377 N/

MSC-IN-65-ED-27

ESTIMATION OF PERCENTAGE POINTS AND THE CONSTRUCTION OF TOLERANCE LIMITS

By

F.M. SPEED

and

A.H. FEIVESON

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

MANNED SPACECRAFT CENTER

HOUSTON, TEXAS

October 18, 1965

MSC INTERNAL NOTE NO. MSC-IN-65-ED-27

Prepared by: Fred M. Speed Alan H. Feiveson

F.M. Speed and A. H. Feiveson
ED13 Theory and Analysis Office

Approved: Eugene L. Davis

Eugene L. Davis, Jr., Chief,
Theory and Analysis Office

Approved: Eugene H. Brock

Eugene H. Brock, Chief,
Computation and Analysis Division

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

MANNED SPACECRAFT CENTER

HOUSTON, TEXAS

October 18, 1965

TABLE OF CONTENTS

ABSTRACT

This paper provides the experimenter with one method of performing several statistical tests, when the data distribution is not normal or is unknown. The method is applied to simulated landing data for a lunar excursion module.

INTRODUCTION

An error frequently committed in statistical analysis of data obtained for reliability studies is to assume that the population from which the data is taken has a normal distribution when, in fact, it does not. One effect of making such an error is that probabilities and tolerance limits obtained by standard statistical techniques are invalid; hence, if the reliability criterion is very stringent, the conclusions reached might lead to disastrous consequences.

This paper is divided into three sections. The first section contains an example of the false conclusions that may be obtained when the data is erroneously assumed to be from a normal distribution. The second section contains four theorems that enable the experimenter to perform a reliability study when the distribution is not normal or is unknown. The third section illustrates the use of the theorems developed in Section II.

SYMBOLS

X, R, $\tilde{V}_{ij}$, $\dot{Z}$	Random variables unless specified otherwise.
n	Sample size.
ξ_p	$(100 \times p)$ the percentage point of the distribution of X.
x_i	ith Sample value of X.
$x(i)$	ith Ordered Sample value of X.
$F(z)$	Cumulative distribution function of X. [i.e., $F(z) = \Pr\{X \leq z\}$]
$\binom{n}{r}$	Binomial coefficient equal to $\frac{n!}{r!(n-r)!}$
p, α, β	Probabilities.
All lower-case letters	Constants, unless specified otherwise.
μ	Mean.
σ^2	Variance.
$\Phi(x)$	Cumulative distribution function of a standardized normally-distributed random variable.

$f(x)$	Probability density function of X.
S	Total number of observations $\leq z_0$.
$I_\beta(k,m)$	Incomplete Beta function with parameters k and m.

SECTION I - EXAMPLE OF ERROR

In many cases, reaction times have a log normal distribution[1] with parameters μ and σ^2; i.e., their logarithms are normally distributed with mean μ and variance σ^2. If an experimenter observes a sample of reaction times, R, and estimates probabilities of R exceeding given values, he incorporates serious errors into his estimates by assuming that R is normally distributed. The magnitude of the error can be best illustrated by the following example.

Table I shows 150 observations of a random variable, R, having the log normal distribution, arranged in ascending order. A number, t, is desired such that the probability of R exceeding t is small, for instance, $1-\beta$, where β is a number close to 1.

If R is normally distributed and β equals .9986, t would be estimated by the familiar expression:

$$t_{est} = \bar{R} + S_R \qquad [1]$$

where $\bar{R}$ and S_R are the sample mean and standard derivations of the data. However, R is not normally distributed, and estimation of t by equation [1] is erroneous. If R is incorrectly assumed to be normally distributed, one would obtain

