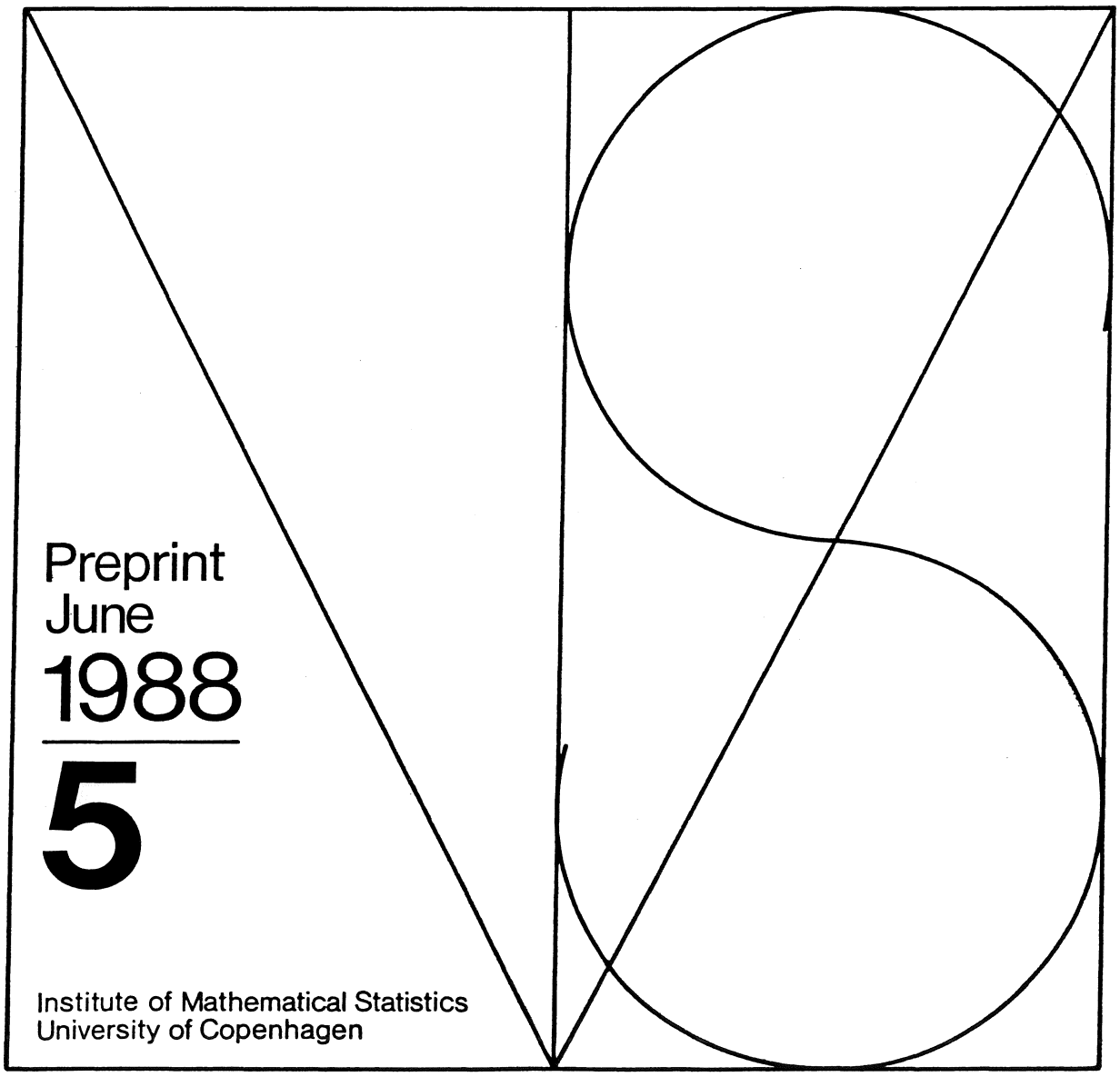
Tue Tjur

Statistical Tables for Personal Computer Users



Preprint
June
1988
5

Institute of Mathematical Statistics
University of Copenhagen

ISSN 0902-8846

Tue Tjur

STATISTICAL TABLES FOR PERSONAL COMPUTER USERS

Preprint 1988 No. 5

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

June 1988

Summary

The main contents of the present preprint is on the floppy disk enclosed. This contains a menu driven, self explanatory program T.EXE replacing a standard collection of statistical tables, and a Turbo Pascal 4 unit DISTR.PAS containing the numerical procedures referenced by this program. The following pages contain a detailed description of the numerical algorithms.

