Mathematical Principles of Statistical Quality Control

by

J. Michael

A thesis submitted to
McGill University
in partial fulfilment
of the requirements
for the degree of
Master of Science
in Mathematics

Montreal, Canada August, 1955

ACKNOWLEDGEMENT

I would like to thank Professor M. Kozakiewicz for his valuable suggestions, guidance, and assistance during the preparation of this thesis

TABLE OF CONTENTS

Introduction

In an industrial process we usually set up a standard for the quality of a given kind of product. That is, we lay down specifications for weight, thickness, diameter, breaking strength, finish, etc. by which an article can definitely be classed as conforming or nonconforming, even if in many cases the specifications are partly arbitrary. We then try to make all units of the product conform with this standard. However, it is impossible to make all units exactly alike. Therefore, there is bound to be some variation in the quality of the product. The problem then is: how much may the quality of a product vary and yet be controlled? We say that the quality is in statisbical control if all of the observed variations lie within certain limits. Thus we see that a controlled quality is not a constant quality but a variable quality.

We recognize two more or less distinct types of causes of variability in the quality of a manufactured product. These are random, or chance, causes and assignable causes. By random, or chance, causes we mean the whole host of small influences lying behind the particular measurement or result we happen to obtain. These causes are very large in number and the effect of each on the industrial process is very slight. It is not possible to track down and eliminate these chance causes. On the other hand, assignable causes are those which come in intermittently or perhaps permanently to make changes in the process of such magnitude as to be of practical importance. Assignable causes, if they exist, are very few in number and the effect of each on the industrial process is marked. These causes may be found and eliminated.

The control chart, by helping us to locate and eliminate assignable causes, is a most powerful tool in achieving a state of statistical control in the various stages of the industrial process. The control chart was discovered and developed in 1924 by W. A. Shewhart of the Bell Telephone Laboratories. He realized that some of the observed variation in performance was natural to a process and unavoidable. But from time to time there would be variations which could not be so explained. He reached the conclusion that it would be desirable and possible to set limits upon the natural variation of any process. Fluctuations within these limits could be readily explained by chance causes, but any variation outside these limits would indicate the presence of an assignable cause. The development of the control chart followed, which provides a reasonable test for determining when a process can be considered to be in control.

There are many advantages to be gained through control. As we proceed to eliminate assignable causes of variability, the quality of the product usually approaches a state of stable equilibrium. As the quality approaches this comparatively stable state, the need for inspection is reduced. Thus there is a reduction in the cost of inspection. Furthermore, we have a more standard product since the quality of the finished product will exhibit minimum variability. Finally, by eliminating assignable causes of variability, we reduce the proportion of defectives to a minimum with a resulting reduction in the cost of rejection.

Once a state of statistical control has been achieved in the various stages of the industrial process, as evidenced by the control charts, we can be quite certain about the quality of the finished product. Nevertheless, a final verification of quality may be

desirable. Procedures such as single and double sampling inspection and the sequential method afford calculated protection to producer and consumer independently of the state of control in the industrial process. These are methods whereby lots of merchandise are accepted or rejected on the basis of a sample drawn from the lot. This practice arises from the fact that it is often more economical to tolerate a small percentage of defectives than to bear the cost of 100 per cent inspection.

In the statistical methods discussed above, we have assumed that the observations constitute a random sample from a fixed population. If any doubt exists concerning the randomness of a set of observations it is necessary to test the randomness of the observations before the usual statistical methods based on randomness can be applied. The theory of runs provides us with a method for testing randomness which is based on frequency functions of runs. This method does not depend on the frequency function of the basis variable and is therefore known as a nonparametric method.

Chapter I

Some Mathematical Definitions and Theorems

We assume a knowledge of mathematical statistics at the undergraduate level. However, we introduce the following definitions and theorems for the sake of clarity of terminology and because some of these notions are not usually covered in most texts on statistics.

Frequency Function

(a) Discrete Variable

A function $f(x)$ that yields the probability that the discrete random variable x will assume any particular value in its range is called the frequency function of the discrete random variable x.

(b) Continuous Variable

A frequency function (probability density) for a continuous random variable x is a function* $f(x)$ that possesses the following properties:

(i) $\quad f(x) \geq 0$

(ii) $$\int_{-\infty}^{\infty} f(x)\,dx = 1 \qquad (1)$$

(iii) $$\int_{a}^{b} f(x)\,dx = P(a < x < b)$$

where a and b are any two values of x, with $a < b$ and $P(a < x < b)$ is the probability that x will assume a value between a and b.

Joint Continuous Frequency Function

A frequency function for n continuous random variables $x_1, x_2, \ldots, x_n$ is a function $f(x_1, x_2, \ldots, x_n)$ that possesses the following properties:

$$\text{(i)} \qquad f(x_1, x_2, \ldots, x_n) \geq 0$$

$$\text{(ii)} \qquad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n = 1 \qquad (2)$$

$$\text{(iii)} \qquad \int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n = P(a_1 < x_1 < b_1, \ldots, a_n < x_n < b_n)$$

This is a straightforward generalization of a frequency function for one variable.

Cumulative Distribution Function

(a) Discrete Variable

The cumulative distribution function $F(x)$ is closely related to the frequency function $f(x)$. It is defined by the relation

$$F(x) = \sum_{t \leq x} f(t) \qquad (3)$$

where the summation occurs over all those values of the random variable that are less than or equal to the specified value of x. $F(x_0)$ gives the probability that the random variable x will assume a value less than or equal to x_0, as contrasted to $f(x_0)$ which gives the probability that x will assume the particular value x_0.

(b) Continuous Variable

The cumulative distribution function, $F(x)$, for the continuous random variable x is defined by

$$F(x) = \int_{-\infty}^{x} f(t)\, dt \qquad (4)$$

In this case $F(x_0)$ gives the probability that the random variable x will assume a value less than x_0.

Change of Variable

If $x = h(y)$ is a strictly monotonic function of y and if $f(x)$ is the frequency function of continuous variable x, then $g(y)$, the frequency function of y is given by the formula

$$g(y) = f\left[h(y)\right] \left| h'(y) \right| \tag{5}$$

If $f(u,v)$ is the joint frequency function of u,v and if $z = g(u,v), w = h(u,v)$ are functions of u,v then the joint frequency function, $k(w,z)$, of w,z is given by

$$k(w,z) = f(u,v) \left| \frac{\partial(u,v)}{\partial(w,z)} \right| \tag{6}$$

where u,v in $f(u,v)$ are expressed in terms of w,z. It is assumed that $\left| \frac{\partial(u,v)}{\partial(w,z)} \right| \neq 0$.

Two Types of Errors

Consider the random variable x whose frequency function $f(x;\theta)$ depends upon the parameter θ. Suppose we wish to test, on the basis of one observation, the hypothesis that the parameter θ has the value θ_0 against the alternative hypothesis that it has the value θ_1. We assume that there is only one alternative. Let H_0 be the hypothesis that $\theta = \theta_0$ and let H_1 be the alternative hypothesis that $\theta = \theta_1$. Rejection of H_0 is equivalent to acceptance of H_1.

To test H_0 we choose a number A and make an observation x_1. If $x_1 < A$ we accept H_0 and if $x_1 > A$ we reject H_0, that is we accept H_1.

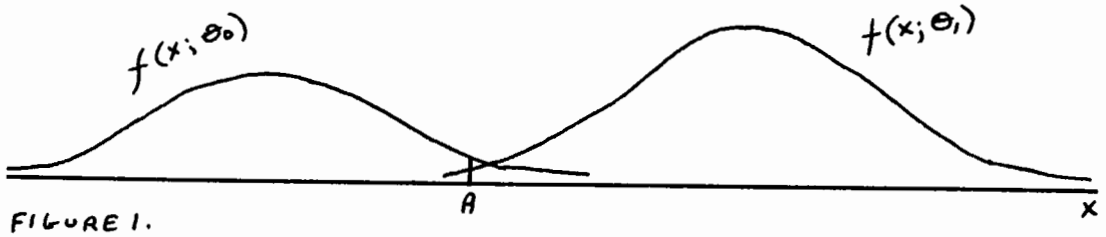

FIGURE 1.

The interval $x > A$ is called the critical region of the test. This is the region which corresponds to the rejection of the hypothesis H_0. To construct the test we have divided the x-axis into two regions, and this can be done quite arbitrarily. As a critical region we could have chosen a finite interval on the x-axis or some other region which would depend on the type I and type II errors discussed below.

There are two kinds of errors possible in this test. We may reject H_0 when it is in fact true; that is, the parameter θ may have the value θ_0 even though the observed value of x did exceed A. This is called the type I error of the test. The size of the type I error is the probability that the sample point will fall in the critical region when H_0 is true. This probability is given by

$$\alpha = \int_A^\infty f(x;\theta_0)\,dx \tag{7}$$

A second possible error is the acceptance of H_0 when it is false; that is, the observed value of x may be less than A even though the true value of θ is θ_1. This is called the type II error of the test. The size of the type II error is the probability that the sample point will fall in the noncritical region when H_1 is true. This probability is given by

$$\beta = \int_{-\infty}^A f(x;\theta_1)\,dx \tag{8}$$

A good test is considered to be one which minimizes the sizes of both errors. However, it is impossible to reduce both errors simultaneously with a single observation. The common procedure is to fix the type I error arbitrarily and then choose the critical region so as to minimize the size of the type II error.

