STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 549–558, DOI 10. 21307

REMARKS ON THE ESTIMATION OF POSITION PARAMETERS

Czesław Domański[1]

ABSTRACT

The article contains some theoretical remarks about selected models of position parameters estimation as well as numerical examples of the problem. We ask a question concerning the existence of possible measures of the quality of interval estimation and we mention some popular measures applied to the task. Point estimation is insufficient in practical problems and it is rather interval estimation that is in wide use. Too wide interval suggests that the information available is not sufficient to make a decision and that we should look for more information, perhaps by increasing the sample size.

Key words: estimation, the positional parameters, statistical models

1. Introduction

When it is impossible to state what the level of accuracy of estimation of random variable parameter is, the question arises whether there are any methods which help to determine the distance between the estimator assessment and the real value of parameter. The answer to this question is provided by J. Neyman – the author of the interval estimation (1937). Sometimes the interval we obtain is too wide. Too wide intervals allow us to draw a conclusion that the available information is not sufficient to take a decision, and therefore we need to search for more information, either by widening the scope of research or by running another series of experiments.

The interval estimation includes almost all types of statistical analyses. In public opinion polls, for instance, when we state that 58% of citizens of the Republic of Poland trust the president usually a footnote should be added stating that the poll is biased with „an error of plus or minus 3%". This means that 58% of the interviewees trust the president. As the research was based on a representative sample, the parameter sought is the percentage of all people who think in this way. Due to a small sample size a reasonable "guess" is that the

[1] University of Lodz, Chair of Statistical Methods. E-mail: czedoman@uni.lodz.pl

parameter can encompass the interval 55% (54% minus 3%) to 61% (57% plus 3%).

How should the results of the interval estimation be interpreted? Can probability assumptions be made on the basis of interval estimation? How certain is the researcher that the parameter searched for will be included in a given interval?

Neyman (1935) proposed an accessible way of constructing interval estimation, defining how accurate the estimation is and calling the new procedure „confidence intervals", and the ends of confidence intervals – „confidence limits".

Neyman (1937) went back to the frequency definition of a real probability. In his later works he provided a more detailed explanation of confidence intervals stating that they should be perceived not as an individual conclusion but rather as a process. In the long term the statistician who always calculates 95% confidence intervals will see that in 95% of cases the real value of parameter can be found in the determined intervals. It is worth mentioning that Neyman was right saying that the probability connected with confidence interval was not a probability. It rather represents the frequency of correct conclusions drawn by a statistician using this method over a longer period of time but says nothing about the „accuracy" of the current estimation.

Majority of researchers find 90% or 95% confidence limits and continue as if they were certain that the interval encompassed the real value of parameter.

2. Statistical models

Every statistical analysis of a certain real phenomenon must be based on a mathematical model (i.e. a model expressed in the form of mathematical dependencies where the way of obtaining information was taken into account).

The researcher should aim at a situation where the applied model is a modest description of nature. This means that the functional form of the model should be simple and the number of its parameters and elements as small as possible.

As we know there are no perfect models which perfectly copy the behaviour of the modelled object. Each new observation and an analysis of the discrepancy between the mathematical model and the real object leads to new, more accurate mathematical models. The main reasons for the discrepancy between the model and the modelled phenomenon are as follows (Domański et al. (2014)):

1) the present state of knowledge on the examined phenomenon;
2) high level of dependence of the modelled phenomenon, which prevents the application of the mathematical model encompassing all qualities of the object;
3) variety and changeability of the object's environment where modelling of the real reasons for the object's condition becomes impossible;
4) costs related to the model's application can become a barrier to the model's complexity. It may occur that a simpler model despite being less accurate

turns out to be better, as the profits connected with giving up complicated measurement often exceed the losses resulting from using a less accurate model.

The starting point for our discussions is always a certain random element X (random variable, finite or non-finite series of random variables). Most frequently it will be called an experiment result, a measurement result, an observation result or, simply an observation. The set of all values of the random element X will be denoted by $\mathcal{X}$ and called space $\mathcal{X}$ of the sample. Space $\mathcal{X}$ will be a finite or a countable set, or a certain area in a finite dimensional space R^n.

Let Ω be a set of elementary events and let $\aleph$ be σ - a body of subsets of the set Ω. An ordered triple $(\Omega, \aleph, P)$ is called a **probabilistic space**, where P denotes probability.

