CONFIDENCE LIMITS FOR THE FRACTION OF A NORMAL POPULATION WHICH LIES BETWEEN TWO GIVEN LIMITS[1]

By J. Wolfowitz

Columbia University

Summary. Let μ and σ^2 be the unknown mean and variance, respectively, of a normally distributed population on which N independent observations $x_1, \cdots, x_N$ have been made. Let L_1 and L_2, $L_1 < L_2$, and α, $0 < \alpha < 1$, be given constants. We define the following symbols:

(a) $$\gamma = (\sqrt{2\pi}\,\sigma)^{-1} \int_{L_1}^{L_2} \exp\left\{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right\} dy$$

(b) $$\bar{x} = N^{-1}\Sigma x_i$$

(c) $$s^2 = (N-1)^{-1}\Sigma(x_i - \bar{x})^2$$

(d) $\chi^2_{1-\alpha}$ as that number for which $P\{\chi^2 < \chi^2_{1-\alpha}\} = 1 - \alpha$ where χ^2 has $N - 1$ degrees of freedom.

(e) $$w = \sqrt{N-1}\,\frac{s}{\chi_{1-\alpha}}$$

(f) $$D = (2\pi)^{-1/2} \int_{(L_1-\bar{x})/w}^{(L_2-\bar{x})/w} \exp\left\{-\frac{1}{2}y^2\right\} dy$$

It is proved that, under restrictions stated precisely below, and before the observations are made, the probability that $D \leq \gamma$ differs from α by a number which can be made arbitrarily small by making N sufficiently large. Thus an approximate (large sample) lower confidence limit for γ is obtained. Similar methods can be applied to obtain upper and two-sided confidence limits.

A problem raised by the present paper (but not attacked here) is to investigate the rapidity of approach to α of $P\{D \leq \gamma\}$. It would perhaps be useful to obtain a series for the latter in powers of $N^{-\frac{1}{2}}$; the first term of such an expansion is obtained here.

---

[1] Formula (5.1) of the present paper was given without proof by the author in July, 1945, in solution of a problem put to him by Dr. M. A. Girshick. At the time, both were members of the Statistical Research Group, formed in the Division of War Research of Columbia University under contract with the National Defense Research Committee of the Office of Scientific Research and Development. The validation of formula (5.1) in all rigor as it is given in the present paper was constructed by the author after he was no longer a member of the Statistical Research Group.

In January, 1945, Professor A. Wald, then a consultant to the Statistical Research Group, and the present author jointly submitted to the Group an unpublished memorandum (#410) entitled "Acceptance Regions Which Involve the Normal Distribution and Large Sample Sizes." While this memorandum dealt with a different problem, its ideas were logically antecedent to formula (5.1). The present author wishes to express his indebtedness to this memorandum and to his colleague Professor Wald.

1. The problem. Let μ and σ^2 be the unknown mean and variance, respectively, of a normally distributed population on which the N independent observations $x_1, x_2, \cdots, x_N$ have been made. Let L_1 and L_2 be given constants with $L_1 < L_2$. We then have that

$$\gamma = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{L_1}^{L_2} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\} dy$$

is the fraction of the normal population which lies between L_1 and L_2. The problem considered in this paper is to construct a lower confidence limit for the unknown γ, when N is large. An upper confidence limit or two-sided confidence limits may be constructed in a manner very similar to that described in the present paper. Since the construction of a lower limit is the problem which occurs most often in practice the discussion will be centered on it.

A lower (confidence) limit on γ with confidence coefficient α is a function $D(x_1, \cdots, x_N)$ of the observations $x_1, \cdots, x_N$ with the property that, *before* the observations are made, the probability is α that $D(x_1, \cdots, x_N) \leq \gamma$. In any specific application it is unknown whether this last inequality holds, because γ is unknown. However, one who proceeds as if this inequality were true is using a procedure which will give correct results $100\alpha\%$ of the time in the long run.

When either $L_1 = -\infty$ or $L_2 = +\infty$ the solution, by use of the non-central t distribution, is well known. For a description of the procedure and necessary tables the reader is referred to [1].

