ZASTOSOWANIA MATEMATYKI
V (1960)

M. JÍLEK and O. LÍKAŘ (Prague)

TABLE OF RANDOM SAMPLE SIZES, NEEDED FOR OBTAINING NON-PARAMETRIC TOLERANCE REGIONS

In research and production questions we often have to solve the problem of determining, on the basis of a random sample, a region within which with given probability lies at least the given proportion of population. Various methods for the determination of these so-called tolerance regions are known for the case of normal distribution (see e. g. [6], [10] on tolerance regions for univariate normal distribution). But often the assumption of normality is not fulfilled and for the determination of tolerance regions it is necessary to choose another method. The present contribution relates to one of those methods.

Wilks [16] has found with the help of order statistics the solution of the problem of the determination of a one-dimensional tolerance region, within which with the probability P lies at least the proportion p of the population, assuming that the distribution is continuous. Wilks has found that if a random sample of size n is taken and ordered, and if for the lower limit of the tolerance region (lower tolerance limit) the pth smallest and for the upper limit of the tolerance region (upper tolerance limit) the sth largest value of random sample are chosen, then we have the following relation between r, s, n, p and P:

$$I_{1-p}(m, n-m+1) = P, \tag{1}$$

where $I_x(a, b)$ is Pearson's incomplete Beta function and $m = r+s$ (1).

This fact may be used for the determination of the tolerance region so that for the p, P and n given beforehand we shall determine m with the help of relation (1), or for the p, P and m given beforehand we shall determine n with the help of relation (1); in practice the second case will probably be the most frequent. But we cannot explicitly solve equation (1) for n except for the simplest case where $m = 1$ (the tolerance

(1) Consequently the solution does not depend on the individual values r and s but only on their sum.

region is limited on one side by the smallest or largest value of the random sample), i. e. where equation (1) becomes

$$1-p^n = P, \tag{2}$$

whence

$$n = \frac{\log(1-P)}{\log p}. \tag{3}$$

In other cases it is possible to use for the determination of n the Scheffé-Tukey approximative formula [11]:

$$n \doteq \frac{1}{4}\chi^2_{1-P,2m}\frac{1+p}{1-p}+\frac{1}{2}(m-1), \tag{4}$$

where $\chi^2_{a,f}$ is the $100a$ per cent critical value of the χ^2 distribution with f degrees of freedom.

Wald [15] generalizes Wilks' method for the multidimensional case: Let us consider k random variables $\xi_1, \xi_2, \ldots, \xi_k$, about which we assume only that they have a continuous joint probability density. Let us take a random sample of size n and order it first according to the sample values of the first variable, and for the tolerance limits of the first variable let us choose the r_1th smallest and the s_1th largest observed sample value ($r_1+s_1 \leqslant n$). Let us order $n_1 = n-(r_1+s_1)$ random sample elements for which the first variable lies within the tolerance limits just determined; according to the second random variable and as tolerance limits for that variable let us choose the r_2th smallest and the s_2th largest observed value ($r_2+s_2 \leqslant n_1$). In a similar way let us determine the tolerance limits for the remaining $k-2$ random variables. The tolerance region constructed in this way forms a k-dimensional block. Wald [15] proves that this method of determination of the tolerance region leads to the same result as Wilks' method for the one-dimensional case, i. e. relation (1) is valid also for the multidimensional case, where of course

$$m = \sum_{j=1}^{k}(r_j+s_j).$$

Some authors ([14], [4], [3], [2], [7]) have generalized Wald's method, where again between p, P, m, and n relation (1) is valid, and where it is possible to obtain smaller tolerance regions than by Wald's original method.

Several devices had been constructed for the determination of those non-parametric tolerance regions, of which let us quote especially Murphy's nomograms [9] and Somerville's table [12], which are best suited for the determination of m if p, P, and n are given. Although for practical purposes tables of values n seem to be most needed (one of authors felt this lack when solving a research problem [8]) we do not know whether such tables have been published anywhere. For this reason we have computed a table of random sample sizes n needed for the determination of the non-parametric tolerance regions for given

$$m = 1;\ 2(2)10; \quad p = 0{,}50;\ 0{,}75;\ 0{,}90;\ 0{,}95;\ 0{,}975;\ 0{,}99;\ 0{,}995;\ 0{,}999$$

and

$$P = 0{,}50;\ 0{,}75;\ 0{,}90;\ 0{,}95;\ 0{,}975;\ 0{,}99;\ 0{,}995;\ 0{,}999.$$

The values for $m = 1$ were calculated with the help of nine-figure [1] decimal logarithms. The other values were calculated with the help of six-figure critical values of χ^2 distribution [13] with the exception of the values for $P = 0{,}999$, which were calculated with the help of the Cornish-Fisher approximation [5].

