

TOLERANCE INTERVALS FOR LINEAR REGRESSION

W. ALLEN WALLIS
UNIVERSITY OF CHICAGO

1. Introduction

Elementary textbooks frequently give the impression that lines drawn parallel to a least squares linear regression at a distance, measured in the direction of the dependent variable, equal to the standard error of estimate will include about 68 per cent of future observations from the same population, that lines at a distance equal to three times the standard error of estimate will include 99.7 per cent, and so forth.

More specifically, let y be a normally distributed random variable whose variance is σ^2 and whose mean ψ is a linear function of a second variable, x :

$$(1) \quad \psi = a + \beta x.$$

From a sample of N independent observations, (x_i, y_i) , maximum likelihood estimates of a and β are:

$$(2) \quad a = \bar{y} - b\bar{x},$$

$$(3) \quad b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

where $\bar{y} = \sum y/N$, $\bar{x} = \sum x/N$, and the summations, like all that follow in this paper, run over all N values of x or y . Then the estimated mean Y of y for any value of x is given by the regression line

$$(4) \quad Y = a + bx.$$

The estimate of σ , called the standard error of estimate and sometimes denoted by $s_{y,x}$, is given by

$$(5) \quad s^2 = \frac{\sum (y - Y)^2}{N - 2} = \frac{\sum y^2 - N\bar{y}^2 - b \sum (x - \bar{x})(y - \bar{y})}{N - 2}.$$

In these terms, the implication often given by elementary textbooks is that, whatever Y and s may be,

$$(6) \quad A = Pr(Y + K_\epsilon s > y > Y - K_\epsilon s) = \epsilon$$

where K_ϵ is that number which a unit normal deviate exceeds in absolute value with probability $1 - \epsilon$; that is, K_ϵ is defined by

$$(7) \quad \frac{1}{\sqrt{2\pi}} \int_{-K_\epsilon}^{+K_\epsilon} e^{-t^2/2} dt = \epsilon,$$

ϵ being between 0 and 1. But since A , through Y and s , is a function of the N random observations (x_i, y_i) , and is therefore a random variable, it is clear that neither (6) nor any statement which puts A equal to a constant can be correct. A correct statement analogous to (6) could, of course, be made by replacing the sample statistics a , b , and s by the population parameters α , β , and σ :

$$(8) \quad Pr(\psi + K_*\sigma > y > \psi - K_*\sigma) = \epsilon.$$

The error underlying (6) is one whose persistence is surprising, considering that this year marks a full quarter of a century since the appearance of Fisher's *Statistical Methods for Research Workers*, which so strongly emphasizes the necessity of distinguishing clearly between population parameters and sample estimates of them.

Pairs of limits within which a specified proportion of the observations in some population may be expected to lie have come to be called tolerance limits, following Shewhart,¹ though I prefer the term *tolerance intervals* if both limits are finite. It is clear that when such intervals are estimated on the basis of a random sample the proportion included is a random variable, and the most that can be asserted is something about the distribution of this random variable, for example, that there is a certain probability of including at least a specified proportion of the population. Thus, if

$$(9) \quad A = Pr(Y + ks > y > Y - ks)$$

where k is some constant, the most we can hope is to select k in such a way that

$$(10) \quad Pr(A \geq P) = \gamma$$

where γ is a specified confidence coefficient and P is the proportion of the population we desire to include within the interval.

2. The Wald-Wolfowitz approximation

Wald and Wolfowitz [7] have shown how values of k may be determined to an extremely good approximation when P and γ are specified. Bowker [1] has given an approximation to their formula which simplifies computations, and he has provided [2] an extensive table of values of k , called tolerance factors, for the case of a simple normal population,² that is, the situation analogous to (1) but with β known to be zero.