2. Acceptance regions. Let γ_0 be any value of the parameter γ. To γ_0 there correspond infinitely many couples (μ, σ) with the property that the normal distributions characterized by these couples all have a fraction γ_0 lying between L_1 and L_2; we may write this symbolically by saying that the couples (μ, σ) satisfy

$$\gamma(\mu, \sigma) = \gamma_0. \tag{2.1}$$

The construction of confidence regions is equivalent to the construction, for every γ_0, of an acceptance region $R(\gamma_0)$ in the N-dimensional Euclidean space, with the property that every normal distribution whose parameters μ and σ satisfy (2.1) assigns to $R(\gamma_0)$ the constant probability α. While this property of similarity (cf. [2]) is sufficient for the construction of confidence regions, additional properties of the acceptance regions $R(\gamma_0)$ are needed in order that the confidence region be an interval or that the upper confidence limit be always one (i.e., that the confidence limits turn out to be a lower limit only), or to insure other features deemed desirable.

It is easy to construct acceptance regions which will fulfill the condition of similarity. As an example, consider the case $N = 3$ for convenience. Let b_1, b_2, b_3 be a number triple such that $b_1 + b_2 + b_3 = 0$. Let $R(\gamma_0)$, for any given γ_0, $0 < \gamma_0 < 1$, consist of all the points x_1, x_2, x_3 which are such that the absolute value of the angle $\psi(-\pi \leq \psi \leq \pi)$ between the vector (b_1, b_2, b_3) and the vector

$(x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x})$ does not lie between $\pi\alpha\gamma_0$ and $\pi + \pi\alpha(\gamma_0 - 1)$. (We define, in general, $\sum_1^N x_i = N\bar{x}$. The points (x_1, x_2, x_3) for which $x_1 = x_2 = x_3$ may be disregarded, since their probability is zero when the distribution is continuous.) One readily verifies that the probability of $R(\gamma_0)$ for any γ_0 is α, no matter what μ and σ are, and hence this is true in particular for the pairs which satisfy (2.1).

The above method of constructing acceptance regions yields confidence regions which, while they cover the unknown γ with confidence coefficient α, are not very meaningful otherwise. The fact that the probability of $R(\gamma_0)$ is α whether or not (μ, σ) satisfies (2.1) is already indicative of their lack of discrimination. Since $\bar{x}$ and s (where s is defined by

$$ns^2 = \sum_1^N (x_i - \bar{x})^2$$

and $n = N - 1$) are sufficient estimates of μ and σ, which in turn determine γ, it is clear that desirable confidence regions should be functions only of $\bar{x}$ and s. Consequently our first task must be to construct the acceptance regions $R(\gamma_0)$ in the $\bar{x}$, s plane. In the present paper we construct in the $\bar{x}$, s plane regions $R(\gamma_0)$ which have the property that their probability, under any normal distribution whose parameters satisfy (2.1), differs from the prescribed α by a quantity which is bounded in absolute value for all γ_0, in such a way that the bound approaches zero as N increases. Thus when the sample number is sizeable we can obtain confidence regions for γ which correspond to a confidence coefficient which differs little from α. Finally, the acceptance regions $R(\gamma_0)$ which we shall construct will be such that the confidence region will be always an interval, and the upper limit will always be 1, i.e., we will construct a lower confidence limit for γ.

3. Construction of regions $R(\gamma_0)$ in the $\bar{x}$, s plane. First we describe two assumptions which we shall make. It is believed that these are reasonable from the practical standpoint and are satisfied in most actual investigations where the present problem arises. Mathematically their purpose is to enable us to secure a *uniform* bound on the difference between α and the probability of $R(\gamma_0)$ (for *all* γ_0) under all couples (μ, σ) which satisfy (2.1).

ASSUMPTION 1: *There exists a positive d such that*

$$L_1 + d < \mu < L_2 - d.$$

In most practical cases where the present problem will occur γ will be larger than $\frac{1}{2}$. If the latter is the case and μ were very near either L_1 or L_2, then σ would have to be very small. In that case other methods would have to be used in the solution of the practical problem. The present paper deals with the situation, unfortunately only too common in practice, where σ is not too small. Assumption 1 puts a lower bound on σ for any given value γ_0. (The bound is a function of γ_0).

