INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 27, číslo 4, prosinec 2016

STATISTICAL ANALYSIS AND MODELLING OF SUBMISSIONS TO MUNICIPALITIES IN THE CZECH REPUBLIC

STATISTICKÁ ANALÝZA A MODELOVÁNÍ PŘÍCHOZÍ KORESPONDENCE NA ÚŘADY V ČESKÉ REPUBLICE

Anna Pidnebesna[1], Kateřina Helisová[1], Jiří Dvořák[2], Radka Lechnerová[3], Tomáš Lechner[4]

Address: [1]Department of Mathematics, Faculty of Electrical Engineering, Czech Technical University in Prague, Zikova 4, 166 27 Praha 6
[2]Department of Probability and Mathematical Statistics, Fac. of Mathematics and Physics, Charles University in Prague, Sokolovská 83, 186 75 Praha 8
[3]Department of Mathematics and Information Technologies, Private College of Economic Studies in Prague, Lindnerova 1, 180 00 Praha 8 – Libeň
[4]Department of Law, Faculty of Economics and Public Administration, University of Economics in Prague, nám. W. Churchilla 4, 130 67 Praha 3

E-mail: pidneann@fel.cvut.cz, helisova@math.feld.cvut.cz, dvorak@karlin.mff.cuni.cz, radka.lechnerova@svses.cz, tomas.lechner@gmail.com

Abstract: In the recent years, space-time point processes were recognised as a valuable tool in modelling different random events in fields such as biology, medicine, material sciences or economy. Nevertheless, new applications occur frequently. This paper concerns the use of the space-time point processes for modelling of submissions to municipalities in the Czech Republic. The positions and times of submissions are considered to be coordinates in the subset of Euclidean space, so they form a space-time point pattern which is to be analysed. Both continuous and discrete approaches to statistical inference and modelling are employed and their advantages and disadvantages are described. Finally, the most appropriate model is chosen, fitted to the data and its suitability is justified through classical methods of spatial statistics using mainly simulations.

Keywords: intensity of a point process, K-function, Poisson process, space-time point process, submission.

Abstrakt: V posledních letech se časoprostorové bodové procesy ukázaly být užitečným nástrojem při modelování různých náhodných jevů, např. v oblasti biologie, medicíny, materiálových věd či ekonomie, a proto se nyní často vyskytují jejich nové aplikace. Tento článek se zabývá použitím časoprostorových bodových procesů pro modelování korespondence přícházející úřadům

v různých obcích České republiky. Geografické polohy a doby podání předmětů korespondence jsou uvažovány coby souřadnice v podmnožině euklidovského prostoru, což znamená, že tvoří právě časoprostorový bodový proces, který je analyzován. K analýze je použitý jak spojitý tak diskrétní přístup, přičemž jsou popsány jejich výhody a nevýhody. Na závěr je vybrán nejvhodnější model, jehož vhodnost je testována pomocí klasických metod prostorové statistiky s využitím simulací.

Klíčová slova: časoprostorový bodový proces, intenzita bodového procesu, K-funkce, Poissonův proces, příchozí korespondence.

1. Introduction

The motivation for modelling submissions to municipalities is the following. Not all municipalities in the Czech Republic provide the same extent of delegated powers of government. Spatial distribution of delegated powers is based on the historical context of the territorial distribution of government into districts and other local historical contexts. Nevertheless, there exists no precise statistical mapping of the extent of catchment areas in terms of the extent of performance of separate agendas. That is why it is useful to analyse temporal development and spatial distribution of applicants who send submissions. The data provided by electronic records management system could be used for that. Obtained results can help create realistic description of catchment area and its development trends. Note that although we could study also outgoing communication of the municipalities, we concentrate on modelling of submissions, because we suppose that their behaviour is strongly correlated from the nature of the problem, and thus the analysis of the outgoing communication would not yield more information.

From the nature of the data it is obvious that methods of space-time modelling are appropriate tools for such analysis. Modelling of space-time point patterns has become very popular in the recent years (see e.g. [2], [6], [7], [9], [11] or [12]). The main reason is that the increased availability of computational power allows to analyse large datasets as well as to use difficult and time-consuming procedures needed in this field. Such an analysis is what we are dealing with in this paper. We have a large amount of data describing submissions to municipalities in the Czech Republic observed over a long period of time, more precisely, we have applicants' approximate geographical positions and dates of submissions during several years (see Section 2 for more details). Thus the data form an inhomogeneous space-time point pattern consisting of thousands of points.

While the theory of spatial point processes is well developed (see e.g. [4], [5] or [10]), space-time modelling is a newer discipline which is still under development, however there already exist many results which can be useful for our purposes. In [7], the second-order methods for inhomogeneous spatial point pattern data are provided while the inhomogeneous K-function proposed earlier in [1] in the spatial case is extended to the space-time setting. The authors of [9] work with the second-order characteristics, namely the K-function and the pair-correlation function of general inhomogeneous space-time processes. They study space-time separability properties using the first- and second-order intensities and reweighted pair-correlation functions of both the space-time process and its spatial and temporal projections, respectively. In [11], there is introduced a two-step estimation procedure for a flexible class of inhomogeneous space-time shot-noise Cox process models, where in the first step, the Poisson score estimation equation is used for estimating the inhomogeneity parameters, and in the second step, the minimum contrast estimation based on second-order moment characteristics is used for estimation of the interaction parameters. The available literature focuses mainly on the theory and simulation studies while the applications to real datasets are scarce.

A significant part of the analysis presented here is based on the works cited above. However, the introduced methods concern modelling in continuous domains, while in our case, the data are roughly discretised both in time and space as mentioned in Section 2. In order to provide as relevant analysis as possible, we apply the continuous approach as well as a discrete one. Namely, we consider the data to be a realisation of a point process in a connected subset of Euclidean space or a point pattern on a discrete lattice, respectively. We introduce the advantages and disadvantages of the different approaches and choose the most suitable one.

The paper is organised as follows. The data are described in Section 2. In Section 3, the continuous approach to modelling is introduced and the complications which occur in application to the considered dataset are discussed. Section 4 concerns discrete approach, where the data are seen as points lying on a discrete lattice both in time and space. Here, we combine the point process theory with methods for time-series analysis (see e.g. [8]) while we pay a special attention to the problem of multiplicity of events in the dataset which occur in the same lattice points. The results are summarised in Section 5.

2. Data description

The municipalities in the Czech Republic provide both local government and delegated powers of government. The scope of delegated powers is different

for different municipalities. They are divided into three categories depending on this scope: municipalities performing basic scope of delegated powers (the first type), municipalities with authorised municipal office (the second type; performing in addition for example environment and landscape protection and including registry office and building authority), and municipalities with extended powers (the third type; including in addition for example trade licensing office).

The data stemming from the electronic records management systems kept by the municipalities were examined. The data consists of dates, applicants' addresses, agenda and types of communication (electronic, post, personal etc.). However, the information was anonymised by the provider so that we have no specific addresses but only postcodes (ZIP codes) at our disposal. Therefore we identify the spatial position of the communicating subject by position of appropriate post office. We used the coordinate system of unified trigonometric cadastral network (S-JTSK) converted into kilometres where we shifted the origin of the coordinate system in order to optimise numerical calculations.

Thus, we have series containing the date of the incoming submittions and spatial identification within the territory of the Czech Republic. The aim is to use this data to analyse the spatial behaviour and its evolution over time. In accordance with this aim we randomly chose a municipality of the second type with the registry office and building authority. It is a municipality located about 50 km from Prague in the North-West direction in a village having about 2.8 thousand of inhabitants (called the selected municipality in the sequel). The methods and results are illustrated on this municipality while in the Conclusion section, the results for other municipalities are discussed.

The dataset includes 6205 space-time events corresponding to individual submissions to the selected municipality. The data were recorded in the time interval from 27th October 2009 to 20th April 2011. During this time interval 370 workdays took place (i.e. we disregard weekends and holidays). The submissions came from 214 different ZIP codes.

3. Continuous domain modelling

3.1. Basic terms

In this section, we recall basic terms from the theory of point processes.

Definition. *Consider* $(\Omega, \mathcal{F}, P)$ *a probability space. The point process* X *is a measurable mapping from* $(\Omega, \mathcal{F})$ *to* $(N, \mathcal{N})$*, where* N *the system of locally finite subsets of* $\mathbb{R}^d$ *with the* σ*-algebra* $\mathcal{N} = \sigma(\{\mathbf{x} \in N : \sharp(\mathbf{x} \cap A) = m\} :$

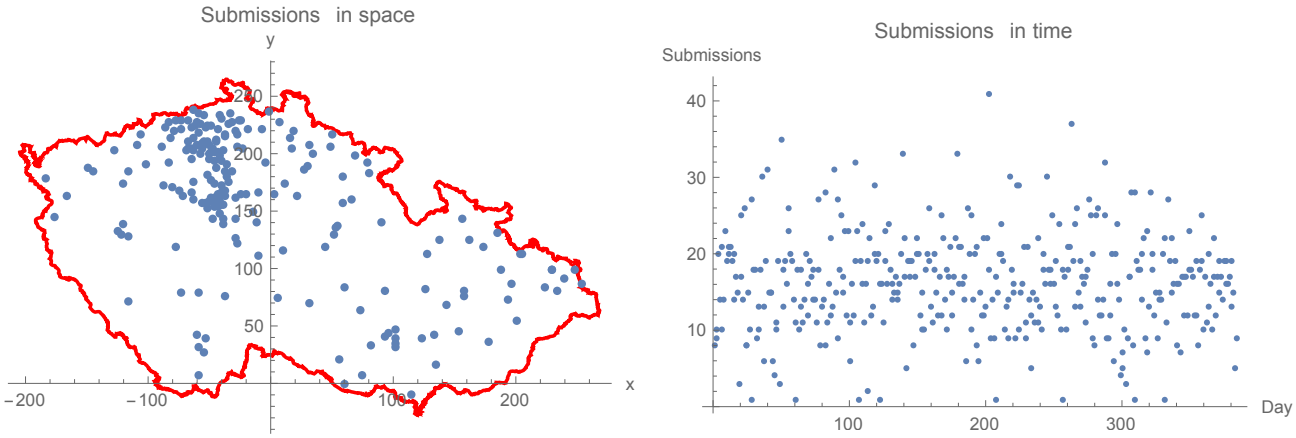


Figure 1: The map of spatial coordinates of submissions (left) and the temporal evolution of the number of submissions (right).

$A \in \mathcal{B}$, $m \in \mathbb{N}_0$), *where* $\mathcal{B}$ *denotes the system of all bounded Borel sets in* $\mathbb{R}^d$, $\sharp(A)$ *is the cardinality of the set* A, $X(A)$ *the number of points of the process* X *occurring in the set* A *and* $|A|$ *the Lebesgue measure of the appropriate dimension of the set* A.