We may generalize these results to samples of size n. The sample observation $(x_1, x_2, \ldots, x_n)$ may be plotted as a point in an n-dimensional space. The sample space is divided into two regions, the critical region R and the acceptance region A. If the sample point falls in R, H_o is rejected; otherwise H_o is accepted. The probability of a type I error is

$$\alpha = \int_R f(x_1;\theta_o) f(x_2;\theta_o) \ldots f(x_n;\theta_o) dx_1 dx_2 \ldots dx_n \qquad (9)$$

The probability of a type II error is

$$\beta = \int_A f(x_1;\theta_1) f(x_2;\theta_1) \ldots f(x_n;\theta_1) dx_1 dx_2 \ldots dx_n \qquad (10)$$

Power Function

The power function is defined as

$$P(\theta) = \int_R f(x_1;\theta) f(x_2;\theta) \ldots f(x_n;\theta) dx_1 dx_2 \ldots dx_n \qquad (11)$$

It is easy to notice that $P(\theta_o)$ is type I error and $P(\theta_1)$ is 1-type II error.

Likelihood Function

Consider the random variable x whose frequency function $f(x;\theta)$ depends upon the parameter θ. Let $x_1, x_2, \ldots, x_n$ denote the n random variables corresponding to n observations of the variable x. Then the function given by

$$L(x_1, x_2, \ldots, x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta) \qquad (12)$$

defines a function of the random variables $x_1, x_2, \ldots, x_n$ and the parameter θ which is known as the likelihood function.

Suppose that the observations are obtained from n independent trials of an experiment for which $f(x;\theta)$ is the frequency function of a discrete random variable x. Then, for any particular set of

values the likelihood function gives the probability of obtaining that set of values, including their order of occurrence. If, however, x is a continuous variable, the likelihood function gives the probability density at the sample point $(x_1,x_2,....,x_n)$, where the sample space is n dimensional.

Expected Value (Mean Value)

(a) Discrete Variable

The expected value of the function h(x) of the random variable x whose frequency function is f(x)is given by

$$E\left[h(x)\right] = \sum_x h(x)f(x) \tag{13}$$

where the sum is taken over the whole range of x.

(b) Continuous Variable

The expected value of the function g(x) of the continuous random variable x whose frequency function is f(x) is given by

$$E\left[g(x)\right] = \int_{-\infty}^{\infty} g(x)f(x)dx \tag{14}$$

It can be proved that the expected value has the following properties:

$$\begin{aligned}
&\text{(i)} && E(x+y) = E(x)+E(y)\\
&\text{(ii)} && E(xy) = E(x)E(y) \text{ when x,y are independent}\\
&\text{(iii)} && E(ax) = aE(x), \text{ a constant}\\
&\text{(iv)} && E(a) = a, \text{ a constant}
\end{aligned} \tag{15}$$

Unbiased Estimate

Consider a random variable x whose frequency function $f(x;\theta)$ depends upon a parameter θ. Let $x_1,x_2,....,x_n$ represent a random sample of size n from the corresponding population and let $t(x_1,x_2,....x_n)$ be any statistic being contemplated as an estimator of θ. The statistic $t = t(x_1,x_2,....,x_n)$ is called an unbiased estimate of the parameter θ if $E(t) = \theta$. This means that the random

variable t possesses a distribution whose mean is the parameter θ being estimated.

(a) Unbiased Estimate of the Population Mean μ

Let $x_1, x_2, \ldots, x_n$ represent a random sample of size n from a population with mean μ and variance σ^2. By properties (i) and (iii) of the expected value we have

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \qquad (16)$$

Thus the sample mean $\bar{x}$ possesses a distribution whose mean is the population mean μ. Consequently, we may use $\bar{x}$ as an unbiased estimate of μ.

(b) Unbiased Estimate of the Population Variance σ^2

Consider the expected value of a sample variance s^2 based on a random sample of size n.

$$\begin{aligned} E(s^2) &= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] \\ &= E\left[\frac{1}{n}\sum_{i=1}^{n}\left\{(x_i-\mu)-(\bar{x}-\mu)\right\}^2\right] \\ &= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2-(\bar{x}-\mu)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}E(x_i-\mu)^2 - E(\bar{x}-\mu)^2 \qquad (17) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 - \sigma_{\bar{x}}^2 \\ &= \sigma^2 - \frac{\sigma^2}{n} \\ &= \frac{n-1}{n}\sigma^2 \end{aligned}$$

Thus S^2 is not an unbiased estimate of σ^2.

Now $E(\frac{n}{n-1}S^2) = \frac{n}{n-1}E(S^2) = \sigma^2$.

Therefore, we may use $\frac{n}{n-1}S^2$ as an unbiased estimate of σ^2. If the sample is very large, we may estimate σ^2 by S^2 since $\frac{n}{n-1} \cong 1$. Since

$$\frac{n}{n-1}s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

we can avoid the bias in estimating variances by dividing the sum of the squared deviations by n-1 rather than by n.

Moments

(a) Discrete Variable

The kth moment about the origin of a discrete random variable x with frequency function f(x) is given by

$$\mu'_k = E(x^k) = \sum_{x=0}^{\infty} x^k f(x) \qquad (18)$$

The kth moment about the mean of a discrete random variable x with frequency function f(x) is given by

$$\mu_k = E\left[(x-\mu)^k\right] = \sum_{x=0}^{\infty}(x-\mu)^k f(x) \qquad (19)$$

where μ is the mean of the distribution.

(b) Continuous Variable

The kth moment about the origin of a continuous random variable x with frequency function f(x) is defined by

$$\mu'_k = E(x^k) = \int_{-\infty}^{\infty} x^k f(x)\,dx \qquad (20)$$

The kth moment about the origin of a function g(x) of a continuous random variable x with frequency function f(x) is defined by

$$\mu'_{k:g(x)} = E\left[g^k(x)\right] = \int_{-\infty}^{\infty} g^k(x) f(x)\,dx \qquad (21)$$

If $g(x) = x - \mu$ then the kth moment about the origin of $g(x)$ would be the kth moment of x about its mean.

Moment Generating Function

(a) Discrete Variable

The moment generating function of a discrete random variable x with frequency function $f(x)$ is given by

$$M_x(\theta) = E\left(e^{\theta x}\right) = \sum_{x=0}^{\infty} e^{\theta x} f(x) \qquad (22)$$

This series is a function of the parameter θ only. The subscript is placed on $M(\theta)$ to show what variable is being considered.

(b) Continuous Variable

The moment generating function of a continuous random variable x with frequency function $f(x)$ is given by

$$M_x(\theta) = E\left(e^{\theta x}\right) = \int_{-\infty}^{\infty} e^{\theta x} f(x)\, dx \qquad (23)$$

It can be proved that the moment generating function, $M_X(\theta)$, considered as a function of a real variable, possesses derivatives at $\theta = 0$, if it exists in a neighbourhood including the origin. It can also be proved that all moments exist and that $M_X(\theta)$ can be expanded in a Maclaurin's series. We shall always assume that $M_X(\theta)$ exists in some open interval about the origin. It can also be proved that, when $M_x(\theta)$ exists, differentiation under the integral sign is permissible.

If we differentiate the members of the above relation k times with respect to θ and evaluate the resulting derivative at $\theta = 0$, we have

$$M_x^{(k)}(\theta)\Big|_{\theta=0} = \left[\frac{d^k}{d\theta^k}\int_{-\infty}^{\infty} e^{\theta x} f(x)\,dx\right]_{\theta=0}$$

$$= \int_{-\infty}^{\infty}\left[\frac{d^k}{d\theta^k} e^{\theta x}\right]_{\theta=0} f(x)\,dx \qquad (24)$$

$$= \int_{-\infty}^{\infty} x^k f(x)\,dx = E(x^k) = \mu_k'$$

Thus the moments of a distribution may be obtained from the moment generating function by differentiation.

Properties of the Moment Generating Function $M_X(0)$

If a,b are constants, then

(i) $M_{aX}(\theta) = M_X(a\theta)$

(ii) $M_{aX+b}(\theta) = e^{b\theta}M_X(a\theta)$

(iii) $M_{X_1+X_2+\cdots+X_n}(\theta) = M_{X_1}(\theta)M_{X_2}(\theta)\cdots M_{X_n}(\theta)$

where $X_1, X_2, \ldots, X_n$ are independent variables.

We state the uniqueness theorem and continuity theorem without proof.

Uniqueness Theorem

If F(x) has the moment generating function M(θ), and M(θ) exists for $|\theta| \le h, h>0$, and if the cumulative distribution function G(x) has the same moment generating function, then G(x) = F(x).

Continuity Theorem

Let $F_n(x)$ and $M_n(\theta)$ be respectively the cumulative distribution function and moment generating function of a random variable $X_n (n = 1,2,3,....)$. If $M_n(\theta)$ exists for $|\theta| < h$ for all n and if there exists a function $M(\theta)$ such that $\lim_{n\to\infty} M_n(\theta) = M(\theta)$ for $|\theta| < h'$, then $\lim_{n\to\infty} F_n(x) = F(x)$, where $F(x)$ is the cumulative distribution function of a random variable X with moment generating function $M(\theta)$.