This case, in which a random sample of N is drawn from a single normal population of unknown mean and variance, is the only one considered by Wald and Wolfowitz, but it is a simple matter to extend their results to cover any normally distributed variable for whose mean we have a normally distributed estimate with variance σ^2/N' and for whose variance we have an estimate independently distributed as $\sigma^2\chi^2/n$ for n degrees of freedom. We shall call N' the *effective number of observations*; thus, the effective number of observations for a certain statistic is that

¹ Bowker [2] discusses the meaning of this term, contrasts tolerance limits with various other limits commonly estimated in statistics, and gives a good bibliography on tolerance limits.

² Bowker's table [2] shows tolerance factors to 3 decimal places for $P = 0.75, 0.90, 0.95, 0.99$ and 0.999 ; $\gamma = 0.75, 0.90, 0.95$ and 0.99 ; and $N = 2(1)102(2)180(5)300(10)400(25)750(50)1000$.

number which, when divided into the variance of an observation, gives the variance of the statistic.

Without assuming any connection between N' and n , the Wald-Wolfowitz derivation of tolerance factors may be carried through with negligible alterations. For expository purposes, we summarize the derivation, referring those interested in more detail to the original paper [7]. The summary will be in general terms and we will return later to the case of linear regression as an application.

Given a statistic Z having the following characteristics:

- (i) It is normally distributed,
 - (ii) Its expected value ζ is regarded as the mean of a normal population of unknown variance σ^2 ,
 - (iii) It has sampling variance σ^2/N' , where N' is known,
- and given an independent estimate s^2 of σ^2 which is distributed as $\sigma^2\chi^2/n$ for n degrees of freedom, the problem is to find that value of k for which

$$(11) \quad Pr(A \geq P) = \gamma$$

where γ is the required confidence coefficient, P is the proportion of the population required to be included within the interval $Z \pm ks$, and A is the proportion of the population actually included in a given interval:

$$(12) \quad A = \frac{1}{\sigma \sqrt{2\pi}} \int_{Z-ks}^{Z+ks} e^{-(z-\zeta)^2/2\sigma^2} dz.$$

The distribution of A is clearly independent of ζ and σ , since ζ merely determines the point about which Z will be distributed and the sampling variance of s is proportional to σ ; so we may without loss of generality take $\zeta = 0$ and $\sigma = 1$ in our further computations.

The probability of A 's exceeding P depends on P , k , N' , and n ; to emphasize the dependence on P and k for given N' and n , we write

$$(13) \quad F(P, k) = Pr(A \geq P).$$

Also, we denote the conditional probability of A 's exceeding P for a particular value of Z by $F(P, k|Z)$.

If $F(P, k|Z)$ is known, $F(P, k)$ may be found by forming the product $F(P, k|Z) \cdot \sqrt{N'/2\pi} \exp(-\frac{1}{2}N'Z^2)dZ$, representing the probability that Z will lie in a particular interval of length dZ and A will exceed P , and integrating out Z ; the result is, of course, also equal to the expectation of $F(P, k|Z)$, since the random variable $F(P, k|Z)$ has been multiplied by its probability density and integrated. In other words,

$$(14) \quad F(P, k) = \sqrt{\frac{N'}{2\pi}} \int_{-\infty}^{\infty} F(P, k|Z) e^{-N'Z^2/2} dZ = EF(P, k|Z).$$

$F(P, k)$ can be approximated, therefore, by expanding $F(P, k|Z)$ in a Taylor's series and taking expectations. (Wald and Wolfowitz have verified the validity of the Taylor's expansion.)

Since $F(P, k|Z)$ is an even function of Z , its odd derivatives are zero, and the Taylor's expansion about $Z = 0$ is

$$(15) \quad F(P, k|Z) = F(P, k|0) + \frac{Z^2}{2!} \frac{\partial^2 F}{\partial Z^2} + \frac{Z^4}{4!} \frac{\partial^4 F}{\partial Z^4} + \dots$$

all derivatives to be evaluated at $Z = 0$. Taking expectations, we find

$$(16) \quad F(P, k) = EF(P, k|Z) \\ = F(P, k|0) + \frac{1}{2N'} \frac{\partial^2 F}{\partial Z^2} + \frac{1}{8N'^2} \frac{\partial^4 F}{\partial Z^4} + \dots$$

since the second and fourth moments of Z , which is normally distributed with mean 0 and variance $1/N'$, are $1/N'$ and $3/N'^2$, respectively.