Definition. *A locally finite diffusion measure* μ *on* $\mathcal{B}$ *satisfying* $\mu(A) = \mathbb{E}X(A)$ *for all* $A \in \mathcal{B}$ *is called the intensity measure of the process* X. *If there is a function* $\lambda(x)$, $x \in \mathbb{R}^d$, *such that* $\mu(A) = \int_A \lambda(x)\,\mathrm{d}x$, *then* $\lambda(x)$ *is called the intensity function. If* $\lambda(x) = \lambda$ *is constant then* λ *is called intensity.*

Definition. *The point process* X *with constant intensity* λ *is called homogeneous. Otherwise, it is called inhomogeneous.*

Definition. *The point process* X *is called stationary if its distribution is invariant under translations, i.e. for all* $v \in \mathbb{R}^d$, *the distribution of* $X + v = \{u + v, u \in X\}$ *is the same as that of* X.

Definition. *The Poisson point process is the point process* Φ *satisfying:*

- *for any finite collection* $\{A_n\}$ *of disjoint sets in* $\mathbb{R}^d$, *the numbers of points in these sets,* $\Phi(A_n)$, *are independent random variables,*
- *for each* $A \subset \mathbb{R}^d$ *such that* $\mu(A) < \infty$, $\Phi(A)$ *has the Poisson distribution with parameter* $\mu(A)$.

Definition. *Consider an arbitrary Borel set* A *and suppose that* $\mathcal{K}$ *is a measure on Borel sets such that*

$$\mathcal{K}(B) = \frac{1}{|A|}\mathbb{E}\sum_{x,y\in X}^{\neq} \frac{\mathbb{I}_{\{x\in A,\, x-y\in B\}}}{\lambda(x)\lambda(y)}, \quad B \in \mathcal{B},$$

does not depend on A, where $\sum_{x,y\in X}^{\neq}$ denotes the sum over all pairs of mutually different points. Then for a stationary point process, the K-function is defined as

$$K(r) = \mathcal{K}(b(0,r)), \quad r \in (0,\infty),$$

where $b(0,r)$ denotes a ball with centre in the origin 0 and radius $r > 0$.

In other words, $\lambda K(r)$ is the mean number of further points in the distance less or equal to r from the origin given that X has a point in the origin. In special cases, K-function can be equivalently expressed using the so-called pair-correlation function.

Definition. *Assume that for the point process X, the first and second order intensity functions $\lambda(u)$ and $\lambda^{(2)}(u,v)$ exist, respectively, i.e. it holds*

$$\int h_1(u)\lambda(u)\,\mathrm{d}u = \mathbb{E}\sum_{u\in X} h_1(u),$$

$$\iint h_2(u,v)\lambda^{(2)}(u,v)\,\mathrm{d}u\,\mathrm{d}v = \mathbb{E}\sum_{u,v\in X}^{\neq} h_2(u,v),$$

for any non-negative Borel functions $h_1 : \mathbb{R}^d \to \mathbb{R}$ and $h_2 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Then the pair-correlation function is defined as

$$g(u,v) = \frac{\lambda^{(2)}(u,v)}{\lambda(u)\lambda(v)}, \quad u,v \in \mathbb{R}^d.$$

Remark. *If there exists a function $\widetilde{g} : \mathbb{R}^d \to \mathbb{R}$ such that $g(x,y) = \widetilde{g}(y-x),\ x,y \in \mathbb{R}^d$, then the K-function is*

$$K(r) = \int_{b(0,r)} \widetilde{g}(u)\,\mathrm{d}u, \quad r > 0.$$

Definition. *Let X be a space-time point process on $\mathbb{R}^2 \times \mathbb{R}^+$ (i.e. the process whose first two coordinates correspond to the position in the plane and the third coordinate is the temporal item). Assume that $g((u,s),(v,t)) = \widetilde{g}(u-v, s-t),\ (u,s),(v,t) \in \mathbb{R}^2 \times \mathbb{R}$. Then the space-time K-function defined as*

$$K(r,t) = \int_{\mathbb{R}^2\times\mathbb{R}} \mathbb{I}\{\|u\| \le r, |s| \le t\}\widetilde{g}(u,s)\,\mathrm{d}(u,s), \quad r > 0, \quad t > 0,$$

where $\|u\|$ denotes the length of the vector u, $|s|$ denotes the absolute value of s and $\mathbb{I}$ denotes the indicator function.

Note that for the space-time Poisson process it holds that $g = 1$ and $K(r,t) = 2\pi r^2 t$ (see [9]).

3.2. Methodology

As seen from Figure 1, we may expect point clusters in the process, especially in spatial item. Our first approach is to use for modelling an inhomogeneous space-time Neyman-Scott process (see e.g. [10]). It is the Poisson process driven by a random intensity function

$$\Lambda(u,t) = \sum_{(v,s)\in\Phi} \mu\, f(u,t)k_1(v-u)k_2(s-t), \quad (u,t) \in \mathbb{R}^2 \times \mathbb{R}, \tag{1}$$

where Φ is a stationary Poisson process with an intensity $\nu > 0$, $k_1 : \mathbb{R}^2 \to \mathbb{R}$ and $k_2 : \mathbb{R} \to \mathbb{R}$ are probability density functions, $\mu > 0$ is an arbitrary constant and $f : \mathbb{R}^2 \times \mathbb{R} \to [0,1]$ is an inhomogeneity function whose larger values mean higher intensity in the corresponding spatial and temporal coordinates, and vice versa. Since in this model we know the theoretical form of the K-function, which can be estimated non-parametrically from the data (see below), we can estimate the model parameters by the method based on a contrast function, see [10].

There are several methods for estimating spatial and temporal intensity, respectively (see e.g. [4], [5] or [10]). The ones we use in this paper are described below. We use the estimate of space-time intensity function in the product form of estimates of spatial and temporal intensity functions (see [11]), i.e.

$$\widehat{\lambda}(u,t) = \frac{\widehat{\lambda}_{sp}(u)\,\widehat{\lambda}_{tm}(t)}{X(W \times [0,T])}, \tag{2}$$

where W denotes the spatial observation window (i.e. the area of the Czech Republic) and $[0,T]$ is the observed time interval.

In order to estimate the spatial intensity, we use the Voronoi tessellation (i.e. division of the space to cells corresponding to the points of the point pattern so that each cell is formed by the points in space whose distance to the corresponding point of the pattern is smaller than the distance to the other points of the pattern; for more details see e.g. [4]) as follows. Denote

$$X_{sp} = \{u\,|(u,t) \in X\} = \{(x_i, y_i) \in W,\, i \in I_{sp}\}$$

the spatial projections of the process X, where I_{sp} is the index set of spatial points with at least one point of the pattern. In the data, there are multiple points in many spatial coordinates (x_i, y_i), therefore to each coordinates (x_i, y_i), we assign the number $n_i^{(sp)}$ of points lying in these coordinates. Then

the estimate of the spatial intensity is given by

$$\widehat{\lambda}_{sp}(u) = \sum_{i \in I_{sp}} \frac{n_i^{(sp)}}{|C_i|} \mathbb{I}\{u \in C_i\}, \tag{3}$$

where $\{C_i, i \in I_{sp}\}$ is the system of cells of the Voronoi tessellation built over the points X_{sp} on the observation window W.

For estimation of temporal intensity, methods for estimating trends of time series (see [8]) are employed since the data are recorded day by day. Analogously to the approach above, we denote

$$X_{tm} = \{t \,|\, (u,t) \in X\} = \{t_i \in [0,T], i \in I_{tm}\},$$

the temporal projections of the process X, where I_{tm} is the index set of temporal points with at least one point of the pattern, and assign the number of points $n_i^{(tm)}$ to each time t_i. Then we construct the periodogram (see [8]), which is a tool using for finding the importance of frequencies (periods) in the data. From it as well as from the nature of the data, it is obvious that the most important time period in the data is one month and the second important period is one week. Further, when looking at the temporal evolution of the data in details, we can observe different behaviour in different working days caused probably by the fact that the office hours of the selected municipality are on Mondays and Wednesdays. Therefore we divided the days into two groups

$$A = \{\mathit{Monday}, \mathit{Wednesday}\},$$
$$B = \{\mathit{Tuesday}, \mathit{Thursday}, \mathit{Friday}\}.$$

Let $t \in \{0, 1, \ldots, T\}$ and consider functions $m(t)$ and $d(t)$ describing to which month and (working) day t corresponds, namely

$$m(t) = 1 \text{ if } t \text{ is a day in January}, \ldots,$$
$$m(t) = 12 \text{ if } t \text{ is a day in December}, \tag{4}$$

and analogously

$$d(t) = 1 \text{ if } t \text{ corresponds to Monday}, \ldots,$$
$$d(t) = 5 \text{ if } t \text{ corresponds to Friday}. \tag{5}$$

Then denote $M_k = \{t \in \{0, \ldots, T\} : m(t) = k\}$ for $k = 1, \ldots, 12$ the set of days belonging to the month k, and $D_A = \{t \in \{0, \ldots, T\} : d(t) \in A\}$ and

$D_B = \{t \in [0,T] : d(t) \in B\}$ is the sets of days belonging to the day group A and B, respectively. Estimate of the temporal intensity is then

$$\widehat{\lambda}_{tm}(t) = y_m(t) + y_d(t), \quad t \in [0,T], \tag{6}$$

where the functions $y_m(t)$ and $y_d(t)$ are defined as

$$y_m(t) = \sum_{k=1}^{12} \frac{\mathbb{I}\{t \in M_k\}}{\sharp(\{0,\dots,T\} \cap M_k)} \sum_{s=0}^{T} n_s^{(tm)} \mathbb{I}\{s \in M_k\},$$

which denotes the average number of submitions in a single day of the month corresponding to t, and

$$y_d(t) = \frac{\sum_{s=0}^{T}(n_s^{(tm)} - y_m(s))\mathbb{I}\{s \in D_A\}}{\sharp(\{0,\dots,T\} \cap D_A)} \mathbb{I}\{t \in D_A\} + \frac{\sum_{s=0}^{T}(n_s^{(tm)} - y_m(s))\mathbb{I}\{s \in D_B\}}{\sharp(\{0,\dots,T\} \cap D_B)} \mathbb{I}\{t \in D_B\},$$

which can be interpreted as the mean deviation of the day t from the monthly average with respect to the type of day A or B, respectively.

Thus using (3) and (6) for calculation of (2), we get the estimate of space-time intensity $\widehat{\lambda}(u,t)$, and so we can estimate the space-time K-function as

$$\widehat{K}(r,\tau) = \sum_{(u,t),(v,s)\in X}^{\neq} \frac{\mathbb{I}\{\| u - v \| \leq r\}\mathbb{I}\{|t-s| \leq \tau\} p((u,t),(v,s))}{\omega((u,t),(v,s))\widehat{\lambda}(u,t)\widehat{\lambda}(v,s)},$$

where $p((u,t),(v,s))$ is the number of pairs of $(u,t),(v,s)$ in the dataset, and $\omega((u,t),(v,s))$ is the edge correction. There are many possibilities of edge corrections (see e.g. [10]). In this paper we use the translation edge correction $\omega((u,t),(v,s)) = |(W \cap (W+u-v)) \times [0,T]|$, where $W + x = \{w + x | w \in W\}$.

