CZASOPISMO TECHNICZNE M
WYDAWNICTWO POLITECHNIKI KRAKOWSKIEJ

z. 3-M/2008
ISSN 0011-4561
ISSN 1897-6328

RENATA DWORNICKA, JACEK PIETRASZEK*

SIMULATION BASED DETERMINATION OF SAMPLE REQUIRED SIZE FOR ESTIMATION OF RETURNABLE PACKAGE DISTRIBUTION WITH ASSUMED ESTIMATION MAXIMUM ERROR

SYMULACYJNE WYZNACZANIE NIEZBĘDNEJ LICZNOŚCI PRÓBY DLA OSZACOWANIA ROZKŁADU POPULACJI GENERALNEJ OPAKOWAŃ ZWROTNYCH PRZY ZADANYM MAKSYMALNYM BŁĘDZIE OSZACOWANIA

Abstract

Kompania Piwowarska SA has a population of 140 million of returnable beer bottles. These bottles are classified and after positive evaluation are washed and refilled. It is necessary to perform periodic estimation of class distribution in a general population of bottles in view of future new bottles order. The authors proposed computer simulation basing on real information of distribution taken from Kompania's archives. It was very optimistic because it allows reduction of sample size about nine times in comparison with the first estimation of Kolmogorov's distribution.

Keywords: *DoE*, *experimental design*, *industrial statistics*, *statistical estimation*

Streszczenie

Kompania Piwowarska SA wykorzystuje w dystrybucji populację 140 milionów butelek zwrotnych, których stan jest oceniany i po pozytywnej klasyfikacji są myte i ponownie napełniane. Niezbędne jest dokonywanie okresowych ocen liczebności poszczególnych klas butelek, aby zamówić dostawę nowych z odpowiednim wyprzedzeniem. Autorzy zaproponowali symulację komputerową opartą na rzeczywistych danych uzyskanych z archiwów Kompanii. Uzyskane wyniki pozwoliły na niemal dziewięcioletnią redukcję rozmiaru próby w porównaniu z teoretycznym oszacowaniem z rozkładu Kołmogorowa.

Słowa kluczowe: *DoE*, *planowanie doświadczeń*, *statystyka przemysłowa*, *estymacja*

*Renata Dwornicka, MSc, Jacek Pietraszek, PhD, Institute of Applied Informatics, Cracow University of Technology.

1. Introduction

Distribution network of Kompania Piwowarska consists of a chain of warehouses spread throughout Poland divided into five regions: Northern, Eastern, Southern, Western and Central. Kompania has a population of 140 million of returnable beer bottles on the market. These bottles are classified and after positive evaluation are washed and refilled.

The wear of bottle is determined by reference to patterns. Evaluation is quantified into seven ranges with mark 0 – an ideal bottle and mark 6 – the worst. It is necessary to perform periodic estimation of the mentioned class distribution in a general population of bottles in view of future new bottles order. Till now it has been done by sampling of 15 thousand of bottles and comprehensive classification of all bottles. It was assumed that distribution of a general population is almost equal to sample distribution but without any error estimation.

Following the above a question arises: is it possible to reliably estimate an error of general population distribution determination? And next: is it possible to estimate the required size of a sample as a function of an assumed maximum error value?

2. Theoretical approach

An entry point for a theoretical solution is an interval estimation of unknown continuous cumulative distribution $F(x)$ of population distribution, which estimation is calculated based on λ Kolmogorov statistic. The λ statistic is defined by formula [1]

$$\lambda = D_n \sqrt{n} \tag{1}$$

while D_n statistic by formula

$$D_n = \sup_x \left| F_n(x) - F(x) \right| \tag{2}$$

where:

n – sample size,
$F_n(x)$ – empirical cumulative distribution,
$F(x)$ – unknown cumulative distribution of general population.

Kolmogorov showed that asymptotic ($n \to \infty$) distribution of statistic λ is not dependent on theoretical cumulative distribution $F(x)$. Statistical tables contain values of D_n [4] which is dependent on sample size n and significance level α, or values of asymptotic statistic λ [2] which is only dependent on significance level α.

