On the confidence interval of the extreme value of a second sample from a normal universe

Kudo, Akio
Indian Statistical Institute | Kyushu University

https://doi.org/10.5109/12967

出版情報：統計数理研究. 6 (3/4), pp.51-56, 1956-03. Research Association of Statistical Sciences
バージョン：
権利関係：

ON THE CONFIDENCE INTERVAL OF THE EXTREME VALUE OF A SECOND SAMPLE FROM A NORMAL UNIVERSE

By

Akio KUDÔ

(*Received Feb.* 2, 1956)

§ 1. **Introduction.** Let $x_1, x_2, \ldots, x_n$ be a sample of size n from a normal population with unknown mean value and variance. In this paper, we shall be concerned with the problem of making inference on the extreme values of another independent sample $(y_1, y_2, \ldots, y_m)$ from the same population, namely $y_{\max} = \max(y_1, y_2, \ldots, y_m)$. At this point, we shall adopt the formulation of the two sample theory due to T. KITAGAWA [10].

The current theories of statistical inference would be interested mainly in the inferences only on the unknown parameters of the population from which the sample was drawn, and if we are concerned with the problem concerning the observations in future, namely a second sample to be drawn from the same population, the classical formulation would suggest very often the following round-about way.

(i) First make a statistical inference about the population parameter from the first sample.

(ii) Next give a prediction about a second sample to be drawn from the same population, according to the inference about the population parameter obtained from the first sample.

If a statistician is asked to give answer to our problem, he might give an estimate $\hat{y}$ of the maximum, say, the value obtained by the following formulae

$$\Phi(\hat{y}/\bar{x}, s^2) = 1 - \frac{1}{N}, \tag{1.1}$$

where $\bar{x}$ and s^2 are the sample mean and variance of the first sample and $\Phi(y/\mu, \sigma^2)$ is the normal distribution function with mean μ and variance σ^2. In this case we have

$$\hat{y} = \bar{x} + \lambda_N s, \tag{1.2}$$

where λ is defined by $\Phi(\lambda_N/0,1) = 1 - 1/N$. Although this procedure is by no means correct, some statistician might be tempted to use this in their practical situations.

In this paper, we shall give direct formulae for the tabulation of the confidence interval of the extreme value of a second sample of the following form,

$$P(y_{\max} \leq \bar{x} + \lambda_p(n, m)\, s) = p, \tag{1.3}$$

where n and m denote the sizes of the first and the second sample respectively. In this formulation we are concerned with an inference from the first sample directly to a second one, without taking the round-about procedure just above mentioned.

There does not exist however enough theoretical justification for adopting the statistic $x + \lambda_p(n, m)s$ as the one which gives the confidence interval. There are, however, justifications from an intuitional and practical point of view.

The first of all, there is a similarity between (1.2) and the right hand side of the inequality in (1.3), and secondly this is a function of the sufficient statistics of unknown parameters of the population in our concern. These two facts tempt us to assume that we are making full use of the information contained in the first sample. The further investigation at this point is obviously desirable.

In addition to the intuitional justification for our statistic, this statistic has a practical advantage, as this statistic seems to be the least complicated function of the sufficient statistics, namely $\bar{x}$ and s^2 and this will make quick calculation of the interval possible after observing the first sample possible, provided that extensive tables of $\lambda_p(n, m)$ are available at hand. Concerning the calculation of the $\lambda_p(n, m)$, some preparatory tabulation work has been done, namely the tables of $A_{m.2k}(x)$ and $a(n)$ defined in § 2 in a small scale and the further tabulation is now going on. In this paper we shall give a small scale table of $\lambda_p(n, m)$, which gives some representative values only.

Now let us mention the practical implication of our problem. Suppose we have an electric circuit with a number, say m, of electric valves, and we have some justification to assume that these electric valves are of the same nature and they are equally loaded when the circuit is at work. Our main concern is to have some sort of probability statement about the duration of the work of the circuit, namely the length of the hour before breakage and we concentrate our attention to the duration of the m valves, namely the length of the working hour of these valves simultaneously. This is nothing but the minimum of the lives of m electric valves.

As the life of an electric valve is supposed to be distributed in normal (cf. D. J. DAVIS [2]) this problem can be formulated as a problem of making inference on the extreme value of the future observations from the normal population. As a matter of fact this problem was suggested by T. TAGUCHI in connection with his work in the Institute of Japanese Corporation of Telepohone and Telecommunication.

It may be mentioned that the similar results and tables have been obtained for the case of exponential distribution by the author [12].

§ 2. **The confidence interval of the extreme value of a second sample.** Let $(x_1, \ldots, x_n)$ be a sample of size n from a normal population with unknown mean m and variance σ^2, and let $\bar{x}$ and s^2 be the sample mean and the sample variance. In future we shall have m events with characters $(y_1, y_2, \ldots, y_m)$, each of which is expected to be distributed independently in the same distribution $N(m, \sigma^2)$.

