a convenient order of hypotheses, and hence for the power of the multiple procedure.

The theorem is now applied to the comparisons between different time points.

2. NEW PROPOSAL FOR PAIRWISE COMPARISONS OF DEPENDENT SAMPLES

We consider the $p = k(k-1)/2$ pairs $(1,2), \ldots, (k-1, k)$ of different points in time and calculate the corresponding differences $d_{j1} = x_{j1} - x_{j2}, \ldots, d_{jp} = x_{j,k-1} - x_{jk}$ for each sample vector x_j $(j = 1, \ldots, n)$.

Alternatively, one could select special pairs which are important to compare in a study. A typical example is the many-to-one situation. Then p would be smaller. All other statements in the subsequent text are valid for this reduced test situation as well.

For $j = 1, \ldots, n$ the vectors $(d_{j1}, \ldots, d_{jp})'$ are independent from each other and have a multivariate normal distribution with expectation $(\theta_1, \ldots, \theta_p)'$. Under the additional compound symmetry assumption for the vectors x_j, the p components of the vector of differences have also equal variances. Therefore the above theorem is applicable for the hypotheses $\theta_l = 0$, $l = 1, \ldots, p$, resulting in the following procedure:

- Order the p differences of time points for decreasing values of $\sum_{j=1}^{n} d_{jl}^2$, $l = 1, \ldots, p$.
- In this order, carry out the usual one-sample t tests for the difference values or, equivalently, the usual t test for paired samples to test the hypotheses $\theta_l = 0$, $l = 1, \ldots, p$ at the full level α as long as all tests yield significant results. Stop at the first non-significant result. Dependent on the practical problem, the tests may be one-sided or two-sided.

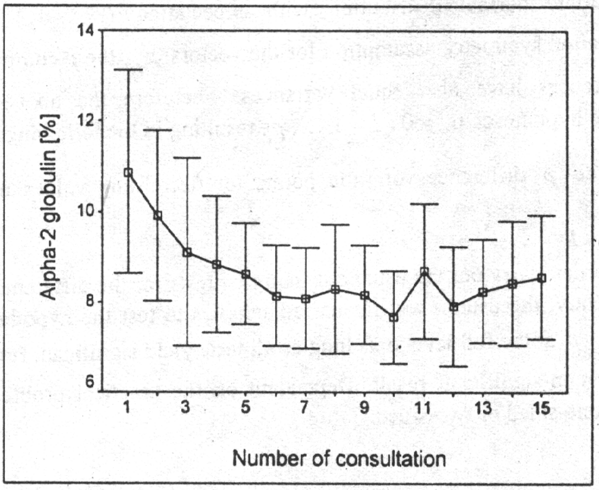
This procedure is referred to as 'sequential t tests' in this paper. It keeps the experimentwise error rate α. The resulting order of hypotheses may be useful because $\sum_{j=1}^{n} d_{jl}^2 / n = \frac{n-1}{n} s_l^2 + \bar{d}_l^2$ for each $l = 1, \ldots, p$. Therefore with equal variances for all differences, the order of hypotheses is mainly determined by the mean differences. Pairs of time points with large mean differences and hence large t values should be in the front part of the ordered sequence of pairs.

If the compound symmetry assumption is not met with the data, then again the procedure keeps the experimentwise error rate α nonetheless, but the power of the procedure may be insufficient.

3. EXAMPLE

The data of the following example are from a prospective clinical study of patients with recent-onset rheumatoid arthritis to examine the relationship between inflammatory disease activity and joint destruction in follow up, and to evaluate prognostic markers for severe joint erosions early in the disease (Schuster, 2000). Patients with a disease duration of less than 2 years were included into the study. Variables of clinical and laboratory disease activity were monitored. Here we show sequential data on the fraction of α_2-globulins in the serum protein electrophoresis (given in % of total serum protein concentration), a parameter associated with the acute phase response in inflammatory diseases.