3.3. Numerical results

On the Figure 2, the estimates of spatial and temporal intensities are shown. They are then used for empirical estimates of the K-function. As seen from Figure 3, the space-time K-function estimated from the data lies above the theoretical K-function for the Poisson process which indicates the presence of clusters in the data. However, the most important jump is in the point (0,0) which corresponds to the distance 0 km in space and 0 days in time.

Thus the conclusion is that the pattern has clusters as expected, but their typical scale is less than 1 km in the space and less than 1 day in time. Since the temporal part of the data is recorded in days and spatial part is given by the coordinates of ZIP codes of the municipalities, typically located in the distance of many kilometres from each other, we cannot model the

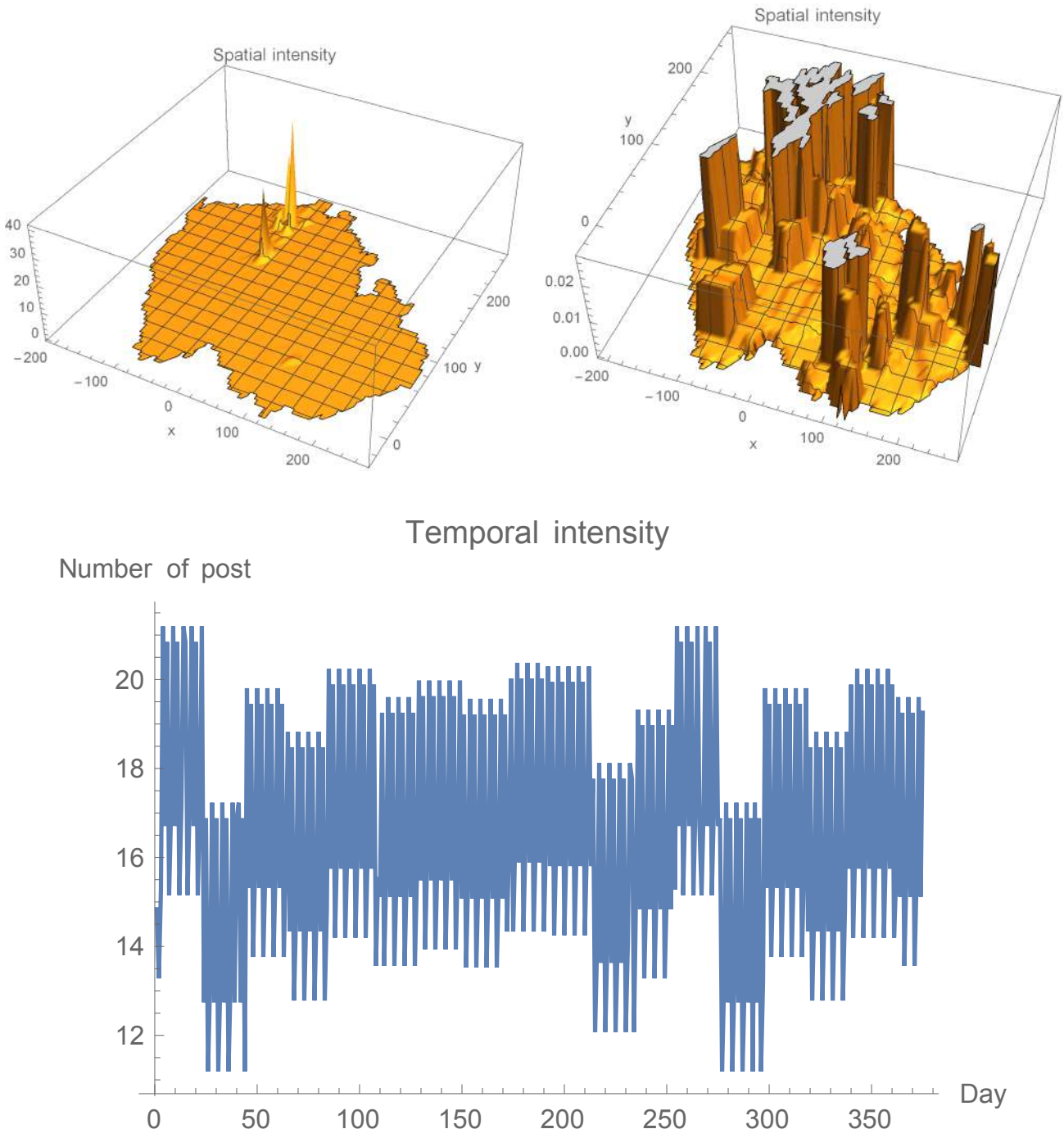


Figure 2: The estimate of spatial intensity in 2 different scales (the upper left figure plots the complete graph, the upper right figure shows details for areas with lower intensity) and the estimate of the temporal intensity (bottom).

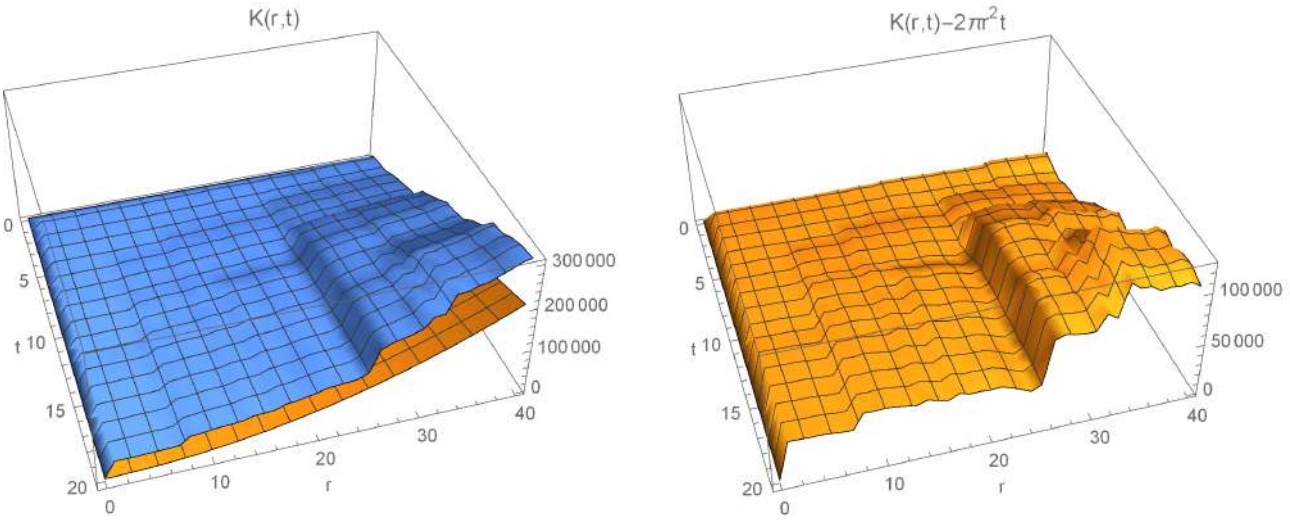


Figure 3: Space-time K-function estimated from the data plotted by blue color compared to the theoretical K-function for the Poisson point process plotted by orange color (left), and the corresponding differences between them (right). Units are given by kilometers and days, respectively.

clusters themselves. The clusters are formed by events with the same spatial and temporal coordinates due to the rather rough discretisation. Hence it is not possible to infer the precise scale of the clusters from the data and the continuous domain modelling is not appropriate for this dataset. Therefore we neither attempt to estimate the parameters of the Neymann-Scott model nor consider another form of the space-time intensity function than (2) as it cannot remedy the jump of the K-function estimate at the point (0,0).

4. Discrete domain modelling

In this section, we focus on discrete modelling. We consider the data to be a realisation of the process of points placed on a discrete lattice where the points of the lattice (called knots in the sequel so that they cannot be interchanged with the points of the process) are given by the cartesian product of the spatial positions (coming from a finite set) and dates of the submissions.

First, we study the dependence between the spatial and temporal parts of the process, and then, we focus on modelling of the number of points in the knots of the lattice.

4.1. Methodology

4.1.1. Testing the independence of spatial and temporal coordinates. Information about independence of the spatial and temporal parts of

the process is of great benefit, since in the case of independence, we can analyse spatial and temporal projections independently on each other. Otherwise we must consider them together.

Lets have the hypothesis H_0 : "Spatial position of points is independent of their temporal coordinate." In order to test it, we consider the process as couples of spatial and temporal coordinates, i.e.

$$X = \{(\xi, t),\, \xi \in S,\, t \in Tm\},$$

where S is the set of spatial positions of ZIP codes and Tm is the set of times $\{0, \ldots, T\}$. The data form a realisation of such process represented as a collection of N points $\xi_i = (x_i, y_i)$ with the associated (random) times t_i, $i = 1, \ldots, N$, so the form the point pattern

$$X = \{(\xi_1, t_1), \ldots, (\xi_N, t_N)\}.$$

For testing the null-hypothesis we use the so-called mark independence test which works as follows. Suppose that the process is characterised by some properties (e.g. the number of lattice knots with more than one point etc.) called characteristics or statistics in the sequel. We simulate permutations of times with respect to the spatial coordinates. The main idea is that in the case of independence of the spatial coordinates and times, both the permuted patterns and the data have similar characteristics. Using mathematical terminology, denote $\pi(1, \ldots, N)$ a uniform random permutation of the numbers $\{1, \ldots, N\}$ and

$$\widetilde{X} = \{(\xi_1, t_{\pi(1)}), \ldots, (\xi_N, t_{\pi(N)})\}$$

the corresponding pattern with permuted times. Further, consider K independent uniform permutations $\pi_1, \ldots, \pi_K$, which produce the patterns $\widetilde{X}_1, \ldots, \widetilde{X}_K$. Using these patterns, we can construct the empirical 95%-confidence intervals for arbitrary statistics under the null hypothesis. Thus, if the value of corresponding statistic calculated from the original dataset lies outside this interval, we reject the null-hypothesis, and vice versa.

For our purposes, we chose the following statistics:

- T_1 – the number of knots of the lattice with at least one point,
- T_2 – the number of knots of the lattice with more than one point,
- T_3 – the average number of points in a knot in the set of knots with at least one point,
- T_4 – the average number of points in a knot in the set of knots with more than one point.

4.1.2. Modelling the number of points by Poisson distribution. As seen from the numerical results below, we cannot work with the space and the time separately, so we still consider the process of points on the lattice.