Key words. Statistical distributions; percentiles; statistical tables.

AMS subject classification 62Q05.

Introduction.

A standard collection of statistical tables, including such things as a chi square percentage point table, a similar table for the F-distribution etc., is still among the everyday tools in applied statistics. Most packages for statistical computing have built-in procedures for calculation of the relevant tail probabilities, but experience shows that statisticians spend a lot of time doing computations which are non-standard in the software, making use of their own programs or paper, pencil and desk calculator; for the evaluation of the results coming out of this, the statistical tables are used.

However, there is no reason why this task should not be taken over by a computer also. If you have easy access to an IBM PC or compatible with a 8087 or higher coprocessor, you will probably find the present program useful. Insert the floppy disk in drive A and start the program by A:T , or copy the file A:T.EXE to the hard disk and start the program by a similar command. The program is arranged like an ordinary printed collection of statistical tables, with 'pages' containing the commonly used classical distributions. Further explanations should be unnecessary.

Numerical algorithms.

There is nothing particularly sophisticated about the numerical algorithms implemented in the program. All formulas referred to are well known from the litterature, to be found e.g. in Abromowitz and Stegun (1965). The high accuracy is mainly due to the fact that the Turbo Pascal type 'extended' with 19-20 significant digits is used for all computations. For those who want these algorithms for other purposes, the Turbo Pascal 4 unit containing them is included as the ascii file A:DISTR.PAS. The algorithms can be used by a Turbo Pascal 4 program referencing this unit. Inclusion of the procedures in programs under Turbo Pascal version 3 is straightforward, and only minor corrections are required in order to include one of the procedures in a program written in some other pascal or algol version. The algorithms and

their accuracy is briefly described below. For completeness, the mathematical results referred to are also explained.

The gamma function.
The procedure lnGamma returns the log gamma function at the half argument, i.e.

$$\text{lnGamma}(f) \quad = \quad \ln \Gamma(f/2) \ .$$

Only integer values of f are allowed. For $f \leq 500$, lnGamma(f) is computed by straightforward summation, as

$$\text{lnGamma}(f) \quad = \quad \left[\ln(\tfrac{f-2}{2}) + \ln(\tfrac{f-4}{2}) + \ldots\right] + R$$

where summation is continued as long as the logarithm is defined (i.e. the last term is either $\ln(\frac{1}{2})$ or $\ln(\frac{2}{2})$), and the remainder term R is zero for f even, $\ln(\pi)/2$ for f odd. For $f > 500$, a wellknown expansion (Stirlings formula with two additional terms) is used:

$$\ln \Gamma(y) \quad = \quad \frac{\ln(2\pi)}{2} + (y-1/2)\ln(y) - y + \frac{1}{12y} - \frac{1}{360y^3} + R$$

where the remainder term R , which is ignored, turns out to be less than 10^{-15} for $y = f/2 > 250$. Thus, the dominating part of the error is due to the fact that the result is rounded to 19 significant digits. For $f \leq 10^6$ this means 12 significant digits after the decimal point. Thus, expressions of the form const/$\Gamma(f/2)$, computed by exponentiation of ln(const) - lnGamma(f), can be trusted up to a multiplicative error term of the form (1 ± 10^{-12}).

The gamma distribution.
Define

$$\text{pGamma}(f,y) \quad = \quad \frac{1}{\Gamma(f/2)} \int_y^\infty e^{-x} \, x^{f/2-1} \, dx$$

= P(X≥y) for X gamma distributed with parameter $\lambda = f/2$.