The uniqueness theorem states that the distribution function of a variable is uniquely determined by its moment generating function when the moment generating function exists. The continuity theorem states that if one variable has a moment generating function which approaches the moment generating function of a second variable, then the distribution function of the first variable approaches that of the second variable.

Central Limit Theorem

For an arbitrary population with mean μ and finite variance σ^2 the variable $z = \frac{(\bar{x}-\mu)\sqrt{n}}{\sigma}$ has a distribution that approaches the standard normal distribution ($\mu=0$, $\sigma=1$) as $n \to \infty$.

Proof: Let $M_X(\theta)$ be the moment generating function of the original distribution. We assume $M_X(\theta)$ exists for $|\theta| < h$, $h > 0$. Put $\tau = x - \mu$.

$$z = \frac{(\bar{x}-\mu)\sqrt{n}}{\sigma} = \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^{n}\tau_i}{\sigma\sqrt{n}}$$

Let $M_z(\theta)$ be the moment generating function of z. By properties (i) and (iii) of (25)

$$M_z(\theta) = M_{\frac{\sum_{i=1}^{n}\tau_i}{\sigma\sqrt{n}}}(\theta) = M_{\sum_{i=1}^{n}\tau_i}\left(\frac{\theta}{\sigma\sqrt{n}}\right) = \left[M_\tau\left(\frac{\theta}{\sigma\sqrt{n}}\right)\right]^n$$

$$M_T\left(\frac{\theta}{6\sqrt{n}}\right) = 1 + \left(\frac{\theta}{6\sqrt{n}}\right)\mu_1 + \frac{1}{\underline{|2}}\left(\frac{\theta}{6\sqrt{n}}\right)^2\mu_2 + \frac{1}{\underline{|3}}\left(\frac{\theta}{6\sqrt{n}}\right)^3\mu_3 + \cdots\cdots$$

where $\mu_1 = 0$, $\mu_2 = 6^2$

Thus $M_T\left(\frac{\theta}{6\sqrt{n}}\right) = 1 + \frac{\theta^2}{2n} + \frac{1}{\underline{|3}\, n^{3/2}}\frac{\theta^3}{6^3}\mu_3 + \cdots\cdots$

$M_Z(\theta)$ can be written

$$M_Z(\theta) = \left[M_T\left(\frac{\theta}{6\sqrt{n}}\right)\right]^n = \left(1 + \frac{w}{n} + \frac{w\lambda_n}{n}\right)^n$$

where $w = \frac{\theta^2}{2}$ and $\lambda_n \to 0$ as $n \to \infty$.

We shall prove the following:

If $\lambda_n \to 0$ as $n \to \infty$, then $\left(1 + \frac{w}{n} + \frac{w\lambda_n}{n}\right)^n \to e^w$ as $n \to \infty$ (θ fixed).

Let $\frac{w}{n}(1 + \lambda_n) = \zeta_n$

$$M_Z(\theta) = \left(1 + \frac{w}{n} + \frac{w\lambda_n}{n}\right)^n = (1+\zeta_n)^n$$

$$= (1+\zeta_n)^{\frac{w(1+\lambda_n)}{\zeta_n}}$$

$$= (1+\zeta_n)^{\frac{w}{\zeta_n}}(1+\zeta_n)^{\frac{w\lambda_n}{\zeta_n}}$$

$$= \left[(1+\zeta_n)^{\frac{1}{\zeta_n}}\right]^w\left[(1+\zeta_n)^{\frac{1}{\zeta_n}}\right]^{w\lambda_n}$$

By definition $e = \lim_{n\to\infty}(1+\frac{1}{n})^n$. Thus for n sufficiently large

$$2 < (1+\zeta_n)^{\frac{1}{\zeta_n}} < 3$$

$$\therefore \quad 2^{w\lambda_n} < \left[(1+\zeta_n)^{\frac{1}{\delta_n}}\right]^{w\lambda_n} < 3^{w\lambda_n} \quad \text{if } w\lambda_n > 0$$

$$\text{and } 3^{w\lambda_n} < \left[(1+\zeta_n)^{\frac{1}{\delta_n}}\right]^{w\lambda_n} < 2^{w\lambda_n} \quad \text{if } w\lambda_n < 0$$

Now as $n \to \infty$, $\lambda_n \to 0$, $\therefore$ $2^{w\lambda_n} \to 1$ and $3^{w\lambda_n} \to 1$.

Consequently $\left[(1+\zeta_n)^{\frac{1}{\delta_n}}\right]^{w\lambda_n} \to 1$ as $n \to \infty$.

$$\therefore \quad M_\zeta(\theta) \to e^{w} = e^{\frac{\theta^2}{2}} \text{ as } n \to \infty.$$

However, $e^{\frac{\theta^2}{2}}$ is the moment generating function of the standard normal variable ($\mu=0$, $\sigma=1$). The central limit theorem then follows from the continuity theorem.

The central limit theorem states that if an arbitrary population has a finite variance σ^2 and mean μ, then the distribution of the sample mean, for large n, is approximately normal with mean μ and variance $\frac{\sigma^2}{n}$. Nothing is assumed about the form of the population distribution function.

Chapter 2

Control Charts

The control chart provides a reasonable test for determining when an industrial process can be considered to be in control. To construct a control chart we take a three standard deviation band about the mean of the statistic in question. We then sample the process periodically and plot the successive sample points on the control chart. The process is said to be in statistical control if all of the sample points lie within the control band.

Consider the control chart for the mean (figure 1). By the central limit theorem we know that, for large samples, the distribution of the sample mean is approximately normal with mean μ (population mean). Now the control band is a three standard deviation band about the population mean μ. Thus the probability that a sample mean, when plotted on the control chart, will fall outside the control band is approximately equal to the probability that a normal variable will assume a value more than three standard deviations away from its mean. This probability is .003. The sample size should, of course, be large (at least 50). Because of this small probability there will be, by chance causes alone, very few sample points outside of the control band. When a sample point does fall outside of the control band, it is reasonable to assume that the production process is no longer behaving properly. That is, points outside of the control band indicate the presence of assignable causes which may be found and eliminated. We thus investigate only those points that fall outside of the control band. The control chart, by helping us to locate and eliminate assignable causes, is a most powerful tool in controlling the industrial process.

We now use the results of the central limit theorem to construct a control chart for the mean. This chart is illustrated below.

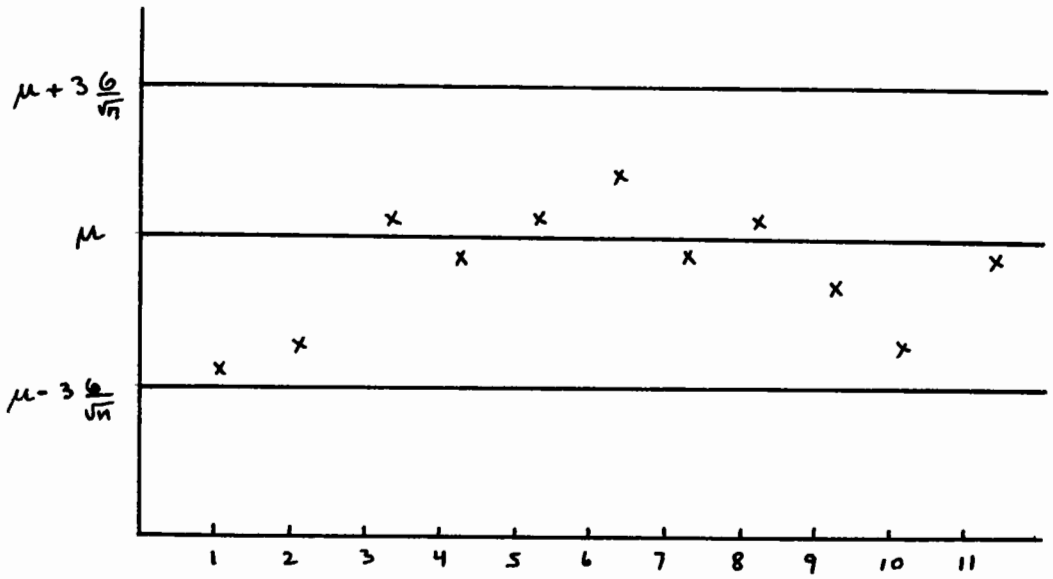

FIGURE 1. CONTROL CHART FOR THE MEAN.

Because of the results that we have established, it is not essential that the basic variable be normally distributed for such charts; consequently they are of wide applicability. The middle line is thought of as corresponding to the process average, although it is usually merely the mean of past sample means and by (16) of chapter 1 is expected to be a very good estimate of the process average. Similarly, by (17) of chapter 1 we may arrive at a good estimate of σ by taking the mean of past sample standard deviations. The other two lines serve as control limits for the sample means. It will be observed that these two control lines are spaced three standard deviations from the mean line. Time units for successive samples

are recorded along the x-axis. By the argument presented above we know that the process is in statistical control if all of the sample points lie within the control band. The chart in figure 1 shows that this process is in statistical control.