Figure 1. Alpha-2 globulin values in the sample data (mean ± std dev)



The proposed multiple procedure is demonstrated on a subsample of 12 patients from this study. We selected these patients because they had complete values of α_2-globulin over 15 consultations. During the first 6 months of observation, patients were seen monthly, and later in 8-week intervals.

Figure 1 depicts means and standard deviations of α_2-globulin over the 15 time points. As seen in this graph, the values decrease in the first 6 months and remain, more or less, unchanged later on.

Table 1 shows a selection of all pairwise comparisons. All 105 possible pairs of time points have been included in the procedure, but not all of them are presented in the table. The rows, corresponding to the different pairs of time points, are ordered according to the criterion $\sum_{j=1}^{n} d_{jl}^2$ (last column). In the selected pairs, all mean differences are positive, reflecting the decrease in the means of the first few time points. Combining the mean differences with the corresponding standard errors, the P values in the paired t tests can be derived (5th column).

Table 1. Results of the proposed sequential procedure in the rheumatoid arthritis data

Ser. no.	Time points	Mean difference	Std. error	P value in t test	Order criterion
1	1 - 6	2.75	0.78	0.0047	171.0
2	1 - 10	3.23	0.54	0.0001	163.9
3	1 - 12	2.98	0.65	0.0008	163.4
4	1 - 9	2.73	0.70	0.0026	154.2
5	1 - 8	2.60	0.71	0.0039	148.3
6	1 - 7	2.81	0.60	0.0006	141.7
7	1 - 13	2.64	0.61	0.0012	133.1
8	1 - 5	2.26	0.72	0.0094	129.4
9	1 - 15	2.32	0.69	0.0062	126.6
10	1 - 14	2.47	0.60	0.0017	120.6
11	1 - 3	1.78	0.78	0.0441	119.4
12	1 - 11	2.20	0.68	0.0077	118.5
13	1 - 4	2.05	0.72	0.0154	118.1
14	2 - 10	2.26	0.56	0.0019	102.4
15	2 - 12	2.02	0.63	0.0086	101.4
16	1 - 2	0.97	0.73	0.2144	82.3
17	2 - 13	1.68	0.59	0.0163	79.8
18	3 - 12	1.20	0.68	0.1047	78.1
19	3 - 10	1.44	0.62	0.0396	75.3
...	...	...	...	...	...

Without any adjustment for multiplicity of the pairwise comparisons, 30 of all 105 pairwise tests are significant at the 0.05 error level, while 19 are still significant at the 0.01 level. In the proposed sequential procedure we can state P values less than 0.01 for the first 10 pairs of time points, ensuring the differences at the experimentwise error level 0.01 for these pairs. The 11th pair gives a P value of 0.0441, stopping that series. But until the 15th pair, still the 0.05 level is ensured. The 16th position (pair 1 – 2) gives a P value of 0.2144 and finishes the series of significances. Despite of small P values in later rows, no significance can be shown for the subsequent pairs at the experimentwise error level. Thus, a difference in the alpha-2 globulin level in 15 pairs of time points can be observed. In contrast, with the Bônferroni-Holm procedure only one comparison would give significant differences at the 0.05 error level (pair 1 – 10).