The following identity is easily proved by partial integration:

$$(\Gamma 1) \quad \text{pGamma}(f,y) = \frac{1}{\Gamma(f/2)} y^{f/2-1} e^{-y} + \text{pGamma}(f-2,y).$$

Continuing like this, in each step decreasing the degrees of freedom by 2, we obtain the following result:

Lemma 1. *Define*

$$A_0 = \frac{1}{\Gamma(f/2)} y^{f/2-1} e^{-y}$$

and define A_2 , A_4 , ... *recursively by*

$$A_{2k} = \frac{f-2k}{2y} A_{2k-2}$$

Then

$$\text{pGamma}(f,y) = A_0 + A_2 + A_4 + \ldots + A_{2k} + R$$

where the number of terms is determined by the condition that $f-2k$ *is either* 1 *or* 2, *and the remainder term is*

$$R = \begin{cases} 0 \text{ for } f \text{ even} \\ \text{pGamma}(1,y) \text{ for } f \text{ odd.} \end{cases}$$

This can be used for computation of pGamma(f,y) in situations where f is even, or where f is odd and pGamma(1,y) vanishing. However, we use it only when $y \geq 42$ (which implies $\text{pGamma}(1,y) < 10^{-18}$) and $y \geq f/2$.

For $y < 42$ or $y < f/2$, the following procedure is used. Rewrite ($\Gamma 1$) as follows. Replace f with f+2 , make some rearrangement of terms and add 1 on both sides to obtain

$$(\Gamma 2) \qquad 1\text{-pGamma}(f,y) = \frac{1}{\Gamma(\frac{f+2}{2})}\, y^{f/2}\, e^{-y} + 1\text{-pGamma}(f+2,y).$$

By repeated use of this, in each step increasing the degrees of freedom by 2, we obtain the following result:

Lemma 2. *Define*

$$B_0 = \frac{1}{\Gamma(\frac{f+2}{2})}\, y^{f/2}\, e^{-y}$$

and define B_2 , B_4 , ... *recursively by*

$$B_{2k} = \frac{2y}{f+2k}\, B_{2k-2} \,.$$

Then

$$1 - \text{pGamma}(f,y) = B_0 + B_2 + B_4 + \ldots$$

This follows immediately from the fact that the remainder term 1-pGamma(f+2k,y) tends to 0 as $k \to \infty$.

When $y < f/2$ or $y < 42$, pGamma(f,y) is computed from this formula, with summation continued as long as the remainder (which can easily be approximated by the sum of a geometrically decreasing series) exceeds 10^{-20} . Similarly, when applying the 'downwards' method of lemma 1, summation is stopped when the remainder term is seen to be less than 10^{-20}. More detailed considerations show that the number of terms summed never exceeds 10000 when $f < 10^6$. Thus, a very safe bound on the relative error of a single term is $(1 \pm 10^{-12})(1 \pm 10000 \cdot 10^{-19})$. In this expression, the first factor (which stems from the computation of A_0 or B_0 by the procedure lnGamma described above) is the dominating one. In conclusion, the worst possible absolute error on pGamma(f,y) is $\pm 10^{-12}$. For moderate values of f, the algorithm is

much more accurate.

The beta distribution.

Define

$$\text{pBeta}(f_1,f_2,y) \quad = \quad \frac{1}{B(\lambda_1,\lambda_2)} \int_0^y x^{\lambda_1-1} \quad (1-x)^{\lambda_2-1} \, dx \ ,$$

where $\lambda_1 = f_1/2$, $\lambda_2 = f_2/2$. Thus, $\text{pBeta}(f_1,f_2,\cdot)$ is the cdf. of the beta distribution with parameters (λ_1,λ_2) . The Taylor expansion of $(1-x)^{\lambda_2-1}$ (the binomial series) yields the following expansion of the beta cdf.

$$\text{pBeta}(f_1,f_2,y)$$

$$= \quad \frac{1}{B(\lambda_1,\lambda_2)} \int_0^y x^{\lambda_1-1} \left[\sum_{k=0}^{\infty} \binom{\lambda_2-1}{k} (-x)^k \right] dx$$

$$= \quad \frac{1}{B(\lambda_1,\lambda_2)} \sum_{k=0}^{\infty} \frac{(\lambda_2-1)(\lambda_2-2)\ldots(\lambda_2-k)}{(k+\lambda_1)\ k!} (-1)^k \ y^{k+\lambda_1} \quad .$$