We can apply the central limit theorem to show that the variable $\frac{x-np}{\sqrt{npq}}$ where x is distributed according to the binomial law, has a distribution that approaches the standard normal distribution ($\mu=0$, $\sigma=1$) as $n \to \infty$. We may write

$$\frac{x-np}{\sqrt{npq}} = \frac{\left(\frac{\sum_{i=1}^{n} x_i}{n} - p\right)\sqrt{n}}{\sqrt{pq}} \qquad (1)$$

where $x_1, x_2, \ldots, x_n$ are independently distributed according to the law $f(x) = p^x(1-p)^{1-x}$ ($x = 0$ in case of failure, or 1 in case of success). The mean of this distribution is

$$\mu = E(x) = \sum_{x=0}^{1} x\, p^x (1-p)^{1-x} = p$$

and the variance is

$$\sigma^2 = E(x-p)^2 = \sum_{x=0}^{1} (x-p)^2 p^x (1-p)^{1-x} = pq$$

Thus we see that $\frac{x-np}{\sqrt{npq}}$ has the same form as z in the central limit theorem. The theorem may then be applied.

We may write

$$\frac{x-np}{\sqrt{npq}} = \frac{\frac{x}{n} - p}{\sqrt{pq/n}} \qquad (2)$$

Thus the propertion x/n will be approximately normally distributed with mean p and variance pq/n if n is sufficiently large.

We can use this result to construct a control chart for the fraction defective. This chart is illustrated below.

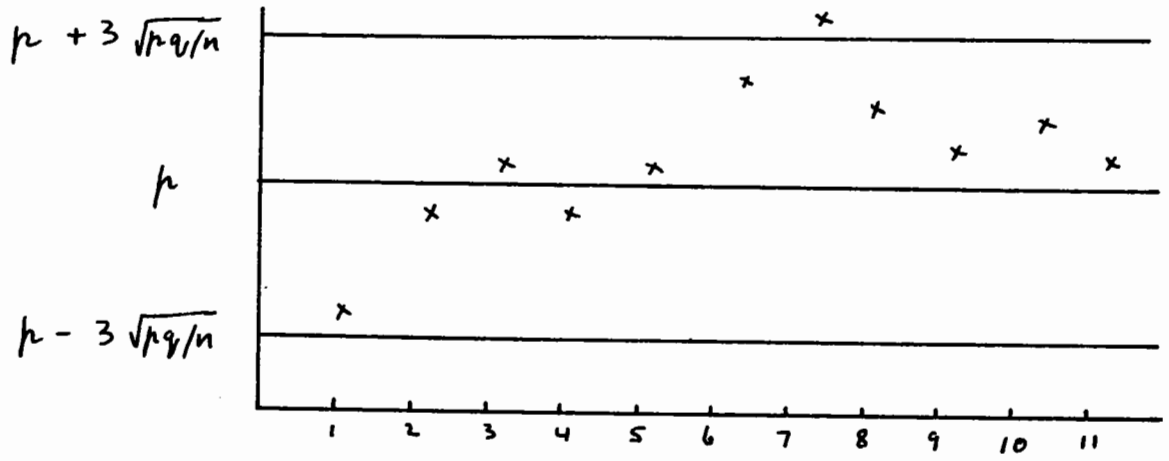

FIGURE 2. CONTROL CHART FOR FRACTION DEFECTIVE

The middle line is thought of as corresponding to the process proportion defective, although it is usually merely the mean of past sample proportions and as such is expected to be a very good approximation of the process proportion defective. The other two lines serve as control limits for sample proportions. These control lines are spaced three standard deviations from the mean line. Along the x-axis are recorded true units for successive samples. By the same argument that we used for the mean we know that the process is in statistical control if all of the sample points lie within the control band. The chart in figure 2 shows that the process is not in statistical control because the seventh sample point lies outside the control band.

The Range

We mentioned earlier in the chapter that we must estimate σ by means of the sample standard deviation. Now, the repeated computation of standard deviations is undesirable because the amount of computation becomes burdensome. It is customary to use the range, which is the difference between the largest and smallest value in the sample, as a substitute for the sample standard deviation in

estimating σ. Not only is the range easy to compute, but it can be shown that, for small samples from a normal population, the range is nearly as efficient for estimating σ as is the sample standard deviation.

We will now investigate the relationship between the range and the standard deviation for a normal distribution. This relationship may be found by calculating the mean of the range R.

$$E(R) = \int_0^{b-a} Rg(R)\,dR \qquad (3)$$

where g(R) is the frequency function of the range and the basic variable x assumes values in the interval (a,b). The distribution of the range is developed in chapter 5. It is clear from (7) of chapter 5 that the evaluation of E(R) will give rise to a complicated double integral. When f(x) is a normal frequency function, these integrations cannot be performed directly for general n; therefore numerical methods of integration are required. Tables are available for the normal variable case which express $E(R) = \mu_R$ in terms of σ for various values of n. The following are a few entries from such a table to indicate the nature of the relationship.

TABLE 1

n	2	3	4	5	10	50	100
$\frac{\mu_R}{\sigma}$	1.128	1.693	2.059	2.326	3.078	4.498	5.015

As an illustration of the use of the above table, consider once more the technique of constructing a control chart for the sample mean $\bar{x}$ as given above. There, a three standard deviation band was constructed for controlling $\bar{x}$. If the range is taken as the measure of variability, σ will be replaced by μ_R/d_n where d_n is the value obtained from the table, that is, the value of the ratio μ_R/σ corresponding to the given value of n. The value of μ_R can be estimated by using the sample mean of the R values obtained for a fairly large number of samples of size n each. n is usually chosen to be an integer near 4. Since μ_R is estimated on the basis of a large number of samples of this size, this estimate is usually quite accurate.

The range has two important disadvantages. First, its value usually increases with n because there is a better chance of obtaining extreme values if a large sample of data is taken than if a small sample is taken. Secondly, the range is usually quite unstable in repeated sampling experiments of the same size when n is large. But if n is chosen less than 10, the estimation of σ by means of the range, rather than the sample standard deviation, is quite accurate. We, therefore, conclude that the range is nearly as good as the sample standard deviation as an estimate of σ for small samples.

Chapter 3

Single and Double Sampling Inspection by Method of Attributes

In a mass production process, suppose articles are produced in lots of N articles each, and suppose each article, upon inspection, can be classified as defective or nondefective. It is often uneconomical to carry out a program of 100 per cent inspection. As an alternative, sampling methods of inspection applicable to each lot have been developed which have the property of guaranteeing that the percentage of defectives remaining after applying the sampling inspection procedure in the long run (that is to a large number of lots) is not more than some preassigned value. Such sampling methods have been developed and put into operation by Dodge and Romig of the Bell Telephone Laboratories. It should be noted that those sampling methods are essentially screening devices for reducing defectives after production, and are not devices for removing the causes of defectives. Dodge and Romig have developed two types of inspection sampling, single sampling and double sampling, which will be considered in turn.

Single Sampling Inspection

Let p be the fraction of defectives in a lot of size N. The number of defectives will be pN. Now let a random sample of size n be drawn from the lot. The probability of obtaining m defectives (and n-m nondefectives) in the sample is

$$P_{m,n,pN,N} = \frac{\binom{pN}{m} \cdot \binom{N-pN}{n-m}}{\binom{N}{n}} \qquad (1)$$

$m = 0,1,2,\ldots,r$ where r is the smaller of n and Np. Let

$$F(c,p,N,n) = P(m \leq c) = \sum_{m=0}^{c} P_{m,n,pN,N} \qquad (2)$$

If any two values of p and p' (pN and p'N being integers) are such that $p < p'$, then it can be shown that

$$F(c,p,N,n) > F(c,p',N,n) \qquad (3)$$

Let p_t be the lot tolerance fraction defective, that is the maximum allowable fraction defective in a lot, which is arbitrarily chosen in advance (that is, .01 or .05). Let

$$P_c = F(c,p_t,N,n) \qquad (4)$$

P_c is known as the consumer's risk; it is approximately the probability that a lot with lot tolerance fraction defective p_t will be accepted without 100 per cent inspection. It follows from (3) that if the lot fraction defective p exceeds p_t then the probability of accepting such a lot on the basis of the sample is less than the consumer's risk. The probability of subjecting a lot with fraction defective actually equal to $\bar{p}$ (process average) to 100 per cent inspection is

$$P_p = 1 - F(c,\bar{p},N,n) \qquad (5)$$

which is called producer's risk. It will be noted from (3) that the smaller the value of $\bar{p}$, the smaller will be the producer's risk. The producer's risk and consumer's risk are highly analogous to type I and type II errors, respectively, in the theory of testing statistical hypotheses as developed by Neyman and Pearson.

Suppose we make the following rules of action with reference to a sampled lot where c is chosen for given values of P_c, p_t, N, n:

(a) Inspect a sample of n articles.

(b) If the number of defectives in the sample does not exceed c, accept the lot.

(c) If the number of defectives in the sample exceeds c, inspect the remainder of the lot.

(d) Replace all defectives found by nondefective articles.