Now, we are interested in the distribution of the number of points in the knots of the lattice. The natural beginning of this study is testing the hypothesis that they have the Poisson distribution with parameter given by the intensity of the corresponding knot. Thus, we consider the process represented as

$$X = \{(\xi_1, t_1, \eta_1), \ldots, (\xi_N, t_N, \eta_N)\},$$

where η_i are independent random variables having the Poisson distribution with the parameter $\widehat{\lambda}(\xi_i, t_i)$ and $\widehat{\lambda}(\xi, t)$ is the intensity estimate (2) described in Section 3.

In order to check the model, we apply again the empirical test based on construction of 95% envelopes for the statistics $T_1, \ldots, T_4$ mentioned above. Since in this case, the intensity plays an important role, we moreover observe the statistic T_5 describing the total number of points.

Note that we do not expect acceptance of the hypothesis of Poisson distribution in the knots of the lattice since it is connected with the independence of the behaviour in the space and in the time which is rejected. Therefore in the Section 4.2, we present also the results for specific parts of the lattice, particularly for the knots corresponding to the selected municipality.

4.1.3. Modelling the number of points by empirical distribution. As seen from the numerical results in the Section 4.2, the Poisson distribution is not satisfactory. Therefore in this section, we focus on the empirical distribution of η. We work with the collection of independent random variables $\eta'(\xi, m(t), \widetilde{d}(t))$ where ξ is the spatial position, $m(t)$ is defined above by (4), and $\widetilde{d}(t) = A$ if $t \in A$ and $\widetilde{d}(t) = B$ if $t \in B$ is the type of working day. Thus for each spatial position ξ, 24 empirical distributions (corresponding to 12 months and 2 different groups of day) are to be estimated.

We use two approaches for simulation-based goodness-of-fit testing. The first one is again the construction of confidence intervals for the statistics $T_1, \ldots, T_5$. The second one is construction of the confidence intervals for the following quantities:

- the mean number of points per day in a given month (e.g. in January), expressed as a difference from the overall mean number of points per day (regardless of the month),
- the mean number of points per day for a given day of week (e.g. for Mondays), expressed as a difference from the respective monthly av-

erages (e.g. Mondays in January contribute by the difference from the January average).

Note that the second set of quantities corresponds to the computation of $y_d(t)$ above where we consider five groups (Monday, ..., Friday) instead of two (A, B).

4.2. Numerical results

In order to test the independence of times and spatial coordinates, we simulate $K = 1000$ permutations and construct the 95% confidence interval for the given statistics $T_1, \ldots, T_4$. The results are introduced in Table 1. Independence of space and time was rejected in all four cases, so it is shown that there exist interactions between spatial and temporal coordinates.

Table 1: Testing independence between time and spatial coordinates. The second and the third column introduces the quantiles for the corresponding statistic calculated from 1000 simulated permutations, the fourth column is the value calculated from the data.

Statistics	$\mathbf{q_{0.025}}$	$\mathbf{q_{0.975}}$	**Data**	**Conclusion**
T_1	3393	3452	3141	*Reject*
T_2	1043	1098	1139	*Reject*
T_3	1.78	1.83	1.98	*Reject*
T_4	3.54	3.66	3.69	*Reject*

In Table 2, there are shown the results for testing Poisson distribution of the number of points in the knots of the lattice. The confidence intervals were constructed from only 100 simulations because the calculations are more time-consuming than in the case of testing independence. It is seen that this null-hypothesis was rejected in almost all the cases. The only acceptance occurs in the case of T_5. However, considering the role of the intensity in Poisson distribution, it is expected result.

Nevertheless, we must take into account the fact that the selected municipality may have a specific behaviour (e.g. because of more personal submissions etc.). Therefore we fit the process to the selected municipality and the remaining municipalities separately. From Table 3, we can conclude that neither in the separated cases, the Poisson distribution can be used for the number of submissions.

Table 2: Testing the hypothesis that the number of points in the knots of the lattice has Poisson distribution with the parameter $\widehat{\lambda}(\xi, t)$ given by (2). The second and the third column introduces the quantiles for the corresponding statistic calculated from 100 simulations, the fourth column is the value calculated from the data.

Statistics	$\mathbf{q_{0.025}}$	$\mathbf{q_{0.975}}$	**Data**	**Conclusion**
T_1	3345	3535	3141	*Reject*
T_2	1051	1124	1139	*Reject*
T_3	1.76	1.84	1.96	*Reject*
T_4	3.44	3.64	3.69	*Reject*
T_5	6019	6364	6205	*Not reject*

Table 3: Testing the hypothesis that the number of points in the knots of the lattice has Poisson distribution with the parameter $\widehat{\lambda}(\xi, t)$ given by (2) when the selected municipality is not included. The second and the third column introduces the quantiles for the corresponding statistic calculated from 100 simulations, the fourth column is the value calculated from the data.

Statistics	$\mathbf{q_{0.025}}$	$\mathbf{q_{0.975}}$	**Data**	**Conclusion**
T_1	2992	3166	2784	*Reject*
T_2	700	775	818	*Reject*
T_3	1.45	1.50	1.63	*Reject*
T_4	2.88	3.06	3.13	*Reject*
T_5	4480	4670	4527	*Not reject*

Finally in Table 4, the results based on empirical distributions are represented. Since the simulation is very time-consuming in this case, only 39 simulations were used. Recall that the model is based on the empirical distributions of 24 combinations of the month and the type of working day for each space point. As seen from the obtained results, it fits the data well.

In order to check the model from another point of view we also construct the confidence regions for the mean number of points per day in a given month and for the mean number of points per day for a given day of week, as defined above, see Table 5 and Table 6. The results indicate that the fitted model describes the temporal dynamics of the data rather well.

Table 4: Testing the hypothesis that the number of points in the knots of the lattice has distribution estimated empirically for 24 combinations of the month and the type of working day (12 months and 2 types of working days) through the statistics $T_1, \ldots, T_5$. The second and the third column introduces the quantiles for the corresponding statistic calculated from 39 simulations, the fourth column is the value calculated from the data.

Statistics	$\mathbf{q_{0.025}}$	$\mathbf{q_{0.975}}$	**Data**	**Conclusion**
T_1	3027	3204	3141	*Not reject*
T_2	1077	1177	1139	*Not reject*
T_3	1.92	2.02	1.98	*Not reject*
T_4	3.54	3.83	3.69	*Not reject*
T_5	5994	6343	6205	*Not reject*

5. Conclusions

We realised that behaviour of the process of submissions to municipalities in the Czech Republic is very complicated in the sense that it cannot be described neither by classical models of point processes such as Poisson process or cluster processes nor by their simple modifications. Neither using Poisson distribution for modelling the number of points in the same time and spatial coordinates was successful. Therefore, we suggested the procedure based on empirical approach, which allows us to describe the process of submissions to municipalities and make forecasts of future development. Despite the procedure is time-consuming, because it requires separate calculations for each spatial coordinate and each of the 24 combinations of the month and the type of working day, we can consider it being appropriate because it captures rather well the complicated structure of the observed dataset.

Finally note that the used methods were also applied to the data corresponding to three other randomly chosen municipalities (of the first, the second and the third type, respectively) and the obtained results were very similar.

Acknowledgement

This research was realised with the financial support of the International Visegrad Fund, the grant SGS15/072/OHK3/1T/13, the Czech Science Foundation, project No. 16-037085, and the grant VŠE IGS F5/75/2016.

Table 5: Testing the hypothesis that the number of points in the knots of the lattice has distribution estimated empirically for 24 combinations of the month and the type of working day (12 months and 2 types of working days). The mean number of points per day in a given month, expressed as a difference from the overall mean number of points per day, is considered here. The second and the third column gives the quantiles for the corresponding statistic calculated from 39 simulations, the fourth column shows the value calculated from the data.

Month	$\mathbf{q_{0.025}}$	$\mathbf{q_{0.975}}$	**Data**	**Conclusion**
January	−1.35	1.20	0.28	*Not reject*
February	−2.18	0.72	−0.70	*Not reject*
March	−1.19	2.05	0.71	*Not reject*
April	−1.47	1.61	0.08	*Not reject*
May	−2.44	3.17	0.45	*Not reject*
June	−1.92	2.49	0.04	*Not reject*
July	−2.04	4.12	0.85	*Not reject*
August	−1.53	5.36	0.77	*Not reject*
September	−2.86	0.31	−1.40	*Not reject*
October	−3.02	1.19	−0.20	*Not reject*
November	0.27	3.35	1.67	*Not reject*
December	−3.60	−0.60	−2.29	*Not reject*

Table 6: Testing the hypothesis that the number of points in the knots of the lattice has distribution estimated empirically for 24 combinations of the month and the type of working day (12 months and 2 types of working days). The mean number of points per day for a given day of week, expressed as a difference from the respective monthly averages, is considered here. The second and the third column gives the quantiles for the corresponding statistic calculated from 39 simulations, the fourth column shows the value calculated from the data.

Day	$\mathbf{q_{0.025}}$	$\mathbf{q_{0.975}}$	**Data**	**Conclusion**
Monday	1.80	3.88	2.98	*Not reject*
Tuesday	−3.35	−0.96	−1.51	*Not reject*
Wednesday	1.38	4.21	2.63	*Not reject*
Thursday	−2.82	−0.87	−3.07	*Reject*
Friday	−2.79	−0.81	−1.07	*Not reject*

References

[1] Baddeley, A., Møller, J., Waagepetersen, R. (2000): Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica* **54**(3), 329–350.

[2] Beneš, V., Prokešová, M., Staňková Helisová, K., Zikmundová, M. (2015): Space-time models in stochastic geometry. In *Stochastic Geometry, Spatial Statistics and Random Fields – Models and Algorithms.* Lecture Notes in Mathematics, Ed. Schmidt, V., Chapter 7, pp. 205–232. Springer Verlag.

[3] Breman, M., Diggle, P. (1989): Estimated weighted integrals of the second-order intensity of a spatial point pprocess. *Journal of the Royal Statistical Society B (Methodological)* **51**(1), 81–92.

[4] Chiu, S. N., Stoyan, D., Kendall, W. S., Mecke, J. (2013): *Stochastic Geometry and Its Applications*, 3rd edition. Wiley & Sons, Chichester.

[5] Daley, D. J., Vere-Jones, D. (2003) / (2008): *An Introduction to the Theory of Point Processes I / II.* Springer Verlag, New York.

[6] Dvořák, J., Prokešová, M. (2016): Parameter estimation for inhomogeneous space-time shot-noise Cox point processes. *Scandinavian Journal of Statistics* **43**(4), 939–961.

[7] Gabriel, E., Diggle, P. J. (2009): Second-order analysis of inhomogeneous space-time point process data. *Statistica Neerlandica* **63**(1), 43–51.

[8] Green, W. H. (2011): *Econometric Analysis*, 7th edition. Pearson Education, New York.

[9] Møller, J., Ghorbani, M. (2012): Aspects of second-order analysis of structured inhomogeneous space-time point processes. *Statistica Neerlandica* **66**(4), 472–491.

