BERICHTE

DER

ARBEITSGRUPPE TECHNOMATHEMATIK

FORSCHUNG - AUSBILDUNG - WEITERBILDUNG

BERICHT NR. 3

THE TRIPPSTADT-PROBLEM

W. KRÜGER

UNIVERSITÄT KAISERSLAUTERN
FACHBEREICH MATHEMATIK
ERWIN-SCHRÖDINGER-STR. 48
6750 KAISERSLAUTERN

FEBRUAR 1984

THE TRIPPSTADT-PROBLEM

W. Krüger, Kaiserslautern

Close to Kaiserslautern is the town of Trippstadt, which, together with five other small towns forms a local administrative unit ("Verbandsgemeinde") called "Kaiserslautern-Süd". Trippstadt has its own beautiful public swimming pool, which causes problems though; the cost for the upkeep of the pool is higher than the income and thus has to be divided among the towns belonging to the "Verbandsgemeinde". Because of this problem the administration wanted to find out which fraction of the total number of pool visitors came from the different towns. They planned to ask each pool guest where he came from. They did this for only three days though because the waiting lines at the cashiers became unbearably long and they could see that because of this the total number of guests would decrease. (They would lose patience and not come at all.) Then they wondered how to find a better method to get the same data and that was when I was asked to help with the solution of the problem.

From May 16th to September 12th 1982 (which was the swimming season) I asked about 4000 guests (approximately 5% of the total) and thus approximated the distribution. Before the actual realization of the statistics, I had to find out what a reasonable mathematical model would look like. To do that we had to make sure which data the cashiers could gather without too much effort. It turned out that the swimmingmaster recorded information about the weather and the total number of guests every day. Both pieces of information had been recorded during the past two years and were useful as pre-information for my statistics.

The first assumption for building my model was that the distribution of guests from the various towns is approximately constant over any given fixed day; for different days there is the possibility that it is different. For example, the approxi-

mate proportion of guests from Trippstadt on a rainy day is higher than on sunny days. When one asks on a specific day a random sample of n guests where they come from, then the probability that n_1 of these came from a specific town, is given by

$$P(X = n_1) = \binom{n}{n_1} p^{n_1} (1-p)^{n-n_1},$$

where $p = N/N_G$, N_G is the total number of guests on that day and N is the number of guests coming from the town concerned. The number N_G is registered at the cashiers, so it is enough to have a good estimate of p to be able to calculate N. To make things simpler, we consider the portion of guests that came from one town, that is e.g. Trippstadt (we can handle the other towns analogously).

One possibility for solving the problem would be to question a proportion of the guests every day. But this method would have been too expensive and time-consuming. So we tried to condense the inquiry to as few days as possible. We shall see what this means for the model. Assume that we ask on a randomly chosen unknown day n, randomly chosen guests where they come from, then the probability that k guests come from Trippstadt is given by

$$P(X = k) = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} d\mu(t),$$

where μ is a discrete probability measure on [0,1]. That means μ describes the random choice of a day and p is the proportion of guests from Trippstadt on that day. The rest of the problem is to determine the number of components of μ, which means the number of different "day types" which have different guest distributions and to characterize these distinctive features.

The first step was to analyze the information about the total number of guest per day from the two preceding years. Five different factors for different "day types" were found. All were connected with weather and whether or not it was a weekday, weekend or holiday. Let us look at this model a little closer. Assume that for a certain fixed "day type" all persons from a specific town "ℓ" decide with the same small probability q_ℓ independently of each other whether or not they should visit

the public pool; then, for large populations N_ℓ of respective towns, the number ξ_ℓ of people from town "ℓ" that do visit the pool is almost a Poisson distributed random variable:

$$P(\xi_\ell = r) = \frac{(\beta_\ell)^r}{r!} e^{-\beta_\ell} \quad , \qquad r = 0,1,...$$

where $\beta_\ell = q_\ell \cdot N_\ell$. The total number of guests on this day is given by $\xi = \sum_\ell \xi_\ell$ i.e. also a Poisson distributed random variable:

$$P(\xi = s) = \frac{\beta^s}{s!} e^{-\beta} \quad , \qquad s = 0,1,... \quad , \quad \beta = \sum_\ell \beta_\ell .$$

Different "day types" produce different values q_ℓ for the respective "visitor motivation", so different β_ℓ for the distribution of the ξ_ℓ and with that eventually different β for the distribution of ξ. This shows that the number of different Poisson distributions from which the distribution of the total number of guests per day follows will give some information on the number of components of μ.
I also talked to the swimming-master, trying to identify further characteristics of μ. In my opinion, mathematical abstract analysis and everyday experience are both important and should be combined for all statistic examinations.