In the context of the task in question, the formula [2] should be treated as pessimistically high estimation of maximum error for cumulative distribution of bottle categories. For a practical approach, what is more important is error estimation for particular categories. If a maximum error of category estimation is defined as a difference between empirical sample distribution and distribution of a general population

$$\varepsilon = \sup_x \left| \rho_n(k) - \rho(k) \right| \tag{3}$$

where:

k – number of bottle category,
$\rho_n(k)$ – empirical function of sample distribution,
$\rho(k)$ – function of general population distribution,

then the error is connected with cumulative distribution error by formula

$$\varepsilon = 2D_n \tag{4}$$

In a case of a small sample, formula (4) allows direct estimation of maximum error of distribution (precision of a bar in histogram) calculated from sample size n and assumed significance level α. The reverse task, determining sample size n for the assumed error value and significance level α is not trivial because sample size n in the formula (4) is confounded inside different factors. In this case, tables [4] are very useful due to the shown values of D_n statistic as a function of sample size n and significance level α.

In the case of a large sample the task is easier because D_n statistic may be eliminated from formula (4) by substitution derived from formula (1)

$$\varepsilon = 2\frac{\lambda}{\sqrt{n}} \tag{5}$$

By solving this formula, the required sample size n is gathered directly

$$n = \left[\left(\frac{2\lambda}{\varepsilon}\right)^2\right] \tag{6}$$

The values of λ statistic are published for typical significance levels in statistical tables e.g. [2]. Table 1 below contains error estimates for distribution function (bottle categories) for some large sample sizes and typical significance levels 0,10; 0,05; 0,01. The values of Tab. 1 were calculated from formula (5).

Table 1

Error estimates for bottle categories

Sample size n	Significance level α		
	$\alpha = 0{,}10$	$\alpha = 0{,}05$	$\alpha = 0{,}01$
100	0,2460	0,2720	0,3260
200	0,1739	0,1923	0,2305
300	0,1420	0,1570	0,1882
400	0,1230	0,1360	0,1630
500	0,1100	0,1216	0,1458
600	0,1004	0,1110	0,1331
700	0,0930	0,1028	0,1232
800	0,0870	0,0962	0,1153
900	0,0820	0,0907	0,1087
1000	0,0778	0,0860	0,1031
2000	0,0550	0,0608	0,0729
3000	0,0449	0,0497	0,0595
4000	0,0389	0,0430	0,0515
5000	0,0348	0,0385	0,0461
6000	0,0318	0,0351	0,0421
7000	0,0294	0,0325	0,0390
8000	0,0275	0,0304	0,0364
9000	0,0259	0,0287	0,0344
10000	0,0246	0,0272	0,0326
15000	0,0201	0,0222	0,0266
20000	0,0174	0,0192	0,0231
30000	0,0142	0,0157	0,0188

The results from Tab. 1 should be treated as: the probability of an error higher than table value equals the significance level for the sample of assumed size, e.g. a probability of an error higher than 2,01% (column position) equals 10% (significance level, column title) for the sample of size 15 000 (row title).

It should be noted that these estimates are pessimistic because of the general approach of Kolmogorov thesis.

3. Simulation approach

The last conclusion from section 3 is rather pessimistic and the results presented in Tab. 1 suggest rather large uncertainty of general population distribution. Because of this, the authors propose that simulation experiment may give sharper estimates because of taking into account specific general population data.

The following scheme of simulation was assumed:

a) general population is generated by reverse distribution method [3] based on distribution estimates gathered from Kompania Piwowarska archives,
b) 10 000 samples of assumed size are randomly taken from the general population; the value 10 000 is assumed as the number of samples because it allows easy calculation of quantiles 90%, 95% and 99%,
c) sample distributions are compared with the general population distribution (assumed from archives data) and errors are calculated: maximum in category, maximum cumulative and RMS (root mean square),
d) the collected results are sorted and error levels for quantiles 90%, 95% and 99% are searched.