Let us consider the problem of determining the mean value of the fraction defectives remaining in a lot having fraction defective p, after applying rules (a) to (d). The probability of obtaining m defectives in a sample of size n is given by (1). If these m defectives are replaced by nondefective articles and the sample is returned to the lot, the lot will contain pN-m defectives. The fraction of defectives reamining after applying rules (a) to (d) has the distribution $\frac{pN-m}{N}$ with probability $P_{m,n,pN,N}$ for $m = 0,1,2,\ldots,c$. Thus the mean value of the fraction defectives reamining after applying rules (a) to (d) is

$$\tilde{p} = \sum_{m=0}^{c} \left(\frac{pN-m}{N}\right) P_{m,n,pN,N} \qquad (6)$$

Note that when $m > c$ the fraction of defectives after inspection is equal to zero since all defectives are replaced by nondefectives. The statistical interpretation of (6) is as follows: If a large number of lots each with fraction defective p are inspected according to rules (a) to (d), then the average fraction defective in all of these lots after inspection is $\tilde{p}$. For given values of c,n, and N, $\tilde{p}$ is a function of p, defined for those values of p for which Np is an integer, which has a maximum with respect to p. Denoting this maximum by $\tilde{p}_L$, it is called <u>average outgoing quality limit</u>. It can be shown that the larger the value of p, beyond the

value maximizing $\tilde{p}$ the smaller will be the value of $\tilde{p}$. The reason for this is that the greater the value of p the greater the probability that each lot will have to be inspected 100 per cent.

If the consumer's risk, n, and N are chosen in advance, then c and hence $\tilde{p}_L$ is determined. Thus, we are able to make the following statistical interpretation of those results: If rules (a) to (d) are followed for lot after lot and for given values of c,n,N, the average fraction defective per lot after inspection never exceeds $\tilde{p}_L$ no matter what fractions defective exist in the lots before the inspection.

There are various combinations of values of c and n, each having a $\tilde{p}$ with maximum $\tilde{p}_L$ (approximately) with respect to p.

The mean value of the number of articles inspected per lot for lots having fraction defective p is given by

$$I = n + (N-n)\left[1 - F(c, p, N, n)\right] \qquad (7)$$

since n (the number in the sample) will be inspected in every lot and N-n (the remainder in the lot) will be inspected if the number of defectives in the sample exceeds c.

There are two methods of specifying consumer protection.

(1) Lot Quality Protection

By considering the various combinations of values of c and n corresponding to a given consumer's risk, Pc, and lot tolerance fraction defective, p_t, there is, in general, a unique combination for $p = \bar{p}$ and for given N for which I is minimized.

(2) Average Quality Protection

Similarly by considering the various combinations of values of c and n corresponding to a given average outgoing quality limit, $\tilde{p}_L$, there is, in general, a unique combination for $p = \bar{p}$ and for given N

for which I is minimized.

In both cases the amount of inspection is reduced to a minimum which is valuable from a practical point of view. Extensive tabulations of pairs of values of c and n, for given values of consumer's risk, Pc, and outgoing quality limit, $\tilde{p}_L$, have been prepared by Dodge and Romig.

As an illustration consider a lot of 1000 pieces for which the process average fraction defective is $\bar{p} = .01$ and for which the consumer is willing to assume a risk of Pc = .10 of accepting a lot with a fraction defective of $p_t = .05$. By allowing c to assume small integral values and working numerically by trial and error methods, it will be found that the minimum amount of inspection will occur if a sample of 130 is taken and if the maximum allowable number of defectives is 3. With these values of n and c, it will also be found that the mean number of pieces inspected will be 164 so long as production remains in control. If the consumer requests an average outgoing quality limit of, say, $\tilde{p}_L = .03$, the minimum amount of inspection will occur if c = 2 and n = 44. These results are easily obtained by consulting the Dodge and Romig tables.

Double Sampling Inspection

In double sampling inspection from a given lot of size N, the procedure for taking action regarding a given lot is as follows:

(a) A first sample of size n_1 is drawn from the lot.

(b) If the number of defectives is $\leq c_1$, the lot is accepted without further sampling.

(c) If the number of defectives in the first sample exceeds c_2, inspect the remainder of the lot.

(d) If the number of defectives in the first sample exceeds c_1 but not c_2, inspect a second sample of size n_2.

(e) If the total number of defectives in both samples does not exceed c_2, accept the lot.

(f) If the total number of defectives in both samples exceeds c_2, inspect the remainder of the lot.

(g) Replace all defectives found by nondefective articles.

As in the case of single sampling, we have two kinds of consumer protection: (i) lot quality protection and (ii) average quality protection.

Consumer risk, the probability of accepting a lot with fraction defective p_t without 100 per cent inspection, is given by

$$P_c = \sum_{m=0}^{c_1} P_{m,n_1,p_tN,N} + \sum_{i=1}^{c_2-c_1} \sum_{m=0}^{c_2-c_1-i} \left(P_{c_1+i,n_1,p_tN,N}\right)\left(P_{m,n_2,p_tN-c_1-i,N-n_1}\right) \quad (8)$$

The single run in this formula is simply the probability of accepting the lot on the basis of the first sample and the double sum is the probability of accepting the lot on the basis of the first and second samples combined after having failed to accept on the basis of the first sample alone.

The mean value of the fraction defectives remaining after the defectives have been removed by the double sampling procedure, for lots having fraction defective p originally, is given by

$$\bar{p} = \sum_{m=0}^{c_1} \left(\frac{Np-m}{N}\right) P_{m,n_1,pN,N} + \sum_{i=1}^{c_2-c_1} \sum_{m=0}^{c_2-c_1-i} \left(\frac{Np-(c_1+i+m)}{N}\right)\left(P_{c_1+i,n_1,pN,N}\right)\left(P_{m,n_2,pN-c_1-i,N-n_1}\right) \quad (9)$$

The mean value of the number of articles inspected per lot for lots having fraction defective p is

$$\bar{I} = n_1 + n_2 \left(1 - \sum_{m=0}^{c_1} P_{m, n_1, pN, N}\right) + (N - n_1 - n_2)(1 - P_c) \quad (10)$$

where Pc is the value of the probability given in (8) with p_t replaced by p.

For given values of N, n_1, n_2, c_1, c_2, $\tilde{p}$ is a function of p, defined for those values of p for which Np is an integer, and has a maximum value $\tilde{p}_L$, the average outgoing quality limit. For a given value of N there are many values of n_1, n_2, c_1, and c_2 which will yield the same value of $\tilde{p}_L$ (approximately), or will yield the same consumer risk (approximately) for a given lot tolerance fraction defective. Dodge and Romig have arbitrarily chosen as the basis for the relationship between n's and c's the following rule: To determine n_1 and n_2 such that for given values of c_1 and c_2, n_1 and c_1 provides the same consumer risk (approximately) as n_1+n_2 and c_2. Even after this restriction there is enough choice left for combinations of n_1, n_2, c_1, c_2 to minimize I. To determine the n's and c's under these conditions for given N, for given consumer risk, (or average outgoing quality) involves a considerable amount of computation. Dodge and Romig have prepared tables for double sampling analogous to those for single sampling.

For a given amount of consumer protection, a smaller average amount of inspection is required under double sampling than under single sampling, particularly for large lots and low process average fraction defective $\bar{p}$.

Chapter 4

Sequential Method

In industrial sampling inspection we are interested in minimizing the amount of inspection needed to attain certain objectives. We stated in the preceding chapter that double sampling requires, on the average, fewer observations than single sampling to achieve the same results. However, these methods require that the sample size be fixed in advance. In the sequential method the sample size is not fixed in advance but is determined during the course of the sampling which may terminate at any observation. Using this method we often arrive at a decision with fewer observations, on the average, than the fixed size sample method possessing the same type I and type II errors. The saving in observations is sometimes more than 50 per cent with a resulting decrease in the cost of sampling. Sequential testing has been developed only for the case of testing an hypothesis H_0 against a single alternative H_1. However, in practical problems this restriction is not serious since we can almost always frame the test in terms of a single alternative. The reason for the advantage of the sequential approach over the fixed size sample approach lies in the ability of the sequential method to reach an early decision for samples that are extremely favorable to either H_0 or H_1. This ability to arrive at an early decision is very useful in sampling inspection where it is not uncommon for lots to be very bad when they are bad or very good when they are good.

For the purpose of describing a sequential test, consider a continuous random variable x whose frequency function $f(x;\theta)$ depends upon the parameter θ. Although the sequential test will be described for a continuous variable, it may be applied to either discrete or continuous variables. Suppose we wish to test the hypothesis that $\theta = \theta_0$ against the alternative hypothesis that $\theta = \theta_1$. Let H_0 be the hypothesis $\theta = \theta_0$ and let H_1 be the alternative hypothesis that $\theta = \theta_1$. Observations are denoted by $x_1, x_2, \ldots$ where the subscripts give the order in which the observations are taken.

The sequential test employs the likelihood ratio

$$\lambda_m = \frac{\prod_{i=1}^{m} f(x_i;\theta_1)}{\prod_{i=1}^{m} f(x_i;\theta_0)}, \quad (m = 1, 2, \cdots) \qquad (1)$$

and two positive numbers A and B, with $A > 1$ and $B < 1$. As observations are made, we compute the ratios $\lambda_1, \lambda_2, \lambda_3, \ldots$ and continue taking observations as long as

$$B < \lambda_m < A \qquad (2)$$

If for some m $\lambda_m \leq B$, H_0 is accepted and the test is completed. If $\lambda_m \geq A$ for some m, H_0 is rejected (that is H_1 is accepted) and the test is completed. The procedure then is to continue sampling until λ_m falls outside the interval specified by (2). The sampling then ceases. Thus we must decide at every stage of the sampling whether to accept the hypothesis, to reject the hypothesis, or to continue sampling.