[10] Møller, J., Waagepetersen, R. (2004): *Statistical Inference and Simulations for Spatial Point Processes.* Chapman and Hall/CRC, Boca Raton.

[11] Prokešová, M., Dvořák, J. (2014): Statistics for inhomogeneous space-time shot-noise Cox processes. *Methodology and Computing in Applied Probability* **16**(2), 433–449.

[12] Schoenberg, F. (2005): Consistent parametric estimation of the intensity of a spatial-temporal point process. *Journal of Statistical Planning and Inference* **128**(1), 79–93.

A REVISIT TO TWO-SAMPLE LOCATION TESTS AND A GENERALIZED BEHRENS-FISHER PROBLEM

OPĚTOVNÉ VYHODNOCENÍ DVOUVÝBĚROVÝCH TESTŮ O ROZDÍLU V POLOZE A ZOBECNĚNÉHO BEHRENSOVA-FISHEROVA PROBLÉMU

Tomáš Marcinko, Dagmar Blatná

Adresa: University of Economics in Prague, nám. W. Churchilla 1938/4, 130 67 Praha 3

E-mail: xmart14@vse.cz, dagmar.blatna@vse.cz

Abstract: The aim of this article is to revisit the effects of non-normality and heteroskedasticity on the power functions of several parametric and rank-based two-sample location tests and to overcome the shortcomings of previous studies, which were often contradictory. An extensive Monte Carlo simulation study clearly showed that rank-based methods can lead to significant gain in power and efficiency in case of non-normality as indicated by Pitman asymptotic relative efficiencies, however, these tests are also very sensitive to violations of homoskedasticity. Nevertheless, in cases of concurrent non-normality and homoskedasticity the Fligner-Policello test or an asymptotic test based on the Hodges-Lehmann estimate of shift may be useful, especially for sufficiently large sample sizes.

Keywords: Two-sample location problem, generalized Behrens-Fisher problem, t-tests, rank-based tests, Hodges-Lehmann estimate of shift, bootstrap.

Abstrakt: Cílem tohoto článku je opětovné vyhodnocení vlivu nenormality a heteroskedasticity na silofunkce vybraných parametrických a pořadových dvouvýběrových testů o rozdílu v poloze a překonání částečných nedostatků předchozích studií, které často vedly k protichůdným závěrům. Výsledky rozsáhlé Monte Carlo simulační studie jasně ukázaly, že pořadové metody (zejména Wilcoxonův a van der Waerdenův test) mohou v souladu s Pitmanovou asymptotickou relativní vydatností dosahovat významně větší sílu a eficienci v případě porušení předpokladu normality, zároveň jsou však velmi citlivé na porušení předpokladu homoskedasticity. Při současném porušení předpokladů normality a homoskedasticity se však jeví být užitečný hlavně Flignerův-Policellův test nebo asymptotický test založený na Hodgesově-Lehmannově odhadu posunutí v poloze, a to zejména v případě dostatečně velkého rozsahu výběru.

Klíčová slova: Dvouvýběrové testy o rozdílu v poloze, zobecněný Behrensův-Fisherův problém, parametrické a pořadové testy, Hodgesův-Lehmannův odhad posunutí v poloze, bootstrap.

1. Introduction

Although nonparametric statistical location tests have been present a long time, particularly after a Wilcoxon rank-based alternative to parametric t-tests was proposed in [13], attitudes concerning their usefulness changed a lot in the following decades and their use in practical situations is still being questioned among many statisticians.

Although rank-based statistical methods enjoyed some popularity during the 1950's, especially among educational and psychological researchers, they became even more popular when it was proved in [8] and [3] that the proposed rank-based tests are asymptotically almost as efficient as the traditional methods when the data come from populations that follow a normal distribution and, furthermore, can be even more efficient than the traditional methods when the assumption of normality is not satisfied. The Pitman asymptotic relative efficiencies of the one-sample and two-sample rank-based test procedures with respect to the corresponding t-tests are given in Table 1.

Table 1: Pitman asymptotic relative efficiencies of rank-based tests and t-test

Distribution	Normal	Uniform	Logistic	Laplace	Cauchy
e(sign, t)	0.637	0.333	0.822	2.000	∞
e(Wilcoxon, t)	0.955	1.000	1.097	1.500	∞
e(Waerden, Wilcoxon)	1.047	∞	0.955	0.847	0.708

However, these asymptotic results were soon challenged, especially when widely cited results [2] concluded based on a small simulation study that the parametric t-tests are remarkably robust against violations of their underlying assumptions, thereby making the use of nonparametric tests unnecessary. This conclusion was subsequently accepted by many researchers and the movement towards nonparametric methods was described as unproductive due to sufficient robustness of the traditional parametric tests to nonnormality. Furthermore, this assertion was also often accompanied by an argument that nonparametric tests are less powerful than parametric tests in case of finite populations and that the Pitman asymptotic relative efficiencies

proving higher asymptotic efficiency of rank-based methods are calculated under a rather unrealistic assumption of infinitely large samples [1].

Nevertheless, this argument was not fully supported by other studies, which inter alia concluded that the normal scores test gave the most satisfactory results, followed closely by the Wilcoxon rank-sum test, and even when the populations were normally distributed these tests were only slightly inferior to the t-test, while they were much superior in cases of some non-normal populations [9]. Another comparisons between the power of the Wilcoxon rank-sum test and the power of the two-sample t-test concluded that in general the Wilcoxon statistic held very large power advantages over the t statistic, while asymptotic relative efficiencies were reasonably good indicators of the relative power of these two tests ([1], [10], [4]). However, results obtained from smaller samples were often markedly different from the results obtained from larger samples. Unfortunately, all of these studies deal mainly with a violation of normality, while the problem of potential heteroskedasticity is often omitted. In case of heteroskedasticity the use of Wilcoxon rank-sum test cannot be recommended due to its dramatic lack of robustness not only to the difference between population variances, but also to the population differences other than between means, and in this sense it cannot be deemed as a proper nonparametric analogue of the t-test, and therefore the well-known Welch t-test should be preferred [12]. More thorough description of historical developments in the evaluation of the robustness of the two-sample t-tests and its non-parametric alternatives are also summarized in [12].

The aim of this article is to revisit the problem of assumption violations of the two-sample location tests and to overcome the shortcomings of the aforementioned studies. Firstly, in case of homoskedasticity we will try to establish, if the asymptotic relative efficiencies are in fact good indicators of the relative power between the compared tests and how many observations we need for rank-based tests to become more powerful than the parametric t-tests. Secondly, this article will focus also on the violation of homoskedasticity assumption and will compare the results obtained by the Welch t-test with an appropriate modification of the Wilcoxon rank-sum test. Finally, an asymptotic test based on the robust Hodges-Lehmann estimate of shift using bootstrap estimate of its standard error will be considered.

2. Parametric and non-parametric methods

Let $X_1, \ldots, X_{n_1}$ be a random sample with common distribution function F and let $Y_1, \ldots, Y_{n_2}$ be another random sample, independent of the first, with common distribution function G. Consider a standard two-sample problem

of testing a null hypothesis that the location parameters of two populations are the same.

Undoubtedly the most popular tests regarding this problem are the two-sample Student's t-test and the approximate two-sample Welch's t-test, both of which can be used to test the null hypothesis of the form $H_0 : \mu_1 = \mu_2$, where μ_1 and μ_2 denote population means. Needless to say, these parametric t-tests were derived under the assumption that the observations are independent and identically distributed and follow a normal distribution. Student's t-test, which also assumes equal population variances, is of course a uniformly most powerful unbiased test of the null hypothesis under normality, but can lose its optimality when the assumption of normality is violated.

As the theoretical approximation of the power functions of the t-tests under non-normality obtained by the terms of the Edgeworth series in [11] can be cumbersome and unpractical, power functions of the relevant tests were estimated by means of Monte Carlo simulations. Moreover, the estimated power functions of these parametric tests were subsequently compared with the power functions of appropriate non-parametric alternatives.

Several two-sample rank-based tests were considered for the purpose of this article, mainly the Wilcoxon rank-sum test, the van der Waerden test and the Brown-Mood median test, all of which are included in R and provided by functions `wilcox_test`, `normal_test` and `median_test` of the `coin` package. It is important to notice that all of these tests are in fact testing a slightly different null hypothesis of the form $\mathsf{H_0} : F(x) = G(x)$ for every x, which asserts that the compared populations have the same probability distribution, the common distribution not being specified. The alternative hypothesis is typically of the location-shift form $\mathsf{H_1} : F(x) = G(x - \Delta)$ for every x, i.e. it says that compared populations are the same except for a shift by the amount Δ. For this hypothesis the Wilcoxon test also constitutes a locally most powerful rank test in case of densities of logistic type, while the van der Waerden and median tests are asymptotically optimal for densities of normal and double exponential type respectively [6].

Since the null hypothesis of these nonparametric alternatives to the t-tests implicitly assumes that population variances are the same, these tests may fail in case of heteroskedasticity. However, if one needs to test the null hypothesis $\mathsf{H_0} : \theta_1 = \theta_2$, where θ_1 and θ_2 denote population location parameters (e.g. population medians), in case of different population variances there is a possibility of utilizing the Fligner-Policello modification of the Mann-Whitney test proposed in [5], which is suitable specifically for this null hypothesis and is provided by a function `pFligPoli` of `NSM3` package. Moreover, as the Fligner-Policello test does not assume assumptions of nor-

mality or homoskedasticity, it can be deemed as a proper nonparametric alternative to the Welch's t-test for a generalized Behrens-Fisher problem (i.e. a generalization of the Behrens-Fisher problem when the populations are not assumed to be normally distributed) in case of symmetrically distributed populations [7]. The last procedures, which will be considered in this article, is an asymptotic test based on asymptotic normal distribution of the Hodges-Lehmann estimate of shift in location between the populations of the form $\widehat{\Delta}_{HL} = \text{med}\,\{X_i - Y_j : 1 \leq i \leq n_1, 1 \leq j \leq n_2\}$ with standard error being estimated by bootstrap, which can be useful in case of a violation of both normal and homoskedasticity assumption, and a Fisher-Pitman permutation test based on the difference in population means, which may be useful in case of a violation of normality when population variances are equal.