The till-now estimates of general population gathered from Kompania Piwowarska archives are presented in Tab. 2.

Table 2

The till-now estimates of bottle categories in general population

Category number	Percent of share
0	0,10
1	0,13
2	0,13
3	0,20
4	0,26
5	0,10
6	0,08

The small simulation program (C# language) was prepared that generated a general population and series of samples were taken to error estimate. The result was a large set of raw data. The data set was aggregated and performed to calculate errors' values for typical significance levels (Tab. 3).

Table 3

Errors estimations of bottle categories from simulation

Sample size	Maximum error ε_{max}			Cumulative error ε_{kum}			RMS error ε_{RMS}		
	$\alpha = 0{,}10$	$\alpha = 0{,}05$	$\alpha = 0{,}01$	$\alpha = 0{,}10$	$\alpha = 0{,}05$	$\alpha = 0{,}01$	$\alpha = 0{,}10$	$\alpha = 0{,}05$	$\alpha = 0{,}01$
10	0,2572	0,3571	0,3572	0,3143	0,3572	0,4286	0,1400	0,1591	0,1761
20	0,2071	0,2072	0,2572	0,2215	0,2642	0,3144	0,1015	0,1116	0,1294
30	0,1571	0,1572	0,2238	0,1810	0,2048	0,2619	0,0849	0,0921	0,1050
40	0,1322	0,1571	0,1822	0,1571	0,1786	0,2214	0,0728	0,0798	0,0923
50	0,1172	0,1371	0,1571	0,1429	0,1544	0,1944	0,0654	0,0713	0,0824
60	0,1072	0,1238	0,1429	0,1286	0,1476	0,1809	0,0597	0,0654	0,0755
70	0,1000	0,1143	0,1428	0,1143	0,1428	0,1714	0,0556	0,0606	0,0704
80	0,0946	0,1071	0,1321	0,1143	0,1286	0,1536	0,0521	0,0566	0,0654
90	0,0905	0,1016	0,1238	0,1063	0,1190	0,1508	0,0491	0,0533	0,0624
100	0,0871	0,0929	0,1129	0,1014	0,1143	0,1414	0,0465	0,0506	0,0587
200	0,0621	0,0671	0,0821	0,0714	0,0808	0,1007	0,0330	0,0359	0,0420
300	0,0495	0,0538	0,0662	0,0581	0,0657	0,0814	0,0270	0,0293	0,0335
400	0,0422	0,0471	0,0547	0,0493	0,0564	0,0690	0,0232	0,0252	0,0289
500	0,0388	0,0428	0,0508	0,0457	0,0515	0,0631	0,0210	0,0227	0,0261
600	0,0345	0,0379	0,0455	0,0407	0,0464	0,0581	0,0190	0,0206	0,0237
700	0,0315	0,0357	0,0429	0,0385	0,0429	0,0542	0,0176	0,0193	0,0222
800	0,0303	0,0334	0,0397	0,0357	0,0406	0,0506	0,0166	0,0181	0,0210
900	0,0283	0,0316	0,0373	0,0337	0,0376	0,0470	0,0155	0,0169	0,0194
1000	0,0271	0,0299	0,0351	0,0317	0,0363	0,0444	0,0148	0,0162	0,0186
2000	0,0188	0,0208	0,0247	0,0223	0,0253	0,0314	0,0103	0,0112	0,0131
3000	0,0155	0,0172	0,0202	0,0183	0,0206	0,0256	0,0086	0,0093	0,0107
4000	0,0134	0,0149	0,0176	0,0158	0,0180	0,0224	0,0074	0,0080	0,0094
5000	0,0120	0,0132	0,0158	0,0142	0,0162	0,0202	0,0066	0,0072	0,0083
6000	0,0110	0,0121	0,0144	0,0129	0,0146	0,0181	0,0060	0,0065	0,0075
7000	0,0102	0,0113	0,0134	0,0121	0,0136	0,0169	0,0056	0,0061	0,0070
8000	0,0094	0,0104	0,0124	0,0113	0,0127	0,0156	0,0052	0,0056	0,0065
9000	0,0090	0,0099	0,0119	0,0105	0,0119	0,0146	0,0049	0,0054	0,0062
10000	0,0084	0,0093	0,0113	0,0099	0,0112	0,0143	0,0046	0,0050	0,0059
15000	0,0069	0,0077	0,0091	0,0083	0,0094	0,0115	0,0038	0,0042	0,0048
20000	0,0061	0,0067	0,0080	0,0071	0,0081	0,0100	0,0033	0,0036	0,0042
30000	0,0049	0,0054	0,0064	0,0058	0,0066	0,0082	0,0027	0,0030	0,0034