We will now show that the sampling cannot go on indefinitely. Let

$$z = \log \left[\frac{f(x;\theta_1)}{f(x;\theta_0)} \right] \qquad (3)$$

Then z will have some frequency function, say $g(z)$, which is determined by the frequency function of x. The sequence of observations $x_1, x_2, \ldots$ determines a sequence of z observations $z_1, z_2, \ldots$ The inequality (2) becomes

$$\log B < \sum_{i=1}^{m} z_i < \log A \qquad (4)$$

where log B is negative and log A is positive (since $B < 1$ and $A > 1$). Let $c = \log A - \log B$ and let p be the area under $g(z)$ between -c and c

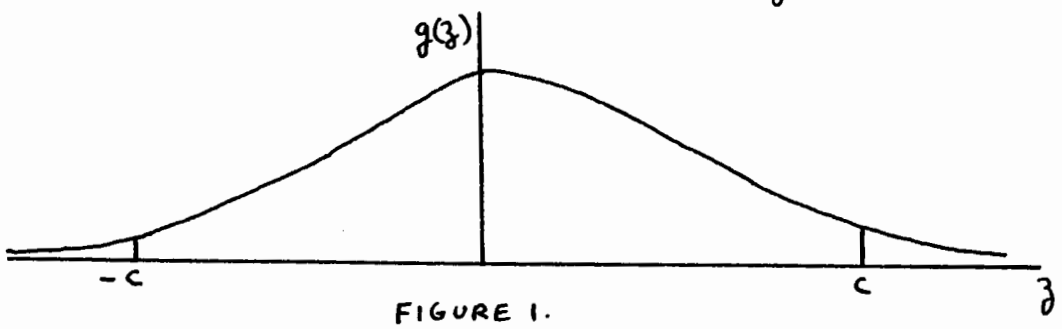

FIGURE 1.

If any one of the z_i falls outside the interval -c to c, the inequality (4) will be violated. Of course the inequality (4) may be violated even though all the z's do fall in the interval -c to c. Thus if (4) is to hold for all m, at the very least every z_i must fall between -c and c. The probability that every z_i falls in the interval is p^m for the first m observations. This probability approaches zero as m increases, since p is less than 1. Thus (4) cannot remain true indefinitely. It follows that the sampling will terminate after a finite number of observations. However, we never know how large a sample will be required to arrive at a decision because n, the sample size, is a random variable. A general formula

does exist for calculating the mean value of n, so that we can determine in advance how large n is likely to be. We state this formula without proof.

$$E(n) \cong \frac{P(\theta) \log A + [1 - P(\theta)] \log B}{E(z)} \qquad (5)$$

where $P(\theta)$ is the power function ((11) of chapter 1) of the test and z, A, B are defined by formulas (3), (6), (7) respectively.

The exact values of A and B are not available. However, excellent approximations are given by choosing

$$A \cong \frac{1-\beta}{\alpha} \qquad (6)$$

$$B \cong \frac{\beta}{1-\alpha} \qquad (7)$$

where α is the type I error and β is the type II error. It is beyond the scope of this thesis to discuss in detail the derivation of formulas (6) and (7). However, we can indicate briefly, if not completely, how these formulas are obtained. Suppose λ_m were a continuous function of a continuous variable m so that λ_m could be plotted as a curve against m. Suppose the test were performed by moving out along the m axis until λ_m first equaled A or B. That is, the test is continued as long as (2) is true and ceases when either $\lambda_m = B$ (H_0 accepted) or $\lambda_m = A$ (H_1 accepted). At all points of the sample space where H_0 is accepted, the likelihood of H_1, say L_1, is exactly B times the likelihood L_0 of H_0, since $\lambda = L_1/L_0 = B$ at these points. Therefore, the integral of L_1, over these points is exactly equal to B times the integral of L_0 over those points. But the first integral is β, by (10) of chapter 1, and the second is $1-\alpha$ (the probability of accepting H_0 when it is true). So we

would have β exactly equal to $B(1-\alpha)$ if continuous sampling were possible, and (7) would hold exactly. By a similar argument at $\lambda_m = A$, (6) would be an exact equality if m were a continuous variable. Since m is a discrete variable, formulas (6) and (7) are approximations. Investigations show that the error in using formulas (6) and (7) is quite small when both α and β are less than one-half.

Equations (6) and (7) make the actual performance of a sequential test very simple. We merely select α and β arbitrarily, compute A and B, and proceed with the test. The sequential test may be summarized as follows. To test the hypothesis H_0 against the alternative hypothesis H_1, calculate the likelihood ratio λ_m and proceed as follows:

$$\begin{aligned}
&\text{(i) if } \lambda_m \leq \frac{\beta}{1-\alpha}, \text{ accept } H_0 \\
&\text{(ii) if } \lambda_m \geq \frac{1-\beta}{\alpha}, \text{ reject } H_0 \text{ (accept } H_1) \qquad (8) \\
&\text{(iii) if } \frac{\beta}{1-\alpha} < \lambda_m < \frac{1-\beta}{\alpha}, \text{ take an additional observation}
\end{aligned}$$

Using this method we can decide in advance what size type I and type II errors to tolerate, rather than fix the type I error and then be forced to calculate the type II error as is usually done in fixed size sample tests.

As an example of how a sequential test is constructed, consider the problem of testing whether the mean of a normal variable with variance 1 has the value θ_0 or the value θ_1. H_0 is the hypothesis that $\theta = \theta_0$ and H_1 is the alternative hypothesis that $\theta = \theta_1$.

$$f(x;\theta) = \frac{e^{-\frac{1}{2}(x-\theta)^2}}{\sqrt{2\pi}}$$

(1) becomes

$$\lambda_m = \frac{\prod_{i=1}^{m} e^{-\frac{1}{2}(x_i-\theta_1)^2}}{\prod_{i=1}^{m} e^{-\frac{1}{2}(x_i-\theta_0)^2}} = \frac{e^{-\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta_1)^2}}{e^{-\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta_0)^2}}$$

$$= e^{(\theta_1-\theta_0)\sum_{i=1}^{m} x_i + \frac{m}{2}(\theta_0^2-\theta_1^2)}$$

Now property (iii) of (8) is equivalent to

$$\log \frac{\beta}{1-\alpha} < \log \lambda_m < \log \frac{1-\beta}{\alpha}$$

For this problem these inequalities become

$$\log \frac{\beta}{1-\alpha} + \frac{m}{2}(\theta_1^2-\theta_0^2) < (\theta_1-\theta_0)\sum_{i=1}^{m} x_i < \log \frac{1-\beta}{\alpha} + \frac{m}{2}(\theta_1^2-\theta_0^2)$$

If $\theta_1 > \theta_0$ this is equivalent to

$$\frac{1}{\theta_1-\theta_0}\log \frac{\beta}{1-\alpha} + \frac{m}{2}(\theta_0+\theta_1) < \sum_{i=1}^{m} x_i < \frac{1}{\theta_1-\theta_0}\log \frac{1-\beta}{\alpha} + \frac{m}{2}(\theta_0+\theta_1)$$

For $\theta_1 < \theta_0$ these inequalities would be reversed.

As a numerical illustration, suppose that

$$\alpha = .05 \, , \quad \beta = .10 \, , \quad \theta_0 = 9.5, \quad \theta_1 = 10$$

The last inequality then becomes

$$-4.50 + 9.75\, m < \sum_{i=1}^{m} x_i < 5.78 + 9.75\, m$$

The test now proceeds as follows:-

(i) If $\sum_{i=1}^{m} x_i \leq -4.50 + 9.75\, m$ accept $\theta = 9.5$

(ii) If $\sum_{i=1}^{m} x_i \geq 5.78 + 9.75\, m$ accept $\theta = 10$

(iii) If neither inequality is satisfied take another observation.

As a second example, consider the problem of testing whether $p = p_0$ or $p = p_1$ for a binomial distribution. If we choose $x = 1$ for success and $x = 0$ for failure, $f(x;\theta)$ will be given by $f(1;p) = p$ and $f(0;p) = q$. Suppose that the first m trials of the event produced d_m successes. Then the likelihood function $\prod_{i=1}^{m} f(x_i;\theta)$ will consist of the product of p's and q's, a p occurring as a factor whenever a success occurred and a q otherwise. The likelihood ration (1) then becomes

$$\lambda_m = \frac{p_1^{d_m} q_1^{m-d_m}}{p_0^{d_m} q_0^{m-d_m}}$$

If we substitute this expression in the sequential test and assign numerical values to p_0, p_1, α, β, we may proceed as we did in the previous example.