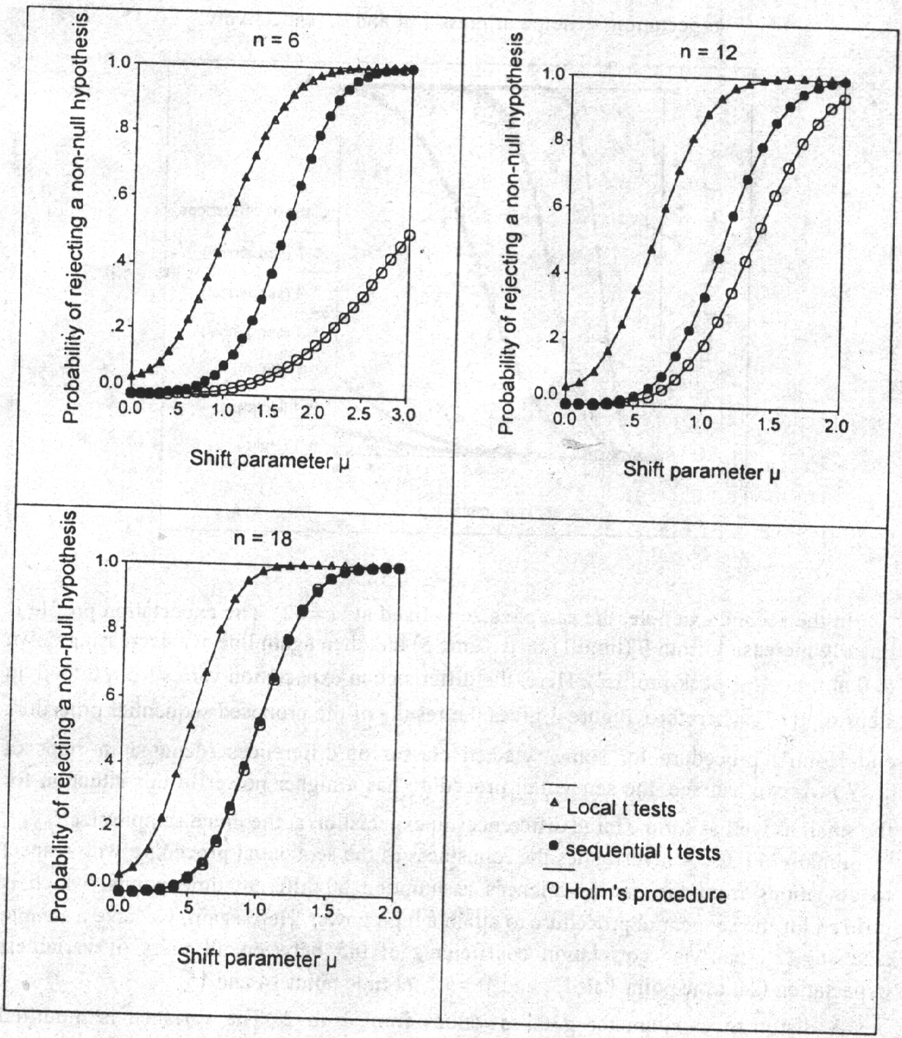
4. SIMULATION EXPERIMENTS

Of course, the data of the above example may be influenced by chance. In order to check the preferences of methods as found in the example, simulation studies have been carried out taking over some basic parameters of the example but not all. In all simulation series, a variable has been measured at 15 time points for each individual. The variance at each time point is 1. A correlation coefficient of 0.5 is assumed for all pairs of measurements from different time points.

In the first series, the sample size varies from 6 through 12 up to 18 (Figure 2). The expectation of the variable of interest has been set to 0 for the first 13 measurements in time and to μ for the other two ones ('two-stage profile'), where μ varies from 0.0 to 3.0 (for $n=6$) or to 2.0 (for $n=12$ and for $n=18$) in steps of 0.1 on the axis of abscissas of the figures. The probability of rejection for a non-null hypothesis at the 0.05 α level is given on the ordinate. In this depicted constellation, 26 out of the 105 pairs of dependent variables have a difference of expectations unequal to zero (as long as $\mu > 0$)). Because of the equal pairwise correlations among all variables, all of these 26 pairs should be detected with the same probability. Therefore, this probability is estimated as number of detected pairs out of these 26 pairs, divided by 26 and averaged over 100,000 simulation runs. The corresponding results are given for the 'local t tests' (i.e., without any adjustment for multiple testing), for the 'sequential t tests' proposed here and for t tests embedded in 'Holm's procedure'.