ASSUMPTION 2: *The standard deviation σ is less than a positive number C.*

In most practical problems such an upper bound can reasonably be set. Naturally, the larger d and the smaller C the more a priori information is at our disposal, the closer are our approximations and the narrower our limits. The effect of Assumptions 1 and 2 is to place a lower limit G on γ where

$$G = \gamma(L_1 + d, C) = \gamma(L_2 - d, C).$$

Let γ_0 be any positive number such that $G < \gamma_0 < 1$. For an $\bar{x}$ such that $L_1 < \bar{x} < L_2$, let $r(\bar{x}, \gamma_0)$ be the positive number such that

$$\gamma(\bar{x}, r(\bar{x}, \gamma_0)) = \gamma_0 .$$

We define $\chi^2_{1-\alpha}$ to be that number for which

$$P(\chi^2 < \chi^2_{1-\alpha}) = 1 - \alpha,$$

where χ^2 has n degrees of freedom and P is the probability of the relation in parentheses. The number $\chi^2_{1-\alpha}$ may be found in tables of the χ^2-distribution if the value of α is one of those in common use. Finally define

$$\varphi(\bar{x}, \gamma_0) = r(\bar{x}, \gamma_0) \sqrt{\frac{\chi^2_{1-\alpha}}{n}}$$

The acceptance regions $R(\gamma_0)$, $G < \gamma_0 < 1$, which we shall employ, are defined as follows for any γ_0, $G < \gamma_0 < 1$:

$$L_1 \leq \bar{x} \leq L_2$$

$$s \geq \varphi(\bar{x}, \gamma_0).$$

4. Proof that $P\{R(\gamma_0)\} \sim \alpha$. This section will be devoted to a proof of the following:

THEOREM. *Let $R(\gamma_0)$ be as defined in Section 3 for $G < \gamma_0 < 1$. Let the assumptions 1 and 2 of Section 3 be fulfilled. Then the absolute value of the difference between α and the probability of $R(\gamma_0)$ under any couple (μ, σ) which satisfies (2.1) is less than any arbitrarily small positive ϵ when N is sufficiently large, i.e., when N is sufficiently large,*

$$| P\{R(\gamma_0)\} - \alpha | < \epsilon$$

uniformly for all (μ, σ) which satisfy (2.1) with $G < \gamma_0 < 1$, and which fulfill Assumptions 1 and 2.

LEMMA 1. $\dfrac{\partial r(\bar{x}, \gamma_0)}{\partial \bar{x}}$ *exists in the open interval $L_1 < \bar{x} < L_2$.*

PROOF: We have

$$\gamma_0 = \frac{1}{\sqrt{2\pi} r(\bar{x}, \gamma_0)} \int_{L_1}^{L_2} \exp\left\{-\frac{1}{2}\left(\frac{y - \bar{x}}{r(\bar{x}, \gamma_0)}\right)^2\right\} dy$$

$$= \frac{1}{\sqrt{2\pi}} \int_{(L_1 - \bar{x})/r}^{(L_2 - \bar{x})/r} \exp\left\{-\frac{1}{2} y^2\right\} dy.$$

Differentiating with respect to $\bar{x}$ we obtain, since $r > 0$,

$$e^{-R^2/2}\left(1 + R\frac{\partial r}{\partial \bar{x}}\right) = e^{-T^2/2}\left(1 + T\frac{\partial r}{\partial \bar{x}}\right)$$

with

$$R = \frac{L_2 - \bar{x}}{r} \qquad T = \frac{L_1 - \bar{x}}{r}.$$

Hence

$$\frac{\partial r}{\partial \bar{x}} = -\left(\frac{e^{-R^2/2} - e^{-T^2/2}}{Re^{-R^2/2} - Te^{-T^2/2}}\right). \tag{4.1}$$

Since $R > 0$ and $T < 0$ within the open interval $L_1 < \bar{x} < L_2$, it follows that $\frac{\partial r}{\partial \bar{x}}$ exists in the entire open interval.

LEMMA 2. *In the open interval* $L_1 < \bar{x} < L_2$,

$$\frac{\partial P\,\{s \geq \varphi(\bar{x}, \gamma_0)\}}{\partial \bar{x}}$$

exists.

PROOF: We have, with k a suitable constant,

$$P = k\int_{\sqrt{n}\varphi/\sigma}^{\infty} y^{n-1}e^{-y^2/2}\,dy = k\int_{(r(\bar{x},\gamma_0)/\sigma)\chi_{1-\alpha}}^{\infty} y^{n-1}e^{-y^2/2}\,dy.$$

Hence

$$\frac{\partial P}{\partial \bar{x}} = \frac{-k\chi_{1-\alpha}}{\sigma}\frac{\partial r}{\partial \bar{x}}\left(\frac{r\cdot\chi_{1-\alpha}}{\sigma}\right)^{n-1}\exp\left(\frac{-r^2\chi_{1-\alpha}^2}{2\sigma^2}\right). \tag{4.2}$$