On comparing the right hand sides of (15) and (16), we see that (15) will become identical with (16), except for terms involving the second and higher even powers of $1/N'$, if in (15) we set $Z = 1/\sqrt{N'}$; that is,

$$(17) \quad F\left(P, k \mid \frac{1}{\sqrt{N'}}\right) \simeq F(P, k).$$

Unpublished calculations made at the Statistical Research Group, Columbia University in 1945 indicated, according to my recollection, that the approximation is remarkably good even for N' as small as 2, at least within the range of P and γ likely to be used in practice and with $n = N' - 1$, $N' = 2$ being the smallest value tested.

3. Comments on the approximation

We digress to comment on this result before turning to its application to regression functions.

Consider a normal distribution with zero mean; and consider an interval determined by two random, independent components, the location of its center, Z , the mean of which is zero, and its length, $2ks$, k being a constant.

First, we will assume that s has some particular value, and consider the effects of random variations in Z on A , the proportion of the normal distribution included in the interval. The maximum value of A will occur if Z coincides with the origin. If Z is negative, less is included in A from above the origin than for the maximizing interval, but this is partly (but not fully) compensated for by the inclusion of more from below the origin than for the maximizing interval. Similarly, if Z is positive, less is included from below the origin than in the maximizing interval, but this is partly compensated for by the inclusion of more from above the origin. The compensation will be particularly effective if the interval is long enough (that is, if ks is large enough relative to σ) so that both ends of the interval fall in the flat tail-ports of the normal curve. Thus, A is relatively stable with respect to fluctuations in Z , and it turns out that in calculations concerning A we will not do badly if we assume that Z always has a value for which A takes approximately its average value, and then ignore sampling fluctuations in Z . The possibility of doing this successfully depends, of course, upon the proper value of Z not being too sensitive to the interval length.

With respect to variations in the length of the interval, caused by sampling fluctuations in s , there is no such compensation, but rather a reinforcement. Assume that Z has some particular value and consider the effects on A of variations in s . The smaller s , the less is included in the interval from above Z and the less is included from below Z . Thus, A is relatively sensitive to variations in s , so sampling fluctuations in s must be taken into account.

4. Calculation of tolerance factors

To evaluate $F(P, k | 1/\sqrt{N'})$, Wald and Wolfowitz point out that there is a unique value r such that

$$(18) \quad \frac{1}{\sqrt{2\pi}} \int_{1/\sqrt{N'}-r}^{1/\sqrt{N'}+r} e^{-t^2/2} dt = P$$

since the left side is a monotonic increasing function of r . This r , of course, corresponds with the half length ks of an interval centered at $1/\sqrt{N'}$ for which $A = P$, and our problem is to select k large enough, in the light of the sampling distribution of s , to make the probability γ that ks will be at least r . Thus,

$$(19) \quad \begin{aligned} F\left(P, k \mid \frac{1}{\sqrt{N'}}\right) &= Pr\left(s \geq \frac{r}{k}\right) \\ &= Pr\left[\chi^2(n) \geq \frac{n r^2}{k^2}\right] \end{aligned}$$

since $\chi^2(n) = ns^2/\sigma^2$ and here $\sigma = 1$. This probability can be found from tables of the chi-square distribution [6], after first finding r from tables of the normal distribution [4].

If we are given P and γ , and require the appropriate k , we solve for k in

$$(20) \quad \chi_\gamma^2(n) = \frac{n r^2}{k^2}$$

where $\chi_\gamma^2(n)$ is that number which χ^2 for n degrees of freedom has probability γ of exceeding; that is,

$$(21) \quad k = r \sqrt{\frac{n}{\chi_\gamma^2(n)}}.$$

Bowker [1] has given the following approximation to r which for many practical purposes is satisfactory enough if $N' \geq 1$:

$$(22) \quad r = K_P \left(1 + \frac{1}{2N'} - \frac{2K_P^2 - 3}{24N'^2}\right)$$

where K_P is that number which a unit normal deviate has probability $1 - P$ of exceeding in absolute value, as defined by (7).

