

SHORTER CONTRIBUTIONS

551. 509. 3

A Diagram for a Simple Estimation of Correlation Coefficient

by

M. Ogawara, T. Ozawa and T. Fujita

Meteorological Research Institute

(Received April 2, 1955)

As a simple method for estimating the correlation between two variables x_i and y_i ($i=1, 2, \dots, N$), we often use the relative frequency

$$(1) \quad \hat{p} = k/N,$$

of the coincidence of signs of $x_i - \bar{x}$ and $y_i - \bar{y}$, where \bar{x} and \bar{y} is the sample mean value or median of x_i and y_i respectively. If one of the $x_i - \bar{x}$ and $y_i - \bar{y}$ is zero, half of the number of such couples k_0 should be added to the k , and

$$(2) \quad \hat{p} = (p + k_0/2)/N.$$

For a two-variate normal distribution

$$(3) \quad f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \times \left\{ \frac{(x-m_x)^2}{\sigma_x^2} + \frac{(y-m_y)^2}{\sigma_y^2} - \frac{2\rho(x-m_x)(y-m_y)}{\sigma_x\sigma_y} \right\}\right],$$

it is obvious that there is a one-to-one correspondence between the correlation coefficient ρ and the probability p that $(x-m_x)(y-m_y) > 0$; the relation $\rho = \varphi(p)$ is a monotone increasing function of p and can be numerically determined by K. PEARSON'S table [1]. Therefore, we can estimate the ρ by

$$(4) \quad \hat{\rho} = \varphi(\hat{p}),$$

and the confidence limits of ρ , with a confidence coefficient ε , based on the relative frequency k/N , are given by $\rho_1 = \varphi(p_1)$ and $\rho_2 = \varphi(p_2)$, where $p_1 = p_1(k/N)$ and $p_2 = p_2(k/N)$ are the confidence limits of p with the confidence coefficient ε , which can be graphically determined by the diagram already prepared in our laboratory [2]. Thus, the ρ_1 and the ρ_2 are functions of \hat{p} and N , i.e.

$$(5) \quad \rho_1 = \psi_1(\hat{p}, N), \quad \rho_2 = \psi_2(\hat{p}, N).$$

In Fig. 1, a thick curve is $\hat{\rho} = \varphi(\hat{p})$ and two sets of thin curves are $\rho_1 = \psi_1(\hat{p}, N)$ (upper set) and $\rho_2 = \psi_2(\hat{p}, N)$ (lower set) corresponding to various values of N and the confidence coefficient $\varepsilon = 90\%$.

Fig. 2, which is constructed similarly to Fig. 1 excepting that the confidence coefficient is $\varepsilon = 60\%$, is used to test the significance of the difference between two correlation coefficients*. Let the two correlation coefficients in population

* This diagram is based on the diagram of 60% confidence limits of occurrence probability prepared by Mr. E. SUZUKI in our laboratory.

be ρ and ρ' , and the independently observed relative frequencies of sign coincidence for deviations be $\hat{p}=k/N$ and $\hat{p}'=k'/N'$ and suppose, without loss of generality, that $\hat{p}<\hat{p}'$. Using Fig. 2 we can find the upper limit ρ_2 of ρ and the lower limit ρ_1' of ρ' , then if $\rho_2<\rho_1'$ we can conclude that $\rho<\rho'$ on about 4% level of significance. The reason is as follows. Let us denote the probability of an event A when B is true by $P(A|B)$, and let $R=\varphi(X/N)$ and $R'=\varphi(X'/N')$ be random variables and $\hat{\rho}=\varphi(k/N)$ and $\hat{\rho}'=\varphi(k'/N')$ be observed constant values. Then, if $\hat{\rho}\leq\rho_0\leq\hat{\rho}'$, by the definition of confidence limits,

$$(6) \quad P(R\leq\hat{\rho}, \hat{\rho}'\leq R' | \rho=\rho'=\rho_0) = P(R\leq\hat{\rho} | \rho=\rho_0)P(\hat{\rho}'\leq R' | \rho'=\rho_0) \\ \leq P(R\leq\hat{\rho} | \rho=\hat{\rho})P(\hat{\rho}'\leq R' | \rho'=\hat{\rho}') = \alpha \cdot \alpha = \alpha^2,$$

where $\alpha=(1-\varepsilon)/2$ and since $\varepsilon=60\%$, $\alpha^2=4\%$. Thus, when $\hat{\rho}\leq\rho_0\leq\hat{\rho}'$ the null hypothesis $\rho=\rho'=\rho_0$ is rejected on 4% level of significance. For a value of ρ_0 such that $\rho_0<\hat{\rho}$ or $\rho_0>\hat{\rho}'$, we can prove that the inequality (6) holds also, at least approximately, except in cases where $N\gg N'$ or $N\ll N'$ [3], [4].