LEMMA 3. *Let δ be any arbitrarily small positive number. The function* $\left|\frac{\partial P}{\partial \bar{x}}\right|$ *of $\bar{x}$ and γ_0 is bounded for $L_1 + \delta \leq \bar{x} \leq L_2 - \delta$, $G < \gamma_0 < 1$.*

PROOF: From (4.1) we have

$$\left|\frac{\partial r}{\partial \bar{x}}\right| < \frac{e^{-R^2/2} + e^{-T^2/2}}{Re^{-R^2/2} - Te^{-T^2/2}} \leq \max.\left(\frac{1}{R}, \frac{-1}{T}\right) = \max.\left(\frac{r}{L_2 - \bar{x}}, \frac{r}{\bar{x} - L_1}\right) \leq \frac{r}{\delta}.$$

Therefore from (4.2) we have that $\left|\frac{\partial P}{\partial \bar{x}}\right|$ is less than a constant multiplied by $\left(\frac{r}{\sigma}\right)^n \exp\left(\frac{-r^2\chi_{1-\alpha}^2}{2\sigma^2}\right)$ and is therefore bounded.

PROOF OF THE THEOREM: From Lemma 3 and the Theorem of the Mean it follows that, in the closed interval

$$L_1 + \frac{d}{2} \leq \bar{x} \leq L_2 - \frac{d}{2},$$

the function $P\{s \geq \varphi(\bar{x}, \gamma_0)\}$ is uniformly continuous in $\bar{x}$ uniformly for all (μ, σ) which satisfy (2.1) with $G < \gamma_0 < 1$. Hence for every positive ϵ_1 there exists a positive $\eta < \frac{d}{2}$ such that $| l_1 - l_2 | < \eta$,

$$L_1 + \frac{d}{2} \leq l_1, l_2 \leq L_2 - \frac{d}{2},$$

implies

$$| P\{s \geq \varphi(l_1, \gamma_0)\} - P\{s \geq \varphi(l_2, \gamma_0)\} | < \epsilon_1 .$$

For fixed arbitrary $\epsilon_2 > 0$ we have, when N is sufficiently large,

$$P\{ | \bar{x} - \mu | < \eta\} > 1 - \epsilon_2 ,$$

from Assumption 2 and the stochastic convergence of $\bar{x}$. Now

$$P\{s \geq \varphi(\mu, \gamma_0)\} = \alpha.$$

Hence, when N is sufficiently large,

$$| P\{R(\gamma_0)\} - \alpha | \leq \epsilon_1(1 - \epsilon_2) + \epsilon_2 \leq \epsilon_1 + \epsilon_2 .$$

Since ϵ_1 and ϵ_2 are arbitrarily small, this proves the desired result.

5. Construction of large sample confidence regions. The acceptance regions $R(\gamma_0)$ whose size never differs from α by more than a uniform bound which approaches zero as N increases, readily yield a lower confidence limit for γ (within the approximation involved). The confidence region consists of all the γ_0 for which $R(\gamma_0)$ contains the observed $\bar{x}$, s. Our acceptance regions $R(\gamma_0)$ are so constructed that, if $\gamma_1 < \gamma_2$, $R(\gamma_1)$ is entirely contained within $R(\gamma_2)$. Hence the confidence region is an interval, one end of which is always unity, as was desired. The rule for constructing the lower confidence limit D is, therefore, as follows:

a) if $\bar{x} < L_1$ or $\bar{x} > L_2$, then $D = G$

b) if $L_1 \leq \bar{x} \leq L_2$, then

$$D = \frac{1}{\sqrt{2\pi}} \int_{(L_1 - \bar{x})/w}^{(L_2 - \bar{x})/w} \exp\{-\tfrac{1}{2}y^2\}\, dy \tag{5.1}$$

where

$$w = \sqrt{N-1} \cdot \frac{s}{\chi_{1-\alpha}}.$$

(The value of D may be found in a table of the normal distribution. It is easy to see that $s = \varphi(\bar{x}, D)$, i.e., D is the smallest value of γ_0 for which $\bar{x}$, s will still lie in $R(\gamma_0)$).

If the statement $D \leq \gamma$ is made in a large number of cases, where the assumptions are fulfilled and the sample size is large, the proportion of correct statements will be close to α.

REFERENCES

[1] N. L. Johnson and B. L. Welch, "Applications of the non-central t-distribution," *Biometrika*, Vol. 31, Parts III and IV (1940), pp. 362–389.

[2] J. Neyman, "Outline of a theory of estimation," *Phil. Trans. Roy. Soc. London*, series A, Vol. 236 (1937), pp. 333–80.