$$t_{incorrect} = .435 + 3(.219) = 1.092$$

TABLE I - VALUES OF R ARRANGED IN ASCENDING ORDER

.1420	.2572	.3356	.4214	.5997
.1423	.2578	.3398	.4276	.6062
.1459	.2585	.3433	.4301	.6079
.1477	.2649	.3566	.4411	.6237
.1503	.2658	.3570	.4447	.6361
.1546	.2730	.3600	.4477	.6398
.1558	.2771	.3604	.4620	.6442
.1948	.2779	.3613	.4655	.6475
.1982	.2805	.3621	.4678	.6479
.2010	.2855	.3634	.4698	.6530
.2056	.2885	.3635	.4807	.6601
.2100	.2921	.3704	.4828	.6666
.2127	.2921	.3708	.4835	.6681
.2175	.2927	.3810	.4866	.6706
.2183	.2935	.3812	.4936	.6780
.2218	.2936	.3824	.4971	.6839
.2321	.2981	.3827	.4993	.6945
.2360	.3006	.3832	.5055	.8602
.2373	.3028	.3912	.5076	.8624
.2378	.3028	.3919	.5233	.8747
.2378	.3052	.3934	.5379	.8825
.2398	.3108	.4024	.5455	.8879
.2414	.3139	.4066	.5460	.9177
.2421	.3139	.4085	.5470	.9263
.2429	.3149	.4091	.5564	.9456
.2449	.3172	.4115	.5721	.9632
.2456	.3173	.4128	.5803	1.0351
.2504	.3268	.4146	.5837	1.1202
.2508	.3333	.4158	.5854	1.1390
.2512	.3347	.4193	.5954	1.1928

$$\bar{R} = .435$$

$$S_R = .219$$

The magnitude of the error can be shown in two ways: First, consider the true probability (not .9986) of R exceeding $t_{incorrect}$.

Since $\log R \sim N(\mu, \sigma^2)$, it follows that

$$\Pr\{R < t\} = \Phi\left(\frac{\log t - \mu}{\sigma}\right)$$

where $\Phi(\cdot)$ is the standardized normal distribution function.

The 150 observations in Table I came from a log normal distribution with $\mu = -1$ and $\sigma = \frac{1}{2}$. Therefore,

$$\Pr\{R < t_{incorrect}\} = \Phi\left(\frac{\log t_{incorrect} - (-1)}{1/2}\right)$$

$$= \Phi(2.176) = .9852,$$

as compared with .9986. The probability, .9852, is corroborated by the data. Note that, of the 150 observations, 3 exceed $t_{incorrect}$. If the actual probability of R exceeding $t_{incorrect}$ were $1 - .9986 = .0014$, it is extremely unlikely that this event would occur as many as 3 times out of 150 trials.

Another way of determining the magnitude of error is to compute the number t_{true} such that $\Pr\{R < t_{true}\}$ actually is equal to .9986. Thus, t_{true} must satisfy

$$\Phi\left(\frac{\log t_{true} - (-1)}{1/2}\right) = .9986. \qquad [2]$$

Solving [2] yields $t_{true} = e^{1/2} = 1.6487$, a number considerably higher than $t_{incorrect}$.

SECTION II - THEOREMS

Suppose X is an observable random variable. From the failure analysis viewpoint, it might be desirable to estimate percentage points and tolerance limits for X. A percentage point, ξ_p, is a number such that the probability of X exceeding ξ_p is equal to 1-p. Tolerance limits for X define an interval $[x(i), x(j)]$. This interval is such that, at least 100 β percent of the time, the probability is $1-\alpha$ that $x(i) \leq X \leq x(j)$, where $1-\alpha$ is the chosen level of confidence, and β is any arbitrary positive number less than 1.

Let $x_1, x_2, \ldots, x_n$ be a sample of n independent observations of X; and suppose F(z), the cumulative distribution function of X, is continuous and strictly increasing over the range of interest. If $x(1), x(2), \ldots,$ $x(n)$ denotes the observed sample arranged in ascending order (that is, $x(i) \leq x(j)$ for $i < j$), then the four following theorems hold:

<u>THEOREM</u> 1: If z is any real number, then $\Pr\left\{x(i) \leq z\right\} = \sum_{r=i}^{n} \binom{n}{r} [F(z)]^r \, [1-F(z)]^{n-r}$

<u>Proof</u>:

For a given observation of X, the event $\{X \leq z\}$ has probability F(z). Let S equal the total number of observations of X less than n equal to z. Then, X has the binomial distribution with parameter F(z). Thus,

$$\Pr\{S \geq i\} = \sum_{r=i}^{n} \binom{n}{r} [F(z)]^r [1-F(z)]^{n-r}.$$

But, $S \geq i$ means that there are at least i observations less than n equal to z. This is equivalent to starting $x(i) \leq z$.