Hence we have

Lemma 3. *Define*

$$C_0 \quad = \quad \frac{y^{\lambda_1}}{\lambda_1 B(\lambda_1,\lambda_2)}$$

$$= \exp\left(\frac{f_1}{2} \log(y) + \text{lnGamma}(f_1+f_2) - \text{lnGamma}(f_1+2) - \text{lnGamma}(f_2)\right)$$

and define C_2, C_4, ... *recursively by*

$$C_{2k} \quad = \quad - \frac{y(f_2-2k)(f_1+2k-2)}{2k(f_1+2k)} C_{2k-2} \ .$$

Then $\text{pBeta}(f_1,f_2,y) \quad = \quad C_0 + C_2 + C_4 + \ldots$.

However, this result is computationally relevant only for

small values of f_1 and f_2. Despite the fact that C_{2k} converges rapidly towards zero as $k \to \infty$ (with the same speed as y^k), the first $f_2/2$ terms of the series are of alternating signs, which may lead to the subtraction of very large and approximately equal numbers. We apply the method of lemma 3 only for $f_1+f_2 < 41$ and $y \leq 0.5$, where all terms of the sum turn out to be less than 10^5. Since the number of terms to be taken into account is small (less than 100, when summation is continued only until the remainder is seen to be less than 10^{-20}), the result is reliable up to an absolute error of appr. $\pm\ 10^{-19+5+2}$ or $\pm\ 10^{-12}$ (the proportional error stemming from the computation of the first term, i.e. from lnGamma, can be ignored here because f_1 and f_2 are small).

For $y > 0.5$, $f_1+f_2 < 41$, the obvious interchange-formula

$$\text{pBeta}(f_1,f_2,y) = 1 - \text{pBeta}(f_2,f_1,1-y)$$

reduces the problem to one of the type already discussed.

For $f_1+f_2 \geq 41$, a quite different method is used. For $y \leq 1/2$, the beta-cdf. can be rewritten as follows (again using the binomial series).

$$\text{pBeta}(f1,f2,y) = \frac{1}{B(\lambda_1,\lambda_2)} \int_0^y x^{\lambda_1-1} (1-x)^{\lambda_2-1}\, dx$$

$$= \frac{y^{\lambda_1-1}}{B(\lambda_1,\lambda_2)} \int_0^y \left(\frac{x}{y}\right)^{\lambda_1-1} \left[(1-y) - (x-y) \right]^{\lambda_2-1} dx$$

$$= \frac{y^{\lambda_1-1}\,(1-y)^{\lambda_2-1}}{B(\lambda_1,\lambda_2)} \int_0^y \left(\frac{x}{y}\right)^{\lambda_1-1} \left(1 + \frac{y-x}{1-y}\right)^{\lambda_2-1} dx$$

$$= \frac{y^{\lambda_1-1}(1-y)^{\lambda_2-1}}{B(\lambda_1,\lambda_2)} \int_0^y (\tfrac{x}{y})^{\lambda_1-1} \sum_{k=0}^{\infty} \left[\begin{matrix} \lambda_2-1 \\ k \end{matrix} \right] (\tfrac{y-x}{1-y})^k \, dx$$

$$= \frac{y^{\lambda_1-1}(1-y)^{\lambda_2-1}}{B(\lambda_1,\lambda_2)} \sum_{k=0}^{\infty} \left[\begin{matrix} \lambda_2-1 \\ k \end{matrix} \right] y \, (\tfrac{y}{1-y})^k \int_0^1 u^{\lambda_1-1} (1-u)^k \, du$$

$$= \sum_{k=0}^{\infty} \left[\begin{matrix} \lambda_2-1 \\ k \end{matrix} \right] \frac{y^{\lambda_1+k}(1-y)^{\lambda_2-k-1} B(\lambda_1,k+1)}{B(\lambda_1,\lambda_2)} \, .$$

(Notice that the condition $y \leq 1/2$ is necessary here, otherwise we might have $|\frac{y-x}{1-y}| > 1$, in which case the binomial series would not converge). Straightforward computation of the terms on recursive form yields the following result.