Suppose $p_0 = .5$, $p_1 = .7$, $\alpha = .10$, $\beta = .20$

$$\frac{\beta}{1-\alpha} = \frac{2}{9} \quad , \quad \frac{1-\beta}{\alpha} = 8$$

$$\lambda_m = \frac{(.7)^{d_m}(.3)^{m-d_m}}{(.5)^{d_m}(.5)^{m-d_m}} = \left(\frac{3}{5}\right)^m \left(\frac{7}{3}\right)^{d_m}$$

Inequality (i) in the sequential test gives

$$\lambda_m \leq \frac{\beta}{1-\alpha}$$

$$\text{i.e. } \left(\frac{3}{5}\right)^m \left(\frac{7}{3}\right)^{d_m} \leq \frac{2}{9}$$

This can be written more conveniently in the form

$$d_m \leq \frac{\log \frac{2}{9}}{\log \frac{7}{3}} + m \frac{\log \frac{5}{3}}{\log \frac{7}{3}}$$

In a similar manner inequality (ii) in the sequential test gives

$$\lambda_m \geq \frac{1-\beta}{\alpha}$$

$$\text{i.e. } d_m \geq \frac{\log 8}{\log \frac{7}{3}} + m \frac{\log \frac{5}{3}}{\log \frac{7}{3}}$$

If these logs are evaluated, the test proceeds as follows:

(i) If $d_m \leq -1.78 + .603\,m$, accept $p = .5$

(ii) If $d_m \geq 2.45 + .603\,m$, accept $p = .7$

(iii) If neither inequality is satisfied take another trial

For the purpose of determining when one of the inequalities is satisfied, it is convenient to represent these inequalities graphically (Figure 2). If m, d_m are treated as the coordinates of a point, the straight lines

$$d_m = -1.78 + .603\,m$$

$$d_m = 2.45 + .603\,m$$

will serve to divide the m, d_m plane into 3 regions corresponding to the 3 possible decisions at each trial.

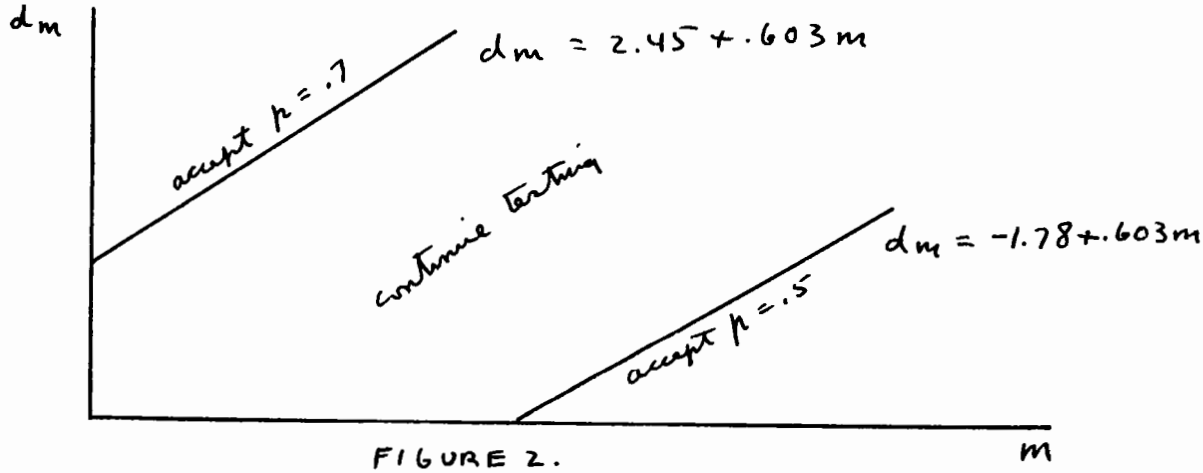

FIGURE 2.

The testing is continued only until the sample point crosses one or other of the two decision boundaries. These boundaries may be infinitely long, but in practice they are usually curtailed to force a decision one way or the other after so many trials. We proved that the probability is 1 that the sequential test will be completed after a finite number of observations.

As we stated earlier, sequential methods often reduce considerably the sample size needed to arrive at a reliable decision. In the preceding example it can be shown that a fixed size sample of approximately 26 will be sufficient to arrive at a decision. It can also be shown, in the theory of sequential analysis, that the average size sample needed to arrive at a decision in this example is approximately 13.

Chapter 5

Range and Tolerance Limits

Range

In chapter 2 we saw that the range was useful as a substitute for the standard deviation as a measure of variability in industrial quality control work. We will now derive an expression for the frequency function of the range.

Consider a random sample $x_1', x_2', \ldots, x_n'$ drawn from a population whose frequency function is $f(x)$, which is assumed to be continuous. Let these sample values be arranged in order of increasing magnitude, and denote the ordered set by $x_1, x_2, \ldots, x_n$. Now consider the problem of finding the probability that the smallest value x_1 and the largest value x_n will fall within specified intervals. The frequency function of the range can be found quite easily by means of this probability.

Let the x-axis be divided into 5 intervals $(-\infty, u)$, $(u, u+\Delta u)$, $(u+\Delta u, v)$, $(v, v+\Delta v)$, $(v+\Delta v, \infty)$, where $u < v$ are any two values of x. The probability that x will fall in any particular one of these intervals is given by the integral of $f(x)$ over that interval; hence the probabilities corresponding to these 5 intervals can be written down even though they cannot be evaluated unless the form of $f(x)$ is known. Let

$$P_2 = \int_u^{u+\Delta u} f(x)\,dx\,, \quad P_3 = \int_{u+\Delta u}^{v} f(x)\,dx\,, \quad P_4 = \int_v^{v+\Delta v} f(x)\,dx \qquad (1)$$

Now let us determine the probability that in a sample of n values of x we will obtain no value in the first interval, at least 1 value in each of the second interval and the fourth interval, and no value in the fifth interval. This procedure is equivalent to finding the probability that the smallest value in the sample will fall between u and $u+\Delta u$ while the largest value falls between v and $v+\Delta v$. The desired probability can be obtained directly from the multinomial distribution by treating x as a discrete variable which can assume only 1 of 5 possible values corresponding to the 5 intervals. If P_1 and P_5 denote the probabilities that x will fall in the first and fifth intervals, respectively, the desired probability is given by the following sum

$$\frac{n!}{0!\,1!\,(n-2)!\,1!\,0!}\,P_1^{0}\,P_2^{1}\,P_3^{n-2}\,P_4^{1}\,P_5^{0} + \sum_{i,j}\frac{n!}{0!\,i!\,(n-i-j)!\,j!\,0!}\,P_1^{0}\,P_2^{i}\,P_3^{n-i-j}\,P_4^{j}\,P_5^{0} \qquad (2)$$

For the last sum of (2) at least one of i, j should be greater than 1, while the first term of (2) corresponds to the case when $i = j = 1$ and consequently $n-i-j = n-2$. Now (2) reduces to

$$n(n-1)\,P_2\,P_4\,P_3^{n-2} + \sum_{i,j}\frac{n!}{i!\,(n-i-j)!\,j!}\,P_2^{i}\,P_3^{n-i-j}\,P_4^{j} \qquad (3)$$

(3) can be simplified somewhat by simplifying the integrals of (1). Since f(x) is assumed to be a continuous function, the mean value theorem for integrals may be applied here. This theorem states that if f(x) is continuous on the interval (α, β), then

$$\int_{\alpha}^{\beta} f(x)\,dx = (\beta-\alpha)\,f(\tau), \qquad \alpha < \tau < \beta$$

a direct application of this theorem to (1) gives,

$$P_2 = \Delta u \cdot f(u + \theta_1 \Delta u) \quad , \qquad 0 \le \theta_1 \le 1$$

$$P_4 = \Delta v \cdot f(v + \theta_2 \Delta v) \quad , \qquad 0 \le \theta_2 \le 1$$

$$P_3 = \int_{u+\Delta u}^{v} f(x)\,dx = \int_{u}^{v} f(x)\,dx - \int_{u}^{u+\Delta u} f(x)\,dx$$

$$= \int_{u}^{v} f(x)\,dx - \Delta u \cdot f(u + \theta_1 \Delta u)$$

If these values for P_2, P_3, P_4 are inserted in (3), it becomes

$$n(n-1) f(u+\theta_1\Delta u) f(v+\theta_2\Delta v)\left[\int_{u}^{v} f(x)dx - \Delta u f(u+\theta_1\Delta u)\right]^{n-2}\Delta u\,\Delta v$$

$$+ \sum_{i,j} T_{ij}(u, v, \Delta u, \Delta v)(\Delta u)^i (\Delta v)^j \qquad (4)$$

where at least one of i, j in the above summation is greater than 1.

This expression is the probability that the smallest value of the sample, x_1, will lie between u and $u+\Delta u$, and at the same time the largest value of the sample, x_n, will lie between v and $v+\Delta v$.