<u>THEOREM</u> 2: If i and j are chosen before observing the data such that $1 \leq i < j \leq n$, then $[x(i), x(j)]$ is a confidence interval, independent of F, for ξ_p, the 100 x p percentage point of the distribution of X. Specifically, the level of confidence equals

$$\Pr\{x(i) \leq \xi_p \leq x(j)\} = \sum_{r=i}^{n} \binom{n}{r} p^r (1-p)^{n-r} - \sum_{r=j}^{n} \binom{n}{r} p^r (1-p)^{n-r}$$

Proof:

Since F is continuous and strictly increasing in the range of interest, ξ_p is uniquely defined for a given p in that range.

$$\begin{aligned}\Pr\{x(i) \leq \xi_p\} &= \Pr\{x(i) \leq \xi_p,\ x(j) < \xi_p\} \\ &+ \Pr\{x(i) \leq \xi_p,\ x(j) \geq \xi_p\} \\ &= \Pr\{x(j) < \xi_p\} + P_r\{x(i) \leq \xi_p,\ x(j) \geq \xi_p\}\end{aligned}$$

since $x(i) \leq x(j)$. Therefore,

$$\Pr\{x(i) < \xi_p\} - \Pr\{x(j) < \xi_p\} = \Pr\{x(i) \leq \xi_p \leq x(j)\}.$$

Since F is continuous,

$$\Pr\{x(i) \leq \xi_p\} - \Pr\{x(j) \leq \xi_p\} = \Pr\{x(i) \leq \xi_p \leq x(j)\}.$$

Hence, from THEOREM 1, it follows that

$$\Pr\{x(i) \leq \xi_p \leq x(j)\} = \sum_{r=i}^{n} \binom{n}{r} [F(\xi_p)^r\ [1-F(\xi_p)]$$

$$- \sum_{r=i}^{n} \binom{n}{r} [F(\xi_p)^r]\ [1-F(\xi_p)]^{n-r}$$

$$= \sum_{r=i}^{n} \binom{n}{r} p^r\ (1-p)^{n-r} - \sum_{r=j}^{n} \binom{n}{r} p^r\ (1-p)^{n-r}$$

since ξ_p is defined so that $F(\xi_p) = p$.

THEOREM 3:[2] Let $f(x)$ be the probability density function of X and let the random variable $\tilde{V}_{ij}$ be the area under $f(x)$ between $x(i)$ and $x(j)$ $(i < j)$. Then $\tilde{V}_{ij}$ equals the probability that X lies between $x(i)$ and $x(j)$ and the density function of $\tilde{V}_{ij}$ is given by

$$h(v_{ij}) = \frac{n!}{(j-i-1)!\ (n-j+1)!}\ v_{ij}^{\ j-i-1}(1-v_{ij})^{n-j+1}$$

THEOREM 4: The probability that 100 β percent, or more, of X will be in the tolerance interval $[x(i), x(j)]$

$\left(\text{that is, } \Pr\{\tilde{V}_{ij} > \beta\}\right)$, is given by

$$1 - \sum_{r=j-i}^{n} \binom{n}{r} \beta^r (1-\beta)^{n-r} .$$

<u>Proof</u>:

Let $\tilde{V}_{ij} = \int_{x(i)}^{x(j)} F(z)\, dz$. Then, $\Pr\{\tilde{V}_{ij} > \beta\} = \int_{\beta}^{1} h(v)\, dv.$

But, by <u>THEOREM</u> 3,

$$h(v_{ij}) = \frac{n!}{(j-i-1)!\,(n-j+i)!}\, v_{ij}^{j-i-1} (1-v_{ij})^{n-j+i} .$$