As we have restricted our attention in the confidence limit of the form $\bar{x} + \lambda s$, our main concern here is to evaluate the probability

$$P(y_{\max} \leq \bar{x} + \lambda s) . \tag{2.1}$$

At first we need the following

Theorem 1. (KUDÔ [11]) *Let $(x_1, \ldots, x_n)$ be distributed normally with a common mean value zero and a common variance σ^2, and a common correlation coefficient ρ, then we have*

$$P(\mathrm{Max}\, x_i < x) = \sum_{k=0}^{\infty} \frac{(2k)!}{k!} \left(\frac{\rho\sigma^2}{2}\right)^k ((1-\rho)\sigma^2)^{k+n} A_{n,2k}\left(\frac{x}{\sqrt{1-\rho}\,\sigma}\right), \tag{2.2}$$

where

$$\begin{aligned} A_{n,k}(x) &= \sum_{\nu_1 + \cdots + \nu_n = k} \frac{1}{\nu_1!} \Phi^{(\nu_1)}(x) \frac{1}{\nu_2!} \Phi^{(\nu_2)}(x) \cdots \frac{1}{\nu_n!} \Phi^{(\nu_n)}(x) \\ &= \sum_{\nu=0}^{k} \frac{1}{\nu!} \Phi^{(\nu)}(x) A_{n-1,k-\nu}(x) , \end{aligned} \tag{2.3}$$

and $\Phi^{(\nu)}(x)$ is the ν-th derivative of the standardized normal distribution function.

Proof. The proof is gven in KUDÔ [11].

The tables of $A_{m,2k}(x)$ $(m = 1(1)10,\ k = 0(1)4,\ x = 1.8(0.1)3.5, 5D)$ were prepared in the Indian Statistical Institute and is included in the same paper of the author.

Our formula for the evaluation of the probability (2.1) can be enunciated in the following

Theorem 2. *In the same notations as in Section 1 we have the following approximation formula,*

$$P(y_{\max} < \bar{x} + \lambda s) \cong \sum_{k=0}^{\infty} \frac{(2k)!}{k!} \left(\frac{1}{n} + \lambda^2(1 - a(n)^2)\right)^k A_{m,2k}(a(n)\lambda) , \tag{2.4}$$

where

$$a(n) = \sqrt{\frac{2}{n}}\, \Gamma\left(\frac{n}{2}\right) \Big/ \Gamma\left(\frac{n-1}{2}\right) . \tag{2.5}$$

Proof. Since s is approximately distributed in a normal distribution with

mean $a(n)$ and variance $1-a(n)^2$ unless n is extremely small, say less than 5 (cf. [9]), $y_i-(x+\lambda s)$ $(i=1, 2, \ldots, m)$ is distributed in a multivariate normal distribution with equal correlation, and we can apply Theorem 1 to our problem. In our present case, the variance is $1+1/n+\lambda^2(1-a(n)^2)$, and the covariance $1/n+\lambda^2(1-a(n)^2)$.

§ 3. **A limiting case.** When the size of the second sample is very large, the evaluation of the probability becomes very difficult, as the tabulation of $A_{m,k}(x)$ becomes difficult when m is very large. And we suggest here another approximation, which is given in

Theorem 3. *Under the same notations as in Theorem 2 we have*

$$P_r(y_{\max} \leq x+\lambda s) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-e^{-A_m(2\sigma(\lambda, n)x+\lambda a(n)-A_m)}} e^{-x^2/2} dx, \tag{3.1}$$

where

$$a(n) = \sqrt{\frac{2}{n}}\, \Gamma\left(\frac{n}{2}\right) \Big/ \Gamma\left(\frac{n-1}{2}\right), \tag{3.2}$$

$$\Phi(A_m) = 1-1/m, \tag{3.3}$$

$$\sigma(\lambda, n)^2 = \frac{1}{n} + \lambda^2(1-a(n)^2). \tag{3.4}$$

Proof. In the first place $\bar{x}+\lambda s$ is asymptotically distributed normally with the mean value $a(n)$ and the variance $1-a(n)^2$, and the asymptotic distribution function of $y_{\max}$ is $\exp[-\exp(-A_m(x-A_m))]$ as stated in [5], [6] and [8], where $F(A_m)=1-1/m$.

Since we have

$$P_r(y_{\max} \leq x+\lambda s) = P_r(y_{\max}-x-\lambda s \leq 0), \tag{3.5}$$

the probability is expressible as the following convolution,

$$\frac{1}{\sqrt{2\pi}\sigma(\lambda, n)} \int_{-\infty}^{+\infty} e^{-e^{-A_m(x-A_m)}} e^{-\frac{1}{2\sigma(\lambda, n)^2}(x+\lambda a(n))^2} dx, \tag{3.6}$$

and this is equal to the right hand side of (3.1).