Wilson and Hilferty [9] have given implicitly the following approximation to $\chi^2(n)$, which is satisfactory for $n \geq 3$:³

$$(23) \quad \chi_\gamma^2(n) = n \left[1 - \frac{2}{9n} + K_{1-2\gamma} \sqrt{\frac{2}{9n}}\right]^3.$$

If $\gamma < \frac{1}{2}$, replace $K_{1-2\gamma}$ by $-K_{2\gamma-1}$.

³ For $n = 1$ and $n = 2$ the following exact formulas are simpler than the approximation:

$$\chi_\gamma^2(1) = K_{1-\gamma}^2 \quad \text{and} \quad \chi_\gamma^2(2) = -2 \log_e \gamma.$$

5. Application to linear regression

To calculate the Wald-Wolfowitz tolerance intervals for a linear regression line as specified by (4), we note that the variance of Y for any value of x is given by the Working-Hotelling formula [10]:

$$(24) \quad \sigma_y^2 = \sigma \left[\frac{1}{N} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2} \right].$$

Hence, the effective number of observations for any value of x is

$$(25) \quad N' = \frac{N \sum (x - \bar{x})^2}{\sum (x - \bar{x})^2 + N(x - \bar{x})^2}.$$

That is, for any value of x the mean value of y is determined as accurately from the regression line as if N' observations had been made at that value of x . Also, we have $n = N - 2$.

Thus, to find an interval within which we can assert with confidence coefficient γ that at least a proportion P of the population lies, we select some value of x and from it compute N' by (25), using the values of N and $\sum (x - \bar{x})^2$ obtained in the sample. Then either a table of the normal distribution [4] or (22) is used to find r ; if (22) is used it is more convenient to use r/K_P at this point. Next, $\chi_\gamma^2(n)$ is found, either from a table of the chi-square distribution [6] or from (23), and the square root of n times its reciprocal is computed—the square root being multiplied by K_P if r/K_P instead of r was used in the preceding step. Now we can obtain k by multiplying r by $\sqrt{n/\chi_\gamma^2(n)}$ or r/K_P by $K_P \sqrt{n/\chi_\gamma^2(n)}$. This is the tolerance factor k . The product of k by s [the standard error of estimate as given by (5)] is added to and subtracted from Y [as given in (4)] to obtain the tolerance interval.

Table I presents the details of a specific calculation, taking $P = 0.90$ and $\gamma = 0.95$; and figure 1 charts the interval, together with the original data, the regression line, the 95 per cent confidence interval for the regression line, and lines 1.64s on either side of the regression line.

6. Conclusion

In conclusion, attention may be called briefly to three points:

(1) Linear regression obviously represents only one of a large class of problems to which the Wald-Wolfowitz method of finding tolerance intervals can be extended. In general, it can be used to find tolerance intervals for any normal population for which we have a normally distributed estimate of the mean and know the ratio between the variance of this estimate and the variance of the population, and also have an independent estimate of the variance of the population which is distributed as $\sigma^2\chi^2/n$ with n degrees of freedom. The only limitation is that when the effective number of observations is too small the Wald-Wolfowitz approximations may not be sufficiently accurate. It seems doubtful that such intervals should be computed when the effective number of observations is less than 1, though this is a matter meriting further investigation.