Hence,

$$\Pr\{\tilde{V}_{ij} > \beta\} = \frac{n!}{(j-i-1)!\,(n-j+i)!} \int_{\beta}^{1} v_{ij}^{j-i-1} (1-v_{ij})^{n-j+i}$$

$$= 1 - I_{\beta}\,[(j-i),\ (n-j+i+1)],$$

where $I_{\beta}\,(k,m)$ is the Incomplete Beta function. The quantity $I_{\beta}\,[(j-i),\ (n-j+i+1)]$ can be obtained from the binomial distribution by the following relationship:(3)

$$I_{\beta}\,[(j-i),\ (n-j+i+1)] = \sum_{r=j-i}^{n} \binom{n}{r} \beta^r (1-\beta)^{n-r}$$

Hence,

$$\Pr\{\tilde{V}_{ij} > \beta\} = 1 - \sum_{r=j-i}^{n} \binom{r}{n} (\beta)^r \, (1-\beta)^{n-r}$$

SECTION III - APPLICATION

For the lunar excursion module to land safely, it is necessary that certain end conditions not be excessive. One of these end conditions is the vertical component of velocity, $\dot{Z}$. Table II gives values of $\dot{Z}$ obtained from 122 independent lunar landing simulations. Statistical tests reject the hypothesis that these values came from a normal or any other, well known distribution. (See Ref. 4; Kolmogorov-Smirnov Goodness of Fit Test.) Therefore, in order to estimate percentage points and tolerance limits of this unknown distribution, it is necessary to use a distribution-free (non parametric) procedure. It is clear that the range of $\dot{Z}$ is an interval on the real line; hence, the conditions of SECTION II are satisfied.

A. ESTIMATION OF ξ_p

Suppose it is desired to estimate $\xi_{.95}$. By Theorem II, any interval of the form $[\dot{Z}(i), \dot{Z}(j)]$ is a confidence for $\xi_{.95}$. However i and j should be chosen so that a reasonable confidence level is attained; that is, it is advantageous to have

$$\Pr\,(\dot{Z}(i) \le \xi_{.95} \le \dot{Z}(j)) = 1 - \alpha$$

TABLE II - VALUES OF $\dot{Z}$ ARRANGED IN ASCENDING ORDER

k	$\dot{Z}_{(k)}$	k	$\dot{Z}_{(k)}$	k	$\dot{Z}_{(k)}$
1	.30	41	3.60	81	5.88
2	.78	42	3.72	82	5.94
3	1.02	43	3.72	83	6.00
4	1.14	44	3.78	84	6.06
5	1.20	45	3.84	85	6.06
6	1.32	46	3.90	86	6.12
7	1.38	47	3.90	87	6.12
8	1.38	48	3.90	88	6.12
9	1.56	49	3.90	89	6.30
10	1.62	50	3.96	90	6.36
11	1.74	51	4.02	91	6.42
12	1.74	52	4.14	92	6.48
13	1.80	53	4.14	93	6.54
14	1.86	54	4.14	94	6.78
15	1.92	55	4.38	95	6.90
16	2.22	56	4.38	96	6.96
17	2.28	57	4.50	97	7.14
18	2.34	58	4.74	98	7.20
19	2.52	59	4.76	99	7.26
20	2.52	60	4.80	100	7.30
21	2.58	61	4.80	101	7.50
22	2.64	62	4.86	102	7.56
23	2.64	63	4.98	103	7.74
24	2.70	64	5.04	104	7.74
25	2.70	65	5.16	105	7.78
26	2.82	66	5.27	106	7.86
27	2.88	67	5.28	107	8.04
28	2.94	68	5.28	108	8.10
29	3.00	69	5.34	109	8.22
30	3.06	70	5.34	110	8.82
31	3.06	71	5.40	111	8.88
32	3.30	72	5.46	112	9.12
33	3.40	73	5.52	113	9.12
34	3.42	74	5.52	114	9.12
35	3.42	75	5.64	115	9.48
36	3.48	76	5.70	116	9.54
37	3.48	77	5.70	117	10.74
38	3.48	78	5.70	118	10.98
39	3.54	79	5.76	119	12.54
40	3.54	80	5.88	120	16.26
				121	16.88
				122	20.82

