STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

S 152

Abstract from the Proceedings of the International Mathematical Mathematical Congres, Amsterdam, Sept. 1954.

V9 David Johannes Stoker

1954

Abstract from the Proceedings of the International Mathematical Congress Amsterdam, Sept. 1954.

AN UPPER BOUND FOR THE DEVIATION FROM NORMALITY OF WILCOXON'S TEST STATISTIC FOR THE TWO-SAMPLE PROBLEM IN THE GENERAL CASE

DAVID JOHANNES STOKER

Given two independent random samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ which are drawn from two populations with (cumulative) distribution functions F and G respectively. Wilcoxon's two-sample test is based on the statistic $\boldsymbol{U}$ defined by the number of pairs (i, j) $(i = 1, \ldots, m; j = 1, \ldots, n)$ with $\boldsymbol{y}_j < \boldsymbol{x}_i$ together with half the number of pairs (i, j) with $\boldsymbol{y}_j = \boldsymbol{x}_i$ (cf. Hemelrijk).

If we define

$$\theta = \mathrm{P}(\boldsymbol{y} < \boldsymbol{x} \mid F, G) + \tfrac{1}{2}\mathrm{P}(\boldsymbol{y} = \boldsymbol{x} \mid F, G) = \tfrac{1}{2}\int_{-\infty}^{\infty} (G(x+0) + G(x-0))\,\mathrm{d}F(x)$$

$$\Gamma_r = \int_{-\infty}^{\infty} \{\tfrac{1}{2}(G(x+0) + G(x-0)) - \theta\}^r\,\mathrm{d}F(x)$$

$$\Lambda_r = \int_{-\infty}^{\infty} \{1 - \tfrac{1}{2}(F(y+0) + F(y-0)) - \theta\}^r\,\mathrm{d}G(y)$$

$$\sigma_1^2 = (m+n)^{-1}\{m\Lambda_2 + n\Gamma_2 + \theta(1-\theta) - (\Lambda_2 + \Gamma_2)\}$$

$$\sigma_2^2 = (m+n)^{-1}(m\Lambda_2 + n\Gamma_2),$$

then we have by means of one of Berry's estimates

$$\sup_{-\infty<\xi<\infty} \left| \mathrm{P}[\boldsymbol{U} - \mathscr{E}(\boldsymbol{U} \mid F, G) \leqq \xi\sigma_{\boldsymbol{U}} \mid F, G] - \Phi(\xi) \right| \leqq \Delta(m, n; F, G)$$

where

$$\Phi(\xi) \overset{\text{def}}{=} (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\xi} \mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x$$

$$\sigma_{\boldsymbol{U}^2} \overset{\text{def}}{=} \mathrm{var}\,(\boldsymbol{U} \mid F, G) = \sigma_1^2 \qquad \text{(cf. D. van Dantzig)}$$

and

$$\Delta(m, n; F, G) \doteq \frac{1.88 \max\,(n\sqrt{\Gamma_4/\Gamma_2};\ m\sqrt{\Lambda_4/\Lambda_2})}{\sqrt{mn(m+n)} \cdot \sigma_2} + \min_{\delta} \left[\frac{\sigma_1^2 - \sigma_2^2}{\delta^2} + \frac{1}{(2\pi)^{1/2}\sigma_2}\{\delta + \max\,(\delta, \sigma_1 - \sigma_2)\} \right].$$

Thus $\Delta(m, n; F, G)$ tends to 0 when m and n both tend to infinity, no matter how, unless σ_2 is equal to 0 or tends to 0.

Although better results have been obtained, we only state here the above upper bound because of its simplicity. Also further improvements of the upper bound have been obtained for values of ξ far from 0.

The method used to obtain an upper bound for the deviation from normality of Wilcoxon's statistic, can also be applied to other statistics of the class of statistics introduced by Hoeffding (1948) and Lehmann (1951).

[1] A. C. BERRY, The accuracy of the Gaussian approximation to the sum of independent variates. Trans. Amer. Math. Soc., **49** (1941), 122—136.

[2] D. VAN DANTZIG, On the consistency and the power of Wilcoxon's two sample test, Proc. Kon. Ned. Ak. van Wetensch., **A54** (1951), 3—10; Indagationes Mathematicae **13** (1951), 1—8.

[3] C. G. ESSEEN, Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law, Acta Math, **77** (1945), 1—125.

[4] J. HEMELRIJK, Note on Wilcoxon's two sample test when ties are present, Ann. Math. Stat. **23** (1952), 133—135.

[5] W. HOEFFDING, A class of statistics with asymptotically normal distributions, Ann. Math. Stat. **19** (1948), 293—325.

[6] E. L. LEHMANN, Consistency and unbiasedness of certain non-parametric tests, Ann. Math. Stat. **22** (1951), 165—179.

LEKSTRAAT 73III,
AMSTERDAM.