Example. A factory producing iodinated salt wishes to determine in what region lies the resulting amount of KI content in the salt produced in that factory with the nominal dosage of 25 mg KI/kg of salt. The task given by the factory is as follows: to determine the limits within which lies the resulting amount of KI in at least 95 per cent of salt packages ($p = 0{,}95$) with the probability $P = 0{,}99$. For the KI content in salt a normal distribution cannot yet be assumed for different production reasons.

Solution. Let us take as the lower limit the smallest and as the upper limit the largest observed value in a random sample—i. e. $n = 2$ — where the sample size n will be found in the published table: $n = 130$.

A random sample of 130 packages was drawn and the smallest amount of KI was found to be 14,5 mg KI/kg of salt and the largest amount was 32,5 mg KI/kg of salt. These amounts form the limits of the region within which with the probability $P = 0{,}99$ lies at least 95 per cent of the population.

Acknowledgements. This contribution has been prepared in connection with the solution of the research work MPPV 01.13/58. The authors wish to express their thanks to the director of the Central Research Institute of Food Industry Mr. F. Voneš and to the director of the Research Institute of Food Industry Economics Mr. J. Habr for their kind agreement to publish this work.

The authors wish also to express their thanks to Mr. J. Łukaszewicz for drawing their attention to a mistake which found his way into their paper.

Table of random sample sizes
needed for obtaining non-parametric tolerance regions

m	p	P 0,50	0,75	0,90	0,95	0,975	0,99	0,995	0,999
	0,50	1	2	4	5	6	7	8	10
	0,75	3	5	9	11	13	17	19	25
	0,90	7	14	22	29	36	44	51	66
1	0,95	14	28	45	59	72	90	104	135
	0,975	28	55	91	119	146	182	210	273
	0,99	69	138	230	299	368	459	528	688
	0,995	139	277	460	598	736	919	1 058	1 379
	0,999	693	1 386	2 302	2 995	3 688	4 603	5 296	6 905
	0,50	4	5	7	8	9	11	12	15
	0,75	7	10	15	18	20	24	27	33
	0,90	17	27	38	46	54	64	72	89
2	0,95	34	54	77	94	110	130	146	181
	0,975	67	107	155	188	221	263	294	368
	0,99	168	269	388	473	555	662	740	922
	0,995	336	538	777	947	1 113	1 325	1 483	1 847
	0,999	1 679	2 692	3 889	4 742	5 570	6 636	7 427	9 251
	0,50	8	10	12	14	15	17	18	22
	0,75	15	20	25	29	33	37	40	48
	0,90	37	51	65	76	85	97	106	126
4	0,95	74	102	132	153	173	198	216	257
	0,975	147	204	266	308	348	398	436	518
	0,99	367	510	667	773	874	1 001	1 094	1 302
	0,995	735	1 021	1 335	1 549	1 751	2 006	2 192	2 609
	0,999	3 672	5 109	6 679	7 752	8 765	10 042	10 974	13 063
	0,50	12	14	17	19	21	23	24	28
	0,75	23	29	35	40	44	49	53	61
	0,90	57	74	91	103	114	128	137	159
6	0,95	114	148	184	208	231	259	279	324
	0,975	227	296	369	418	464	521	562	653
	0,99	567	742	926	1 049	1 164	1 307	1 411	1 641
	0,995	1 134	1 484	1 853	2 100	2 331	2 618	2 826	3 286
	0,999	5 670	7 422	9 273	10 511	11 666	13 105	14 146	16 452
	0,50	16	19	22	24	26	28	30	33
	0,75	31	38	45	50	54	60	64	73
	0,90	77	96	116	129	141	156	167	190
8	0,95	154	193	234	260	285	316	338	387
	0,975	307	387	469	523	574	636	681	779
	0,99	767	968	1 175	1 312	1 439	1 596	1 709	1 957
	0,995	1 534	1 936	2 352	2 627	2 881	3 196	3 422	3 920
	0,999	7 669	9 684	11 769	13 146	14 419	15 996	17 129	19 622
	0,50	20	23	26	29	31	33	35	39
	0,75	39	47	55	60	65	71	75	84
	0,90	97	118	140	154	167	183	195	220
10	0,95	194	237	282	311	338	371	395	447
	0,975	387	476	566	625	680	747	795	900
	0,99	967	1 190	1 418	1 568	1 705	1 874	1 995	2 259
	0,995	1 934	2 382	2 839	3 138	3 413	3 752	3 995	4 525
	0,999	9 669	11 913	14 203	15 702	17 081	18 779	19 994	22 652

References

[1] П. П. Андреев, (P. P. Andrejev), *Математические таблицы*, Москва 1955.