where α is a small probability. In other words, α is the probability that the true value of $\xi_{.95}$ lies outside the interval of estimation. For example, if i is chosen to be 111, and j to be 120, then x(i) = 8.88, x(j) = 16.26, and it follows that

$$\Pr\{8.88 \leq \xi_{.95} \leq 16.26\} = \sum_{r=111}^{122} \binom{122}{r} (.95)^r (.05)^{122-r}$$

$$-\sum_{r=120}^{122} \binom{122}{r} (.95)^r (.05)^{n-r}$$

$$= .9805-.0534 = .9271$$

Since it is of no concern in this particular problem if the true value of $\xi_{.95}$ is <u>less</u> than $\dot{Z}(i)$, the interval in equation [2] may be changed to a <u>one-sided</u> form, $[-\infty, \dot{Z}(j)]$. In this case, equation [2] reduces to

$$\Pr\{\xi_p \leq X(j)\} = 1 - \sum_{r=j}^{n} \binom{n}{r} p^r (1-p)^{n-r} = 1 - \alpha .$$

Returning to the given example, it follows that

$$Pr\{\xi_{.95} \leq 16.26\} = 1 - .0534 = .9466.$$

B. MAXIMUM CONFIDENCE LEVEL

Note that as j increases, α decreases until, the maximum confidence level of $1-p^n$ is attained if $j = n$. For this reason, when p is very close to 1 and n is not very large, any attempt to estimate ξ_p results in a very low confidence level.

A rough estimate of a desirable n for a given p may be obtained using the relation that, for $n > 100$, $1-p^n \doteq 1-e^{-n(1-p)}$. If it is stipulated that the maximum confidence level should be $1-\alpha$, then n must be determined such that $1-e^{-n(1-p)} \doteq 1-\alpha$. In other words, let $n = \frac{(-\log \alpha)}{(1-p)}$.

EXAMPLE:

It is desired to find a sample size that could be used for estimating $\xi_{.9999}$ with a maximum confidence of .99.

SOLUTION:

Let n be approximately equal to

$$\frac{-\log(.01)}{.0001} = 46050.$$

C. TOLERANCE LIMITS

Suppose it is necessary to determine the following sets of tolerance limits for the data given in Table II.

1. Determine i and j such that:

   a. The probability is .90 (that is, $1-\alpha = .90$), that

   b. At least 85% of the time, $\dot{Z}$ lies between x(i) and x(j) ($\beta = .85$).

2. Determine i and j such that:

   a. The probability $(1-\alpha) = .93$, that at least

   b. 90% ($\beta = .90$) of the time $\dot{Z}$ lies between x(i) and x(j).

3. Determine j such that:

   a. The probability is .94, that at least

   b. 85% of the time $\dot{Z}$ will be less than x(j).

4. Determine i and j such that:

a. The probability is .999, that at least

b. 99.865% of the time $\dot{Z}$ will be between x(i) and x(j).

Although these tolerance limits can be obtained by a direct application of Theorem 4, a computer program has been written providing the necessary information in tabular form. The output of this program is presented in Table III. (The computer program that generates Table III is available from the Computation and Analysis Division.)

To construct the set of tolerance limits in example C. 1., read down the .85 Beta column to $1-\alpha = .90$ (or the number closest to .90). Then read the corresponding entry in the J-I columns, which is 109, indicating that the x(i) and x(j) used for the tolerance limits are such that $j-i = 109$. Hence, any of the following sets of x(i) and x(j) could be used to satisfy the desired tolerance limits. C.1.: [x(1), x(110]; [x(2), x(111)]; [x(3), x(112)], etc.