(2) All of the preceding discussion has dealt with two sided tolerance *intervals*, whereas we may frequently be interested in one sided tolerance *limits*. In this case, the intuitive argument of section 3 for disregarding sampling fluctuations in \bar{x} and allowing only for their average effect breaks down; and the mathematical argument

TABLE I

LINEAR REGRESSION TOLERANCE INTERVALS: ILLUSTRATION

Average price of common stock and earnings per share, twelve chemical manufacturers, 1935. From Brumbaugh and Kellogg [3, p. 710].

x = earnings, y = price, $N = 12$, $\bar{x} = 4.43$
 $Y = 11.59 + 16.63x$; $s_{y,x} = 14.58$; $\sum (x - \bar{x})^2 = 52.76$
 Take $P = 0.90$, $\gamma = 0.95$.
 Then $K_P = 1.64485$, $K_P^2 = 2.70554$, $\chi_{\gamma}^2(10) = 3.94030$, $K_P \sqrt{n/\chi_{\gamma}^2(n)} = 2.6204$.

	$\frac{1}{12} + \frac{(x - 4.43)^2}{52.76}$	See Note	$2.6204 \frac{r}{K_P}$	14.58k	11.59 + 16.63x		
x	$1/N'$	r/K_P	k	ks	Y	$Y - ks$	$Y + ks$
- 2.52	1.0000	1.3889	3.639	53.06	- 30.32	- 83.38	22.74
0	0.4553	1.2033	3.153	45.97	11.59	- 34.38	57.56
3	0.1221	1.0595	2.776	40.47	61.48	21.01	101.95
4.43	0.0833	1.0410	2.728	39.77	85.25	45.48	125.02
6	0.1301	1.0633	2.786	40.62	111.37	70.75	151.99
9	0.4792	1.2127	3.178	46.34	161.26	114.92	207.60
11.38	1.0000	1.3889	3.639	53.06	200.84	147.78	253.90

Note: Values in the third column, representing r/K_P , have been found from the WPA normal probability tables [4]. Using the approximation (22) for r , the three central values would be unchanged, but the first two and last two would be 1.3995, 1.2068, 1.2165, and 1.3995, which would not have altered the results appreciably. Formula (22) becomes, in this case, $r/K_P = 1 + 0.5(1/N') - 0.1005(1/N')^2$, since $(2K_P^2 - 3)/24 = 0.1005$.

of section 2 yields only an approximation of order $1/N'$ instead of $1/N'^2$, for $F(P, k|Z)$ is no longer an even function of Z . The following approximate formula for an upper tolerance limit⁴ is based on the assumption that $Z + ks$ is normally distributed with mean $\zeta + k\sigma$ and variance $\sigma^2(1/N' + k^2/2n)$:

$$(26) \quad U = Z + \frac{s}{a} (K_{2P-1} + \sqrt{K_{2P-1}^2 - ab})$$

where

$$(27) \quad a = 1 - \frac{K_{2\gamma-1}^2}{2n},$$

$$(28) \quad b = K_{2P-1}^2 - \frac{K_{2\gamma-1}^2}{N'}$$

If $P < \frac{1}{2}$, replace K_{2P-1} by $-K_{1-2P}$; similarly if $\gamma < \frac{1}{2}$. For a lower limit, L , the sign between the two terms of the right hand side of (26) would be negative. If formula (26) is applied to the data of table I to obtain an upper 95 per cent tolerance limit with confidence coefficient 0.95, the last four tolerance factors (fourth column) become 2.768, 2.849, 3.312, and 3.804. The last factor, for which $N' = 1$, is about $4\frac{1}{2}$ per cent higher than that shown in table I for the 90 per cent tolerance

⁴ Compare Wallis [8, p. 47, formula 91].

intervals with the same confidence coefficient; when $N' > 1$, the discrepancy is less.

(3) Tolerance intervals or tolerance limits may be used to test the hypothesis that further observations are from the same population as an initial sample. That is, the limits would be computed from an initial sample, and each additional observation would be regarded as from the same population if it fell within the limits. Marshall [5] has investigated the power of such a test procedure; as might be expected, the power is rather low.