4. Conclusions

For the task of determining the distribution error two approaches were prepared: the theoretical based on Kolmogorov λ statistic and simulation based on Kompania Piwowarska archives data. Comparing the theoretical and simulation results (Tab. 4) it is easy to point out the simulation estimates are far more optimistic. It is derived from the fact of taking into consideration specific information about general population distribution and lack of necessity of general assumptions being fundamental of Kolmogorov λ statistic.

Table 4

Comparison of maximum error estimates (simulation and theoretical approach)

Sample size	$\alpha = 0{,}10$		$\alpha = 0{,}05$		$\alpha = 0{,}01$	
	simulation	theory	simulation	theory	simulation	theory
100	0,0871	0,2460	0,0929	0,2720	0,1129	0,3260
200	0,0621	0,1739	0,0671	0,1923	0,0821	0,2305
300	0,0495	0,1420	0,0538	0,1570	0,0662	0,1882
400	0,0422	0,1230	0,0471	0,1360	0,0547	0,1630
500	0,0388	0,1100	0,0428	0,1216	0,0508	0,1458
600	0,0345	0,1004	0,0379	0,1110	0,0455	0,1331
700	0,0315	0,0930	0,0357	0,1028	0,0429	0,1232
800	0,0303	0,0870	0,0334	0,0962	0,0397	0,1153
900	0,0283	0,0820	0,0316	0,0907	0,0373	0,1087
1000	0,0271	0,0778	0,0299	0,0860	0,0351	0,1031
2000	0,0188	0,0550	0,0208	0,0608	0,0247	0,0729
3000	0,0155	0,0449	0,0172	0,0497	0,0202	0,0595
4000	0,0134	0,0389	0,0149	0,0430	0,0176	0,0515
5000	0,0120	0,0348	0,0132	0,0385	0,0158	0,0461
6000	0,0110	0,0318	0,0121	0,0351	0,0144	0,0421
7000	0,0102	0,0294	0,0113	0,0325	0,0134	0,0390
8000	0,0094	0,0275	0,0104	0,0304	0,0124	0,0364
9000	0,0090	0,0259	0,0099	0,0287	0,0119	0,0344
10000	0,0084	0,0246	0,0093	0,0272	0,0113	0,0326
15000	0,0069	0,0201	0,0077	0,0222	0,0091	0,0266
20000	0,0061	0,0174	0,0067	0,0192	0,0080	0,0231
30000	0,0049	0,0142	0,0054	0,0157	0,0064	0,0188

In view of the above comparison, the simulation estimates of bottle categories error is about three times smaller than the pessimistic estimates calculated from the theoretical approach. It means that the sample size may be **nine** times smaller (inversely proportional to error square) in comparison to the theoretical estimates at the same uncertainty level, e.g. 1667 bottles in a sample is sufficient in comparison with 15000 formerly.

References

[1] Greń J., *Statystyka matematyczna*, wyd. 1, PWN, Warszawa 1987, 471-473.
[2] Pluciński A., Pluciński E., *Rachunek prawdopodobieństwa. Statystyka matematyczna. Procesy stochastyczne*, WNT, Warszawa 2000.
[3] Zieliński R., *Generatory liczb losowych. Programowanie i testowanie na maszynach cyfrowych*, wyd. 1, WNT, Warszawa 1972, 56-58.
[4] Zieliński R., *Tablice statystyczne*, wyd. 1, PWN, Warszawa 1972, 284.