Lemma 4. *Define*

$$D_0 = \frac{y^{\lambda_1}(1-y)^{\lambda_2-1}}{\lambda_1 B(\lambda_1,\lambda_2)} = 2\frac{y^{f_1/2}(1-y)^{(f_2-2)/2}}{f_1 B(f_1/2,f_2/2)}$$

and define D_2, D_4, ... *recursively by*

$$D_{2k} = \frac{y\,(f_2-2k)}{(1-y)\,(f_1+2k)} D_{2k-2} \, .$$

Then $\text{pBeta}(f_1,f_2,y) = D_0 + D_2 + D_4 + \ldots$.

This algorithm is applied when $f_1 + f_2 \geq 41$ and $y \leq 1/2$. For $y > 1/2$, f_1 and f_2 are interchanged and y is replaced with $1-y$, cfr. the interchange formula on page 6.

Even for $y = 1/2$, the series is convergent, but convergence is very slow if f_1+f_2 is small. This is why we are forced to treat the case $f_1 + f_2 < 41$ seperately.

For $2k > f_2$, the series is alternating with terms of decreasing absolute value. This means that the remainder sum is dominated by the absolute value of the present term. Summation is continued as long as this is $\geq 10^{-25}$. The tail sum - i.e. the alternating part of the series, which is present only for odd values of f_2 - is short (at most 200 terms), and its terms are usually very small. The dominant part of the error stems from the sum from $2k = 0$ to $2[f_2/2]$. In this part of the series, at most 10^6 terms are present, provided that f_1 and f_2 are less than 10^6 (in fact, that many terms will never be summed, because summation is stopped when the remainder sum is seen to be less than 10^{-20}). Thus, ignoring the (proportionally acting) error present in D_0, a very pessimistic estimate of the error is $\pm 10^{-19+6}$ or $\pm 10^{-13}$. It follows that a safe bound on the error is $\pm 3 \cdot 10^{-12}$, based entirely on the multiplicative error coming in via the first term (involving three calls of lnGamma).

In conclusion, the numerical procedures for computation of tail probabilities in the gamma and beta distributions return 11-12 reliable digits after the decimal point (for small degrees of freedom, usually a lot more) when degrees of freedom are kept below 10^6. In the program T.EXE, only 8 are reported, but more are available via transfer to a calculator which is built into the program.

Other distributions.

All distributions available in the program, except the hypergeometric distribution, are handled via wellknown relations to the gamma and beta distributions. These relations will not be given here (they can easily be read off the corresponding procedures on DISTR.PAS, if desired).

The hypergeometric distribution.

The tail probabilities of the hypergeometric distribution are computed by straightforward summation of unnormalized point probabilities, including them all in the denominator and only those of the relevant tail in the numerator. A simple recursive formula for these point probabilities exists. Summation starts somewhere in the middle of the distribution and continues in both directions until the remainder is seen to be less than 10^{-20}. Since at most $2 \cdot 10^6$ terms are summed (under the restrictions on parameter values imposed by the input windows of the program), a safe bound on the error on a tail probability computed in this way is $\pm 10^{-19+7}$ or $\pm 10^{-12}$. This algorithm is part of the main program, not to be found on the unit DISTR.PAS.

Computation of percentiles.

Percentiles are computed by a modified Newton-Raphson method which goes as follows. Suppose we want to find the value x at which a given cdf. F takes a prescribed value $1-p$. At our disposal we have an algorithm which can compute $F(x)$ for any given x, and also the density $f(x) = F'(x)$ and its derivative $f'(x)$ are computable. Let x_k be the result of the k'th iteration. Then x_{k+1} is computed as follows. First, compute a and b such that the exponential curve

$$f_{appr}(x) \quad = \quad \exp(\, a + b(x-x_k) \,)$$

has the same value and first derivative as f at the point x_k. This happens for $a = \log f(x_k)$ and $b = f'(x_k)/f(x_k)$. Define x_{k+1} by the equation