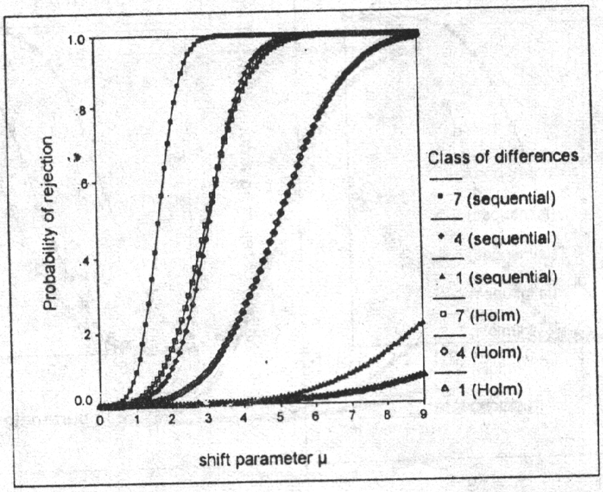
Naturally, the uncorrected t tests have the highest rejection rates, starting at the type I error level for $\mu=0$ and rapidly increasing to values short below 1 for rising values of μ. But the experimentwise type I error rate in the remaining $105-26=79$ pairs of variables (not shown in the figure) is at least 0.8 in all the simulation runs, which is far from any tolerable level (for uncorrelated variables, this rate would even be larger). The experimentwise type I error rate is always maintained with the sequential procedure and with Holm's procedure. As this was clear already from theory, it is not mentioned in the

Figure 2. Results of simulation experiments in the structure with 'two-stage profile' for varying sample size



following examples. For $n = 6$, the proposed sequential procedure has distinctly higher rejection rates for the 26 pairs with 'non-null pairs' than Holm's procedure. This difference is less for $n = 12$. For a sample size of 18, both multiple procedures have nearly identical rejection rates, and for larger n, Holm's procedure would even have better results than the sequential procedure.

Figure 3. Results for simulation series with 'peak profile' and sample size $n = 12$. The rejection rates are given separately for pairs with difference $\frac{1}{7}\mu$, $\frac{4}{7}\mu$, and $\frac{7}{7}\mu$ in expectation, denoted as class 1, 4 and 7, respectively.

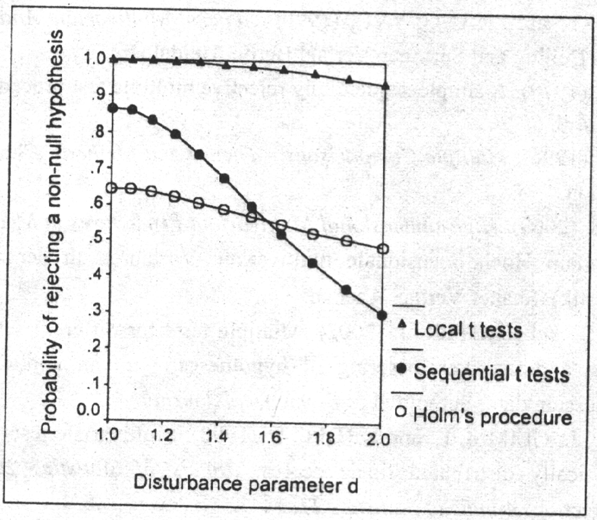


In the second example, the sample size is fixed at $n = 12$. The expectation profile is linearly increasing from 0 (time 1) to μ (time 8) and then again linearly decreasing down to 0 at time 15 ('peak profile'). Here, the difference in expectation varies from 0 to μ in steps of $\mu / 7$. Therefore, Figure 3 gives the results of the proposed sequential procedure and Holm's procedure for some selected classes of differences (denoted in steps of $\mu / 7$). As we can see, the sequential procedure has a higher power in this situation for the small as well as for the large differences in expectation at the given sample size.