Let A be a distinguished σ-body of subsets of the set $X \subset R^n$, and X a measured transformation $(\Omega, \aleph) \to (\chi, A)$. Distribution $P^X(A) = P(X^{-1}(A))$ is a measure on space (χ, A). In statistical problems it is assumed that distribution P belongs to a certain defined class of distributions $\mathcal{P}$ on (χ, A). Knowing the class and having the results of observation of the random variable X, we want to draw correct conclusions about an unknown distribution P. Thus, a mathematical basis for statistical research is a measured space (χ, A) and a family of distributions $\mathcal{P}$. Probabilistic space $(\Omega, \aleph, P)$ plays a subordinate role. The term: a probabilistic space $(\Omega, \aleph, P)$is given, which means that a probabilistic model of a certain phenomenon or experiment is known i.e. we know what are the possible results of the experiment, what events are distinguished and what probabilities are assigned to these events. To sum up, the *a priori* knowledge of the subject of research is given in the form of certain probabilistic models. Probability may result from the very nature of the examined phenomenon or it can be introduced by a researcher.

Let us note that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a family of distributions of probability on a given σ-body of random events in $\mathcal{X}\cdot$

The sample space together with a family of distributions $\mathcal{P}$, i.e. the object:

$$(\chi, \{P_\theta : \theta \in \Theta\}) \tag{1}$$

is called a **statistical model** (statistical space), while representations from $\mathcal{X}$ in R^k – statistics or k-dimensional statistics.

If $\mathbf{X} = (X_1, X_2, ..., X_n)^T$, while $X_1, X_2, ..., X_n$ are independent random variables with a uniform distribution, we will also use a denotation:

$$(\chi, \{P_\theta : \theta \in \Theta\})^n \tag{2}$$

where $\mathcal{X}$ is a set of values of the random variable X (and each of variables $X_1, X_2, \ldots, X_n$) and P_θ is a distribution of the random variable. It is also accepted to use the following terms: $X_1, X_2, \ldots, X_n$ is a sample from distribution P_θ or a sample from population P_θ for a given $\theta \in \Theta$.

3. Confidence intervals for expected value μ

To estimate a certain unknown, real parameter μ we get suitable observations $X_1, \ldots, X_n$ of this value. Each observation $X_j, j = 1, \ldots, n$, was different from μ by a certain random value ε_j (statistical observation error). If nothing is known about the nature of the error ε, then consequently nothing can be said about the size of μ. However, if we can describe the random error ε in terms of the theory of probability, i.e. if we can say something about the distribution of the probability of this random error, then we can in the same terms answer various questions about parameter μ. Thus, the statistical inference becomes a result of the prior knowledge about the parameter and the knowledge obtained from the sample $X_1, \ldots, X_n$.

Let a distribution of random error probability μ be denoted by F; then the sample has a distribution F_μ so that $F_\mu(x) = F(x - \mu)$.

Let us now, on the other hand, analyse four general models of our observations $X_1, \ldots, X_n$.

- Model 1: F is a normal distribution $N(0, \sigma)$ with a known standard deviation σ.
- Model 2: F is a normal distribution $N(0, \sigma)$ with an unknown standard deviation σ.
- Model 3: F is a known distribution with a continuous and strictly ascending distribution function.
- Model 4: F is an unknown distribution with a continuous and strictly ascending distribution function. In this case it seems that „in actual fact we know nothing”, yet it turns out that knowing that the distribution function is continuous and strictly monotonous is sufficient to say something more interesting about the parameter μ, especially when we combine this with data from observation $X_1, \ldots, X_n$.

In the first model the estimation of parameter μ by a mean value from observation

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \tag{3}$$

It is assumed that X has a distribution $N(\mu, \sigma)$, then the mean $\bar{X}_n$ is a random variable with a normal distribution $N(\mu, \sigma/\sqrt{n})$, in other words $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is a random variable with a normal distribution $N(0,1)$ and for an arbitrarily selected $\gamma \in (0,1)$ we get

$$P_\mu\{|\sqrt{n}(\bar{X}_n - \mu)/\sigma| \leq u_{(1+\gamma)/2}\} = \gamma \tag{4}$$

where u_α is a quantile of an order α of a normal distribution $N(0,1)$.

This can be denoted in the form

$$P_\mu\left\{\bar{X}_n - u_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + u_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}\right\} = \gamma \tag{5}$$

and interpreted in the following way: with a selected probability γ, a random interval

$$\left(\bar{X}_n - u_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + u_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}\right) \tag{6}$$

includes the unknown, estimated value of parameter μ.

In the second model the estimation of parameter μ is based on the t Student distribution. In the case under consideration we deal with a random variable

$$\frac{\frac{\bar{X}_n-\mu}{\sigma}\sqrt{n}}{\sqrt{\frac{nS^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}_n-\mu}{S}\sqrt{n-1} \tag{7}$$

with the t Student distribution and with $(n-1)$ degrees of freedom.