[2] D. A. S. Fraser, *Nonparametric tolerance regions*, Ann. Math. Stat. 24 (1953), p. 44-55.

[3] — *Sequentially determined statistically equivalent blocks*, Ann. Math. Stat. 22 (1951), p. 372-381.

[4] — and R. Wormleighton, *Nonparametric estimation IV*, Ann. Math. Stat. 22 (1951), p. 294-298.

[5] H. Goldberg and H. Levine, *Approximate formulae for the percentage points and normalization of t and* χ^2, Ann. Math. Stat. 17 (1946), p. 216-225.

[6] M. Jílek and O. Líkař, *Toleranční meze pro normálni rozdělení*, Aplikace matematiky (in preparation).

[7] J. H. B. Kemperman, *Generalized tolerance limits*, Ann. Math. Stat. 27 (1956), p. 180-186.

[8] O. Líkař and R. Reisenauer, *Tolerence pro obsah* KJ *v kuchyňské soli a jejich hodnocení*, Research Work MPPV 02.19/4-1958.

[9] R. B. Murphy, *Non-parametric tolerance limits*, Ann. Math. Stat. 19 (1948), p. 581-589.

[10] F. Proschan, *Confidence and tolerance intervals for the normal distribution*, Journ. Amer. Stat. Assoc. 48 (1953), p. 550-564.

[11] H. Scheffé and J. W. Tukey, *A formula for sample sizes for population tolerance limits*, Ann. Math. Stat. 15 (1944), p. 217.

[12] P. L. Somerville, *Table for obtaining non-parametric tolerance limits*, Ann. Math. Stat. 29 (1958), p. 599-601.

[13] C. M. Thompson, *Table of percentage points of the* χ^2 *distribution*, Biometrika 32 (1941-1942), p. 187-191.

[14] J. W. Tukey, *Non-parametric estimation II. Statistically equivalent blocks and tolerance regions — the continuous case*, Ann. Math. Stat. 18 (1947), p. 529-539.

[15] A. Wald, *An extension of Wilks' method for setting tolerance limits*, Ann. Math. Stat. 14 (1943), p. 45-55.

[16] S. S. Wilks, *Determination of sample sizes for setting tolerance limits*, Ann. Math. Stat. 12 (1941), p. 91-96.

Praca wpłynęła 14. 4. 1959

M. JÍLEK i O. LÍKAŘ (Praga)

TABLICA LICZNOŚCI PRÓBEK LOSOWYCH, POTRZEBNYCH DO OKREŚLENIA OBSZARÓW NIEZALEŻNYCH OD ROZKŁADU

STRESZCZENIE

W artykule niniejszym jest podana tablica liczności próbek losowych, potrzebnych do określenia obszarów tolerancji, niezależnych od rozkładu, dla m, p i P, gdzie m — ilość elementów n-elementowej próbki leżących na zewnątrz obszaru tolerancji, a P — prawdopodobieństwo, że wewnątrz obszaru tolerancji będzie leżała frakcja p elementów populacji generalnej.

М. ЙИЛЕК и О. ЛИКАЖ (Прага)

ТАБЛИЦА ОБЪЁМА СЛУЧАЙНЫХ ВЫБОРОК, НУЖНЫХ ДЛЯ ОПРЕДЕЛЕНИЯ НЕПАРАМЕТРИЧЕСКИХ ТОЛЕРАНТНЫХ ОБЛАСТЕЙ

РЕЗЮМЕ

В настоящей статье приведена таблица объёма случайных выборок, нужных для определения непараметрических толерантных областей, для m, p и P, где m — число элементов случайной выборки объёма n, лежащих вне толерантной области, и P — вероятность того, что внутри толерантной области будет лежать по крайней мере доля p генеральной совокупности.