Finally, Figure 4 investigates the robustness of the sequential procedure with respect to deviations from the equal variances assumption at different time points, which is utilized for the sequential procedure to attain a high power. Here again, we have a sample size $n = 12$, pairwise correlation coefficients of 0.5 between all pairs of variables, expectation 0 at time point 1 to 13, and $\mu = 1.5$ at time point 14 and 15.

A disturbance parameter d takes values from 1 to 2. The variance is modified according to $\sqrt{\operatorname{Var}(x_i)} = d^{2u-1}$, $i = 0, \ldots, p$ where u is uniformly distributed in [0,1] and is recalculated from one simulation run to the next. Consequently the variance has been set to 1 at the left of the disturbance axis ($d = 1.0$) with increasing variation for increasing magnitude of d. At the right of the disturbance axis ($d = 2.0$), the standard

Figure 4. Results for simulation series with violation of the equal-variances assumption (cf. text)



deviation of a variable can vary from 0.5 to 2. The local *t* tests without adjustment have again the highest rejection probabilities (at the cost of unacceptable experimentwise type I error rates). As already seen in Figure 2, the sequential procedure outperforms Holm's procedure in the undisturbed case. However, with increasing disturbance, the sequential procedure suffers obviously from an ineffective ordering of the pairs of variables in step 1 of the procedure. In the given series, the preference of methods switches at $d \approx 1.55$, corresponding to variances between about 0.4 and 2.4.

We would like to point out that inhomogeneous variances in the pairs of variables may be due both to inhomogeneous variances in the original variable at different times (as in Figure 4), and to an inhomogeneous correlation structure. Both violations of the compound-symmetry assumption have similar effects on the power of the sequential procedure, whereas the type I error is not influenced.

5. ACKNOWLEDGEMENT

We are grateful to Dr. Sylke Kaltenhäuser from the Department of Medicine IV, University of Leipzig for permission to use the rheumatoid arthritis data. We would also like to thank Dr. Serge Provost for his comments to improve the paper.

REFERENCES

1. FANG, K.-T. and ZHANG, Y.-T., (1990). *General Multivariate Analysis*. Science Press Beijing and Springer-Verlag, Berlin, Heidelberg.
2. HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* 6, 65-70.
3. HSU, J.C. (1996). *Multiple Comparisons – Theory and Methods*. Chapman & Hall. London.
4. KROPF, S. (2000). *High-dimensional Multivariate Procedures in Medical Statistics* (German: Hochdimensionale multivariate Verfahren in der medizinischen Statistik). Shaker Verlag, Aachen.
5. KROPF, S. and LÄUTER, J. (2002). Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression data. Submitted to *Biometrical Journal*.
6. LÄUTER, J., GLIMM, E. and KROPF, S. (1998). Multivariate tests based on left-spherically distributed linear scores. *Annals of Statistics* 26, 1972-1988. Correction: *Annals of Statistics* 27, 1441.
7. MAURER, W., HOTHORN, L.A. and LEHMACHER, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays. In: *Biometrie in der chemisch-pharmazeutischen Industrie*. 6. (ed J. Volmar), 3-18. Gustav Fischer Verlag. Stuttgart Jena New York.
8. SCHUSTER, E. (2000). Markov Chain Monte Carlo Methods for Handling Missing Covariates in Longitudinal Mixed Models. In: *Proceedings in Computational Statistics 2000* (eds J. Bethlehem and P. van der Heijden), 439-444. Physica Verlag. Heidelberg.
9. TIMM, N.H. (1980). Multivariate Analysis of Variance of Repeated Measurements. In: Krishnaiah, P.R. (ed.). *Handbook of Statistics. Volume 1 - Analysis of Variance*. North-Holland, Amsterdam, 41-87.