$$\int_{x_k}^{x_{k+1}} f_{appr}(z) \, dz \quad = \quad (1-p) - F(x_k) \quad .$$

(which is easily solved, because f_{appr} is a simple exponential curve). Thus, the idea is to approximate the

density locally by an exponential "tail", and then make the correction to x_k which would make the desired correction to the present value $F(x_k)$ if the approximation was exact. The advantage of this method is that it always works when the density is logarithmically concave, because the approximating exponential curve dominates the density in this case, and so the sequence (x_k) becomes monotone and thereby convergent. The gamma density is log concave for $f \geq 2$, and so are the beta densities when both f_1 and f_2 are ≥ 2. The case $f=1$ for the gamma distribution is handled by transformation to the normal distribution, which has a log concave density. For the beta distribution with either f_1 or f_2 equal to 1, a careful choice of starting value x_0 turns out to be sufficient to ensure that the algorithm converges anyway.

A full discussion of the accuracy of this algorithm, when applied to the beta and gamma distributions, will not be given here. An essential observation is that the dominating source of error is the computation of $F(x)$. This means that the error on x is given by

$$\Delta x = \frac{\Delta p}{f(x)}$$

where Δp is the absolute error on $F(x)$, which is about 10^{-12}, according to the previous sections. But this rough estimate does not take into account that the algorithms applied for the gamma and beta distributions are usually more accurate for points far out in one of the tails. More detailed arguments can be given, and these, together with several empirical tests, comparisons with standard tables etc., are the basis of the decisions made on the number of decimals given by the program. These digits are reliable, except in a few situations where the result is rounded to an integer (the F-distribution with f_2 very small, and the t-distribution with $f = 1$). In these cases, only six digits of the integer result should be trusted. Also, in very extreme cases involving the beta distribution with f_1 or f_2 very large and p very close to 0 or 1, the last digit

reported may be incorrect. In all cases, the number of correct decimals is far beyond what is required in standard applications.

Reference:

Abramowitz, M. and Stegun, I.A. (1965)
Handbook of Mathematical Functions
Dover, New York.

PREPRINTS 1987

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK, TELEPHONE +45 1 35 31 33.

No. 1 Jensen, Søren Tolver and Johansen, Søren: Estimation of Proportional Covariances.

No. 2 Rootzén, Holger: Extremes, Loads, and Strengths.

No. 3 Bertelsen, Aksel: On the Problem of Testing Reality of a Complex Multivariate Normal Distribution.

No. 4 Gill, Richard D. and Johansen, Søren: Product-Integrals and Counting Processes.

No. 5 Leadbetter, M.R. and Rootzén, Holger: Extremal Theory for Stochastic Processes.

No. 6 Tjur, Tue: Block Designs and Electrical Networks.

No. 7 Johansen, Søren: Statistical Analysis of Cointegration Vectors.

No. 8 Bertelsen, Aksel: On the Problem of Testing Reality of a Complex Multivariate Normal Distribution, II.

No. 9 Andersson, S.A. and Perlman, M.D.: Group-Invariant Analogues of Hadamard's Inequality.

No. 10 Hald, Anders: Two Generalizations of the Problem of Points by Bernoulli, de Moivre and Montmort.

No. 11 Andersson, Steen Arne: The Lattice Structure of Orthogonal Linear Models and Orthogonal Variance Component Models.

PREPRINTS 1988

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK, TELEPHONE + 45 1 35 31 33.

No. 1 Jacobsen, Martin: Discrete Exponential Families: Deciding when the Maximum Likelihood Estimator Exists and Is Unique.

No. 2 Johansen, Søren and Juselius, Katarina: Hypothesis Testing for Cointegration Vectors - with an Application to the Demand for Money in Denmark and Finland.

No. 3 Jensen, Søren Tolver, Johansen, Søren and Lauritzen, Steffen L.: An Algorithm for Maximizing a Likelihood Function.

No. 4 Bertelsen, Aksel: On Non-Null Distributions Connected with Testing that a Real Normal Distribution Is Complex.

No. 5 Tjur, Tue: Statistical Tables for Personal Computer Users.