After this examination 12 different "day types" were found and had to be analyzed by the inquiry.
The different characteristics of the "day types" are found in the following chart. Each "day type" has a given number, 3 numbers make up the code for the different "day types".

weekend or holiday	0
weekday	1

school vacation	0
school time	1

sunny	0
overcast	1
rain	2

The execution of the representative inquiry was now to ask the randomly chosen guests on each of the 12 "day types", from which town he came from. The results of this inquiry are realizations of binomially distributed random variables. For the further analysis of the problem we use the following symbols (notation):

n_i : number of guests on "day type" "i" questioned ($i = 1,\ldots,12$).

N : number of all questioned guests ($N = n_1+\ldots+n_{12}$).

p_i : proportion of guests asked that were from Trippstadt on "day type" "i", $p_i \in [0,1]$, $i = 1,\ldots,12$.

λ_i : relative visitors frequency on "day type" "i", $\lambda_i \in [0,1]$, $i = 1,\ldots,12$.
This number is easy to find at the end of the season from the cashiers recorded lists.

x_i : number of guests on "day type" "i" coming from Trippstadt ($i = 1,\ldots,12$).

p : total proportion of people during the season '82 that were from Trippstadt, e.g. $p = \sum_{i=1}^{12} \lambda_i p_i$.

Furthermore, let X_i be $\beta(n_i,p_i)$ - distributed random variables, that is

$$P(X_i = k) = \binom{n_i}{k} p_i^k (1-p_i)^{n_i-k}, \quad k = 0,\ldots,n, \quad i = 1,\ldots,12.$$

The result $x_1,\ldots,x_{12}$ from our inquiry is a realization of $X_1,\ldots,X_{12}$. As the inquiries on the different "day types" are independent from each other, $X_1,\ldots,X_{12}$ are stochastically independent. $P_{(p_1,\ldots,p_{12})}$ is the notation for the common distribution of $X_1,\ldots,X_{12}$. The visitor proportion p of those from Trippstadt can be easily estimated by

$$T = \sum_{i=1}^{12} \frac{\lambda_i}{n_i} X_i.$$

This means that the observation $x_1,\ldots,x_{12}$ gives

$$\hat{p} = \sum_{i=1}^{12} \frac{\lambda_i}{n_i} x_i.$$

Here two more questions have to be clarified:

1. Since the administration only wanted to spend a limited amount for the inquiry, the number N was fixed. How should N be divided among the n_i in a reasonable way, such that $N = \sum_{i=1}^{12} n_i$.

2. How can we construct a suitable confidence region for the estimator T for a given level (for example 95%)?

Both questions are related and will be answered together. Question 2 is the construction of $\varepsilon_1, \varepsilon_2 > 0$ so that

$$P_{(p_1,\ldots,p_{12})}\left(\sum_{i=1}^{12}\lambda_i p_i - \varepsilon_1 \leq T \leq \sum_{i=1}^{12}\lambda_i p_i + \varepsilon_2\right) \geq 1-\alpha \quad \text{for all } p_1,\ldots,p_{12} \in [0,1],$$

where $\alpha \in (0,1)$ and $1-\alpha$ is the given level. One possibility for determining such a confidence interval could be to construct confidence intervals for the different "day types" and to reduce the determination of ε_1 and ε_2 to this. Because of

$$\varepsilon_j = \sum_{i=1}^{12}\lambda_i \cdot \varepsilon_j, \quad j = 1,2,\ldots,12$$