It should be noticed here that the evaluation of the integral is possible by making use of the formula

$$\int_{-\infty}^{+\infty} f(x) e^{-x^2} dx = \sum_{\alpha=1}^{n} \alpha_i^{(n)} f(x_i^{(n)}) + R_n, \tag{3.7}$$

where

$$R_n = \frac{n!}{2^n(2n)!} \pi^{1/2} f^{(2n)}(\xi) \tag{3.8}$$

for some ξ, $-\infty < \xi < \infty$ [14, p 101–102, 369], and $x_i^{(n)}$ and $\alpha_i^{(n)}$ are the zeros and the weight factors of the Hermitian Polynomials, and these values are tabulated by H. E. SALZAR and others [14].

Table of $\lambda_p(n, m)$

$p = 0.90$		
n \ m	5	10
10	2.73	3.05
15	2.51	2 87
20	2.39	2.74
∞	2.04	2.31

$p = 0.95$		
n \ m	5	10
10	3.17	3.57
15	2.98	3.25
20	2.77	3.10
∞	2.06	2.33

§ 4. **Remarks and Acknowledgement.** If the variance of the population were known to us the problem becomes much simpler, and we have

Theorem 4.

$$P_r(y_{\max} < x + \lambda\sigma) = \sum_{k=0}^{\infty} \frac{(2k)!}{k!}\left(\frac{1}{2n}\right) A_{m,2k}(\lambda) . \tag{4.1}$$

Moreover when m is very large, we have approximately

$$P_r(y_{\max} < \bar{x} + \lambda\sigma) \simeq \int_{-\infty}^{+\infty} e^{-e^{-A_m(\lambda+x-A_m)}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx . \tag{4.2}$$

Proof. The proof is the exactly same as to that of Theorem 2 and 3.

As

$$P_r(y_{\max} \leq x + \lambda\sigma) = P_r\left(\frac{y_{\max} - x}{\sigma} \leq \lambda\right), \tag{4.3}$$

and we know the variance-covariance matrix of $(y_1 - x, y_2 - x, \ldots, y_m - x)$, the problem of evaluating the probability,

$$P_r(y_{\max} \leq \bar{x} + \lambda s) = P_r\left(\frac{y_{\max} - x}{s} \leq \lambda\right) \tag{4.4}$$

can be considered to be a special case of the multivariate analogue of Student's t statistic treated by C. W. DUNNRET and M. SOBEL [3], [4]. Along this line the author has got some results which will be discussed elsewhere.

This paper was partly prepared while the author was in the Indian Statistical Institute under its fellowship, and the tables of $A_{n,k}(x)$ were prepared in a small scale by the punched card machines there. The author is deeply greteful to the Indian Statistical Institute for its fellowship and help in computational work made there.

The author also would like to express his thanks to Professor T. KITAGAWA for the suggestions and criticisms given by him.

KYUSYU UNIVERSITY AND INDIAN STATISTICAL INSTITUTE

References.

[1] CRAMÉR, H., Mathematical methods of statistics. Princeton, 1946.
[2] DAVIS, D. J., *An analysis of some failure data.* Jour. Amer. Stat. Ass., **47** (1952), 113-150.
[3] DUNNET, C. W. and SOBEL, M., *A bivariate generalization of Student's t-distribution, with tables for certain cases.* Biometrika, **41** (1954), 153-169.
[4] DUNNET, C. W. and SOBEL, M., *Approximations to the probability integral and percentage points of a multivariate analogue of Student's t-distribution.* Biometrika, **42** (1955), 258-260.
[5] FISHER, R. A., *Limiting form of the frequency distribution of the largest or smallest member of a sample.* Proc. Cam. Phil. Soc., **24** (1928), 180-190.
[6] GUMBEL, E., *Les valueurs extremes des distributions statistique.* Ann. de L'inst. Henri Poincaré, **5** (1935), 115-158.
[7] HARTLEY, H. O., *Studentization.* Biometrika, **33** (1943-46), 173-180.
[8] HOMMA, T., *On the limit distributions of some ranges.* Rep. Stat. Appl. Res., U. J. S. E., **1** (19), 15-26.
[9] JENNET, W. J. and WELCH, B. L., *The control of proportion defective as judged by a single quality charactersitc varying on a continuous scale.* Jour. Roy. Stat. Soc., **6** (1939), 81-88.
[10] KITAGAWA, T., *Successive process of statsitical inference.* 1. Mem. Fac. Sci. Kyushu Univ., Ser. A, **5** (1950), 139-180.
[11] KUDÔ, A., *On the distribution of the maximum value of an equally correlated sample from a normal population.* Submitted to Sankhya.
[12] KUDÔ, A., *On the confidence interval for the future observation from the exponential population.* Submitted to Sankhya.
[13] SALZAR, H. E., ZUCKER, R. and CAPUANO, R., *Table of the zeros and weight factors of the first twenty Hermite Polynomials.* Jour. Res. Nat. Bur. Standard., **48** (1952), 111-116.
[14] SZEGÖ, G., Orthogonal polynomials. Amer. Math. Soc. Colloq. Pub., **23** (1939).