Table 1. Results of the proposed sequential procedure in the rheumatoid arthritis data

Ser. no.	Time points	Mean difference	Std. Error	*P* value in *t* test	Order criterion
1	1 - 6	2.75	0.78	0.0047	171.0
2	1 - 10	3.23	0.54	0.0001	163.9
3	1 - 12	2.98	0.65	0.0008	163.4
4	1 - 9	2.73	0.70	0.0026	154.2
5	1 - 8	2.60	0.71	0.0039	148.3
6	1 - 7	2.81	0.60	0.0006	141.7
7	1 - 13	2.64	0.61	0.0012	133.1
8	1 - 5	2.26	0.72	0.0094	129.4
9	1 - 15	2.32	0.69	0.0062	126.6
10	1 - 14	2.47	0.60	0.0017	120.6
11	1 - 3	1.78	0.78	**0.0441**	119.4
12	1 - 11	2.20	0.68	0.0077	118.5
13	1 - 4	2.05	0.72	0.0154	118.1
14	2 - 10	2.26	0.56	0.0019	102.4
15	2 - 12	2.02	0.63	0.0086	101.4
16	1 - 2	0.97	0.73	**0.2144**	82.3
17	2 - 13	1.68	0.59	0.0163	79.8
18	3 - 12	1.20	0.68	0.1047	78.1
19	3 - 10	1.44	0.62	0.0396	75.3
...	...	...	...	...	...

Figure 1. Alpha-2 globulin values in the sample data (mean ± std dev)

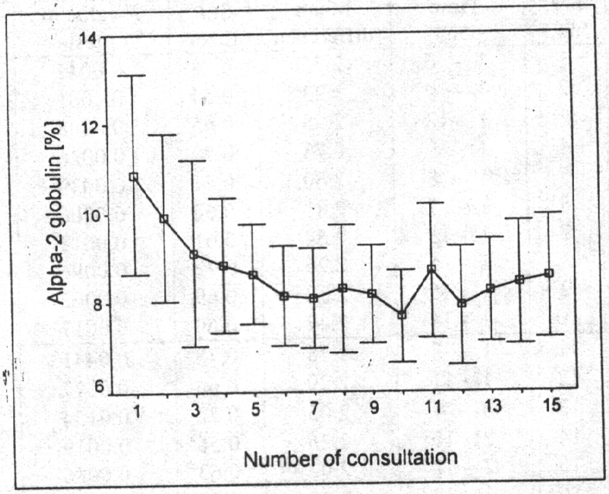


Figure 2. Results of simulation experiments in the structure with 'two-stage profile' for varying sample size