this gives

$$\begin{aligned}
&P_{(p_1,\ldots,p_{12})}\left(\sum_{i=1}^{12}\lambda_i p_i - \varepsilon_1 \leq T \leq \sum_{i=1}^{12}\lambda_i p_i + \varepsilon_2\right) \\
&= P_{(p_1,\ldots,p_{12})}\left(\sum_{i=1}^{12}\lambda_i (p_i - \varepsilon_1) \leq \sum_{i=1}^{12}\lambda_i \frac{X_i}{n_i} \leq \sum_{i=1}^{12}\lambda_i (p_i + \varepsilon_2)\right) \\
&\geq \prod_{i=1}^{12} P_{p_i}\left(p_i - \varepsilon_1 \leq \frac{X_i}{n_i} \leq p_i + \varepsilon_2\right) \\
&= \prod_{i=1}^{12} (1-\alpha) \geq 1 - 12\alpha.
\end{aligned}$$

This result is, however, very unsatisfactory, and so we use another method. Compare [1] and [4].

Definition

Let $((\Omega, \mathfrak{B}), (P_\theta)_{\theta \in \Theta})$ be a decision space and

$$C : \Omega \to \mathfrak{P}(\Theta) = \{A; A \subset \Theta\}$$

so that for $\theta \in \Theta$ $\{\omega \in \Omega; \theta \in C(\omega)\} \in \mathfrak{B}$. Then C is called a confidence region to the level $1-\alpha$, $\alpha \in (0,1)$, if

$$P_\theta(\{\omega \in \Omega; \theta \in C(\omega)\}) \geq 1-\alpha$$

for all $\theta \in \Theta$.

An easy construction of such a confidence region is the following:

For all $\theta \in \Theta$ choose $A(\theta) \in \mathfrak{B}$ so that

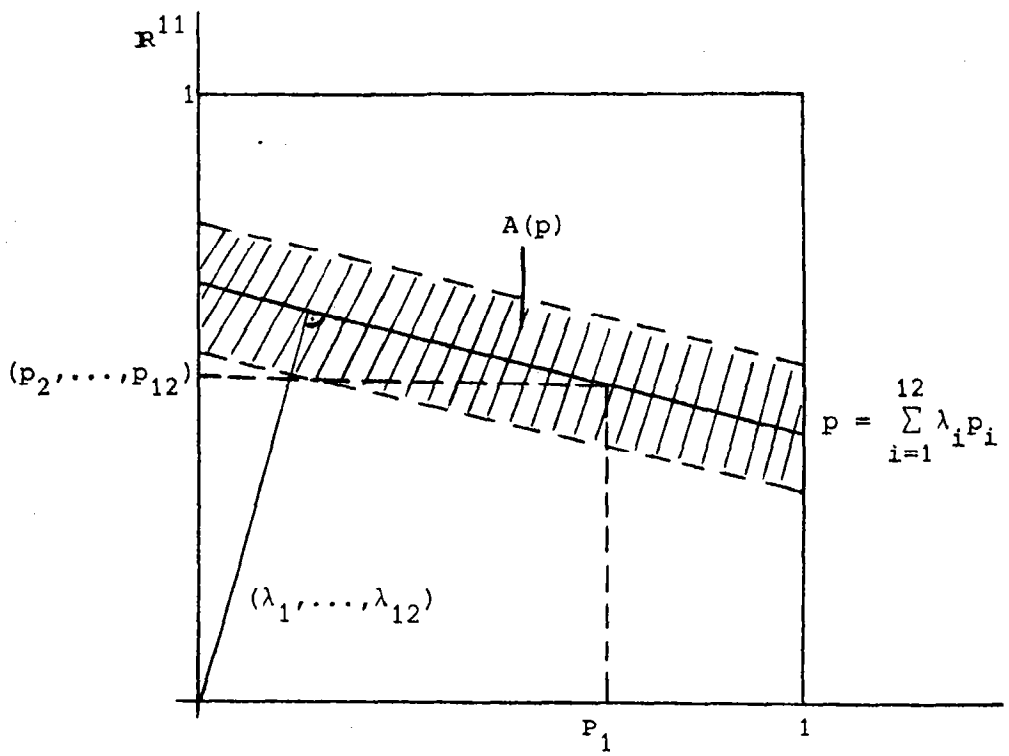
$$P_\theta(A(\theta)) \geq 1-\alpha$$

(*) Then it follows that

$$C(\omega) := \{\theta \in \Theta; \omega \in A(\theta)\}$$

is a confidence region to the level $1-\alpha$.