The possibility of inference on parameter μ changes, because the random variable $\frac{\bar{X}_n-\mu}{S}\sqrt{n-1}$ with the t Student distribution is more dispersed around zero than the random variable $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ with the normal distribution.

Then, for the estimated parameter μ we get a confidence interval at a given level of confidence γ of the form :

$$\left(\bar{X}_n - t_{n-1}\left(\frac{1+\gamma}{2}\right)\frac{S}{\sqrt{n-1}}, \bar{X}_n + t_{n-1}\left(\frac{1+\gamma}{2}\right)\frac{S}{\sqrt{n-1}}\right) \tag{8}$$

where $t_{n-1}(\alpha)$ is a quantile of order α of the t Student distribution with $n-1$ degrees of freedom.

When the standard deviation σ was known like in the first model, the length of the confidence interval (2d) at the confidence level γ could be expressed with the formula 2 $u_{(1+\gamma)/2}\frac{\sigma}{\sqrt{n}}$ and on this basis the required accuracy of the estimation of parameter μ could be obtained. If the unknown standard deviation σ is replaced with its estimation S, then the length of interval calculated in this way will be random. The problem consists in selecting n, in such a way that the random variable never exceeds the pre-assigned number 2d. There are various methods of solving this problem. The simplest and the most transparent method is the so-called two-stage Stein procedure (1956).

In the third model it is the median $M_n$which is the third estimated position parameter. Median μ of the distribution of observations will be estimated with the use of median M_n from a sample $X_1, \dots, X_n$. According to a generally accepted agreement the median M_n from a sample is expressed by the following formula:

$$M_n = \begin{cases} \frac{1}{2}\left(\mathrm{X}_{\frac{n}{2}:n} + \mathrm{X}_{\frac{n}{2}+1:n}\right), \text{for even n}, \\ \mathrm{X}_{\frac{n+1}{2}:n}, \text{for uneven n}. \end{cases} \tag{9}$$

Let us now analyse the problem of the biasedness of estimator M_n. The basic definition where estimator T is called the unbiased parameter θ if $E_\theta T = \theta$ for every θ, cannot be applied here due to the fact that the median M_n cannot have the expected value. We can introduce the notion of median unbiasedness. We say that estimator T is the median-unbiased estimator of parameter θ if for every θ its median is $Med_\theta T = \theta$. In other words, T is the median-unbiased estimator of parameter θ if

$$P_\theta\{T \le \theta\} = P_\theta\{T \ge \theta\} = \frac{1}{2}, \text{ for every } \theta \tag{10}$$

under the assumption that, similarly to the distribution of observation X, also the distribution of estimator T has a continuous and strictly ascending distribution function, that is an unambiguous median.

If the sample $X_1, \dots, X_n$ has an uneven number of elements n, then the median M_n from the sample is a median-unbiased estimator of median μ of distribution F_μ of observation X. It can be noticed that the distribution function of the k-th position statistics $X_{k,n}$, when the sample comes from a distribution with the distribution function F takes the following form:

$$F_{k,n}(x) = \sum_{j=k}^{n} \binom{n}{j} F^j(x)\big(1 - F(x)\big)^{n-j} \tag{11}$$

Let us recall here the formula combining binominal distribution with beta distribution:

$$\sum_{j=1}^{n} \binom{n}{j} x^j (1-x)^{n-j} = B(x; k, n-k+1) \tag{12}$$

Following from (11) and (12) the distribution function of median M_n is given by the formula:

$$P_\mu\{M_n \le x\} = B\left(F(x-\mu); \frac{n+1}{2}, \frac{n+1}{2}\right), \tag{13}$$

therefore

$$P_\mu\{M_n \le x\} = B\left(F(0); \frac{n+1}{2}, \frac{n+1}{2}\right) = B\left(\frac{1}{2}; \frac{n+1}{2}, \frac{n+1}{2}\right) = \frac{1}{2} \tag{14}$$

In the case of the sample $X_1, \dots, X_n$ with the even number of elements the median M_n, which was defined by formula (9), is not the median-unbiased estimator of the median μ and for some distributions F_μ of observations X the difference between the median of estimator M_n and the median μ can be very significant.

Our considerations are now limited to the case of uneven number of observations n in a sample. For the case like this the distribution of the median from a sample is given by the formula (13).