We note the following remarks:

- 1) In the use of our diagram, the fundamental assumption is that the set of observed values (x_i, y_i) , $i=1, 2, \dots, N$ is a random sample from a two-variate

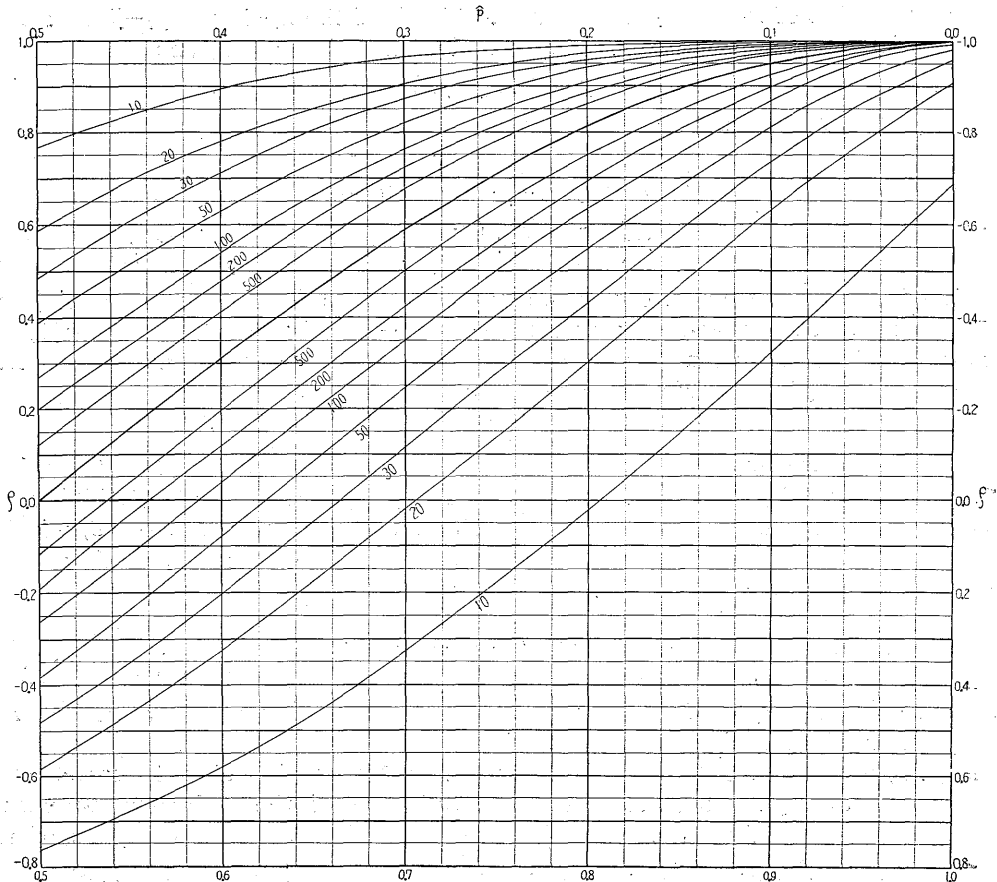


Fig. 1. Estimate and 90% confidence limits of correlation coefficient.