We use this construction principle for the estimator T. The following diagram shows the construction.



The idea here is that for $p_1,\ldots,p_{12} \in [0,1]$ with $p = \sum_{i=1}^{12} \lambda_i p_i$ the same quantity A(p) is used. So, for the concrete observation $x_1,\ldots,x_{12}$ and the related estimated value $\hat{p} = \sum_{i=1}^{12} \lambda_i \frac{x_i}{n_i}$ we get the confidence interval $[p_o(\hat{p}), p_1(\hat{p})]$ with

$$p_o(\hat{p}) = \sup\{p \in [0,1];\ P_{(p_1,\ldots,p_{12})}(T \geq \hat{p}) \leq \tfrac{\alpha}{2} \text{ for all } p_1,\ldots,p_{12} \text{ with } p = \sum_{i=1}^{12} \lambda_i p_i\}$$

$$p_1(\hat{p}) = \inf\{p \in [0,1];\ P_{(p_1,\ldots,p_{12})}(T \leq \hat{p}) \leq \tfrac{\alpha}{2} \text{ for all } p_1,\ldots,p_{12} \text{ with } p = \sum_{i=1}^{12} \lambda_i p_i\}.$$

The values $p_p(\hat{p})$ and $p_1(\hat{p})$ are easy to calculate numerically; one can also estimate them by normal approximation and then use the result to answer the first question. So we get

$$P_{(p_1,\ldots,p_{12})}(T \geq \hat{p}) \approx 1 - \phi\left(\frac{\hat{p} - \sum_{i=1}^{12} \lambda_i p_i}{\left(\sum_{i=1}^{12} \frac{\lambda_i^2}{n_i} p_i(1-p_i)\right)^{\frac{1}{2}}}\right)$$

and

$$P_{(p_1,\ldots,p_{12})}(T \leq \hat{p}) \approx \phi\left(\frac{\hat{p} - \sum_{i=1}^{12} \lambda_i p_i}{\left(\sum_{i=1}^{12} \frac{\lambda_i^2}{n_i} p_i(1-p_i)\right)^{\frac{1}{2}}}\right)$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\, dt$ is the distribution function of the standard normal distribution. So we see that an optimal partition of N into n_i such that $N = \sum_{i=1}^{12} n_i$ is equivalent to the minimization of the variance of the estimator T, because

$$\mathrm{Var}(T) = \sum_{i=1}^{12} \frac{\lambda_i^2}{n_i} t_i(1-t_i).$$

But this is also equivalent to the minimization of the confidence interval. So it is easy to see that the partition

$$n_i = \lambda_i \cdot N \quad \text{for} \quad i = 1,\ldots,12$$

is optimal. Let $n_i = \lambda_i \cdot N$ for $i = 1,\ldots,12$ and $z_\alpha \in \mathbb{R}$ with $1-\alpha = \phi(z_\alpha)$ then it follows from the normal approximation that

$$p_o(\hat{p}) = \hat{p} - \frac{\hat{p}\, z_\alpha^2}{N+z_\alpha^2} \quad \text{and} \quad p_1(\hat{p}) = \hat{p} + \frac{z_\alpha^2(1-\hat{p})}{N+z_\alpha^2}.$$

But this confidence interval is the same as one would get for the decision problem $(\beta(N,p))_{p \in [0,1]}$ of a family of binomial distributions

$$\beta(N,p) = \sum_{k=0}^{N} \binom{N}{k} p^k (1-p)^{N-k} \varepsilon_k \quad \text{where} \quad \varepsilon_k(B) = \begin{cases} 1 & \text{if } k \in B \\ 0 & \text{otherwise} \end{cases}$$

with unknown p. This shows that our confidence interval is optimal in some sense if we have $n_i = \lambda_i \cdot N$ for $i = 1,\ldots,12$.
The only problem here is that at the beginning of the inquiry the λ_i are still unknown. They can only be determined at the

end of the 1982 swimming season. That is why I estimated the values λ_i using the recorded information from the years 1980 and 1981. The result of the inquiry is given in the following charts.