Now, let $x_\gamma(M_n)$ be the quantile of the order γ of estimator M_n, i.e. such a number that

$$P_\mu\{M_n \le x_\gamma(M_n)\} = \gamma \tag{15}$$

On the basis of (13) we get:

$$x_\gamma(M_n) = \mu + F^{-1}\left(B^{-1}\left(\gamma; \frac{n+1}{2}, \frac{n+1}{2}\right)\right) \tag{16}$$

and hence the unilateral confidence interval on the confidence level γ takes the form:

$$\left(M_n - F^{-1}\left(B^{-1}\left(\gamma; \frac{n+1}{2}, \frac{n+1}{2}\right)\right), +\infty\right). \tag{17}$$

Similarly, taking as a basis the relation

$$P_\mu\left\{|M_n| \le x_{\frac{1+\gamma}{2}}(M_n)\right\} = \gamma, \tag{18}$$

we get a bilateral confidence interval at the confidence level γ:

$$\left(M_n - F^{-1}\left(B^{-1}\left(\frac{1+\gamma}{2}; \frac{n+1}{2}, \frac{n+1}{2}\right)\right), M_n + F^{-1}\left(B^{-1}\left(\frac{1+\gamma}{2}; \frac{n+1}{2}, \frac{n+1}{2}\right)\right)\right) \tag{19}$$

where F is a normal distribution $N(0,\sigma)$.

In the fourth model the confidence interval for median is presented. First, we consider constructing the confidence interval for a quantile $x_q = F^{-1}(q)$ of an arbitrary order $q \in (0,1)$, then the confidence interval for the median is a special case for $q = \frac{1}{2}$.

As we analyse the unilateral interval of the form $(X_{i:n}, +\infty)$ with an assumed level of confidence γ, we should choose index $i \in \{1,2,\dots,n\}$ so that $P_F\{X_{i:n} \le x_q\} \ge \gamma$ for every $F \in \mathcal{F}$. As $X_{i:n} < X_{j:n}$, when $i < j$, it is reasonable to choose the biggest number $i = i(n.\gamma)$ which satisfies the given condition. Making use of the distribution of the i –th position statistics from a sample $X_1, \dots, X_n$, of the form (11), we get:

$$\begin{aligned} P_F\{X_{i:n} \le x_q\} &= P_F\{X_{i:n} \le F^{-1}(q)\} \\ &= \sum_{j=i}^{n} \binom{n}{j} \left(F\left(F^{-1}(q)\right)\right)^j \left(1 - F\left(F^{-1}(q)\right)\right)^{n-j} \\ &= \sum_i^n \binom{n}{j} q^j (1-q)^{n-j}. \end{aligned} \tag{20}$$

The solution is the biggest $i = i(n, q)$ so that

$$\sum_{j=i(n,\gamma)}^{n} \binom{n}{j} q^{j}(1-q)^{n-j} \geq \gamma \tag{21}$$

The confidence interval at the level γ for the quantile of the order $q \in (0,1)$ only exists when

$$\sum_{j=i}^{n} \binom{n}{j} q^{j}(1-q)^{n-j} \geq \gamma \tag{22}$$

i.e. when $(1-q)^{n} \leq 1-\gamma$.

As a conclusion we get the unilateral confidence interval for median $(X_{i:n}, +\infty)$, where $i = i\left(n, \frac{1}{2}\right) \in \{1, \dots, n\}$ is the biggest number such that

$$2^{-n} \sum_{s=i(n,\gamma)}^{n} \binom{n}{s} \geq \gamma \tag{23}$$

Due to the discreteness of the distribution the actual confidence interval

$$\gamma^{*} = 2^{-n} \sum_{j=i(n,\gamma)}^{n} \binom{n}{j} \tag{24}$$

can obviously be bigger than the assumed γ.

The bilateral confidence interval $\left(X_{i:n}, X_{j:n}\right)$ takes the form:

$$P_F\left\{X_{i:n} \leq F^{-1}(q) \leq X_{j:n}\right\} = P_F\{X_{i:n} \leq F^{-1}(q)\} - P_F\{X_{j:n} > F^{-1}(q)\}$$

$$= \sum_{s=1}^{j-1} \binom{n}{s} q^{s}(1-q)^{n-s} \tag{25}$$

and the problem of selection of indexes (i, j) arises, so that

$$\sum_{s=i}^{j-1} \binom{n}{s} q^{s}(1-q)^{n-s} \geq \gamma.$$

An attempt of solving this problem was presented in the work of Zieliński (2011). In our research we assume that:

$$P\left\{X_{i:n} \leq F^{-1}(q) \leq X_{j:n}\right\} = \left(\frac{1}{2}\right)^{n} \sum_{s=1}^{j=1} \binom{n}{s} \approx \gamma.$$

Applications of other estimators are given in the monograph of Lehmann (1991).