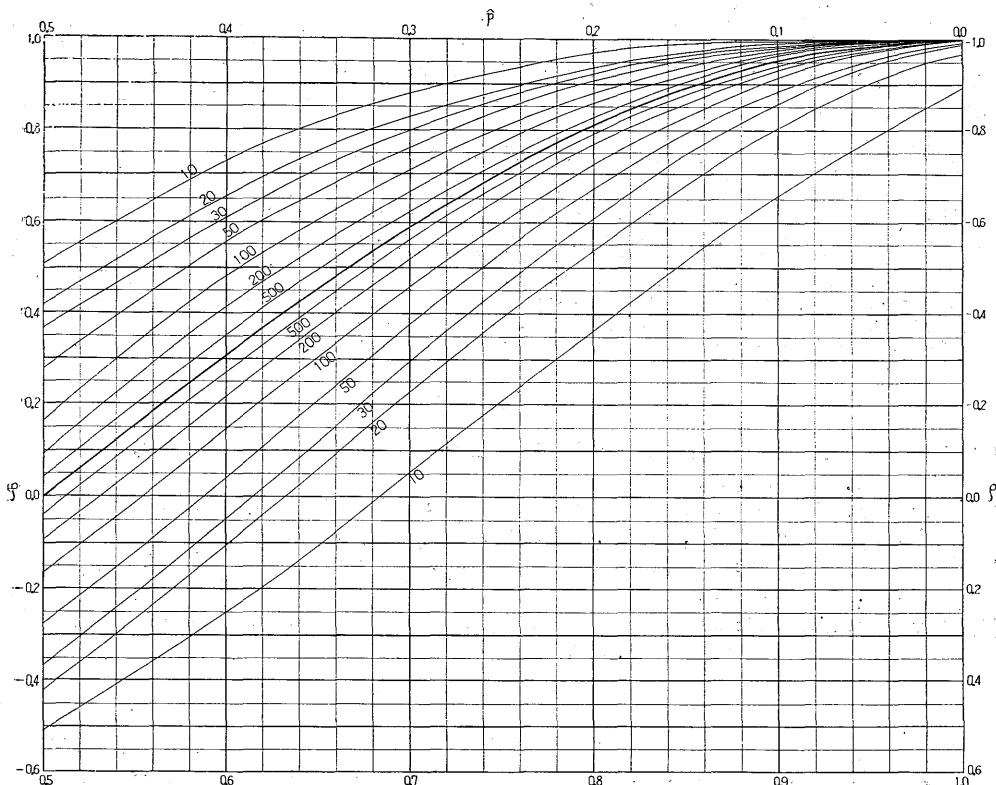


Fig. 2. Diagram for the comparison of two correlation coefficients.

normal population (3).

2) When $\hat{p} \geq 0.50$, the bottom-side and left-hand-side scales should be used and, when $\hat{p} \leq 0.50$, the top-side and right-hand-side scales are used.

3) The error of the curves in the diagrams is, on the whole, 0.01 for both \hat{p} and ρ .

4) It may be a matter of course that the precision of the short-cut estimation by our diagram is inferior to that of the estimation by sample correlation coefficient

$$(7) \quad r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2}}$$

itself; namely, the length of confidence interval of the population correlation coefficient obtained merely by the relative frequency (2) is larger than that obtained from the distribution function of (7) by the whole information of observed values.

Example :

For the monthly mean pressure for January at Haparanda (Sweden) and the monthly mean temperature for July of the same year at Miyako (Japan) we get, for the period 1921~1940 ($N=20$),

$$(8) \quad k=15, k_0=0 \text{ and } \hat{p}=15/20=0.75.$$

From Fig. 1 we have the estimated correlation coefficient $\rho'=0.70$ and the 90%

confidence limits $\rho_1=0.13$, $\rho_2=0.95$.

For comparison, if we calculate the sample correlation coefficient (7) by the same data, we get $r=0.584$, and the 90% confidence limits 0.26, 0.78.

The length of the confidence interval by the simple method is $0.95-0.13=0.82$, while $0.78-0.26=0.52$.

On the other hand, for the period 1887~1920 ($N=34$), we have

$$(9) \quad k'=13, k_0'=2 \text{ and } \hat{\rho}'=(13+1)/34=0.41.$$

From Fig. 1 or Fig. 2 we get $\hat{\rho}'=-0.27$. In order to test the significance of the difference between $\hat{\rho}$ and $\hat{\rho}'$, we use Fig. 2 and we have $\rho_2'=0.08$ and $\rho_1=0.39$.

Since $\rho_2'<\rho_1$, the difference between the correlations in the periods 1887~1920 and 1921~1940 is significant on 4% level of significance.

References

- [1] PEARSON, K., 1931: Tables for Statisticians and Biometricians, Part II. Cambridge Univ. Press. p. 78.
- [2] OGAWARA, M., OZAWA, T. and TOMATSU, K., 1951: Diagrams of Confidence Limits of Occurrence Probability. Journ. Met. Soc. Japan, 29, p. 181.
- [3] OGAWARA, M., 1943: On the Sampling Errors of Correlation Coefficients. Journ. Met. Soc. Japan, 20, p. 298
- [4] OGAWARA, M., 1945: On Confidence Limits. Kō-Kyū-Roku, Institute of Mathematical Statistics, 1, p. 384 and p. 397.

550.35 : 539.16

On the Artificial Radioactivity in the Sea near Japan

by

Y. Miyake, Y. Sugiura and K. Kameda

Meteorological Research Institute

(Received May 28, 1955)

1. Introduction

As a result of Japanese Bikini Expedition aboard the "Shunkotsu-maru" remarkable radioactivity was found around Bikini Atoll [1]. The investigation of radioactivity in sea water near Japan was done next, because the North Equatorial Current on which the main activity flowed has undoubtedly some relation to Kuroshio Current flowing off the coast of Japan.

Sampling of sea water was made during the period from July to September in 1954 by members aboard the following survey vessels: The "Ryōfū-maru" of Central Meteorological Observatory, the "Shumpū-maru" of Kōbe Marine Observatory, the "Meiyō-maru" of Hydrographic Office, the "Shin-yō-maru" of Tokyo University of Fisheries and the "Atsumi-maru" of Maritime Safety Agency.

2. Sampling and measurement

Samples of sea water were collected at positions approximately between 30°N ~35°N and 131°E~140°E. Besides, samples were also collected by several fishing