Code	Number of people questioned	Total number of visitor	Relative number of visitor	Optimal partition
000	565	7061	8,8%	364
001	185	1244	1,5%	62
002	114	514	0,6%	24
010	1025	25346	31,8%	1280
011	127	984	1,2%	50
012	66	407	0,5%	20
100	482	7049	8,7%	360
101	211	4403	5,5%	227
102	77	859	1,1%	45
110	1018	27652	34,3%	1419
111	219	3641	4,5%	186
112	50	1511	1,9%	78
	4139	80651	100 %	4139

Tab. 1: The inquiry.

Code	Kri	Lin	Quei	Scho	Stel	Tri	Kais	others
000	1,95%	0,35%	0,88%	2,48%	8,50%	24,96%	51,68%	9,20%
001	4,86%	0,00%	0,00%	1,08%	5,41%	37,30%	32,97%	18,38%
002	5,26%	0,00%	0,00%	0,00%	5,26%	49,12%	20,18%	20,18%
010	2,05%	0,29%	1,07%	2,05%	2,24%	13,27%	56,78%	22,24%
011	1,87%	0,00%	2,34%	2,11%	2,34%	12,65%	55,74%	22,95%
012	0,00%	0,00%	0,00%	0,00%	6,06%	56,06%	27,27%	10,61%
100	3,53%	0,00%	0,00%	3,53%	6,43%	35,89%	41,08%	9,54%

Code	Kri	Lin	Quei	Scho	Stel	Tri	Kais	others
101	4,27%	0,00%	0,00%	2,48%	8,53%	25,17%	36,49%	23,70%
102	2,60%	0,00%	0,00%	5,19%	5,19%	20,78%	50,65%	15,55%
110	4,21%	0,00%	1,46%	1,29%	6,15%	20,23%	42,88%	23,79%
111	4,39%	0,00%	0,94%	2,19%	8,46%	33,86%	45,14%	5,02%
112	8,00%	0,00%	0,00%	6,00%	10,00%	44,00%	24,00%	8,00%

Tab. 2: Evaluation for the different "day types".

Kri	Lin	Quei	Scho	Stel	Tri	Kais	others
3,30%	0,12%	0,99%	2,08%	5,38%	21,64%	47,13%	19,37%

Tab. 3: Estimated value of the rate of visitors from the different towns.

Kri	Lin	Quei	Scho	Stel	Tri	Kais	others
2,21%	0,00%	0,00%	1,03%	4,22%	20,20%	46,00%	16,92%
4,49%	1,18%	2,08%	3,22%	6,64%	23,13%	48,26%	21,95%

Tab. 4: Confidence region for the level 95%.

To conclude, I would like to make some remarks about generalizations of the problem. The great advantage of this examination was that the recorded results could be put into correspondence with definite "day types", that is, they could be identified with the different components of the mixing measure μ. In more general situations this is certainly not possible. Other information concerning such cases can be found in [2], [3]. Using the methods developed there, the results were finally investigated with boot-strapping arguments. It could be observed that the number of different "day types" had been chosen slightly too large; this, however, had no negative influence on the choice of the confidence region.

References

[1] Krickeberg, K.; Ziezold, H.: Stochastische Methoden. Springer-Verlag, Berlin (1977), p. 34

[2] Krüger, W.: Testing hypotheses on the number of components in a mixture of binomial distributions (to appear)

[3] Krüger, W.: Approximation von Integraldarstellungen und verwandte Fragen. Ph. D. thesis, Kaiserslautern 1983

[4] Witting, H.: Mathematische Statistik. Teubner, Stuttgart (1978)

W. Krüger
Fachbereich Mathematik
Universität Kaiserslautern
Erwin-Schrödinger-Str. 48
D - 6750 Kaiserslautern