Suppose the experimenter desired to use x(5) and x(116) he could assume that the probability is .8915 that at least 85% of the time $\dot{Z}$ would lie between 1.20 and 9.48.

TABLE III - TOLERANCE LIMITS

CONFIDENCE LIMITS (1-α)

N = 122 J-I	BETA .85000	.90000	.95000	.97500	.99865
122	1.00000000	.99999738	.99808452	.95444217	.15711072
121	.99999995	.99996193	.98578505	.81192790	
120	.99999939	.99972358	.94662098	.59084808	
119	.99999541	.99866425	.86417030	.36409954	
118	.99997454	.99516255	.73506989	.19113110	
117	.99988764	.98598032	.57471360		
116	.99958859	.96608552	.41013740		
115	.99871404	.92945379	.26659726		
114	.99649551	.87094480	.15799781		
113	.99153645	.78859880			
112	.98164752	.68520883			
111	.96387916	.56824239			
110	.93487493	.44802688			
109	.89156544	.33500376			
108	.83206039	.23722979			
107	.75645397	.15901061			
106	.66722728				
105	.56904704				
104	.46797915				
103	.37035320				
102	.28162844				
101	.20557864				
100	.14396611				

In example C. 2., the set of tolerance limits is read from the table to be x(i) and x(j) such that j-i = 115. In C. 3., a one-sided case, the x(j) chosen is such that j = 110. This means that the probability is .93 that at least 85% of the time $\dot{Z}$ will be less than 8.82. Note that the last set of tolerance limits (example C. 4.) does not exist for this set of data. That is, there is no i and j such that the probability is .999 that at least 99.865% of the time $\dot{Z}$ will be between x(i) and x(j).

D. SAMPLE SIZE

To find a set of tolerance limits as described in example C. 4., a sample size of approximately 8845 observations would be necessary. The following equation provides an approximation to the number of observations required for a given β and a given confidence level.[5]

$$N = \frac{1}{4} \quad A \quad \cdot \left(\frac{1 + \beta}{1 - \beta}\right) + \frac{1}{2}$$

Where:

A is the (1-α) percentage point of the χ^2 distribution with four degrees of freedom. β is the probability that $\dot{Z}$ will lie between x(i) and x(j). 1-α is the desired confidence level. (In example C. 4., $A = 18.5$ $\beta = .99865$, and $1-\alpha = .999$).

E. POISSON APPROXIMATION TO THE BINOMIAL SUM

For large n and β close to 1, the sum

$$\sum_{r=j-1}^{n} \binom{n}{r} \beta^{r} (1-\beta)^{n-r}$$

can be approximated by

$$\sum_{r=0}^{n-(j-1)} \frac{e^{-\lambda} \lambda^{r}}{r!} \qquad \text{where } \lambda = n(1-\beta)$$

F. TESTING FOR NORMALITY

One method of testing the data for normality is to use the Komogorov-Smirnov test. This test is available in a computer program from the Computation and Analysis Division[6].

REFERENCES

[1] Hald, A.: Statistical Theory with Engineering Applications. John Wiley and Sons, Inc., p. 171.

[2] Mood, Alexander M., and Graybill, Franklin A.: Introduction to the Theory of Statistics. Second Ed., McGraw-Hill Book Company, Inc., p. 405.

[3] Abramowity, Milton, and Stegun, Irene A., Eds: Handbook of Mathematical Functions. National Bureau of Standards, Applied Mathematics Series 55, p. 944.

[4] Siegel, Sidney: Nonparametric Statistics. McGraw-Hill Book Company, Inc.

[5] Hoel, Paul G.: Introduction to Mathematical Statistics. John Wiley and Sons, Inc., p. 284.

[6] Feiveson, Alan H., and Speed, F. M.: Goodness of Fit and Confidence Intervals. MSC-IN-65-ED.

[7] Brunt, H. D.: An Introduction to Mathematical Statistics. New York, Ginn and Company.

[8] Sarhan, Ahmed E., and Greenberg, Bernard G., Ed: Contributions to Order Statistics. John Wiley and Sons Inc.