3. Results

3.1. Two-sample location tests in case of homoskedasticity

To compare the power functions of the considered location tests a simulation study was conducted with the aim of a Monte Carlo estimation of type I and type II errors of the two-sample tests shortly discussed in the previous section. Several symmetric and asymmetric continuous distributions were considered, namely normal distribution $\mathsf{N}(100, 100)$, logistic distribution $\mathsf{Logis}(100, 10\frac{\sqrt{3}}{\pi})$, shifted Student's $t(3)$ distribution $100+10\frac{1}{\sqrt{3}}\,t(3)$, contaminated normal distribution $0.95\,\mathsf{N}(100, 100)+0.05\,\mathsf{N}(100, 10\,000)$, uniform distribution $\mathsf{U}(100-10\sqrt{3}, 100+10\sqrt{3})$, log-normal distribution $\mathsf{LN}(4.6, 0.00995)$, skew normal distribution $\mathsf{SN}(88.417, 15.303, 3)$ and shifted exponential distribution $90 + \mathsf{Exp}(0.1)$. Without the loss of generality, all of these distributions (except for a contaminated normal distribution, in case of which the majority of the population comes from a specified normal distribution, whereas a small proportion of the population ($\epsilon = 0.05$) comes from a normal distribution with the same mean but much larger variance) were calibrated so that they have the population mean 100 and the variance 100. Selected characteristics including inter-quartile range, skewness and kurtosis are given in Table 2. The simulation study was computed in the programming language R, version 2.15.3, and consisted of 20 000 simulated data sets so that the Monte Carlo error was sufficiently low.

In order to examine the mere effects of non-normality on the two-sample location tests, the sample sizes and all the distribution parameters were kept the same for both samples. The estimated type I errors for various two-sample

two-tailed location tests and three different sample sizes are given in Tables 3 and 4 (a significance level of 0.05 was used).

Table 2: Characteristics of the used distributions

Distribution	Variance	IQR	Skewness	Kurtosis
Normal	100	13.49	0.00	3.00
Logistic	100	12.11	0.00	4.20
$t(3)$	100	8.83	NDEF	∞
Contaminated	595	12.52	0.00	42.45
Uniform	100	17.32	0.00	1.80
Log-normal	100	13.40	0.30	3.16
Skew normal	100	13.25	0.67	3.51
Exponential	100	10.99	2.00	9.00

It is obvious that in case of substantial skewness (e.g. exponential distribution), excess kurtosis (e.g. $t(3)$ distribution) or when outliers are present (e.g. contaminated distribution), the Student's t-test becomes conservative, i.e. it has the type I error lower than the significance level. In such cases permutation test may be useful. Otherwise, it holds the type I error reasonably close to significance level and in this sense it can expectedly be deemed quite robust to the violation of normality.

On the other hand, the rank-based tests are insensitive to even extreme violations of normality, even though some of them may be conservative, especially in case of small sample sizes. In such cases an interpolated confidence interval can be recommended. A simple linear interpolated confidence interval was applied to the Wilcoxon rank-sum test for $(n_1, n_2) = (10, 10)$.

It is clear that the Wilcoxon and van der Waerden tests hold the type I error very close to the significance level and can be deemed to be exact. However, the median test proved to be too conservative (or too liberal when utilizing an asymptotic distribution of its test statistic) for practical purposes and it cannot be recommended unless the sample sizes are very large.

The estimated power functions of the two-sample Student's t-test, Wilcoxon rank-sum test and van der Waerden test are given in Figures 1 and 2. It is obvious that the Student's t-test, which is uniformly most powerful and

Table 3: Estimated type I errors of two-sample two-tailed location tests

(n_1, n_2)	Distribution	t-test	Permutation	Wilcoxon**	Waerden	Median
$(10, 10)$	Normal	0.05045	0.05050	0.05125	0.05090	0.02360*
$(10, 10)$	Logistic	0.04895	0.04950	0.05125	0.05090	0.02360*
$(10, 10)$	$t(3)$	0.04320*	0.04890	0.05130	0.05090	0.02360*
$(10, 10)$	Contaminated	0.03125*	0.04965	0.05005	0.05145	0.02340*
$(10, 10)$	Uniform	0.05280	0.05145	0.05115	0.05090	0.02360*
$(10, 10)$	Log-normal	0.04920	0.04955	0.05125	0.05090	0.02360*
$(10, 10)$	Skew normal	0.04910	0.04955	0.05150	0.05090	0.02360*
$(10, 10)$	Exponential	0.04355*	0.05060	0.05130	0.05090	0.02360*
$(30, 30)$	Normal	0.05055	0.05070	0.04885	0.04870	0.01890*
$(30, 30)$	Logistic	0.04935	0.05060	0.04885	0.04870	0.01890*
$(30, 30)$	$t(3)$	0.04495*	0.04860	0.04885	0.04870	0.01890*
$(30, 30)$	Contaminated	0.03115*	0.04780	0.04845	0.04900	0.01905*
$(30, 30)$	Uniform	0.05010	0.04955	0.04885	0.04870	0.01890*
$(30, 30)$	Log-normal	0.04980	0.05070	0.04885	0.04870	0.01890*
$(30, 30)$	Skew normal	0.04945	0.04960	0.04885	0.04870	0.01890*
$(30, 30)$	Exponential	0.04735	0.04960	0.04885	0.04870	0.01890*

* Estimated type I error differs significantly from the significance level 0.05.
** Wilcoxon rank-sum test based on linearly interpolated confidence intervals.

Table 4: Estimated type I errors of two-sample two-tailed location tests

(n_1, n_2)	Distribution	t-test	Permutation	Wilcoxon	Waerden	Median**
$(100, 100)$	Normal	0.05025	0.05075	0.05005	0.05065	0.06655*
$(100, 100)$	Logistic	0.04970	0.04980	0.05005	0.05065	0.06655*
$(100, 100)$	$t(3)$	0.04795	0.05055	0.05005	0.05065	0.06655*
$(100, 100)$	Contaminated	0.04380*	0.04850	0.04945	0.05000	0.06615*
$(100, 100)$	Uniform	0.05140	0.05105	0.05005	0.05065	0.06655*
$(100, 100)$	Log-normal	0.05045	0.05105	0.05005	0.05065	0.06655*
$(100, 100)$	Skew normal	0.04930	0.04975	0.05005	0.05065	0.06655*
$(100, 100)$	Exponential	0.04875	0.04945	0.05005	0.05065	0.06655*

* Estimated type I error differs significantly from the significance level 0.05.
** Brown-Mood median test based on an asymptotic distribution of the test statistic.

unbiased under normality, is only slightly more powerful than the Wilcoxon rank-sum test when the assumption of normality is satisfied, and the difference between the power functions of the t-test and the van der Waerden test under normality is almost negligible, even for smaller sample sizes. This conclusion holds true also for distributions that are quite close to a normal distribution (e.g. log-normal or skew normal). However, in cases of substantially leptokurtic or platykurtic distributions the rank-based methods achieved significantly more power than the t-test with the Wilcoxon rank-sum test being the most powerful for symmetric leptokurtic distributions (e.g. $t(3)$ or contaminated normal) and the van der Waerden test being the most powerful for platykurtic or substantially asymmetric distributions (e.g. uniform or exponential). In case of homoskedasticity the aforementioned results also hold for unequal sample sizes.

3.2. Two-sample location tests in case of heteroskedasticity

Although the classical rank-based location tests studied in the previous section do not assume normal distribution of the random sample, they do implicitly assume homoskedasticity. In order to determine the effect of heteroskedasticity the data were simulated from a normal and a shifted Student's $t(3)$ distribution with the mean 100 and the variance $100\,k_i$, $i = 1, 2$, where $\mathbf{k} = (0.2, 1.8)'$. The estimated type I errors for various two-sample two-tailed location tests and sample sizes are given in Table 5.

The obtained results lead to a conclusion that both the Wilcoxon rank-sum test and the van der Waerden tests fail in case of significant heteroskedasticity, becoming too liberal or too conservative, especially when sample sizes differ. On the other hand, the Welch t-test expectedly proved to be very robust against a violation of homoskedasticity. However, when both assumptions of normality and homoskedasticity are violated, there is a possibility of utilizing the Fligner-Policello modification of the Mann-Whitney test or the asymptotic test based on the Hodges-Lehmann estimate of shift in location. Although both of these tests may be a bit liberal, especially for small sample sizes, they tend to be quite insensitive to heteroskedasticity. Moreover, as is depicted in Figure 3, the power functions of these tests are almost the same for larger sample sizes and they may be more powerful than the parametric t-test in case of concurrent non-normality and heteroskedasticity.

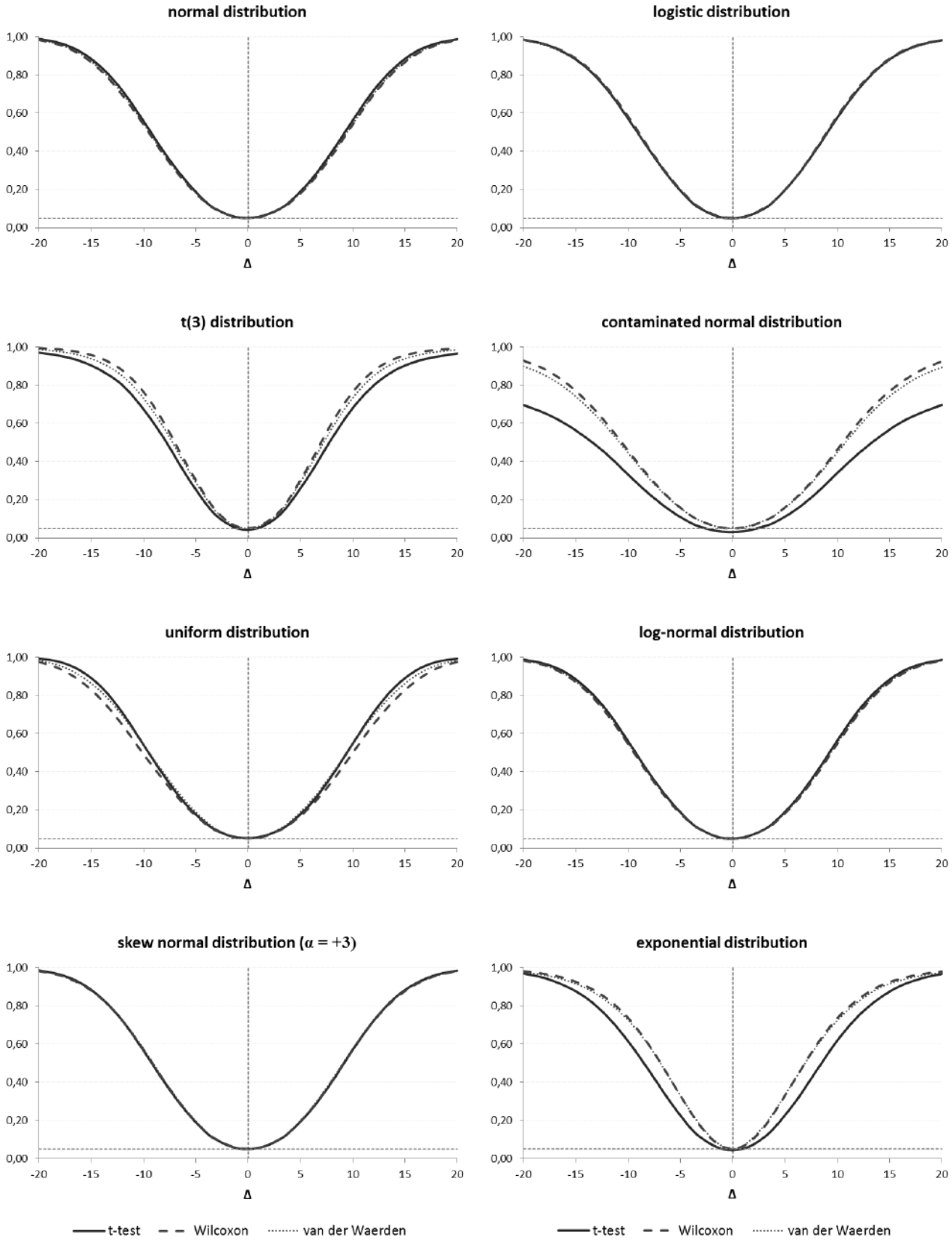


Figure 1: Estimated power functions of the Student's t-test, the Wilcoxon rank-sum test and the van der Waerden test, $(n_1, n_2) = (10, 10)$.