4. Assessment of accuracy of position parameters estimation

Let us now follow the obtained results and assess the accuracy of statistical inference in the four models under consideration. The accuracy of inference will be assessed with the use of the width of confidence interval for μ. Obviously, it depends on the distribution F of error and on the size n of the sample $X_1, \dots, X_n$.

Confidence intervals of models (1) and (3) have a deterministic length depending only on n. Half of their length is denoted by D (1) and D (3), respectively. Intervals (2) and (4) have a random length so for further consideration the expected values of their lengths will be taken and denoted by D (2) and D (4), respectively. Then, we get:

$$D(2) = t_{n-1}\left(\frac{1+\gamma}{2}\right)\frac{E(S)}{\sqrt{n-1}}$$

$$E(S) = \sqrt{\frac{2}{\pi}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}.$$

For $D(4)$ we get:

$$D(4) = \frac{1}{2}\left(E_{N(0,1)}X_{j:n} - E_{N(0,1)}X_{i:n}\right),$$

where by $E_{N(0,1)}X_{j:n}$ we denoted the expected value of the j-th position statistics from the sample $X_1, \dots, X_n$, when the sample comes from the standard normal distribution $N(0,1)$.

Table 1. Assessment of accuracy of position parameters estimation

n	γ	D(1)	D(2)	D(3)	D(4)
15	0.90	0.424699	0.462405	0.524439	0.515701
	0.95	0.506061	0.563081	0.625379	0.714877
	0.99	0.665076	0.781524	0.823391	0.947689
25	0.90	0.328971	0.345613	0.408676	0.408597
	0.95	0.391993	0.416926	0.487204	0.463971
	0.99	0.515166	0.565007	0.641052	0.700479
30	0.90	0.300308	0.312812	0.373624	0.382351
	0.95	0.357839	0.376531	0.445383	0.473288
	0.99	0.470280	0.507456	0.585921	0.672498
50	0.90	0.232617	0.238289	0.290265	0.304216
	0.95	0.277180	0.285621	0.345961	0.356962
	0.99	0.364277	0.380902	0.454954	0.494328
100	0.90	0.164485	0.166455	0.205701	0.214301
	0.95	0.195996	0.198918	0.245140	0.252810
	0.99	0.257583	0.263298	0.322272	0.331143

Source: own calculations

The numbers included in Table 1 clearly show a great significance of both the choice of the statistical model and the statistics, that is the estimator of a suitable position parameter (expected value, median or an arbitrary quantile). The statistics serves as a basis for statistical inference on values which are of interest to the researcher. What is particularly striking are the differences in assessment of accuracy of position parameters for sample sizes $n \leq 30$.

5. Final remarks

In any statistical research we have a set statistical observations and some incomplete information about the distribution of these observations.

It is necessary to analyze the questions which we expect to answer by applying a suitable statistical procedure and the initial assumptions that have to be made so that our answers would be justified. A procedure dependent on some prior assumptions impossible to be verified by the observations collected or logically derived cannot be applied here. Statistical methods, therefore, should be treated not as a tool for a given detailed model but rather as an assisting tool to interpret data for different models.

This article presents certain problems connected with the choice of the procedure appropriate for the assumed statistical model along with the verification of its assumptions on the one hand, and the assessment of the data set and their distribution on the other. It is very important to analyze the behaviour of statistical procedures in very varied conditions.

Acknowledgment

The author would like to thank the anonymous referees for suggestions, which helped to improve the text in several points.

REFERENCES

DOMAŃSKI, CZ., PEKASIEWICZ, D., BASZCZYŃSKA, A., WITASZCZYK, A., (2014). Testy statystyczne w procesie podejmowania decyzji, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

LEHMANN, E.L., (1991). Teoria estymacji punktowej, Wydawnictwo Naukowe PWN, Warszawa.

NEYMAN, J., (1935). On the problem of confidence intervals, The Annals of Mathematical Statistics 6, p. 111.

NEYMAN, J., (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. Phil. Trans. Royal Soc. London, A 236, p. 333

STEIN, C., (1956). Efficient nonparametric testing and estimation, Proc. Third Berkeley Symp. Math. Statist. Probab. 1, pp. 187–195.

ZIELIŃSKI, R., (2011). Statystyka matematyczna stosowana. Elementy, Centrum Studiów Zaawansowanych Politechniki Warszawskiej, Warszawa.