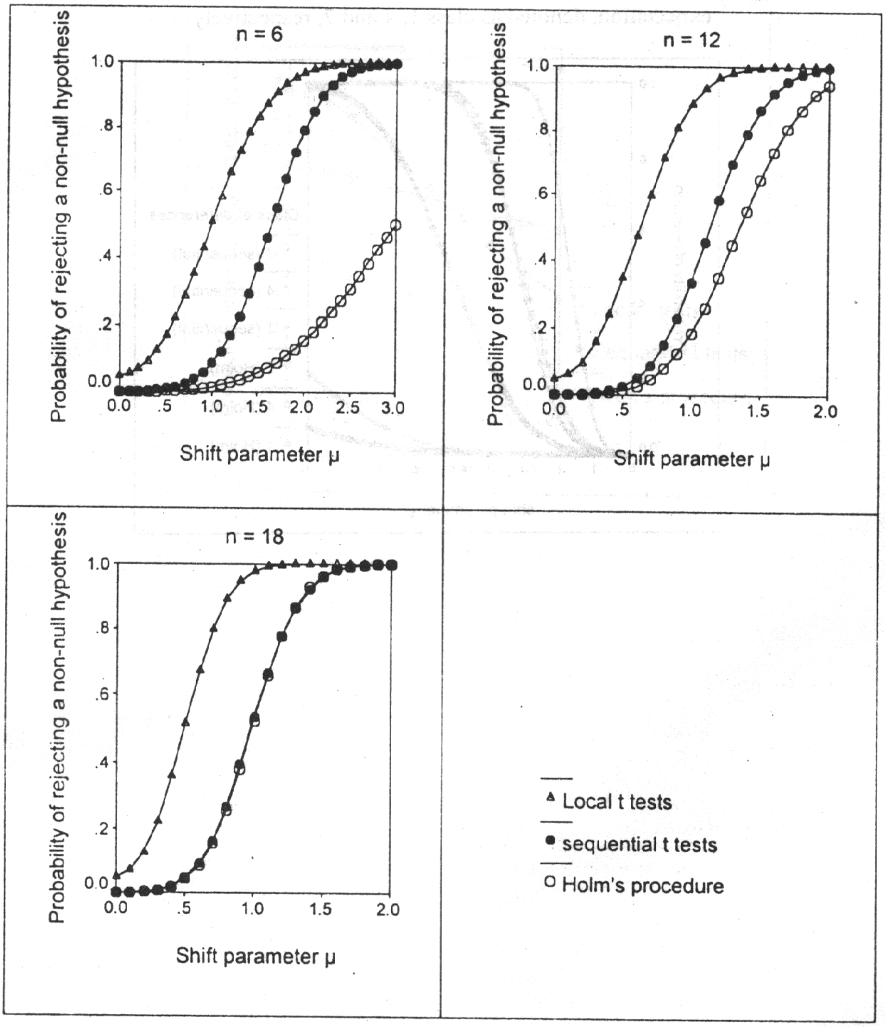


Figure 3. Results for simulation series with 'peak profile' and sample size $n = 12$. The rejection rates are given separately for pairs with difference $\frac{1}{7}\mu$, $\frac{4}{7}\mu$, and $\frac{7}{7}\mu$ in expectation, denoted as class 1, 4 and 7, respectively.

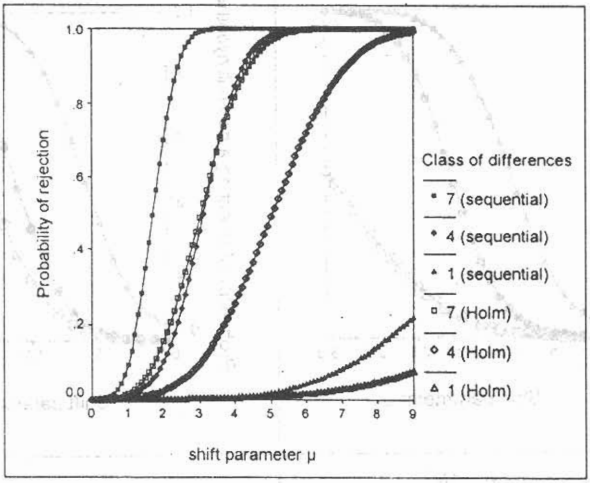
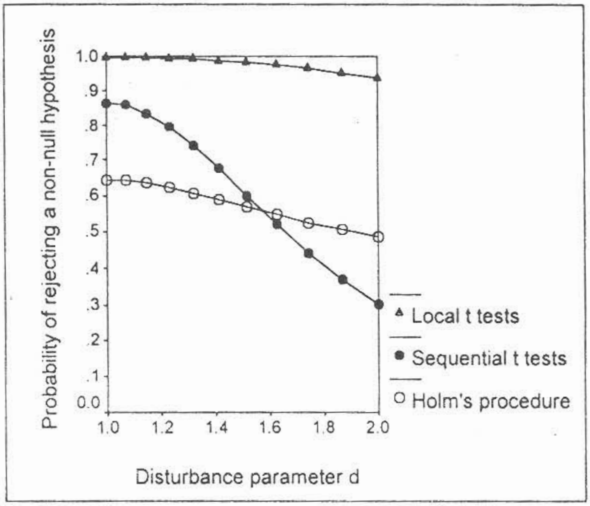


Figure 4. Results for simulation series with violation of the equal-variances assumption (cf. text)


Probability of rejecting a non-null hypothesis
Disturbance parameter d
▲ Local t tests
● Sequential t tests
○ Holm's procedure