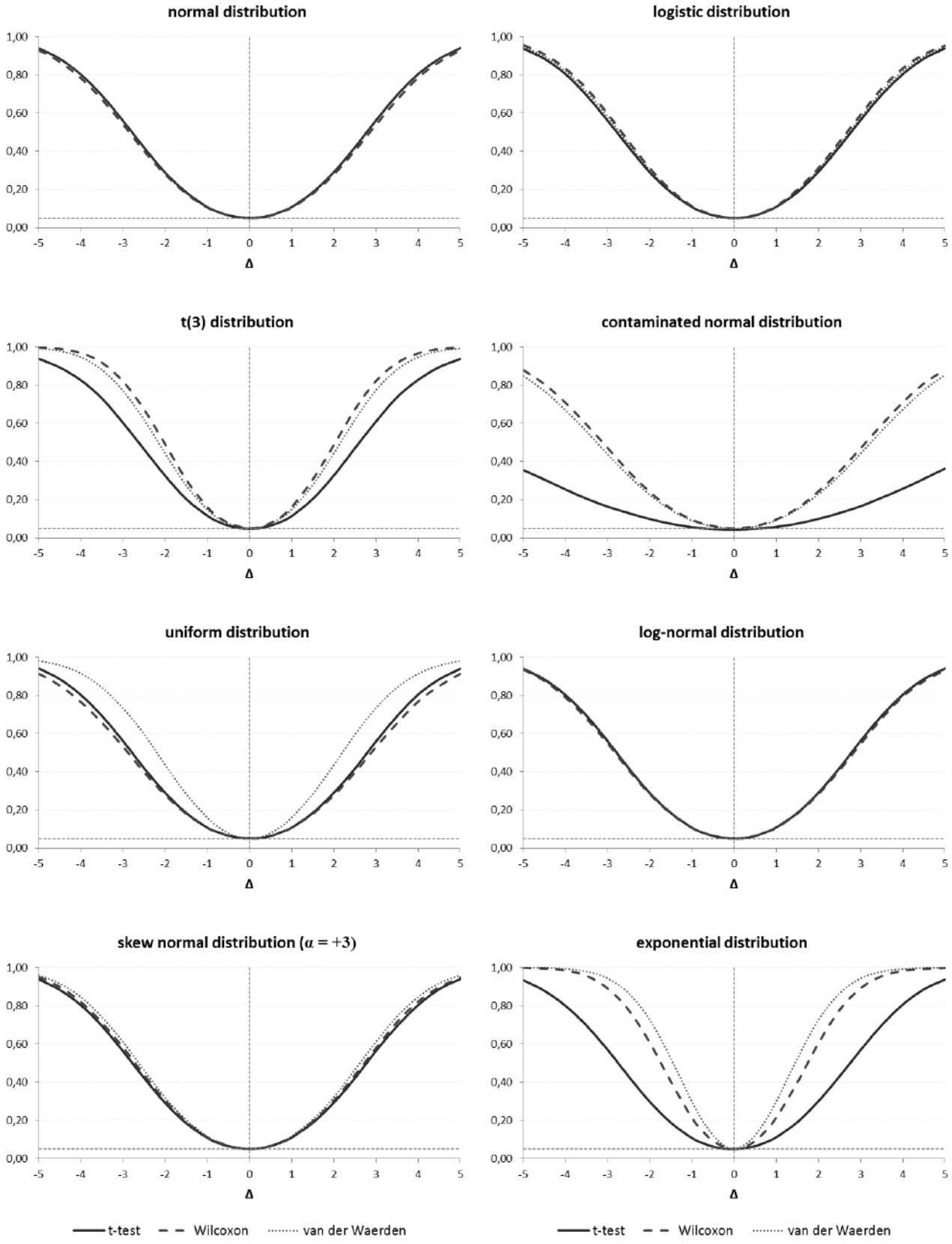


Figure 2: Estimated power functions of the Student's t-test, the Wilcoxon rank-sum test and the van der Waerden test, $(n_1, n_2) = (100, 100)$.

Table 5: Estimated type I errors of various two-sample two-tailed location tests, normal and $t(3)$ distribution, $(\sigma_1^2; \sigma_2^2) = (20, 180)$.

	(n_1, n_2)	Welch	Wilcoxon	Waerden	F-P	Asym. H-L
Normal distribution	(10, 10)	0.05035	0.05985*	0.04980	0.07630*	0.07570*
	(30, 30)	0.04965	0.06820*	0.04110*	0.05765*	0.06605*
	(100, 100)	0.04860	0.06890*	0.03765*	0.05290	0.05510*
	(10, 20)	0.04910	0.03135*	0.01510*	0.05700*	0.06490*
	(50, 100)	0.05025	0.03340*	0.01065*	0.05240	0.05670*
	(20, 200)	0.04945	0.00125*	0.00000*	0.05185	0.05585*
	(20, 10)	0.05025	0.10295*	0.09270*	0.08155*	0.07930*
	(100, 50)	0.04920	0.10690*	0.08465*	0.05440*	0.06720*
	(200, 20)	0.04895	0.16155*	0.21700*	0.06990*	0.07440*
$t(3)$ distribution	(10, 10)	0.04140*	0.05700*	0.05150	0.07375*	0.05620*
	(30, 30)	0.04420*	0.06400*	0.04220*	0.05570*	0.05885*
	(100, 100)	0.04470*	0.06490*	0.04010*	0.05205	0.05090
	(10, 20)	0.04155*	0.03135*	0.01895*	0.05575*	0.05345*
	(50, 100)	0.04725	0.03390*	0.01420*	0.05340*	0.05710*
	(20, 200)	0.04720	0.00285*	0.00055*	0.05410*	0,05225
	(20, 10)	0.04085*	0.09470*	0.08780*	0.07975*	0.05855*
	(100, 50)	0.04485*	0.09870*	0.08535*	0.05450*	0.05780*
	(200, 20)	0.04260*	0.14710*	0.19130*	0.06995*	0,06380*

* Estimated type I error differs significantly from the significance level 0.05.

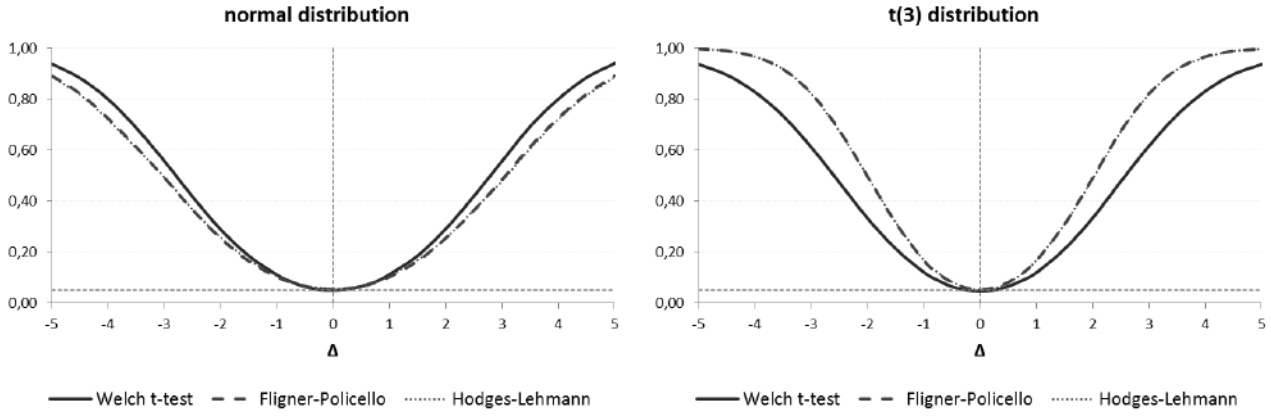


Figure 3: Estimated power functions of the Welch's t-test, the Fligner-Policello test and the asymptotic test based on the Hodges-Lehmann estimate, $(\sigma_1^2, \sigma_2^2) = (20, 180)$, $(n_1, n_2) = (100, 100)$.

4. Conclusion

The results of the simulation study confirmed that the traditional parametric t-tests are relatively robust to a violation of normality and homoskedasticity, however, it was clearly proved that these parametric methods are losing their optimal properties when the underlying assumptions are violated. It was concluded that in many cases of non-normality the appropriate rank-based methods are more powerful, especially the Wilcoxon rank-sum test for symmetric leptokurtic distributions and the van der Waerden test for platykurtic or substantially asymmetric distributions. The Pitman asymptotic relative efficiency can be a good indicator, which rank-based method is suitable.

However, both these well-known rank-based tests proved to be very sensitive to a violation of homoskedasticity and, therefore, these methods should be preferred only in cases, when population variances can be deemed to be equal. Nevertheless, in cases of concurrent non-normality and heteroskedasticity the Fligner-Policello test or the asymptotic test based on the Hodges-Lehmann estimate of shift in location may be useful.

References

[1] Blair, R. C., Higgins, J. J. (1980): A Comparison of the Power of Wilcoxon's Rank-Sum Statistic to That of Student's t-Statistic under Various Nonnormal Distributions. *Journal of Educational Statistics* **5**(4), 309–335.

[2] Boneau, C. A. (1960): The effects of violations of assumptions underlying the t-test. *Psychological Bulletin* **57**(1), 49–64.

[3] Chernoff, H., Savage, I. R. (1958): Asymptotic Normality and Efficiency of Certain Nonparametric Test Statistics. *The Annals of Mathematical Statistics* **29**(4), 972–994.

[4] Fay, M. P., Proschan, M. A. (2010): Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys* **4**, 1–39.

[5] Fligner, M. A., Policello, G. E. (1981): Robust Rank Procedures for the Behrens-Fisher Problem. *Journal of the American Statistical Association* **76**(373), 162–168.

[6] Hájek, J., Šidák, Z., Sen, P. K. (1999): *Theory of Rank Tests.* Academic Press, San Diego, 1999.

[7] Hettmansperger, T. P., McKean, J. W. (2010): *Robust Nonparametric Statistical Methods.* CRC Press, Boca Raton, 2010.