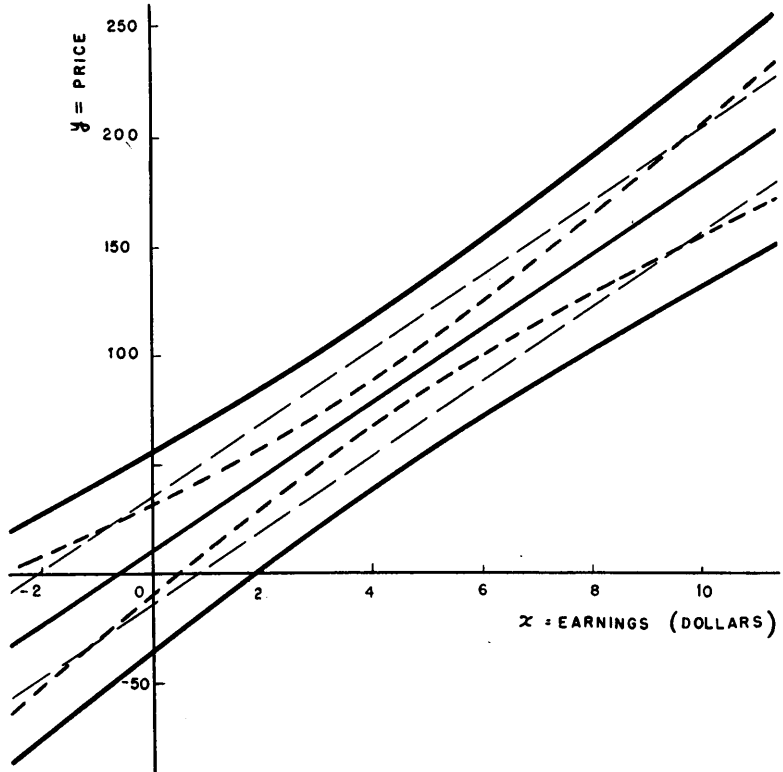






FIGURE 1

Linear regression tolerance intervals: Illustration

Average price of common stock and earnings per share, twelve chemical manufacturers, 1935.
(From Brumbaugh and Kellogg [3].)

-  Linear regression of price on earnings, $Y = 11.59 + 16.63x$
-  95% confidence interval for regression line
-  90% tolerance interval, confidence coefficient 95%
-  Zone $1.64 s_{y,x}$ on each side of regression line, $s_{y,x} = 14.58$

REFERENCES

- [1] A. H. BOWKER, "Computation of factors for tolerance limits on a normal distribution when the sample is large," *Annals of Math. Stat.*, Vol. 17 (1946), pp. 238-240.
- [2] ———, "Tolerance limits for normal distributions," Chap. 2 in Statistical Research Group, Columbia University, *Techniques of Statistical Analysis*, McGraw-Hill, New York, 1947, pp. 95-110.
- [3] M. A. BRUMBAUGH and L. S. KELLOGG, *Business Statistics*, Irwin, Chicago, 1941.
- [4] A. N. LOWAN (Director), *Tables of Probability Functions*, Vol. 2, Federal Works Agency, Work Projects Administration for City of New York, National Bureau of Standards, 1941.
- [5] A. W. MARSHALL, "A note on the power function of the Wald-Wolfowitz tolerance limits for a normal distribution," unpublished memorandum of The Rand Corporation, 1949.
- [6] C. M. THOMPSON, "Table of percentage points of the χ^2 distribution," *Biometrika*, Vol. 32 (1941), pp. 187-191.
- [7] A. WALD and J. WOLFOWITZ, "Tolerance limits for a normal distribution," *Annals of Math. Stat.*, Vol. 17 (1946), pp. 208-215.
- [8] W. A. WALLIS, "Use of variables in acceptance inspection for percent defective," Chap. 1 in Statistical Research Group, Columbia University, *Techniques of Statistical Analysis*, McGraw-Hill, New York, 1947, pp. 3-93.
- [9] E. B. WILSON and M. M. HILFERTY, "The distribution of chi-square," *Proc. Nat. Acad. Sci.*, Vol. 17 (1931), pp. 684-688.
- [10] H. WORKING and H. HOTELLING, "Applications of the theory of error to the interpretation of trends," *Jour. Amer. Stat. Assoc., Suppl.*, Vol. 24 (1929), pp. 73-85.