[8] Hodges, J. L., Jr., Lehmann, E. L. (1956): The Efficiency of Some Nonparametric Competitors of the t-test. *The Annals of Mathematical Statistics* **27**(2), 324–335.

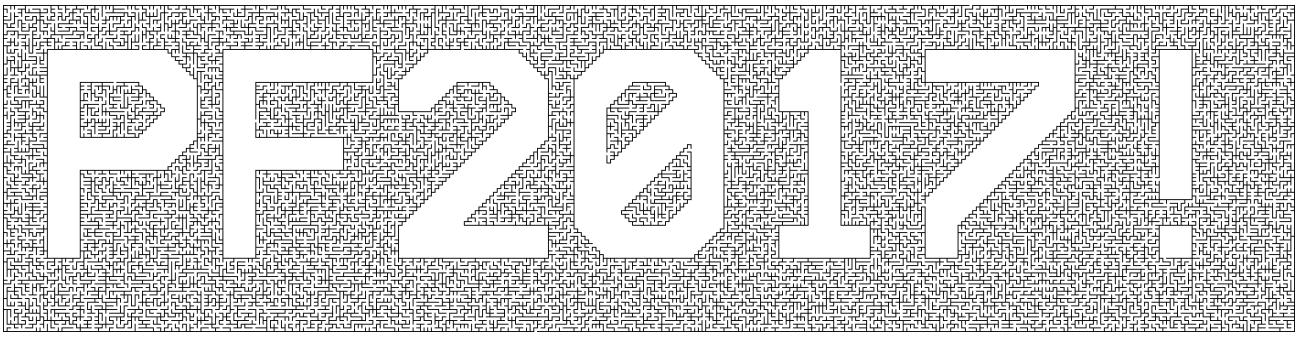
[9] Neave, H. R., Granger, C. W. J. (1968): A Monte Carlo Study Comparing Various Two-Sample Tests for Differences in Mean. *Technometrics* **10**(3), 509–522.

[10] Sawilowsky, S. S., Blair, R. C. (1992): A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin* **111**(2), 352–360.

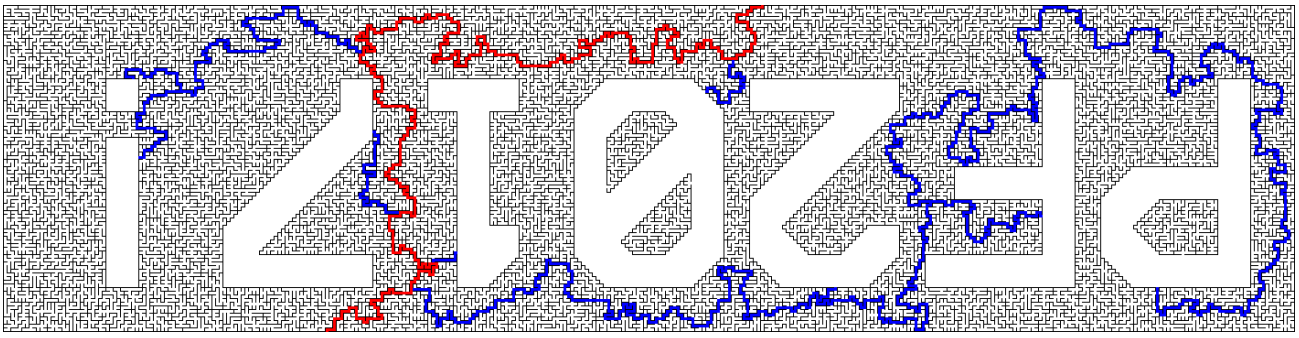
[11] Srivastava, A. B. L. (1958): Effect of non-normality on the power function of t-test. *Biometrika* **45**(3–4), 421–429.

[12] Stonehouse, J. M., Forrester, G. J. (1998): Robustness of the t and U tests under combined assumption violations. *Journal of Applied Statistics* **25**(1), 63–74.

[13] Wilcoxon, F. (1945): Individual comparisons by ranking methods. *Biometrics* **1**(6), 80–83.



Najdete Vy nebo Vaše děti/(pra)vnoučata cestu do/z bludiště?
Najdete cestu k jednotlivým prázdným útvarům?
Řešení je někde v tomto bulletinku.



Řešení k PF2017!

Doc. RNDr. MARTIN JANŽURA, CSc., 1955–2016

Antonín Otáhal

E-mail: otahal@utia.cas.cz

Martin Janžura ukončil studia na MFF UK, obor pravděpodobnost a matematická statistika, v roce 1979. Po ukončení vojenské služby v roce 1980 nastoupil jako řádný aspirant do Ústavu teorie informace a automatizace ČSAV (nyní AV ČR), v němž obhájil disertaci a setrval po celou svou odbornou kariéru.

Velká šíře jeho odborných zájmů zahrnovala oblast teorie pravděpodobnosti, matematické statistiky a teorie informace; zejména šlo o náhodné procesy a náhodná pole (včetně gibbsovských a markovských), princip minimální I-divergence a minimální entropie chyby – ve všech těchto oblastech se zajímal jak o teoretické základy, tak statistickou analýzu.

Řadu let učil markovské procesy na FJFI ČVUT a byl úspěšným školitelem několika studentů doktorského studia. Vedle toho pravidelně přednášel doktorandům KPMS MFF UK.

Kromě výzkumu se intenzivně věnoval organizační práci: spoluvytvořil tradici zimních škol pro mladé pracovníky ÚTIA, významně přispíval – jak odborně, tak organizačně – při organizaci pražských konferencí o teorii informace, statistických rozhodovacích funkcích a náhodných procesech, v devadesátých letech 20. století vedl oddělení stochastické informatiky a později velmi úspěšně působil ve funkci zástupce ředitele ÚTIA. Podílel se na řešení významných výzkumných projektů.

Od roku 1999 byl redaktorem a od roku 2011 šéfredaktorem časopisu *Kybernetika*, kde spolupracoval při vytváření nového redakčního systému a svou prací bezesporu přispěl k zlepšování nejenom impaktního faktoru časopisu, ale především jeho vědecké kvality.

V posledních letech života trpěl bolestivou nemocí, která ho velmi omezovala v chůzi. Navzdory velkému tlaku okolí však odmítal vyhledat lékařskou pomoc a přivedl tento druh svobodného rozhodování až k chmurnému konci – předčasnému úmrtí 2. srpna 2016.

Martin byl přemýšlivý matematik, schopný organizátor, uznávaný kolega a především dobrý přítel. Bude nám chybět.

Podrobněji viz jeho nekrolog v časopise *Kybernetika* na adrese:

http://www.kybernetika.cz/content/2016/4/661

ROBUSTNÍ ZPRÁVA O ROBUSTU 2016

Ondřej Vencálek

E-mail: ondrej.vencalek@upol.cz

Můžu si za to sám. Napsat krátkou zprávu o letošním Robustu jsem slíbil na konci září. Teď je konec listopadu a mně nezbývá než hluboko v paměti lovit ty vzpomínky, které „odolaly“ téměř celý semestr. Na Robustu je však „odolnost“, tedy robustnost, tou nejvyšší ctností. Jiné vzpomínky než ty opravdu robustní sem vlastně ani nepatří.

Vzpomínka 1: příjezd. Krásná zářijová neděle a usměvavý hlavní organizátor! Hned se dal do zpovídání mých dětí (jestli chodí do berušek nebo motýlků či snad koťátek) a navrhl společnou fotografii v legendárním tričku! Trochu mě pak překvapil, když se vyjádřil v tom smyslu, že „tentokrát je Robust na opravdu odlehlém místě“ (volil snad poněkud jiná slova, ale význam byl přesně tento). Nám olomouckým účastníkům se vrchol Pradědu nezdál až zas tak odlehlý. I když uznávám, že dopravní omezení v poslední fázi cesty, tj. na Ovčárnu, mohla cestu značně prodloužit.

Vzpomínka 2: přednášky. Některé si vybavuju velmi ostře, jiné mi splývají. V některých případech, přiznávám, jsem se „nechytl“. Jako vždycky, člověk si vybírá, něco ho zajímá víc a něco míň, někdy ho řečník nadchne i pro téma, nad kterým zatím moc nepřemýšlel. S obdivem jsem sledoval zejména přednášky svých vrstevníků Šárky Došlé, Michala Pešty, Matúše Maciaka i dalších. Myslím, že mají schopnost zaujmout a pozornost udržet, vysvětlit problém i jeho řešení a přitom všem nenudit. Úžasné!

Vzpomínka 3: večery. Tolik milých a blízkých lidí, přátel! S každým z nich jsem chtěl posedět a aspoň na chvíli mluvit. Jenže když chcete s někým opravdu mluvit, tak to přece jen chvíli trvá, než se dostanete přes úvodní fráze. A když už v konverzaci pokročíte, tak přece nepoběžíte zase „o stůl dál“. A přátel je víc, než večerů... A nakonec si stejně stihnete ještě vyrobit nová přátelství, prostě proto, že ti lidi kolem vás vám připadnou prostě skvělí.

Vzpomínka 4: závěr. Poslední večer vyhlašuje hlavní organizátor vítěze studentské soutěže. Oceněných je vícero. Studenti si vyhlašování užívají. Můj spolužák Tomáš Jurczyk, účastník pěti Robustů, uznale pokyvuje hlavou: „Tenhle Robust je naprosto výjimečný! Taková atmosféra, to nemá obdobu!“. A já si v duchu notuju Fabiánovu „robustí melodii“ a zpívám si slova nové sloky: „Němčičky, Jetřichovice, Kurzovní; konečně je ňákej Robust na nejvyšší úrovni“. A říkám si „jo, tenhle Robust byl opravdu na nejvyšší úrovni“.

Dodatek: atmosféru letošního Robustu si můžete připomenout při prohlížení momentek, které pořídil a „vystavil“ Matúš Maciak:

http://www.mmatthew.matfyz.cz/gall_new.php nebo
http://www.karlin.mff.cuni.cz/~maciak (Conference Photos)



















Po řádcích: výlet, přednášky, diskuze a společenský program

Těžká posterová sekce na lehkém uměleckém pozadí

Slavnostní dvacátý Robust 2016, Jeseníky, 11. až 16. září 2016

Obsah

Vědecké a odborné články

Zprávy a informace

Informační bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Vydavatelem je Česká statistická společnost, IČ 00550795, adresa společnosti je Na padesátém 81, 100 82 Praha 10. Evidenční číslo registrace vedené Ministerstvem kultury ČR dle zákona č. 46/2000 Sb. je E 21214. Časopis je na Seznamu recenzovaných neimpaktovaných periodik vydávaných v ČR, více viz server `http://www.vyzkum.cz/`.

The Information Bulletin of the Czech Statistical Society is published quarterly.
The contributions in the journal are published in English, Czech and Slovak languages.

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.