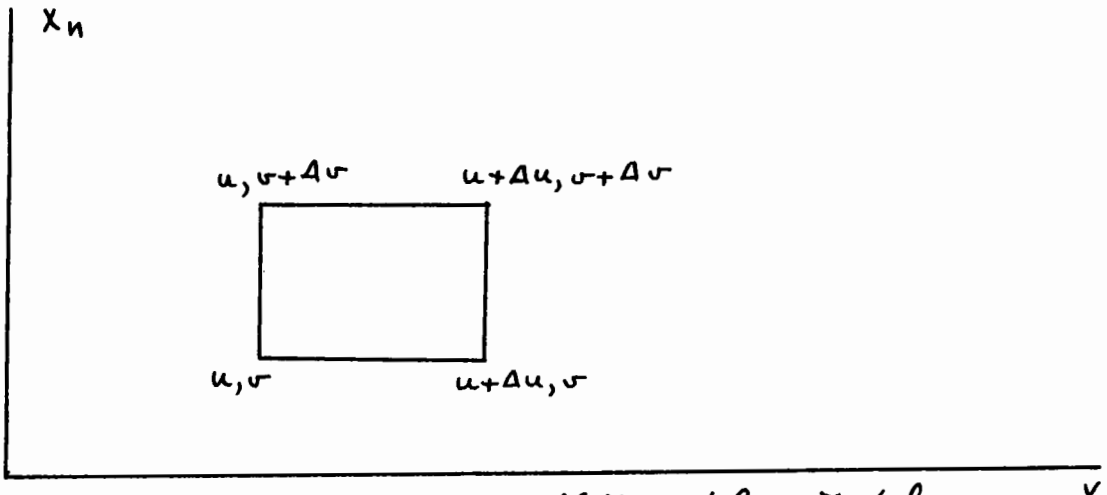

FIGURE 1. Sample space for smallest and largest values

Geometrically (4) gives the probability that the point (x_1, x_n) will lie inside the rectangle sketched above. In order to find the frequency function of the two variables x_1 and x_n at the point (u,v) it is necessary to divide (4) by the area of the rectangle, namely $\Delta u \Delta v$, and take the limit of the resulting quotient as Δu and Δv approach 0. If this frequency function is denoted by $f(u,v)$, it following from (4) that

$$f(u,v) = n(n-1)f(u)f(v)\left[\int_u^v f(x)\,dx\right]^{n-2} \qquad (5)$$

In the second term of (4) at least one of $i, j > 1$. Therefore, as Δu and Δv approach 0 so does this term.

Since $f(u,v)$ is the frequency function of the variables x_1 and x_n at the arbitrary point (u,v), (5) gives the desired joint frequency function of the smallest and largest values of a sample of size n. We may state our results as follows:

If u and v denote the smallest and largest values, respectively, in a random sample of size n from a population with the continuous frequency function $f(x)$, then the joint frequency function of u and v is given by (5).

The frequency function for the range can be obtained very quickly from this result. Let $R = v-u$ represent a change of variable from v to R with u held fixed. Then by (6) of chapter 1 the joint distribution of u and R is given by

$$g(u,R) = f(u,u+R)\left|\frac{\partial(u,v)}{\partial(u,R)}\right| = f(u,u+R)$$

$$g(u,R) = n(n-1)f(u)f(u+R)\left[\int_u^{u+R} f(x)\,dx\right]^{n-2} \qquad (6)$$

In order to obtain the frequency function of R, say g(R), it is necessary to integrate g(u,R) with respect to u over the range of u when R is fixed. If the range of x is from a to b, then the range of u with R fixed will be from a to b-R. This upper limit arises from the fact that u is always R units smaller than v and v cannot exceed the upper limit b for x. We may express these results as follows: If the continuous variable x has the frequency function f(x) and if x assumes values in the interval (a,b) only, then the frequency function of the range, g(R), for a random sample of size n is given by

$$g(R) = n(n-1)\int_{a}^{b-R} f(u)\,f(u+R)\left[\int_{u}^{u+R} f(x)\,dx\right]^{n-2} du \qquad (7)$$

Tolerance Limits

Consider the problem of determining the range of variability of some quality characteristic of a product coming off a production line. The producer is interested in knowing how this characteristic varies, because the consumer may reject a purchased lot if the variation is beyond certain limits. If it is known that the characteristic is approximately normally distributed, normal curve methods based on the sample mean and sample standard deviation can be used to determine an interval within which the characteristic would be expected to lie. Experience might show, however, that the distribution of the characteristic differs considerably from normality. Therefore, we require a method for determining such an interval without the necessity of a normality assumption. We will now discuss a method which does not require a knowledge of the form

of the frequency function. Such a method is called non-parametric.

Let a sample of size n be drawn from a population with the frequency function f(x) which is known to be continuous but otherwise unspecified. Consider two functions of the sample, $L_1(x_1,x_2,\ldots,x_n)$ and $L_2(x_1,x_2,\ldots,x_n)$, such that $L_1 \leq L_2$, and such that a fixed percentage of the population may be expected to lie in the interval (L_1,L_2) regardless of f(x). Functions such as L_1 and L_2 are called tolerance limits. By choosing the functions L_1 and L_2 so that a high percentage of the population will ordinarily be found to lie in the interval (L_1,L_2), the desired interval will be obtained. If L_1 and L_2 are chosen as the smallest and largest values, respectively, that occur in the sample, it will be found that the percentage of the population which can be expected to lie between L_1 and L_2 does not depend on the form of f(x).

Since the variable x possesses the continuous frequency function f(x), we may apply the relationship (5) established in the preceding section. Thus if u and v denote the smallest and largest values, respectively, in a random sample of size n from a population with the continuous frequency function f(x), the joint frequency function of u and v will be given by

$$f(u,v) = n(n-1)f(u)f(v)\left[\int_u^v f(x)\,dx\right]^{n-2}$$

Now the integral $\int_u^v f(x)dx$ is precisely the desired proportion of the population lying between the extreme values of the sample.

Consider the frequency function of w and z, say k(w,z), where

$$z = \int_u^v f(x)\,dx \quad , \quad w = \int_{-\infty}^{u} f(x)\,dx \tag{8}$$

By (6) of chapter 1 it follows that

$$k(w,z) = f(u,v)\left|\frac{\partial(u,v)}{\partial(w,z)}\right| = n(n-1)z^{n-2} \tag{9}$$

Now the frequency function of z, say h(z), may be obtained by integrating k(w,z) with respect to w over the range of w when z is fixed. Observe that w+z is the probability that x will not exceed v. Therefore, $w+z \leq 1$. Since z is fixed, w can assume values from 0 to 1-z only. Thus

$$h(z) = \int_0^{1-z} k(w,z)\,dw = \int_0^{1-z} n(n-1)z^{n-2}\,dw = n(n-1)z^{n-2}(1-z) \tag{10}$$

We may state our results as follows:
If a variable possesses a continuous frequency function and if z denotes the proportion of the population that lies between the extreme values of a random sample of size n drawn from this population, then the frequency function of z is given by (10).

As an example consider the problem of determining how large a sample must be taken in order to be certain with a probability of 0.95 that at least 99% of the population will lie between the extreme values of the sample. The solution is given by determining the value of n that satisfies the equation

$$\int_{.99}^{1} h(z)\,dz = .95$$

If the value of h(z) given in (10) is inserted and the integration is performed, this equation becomes

$$n(n-1)\left[\frac{1}{n-1} - \frac{1}{n} - \frac{(.99)^{n-1}}{n-1} + \frac{(.99)^{n}}{n}\right] = .95$$

which simplifies to

$$(.99)^{n} = \frac{4.95}{n+99}$$

It will be found by trial and error methods that the integer that most nearly satisfies this equation is n = 473. Thus a sample of size 473 is required in order to be certain with a probability of 0.95 that at least 99% of the population will lie between the extreme values of the sample. It is clear from this example that a very large sample is necessary before the extreme values will suffice to set limits within which practically all the population would be expected to lie.

The transcendental equation that arises in determining the value of n for problems of this type is not easy to solve. Consequently a simple approximate solution is highly desirable. Such an approximation, which is surprisingly good, is given by the formula

$$n \cong \frac{1}{4}\chi^2_\alpha \cdot \frac{1+\delta}{1-\delta} + \frac{1}{2} \qquad (11)$$

Where δ is the proportion of the population to be covered by the sample range, α is 1 minus the desired probability, and χ^2_α is the value of χ^2 for 4 degrees of freedom for which $P(\chi^2 > \chi^2_\alpha) = \alpha$. If formula (11) is applied to the above problem, we get

$$n \cong \frac{1}{4}(9.488)\frac{1.99}{.01} + \frac{1}{2} = 473$$

Chapter 6

Theory of Runs

In the statistical methods that we have considered in the preceding chapters, we have assumed that the observations constitute a random sample from a fixed population. However, we may suspect, when we take a set of observations over some time interval, that these observations do not behave like a random sample. It then becomes necessary to test the randomness of the sample before the usual statistical methods based on randomness can be applied. The object of this chapter is to discuss a method for testing randomness which is based on frequency functions of runs. This method does not depend on the frequency function of the basic variable and is therefore known as a nonparametric method.

Consider an arbitrary sequence of n elements, each element being one of several mutually exclusive kinds. Each sequence of elements of one kind, bounded by elements of another kind or no element, is called a run. The simplest case is that in which there are two kinds of elements. We shall consider this case in detail, and also briefly mention some results for the case of several kinds of elements.

Two Kinds of Elements

Suppose we have n_1 a's and n_2 b's ($n_1+n_2=n$). Let r_{1j} denote the number of runs of a's of length j and r_{2j} denote the number of runs of b's of length j. For example, if the arrangement is

aaabbaabaabbbab

then $r_{11}=1$, $r_{12}=2$, $r_{13}=1$, $r_{21}=2$, $r_{22}=2$ and the other r's are zero.

Observe that $\sum_j j r_{1j} = n_1$, the number of a's, and $\sum_j j r_{2j} = n_2$, the number of b's. Let $r_1 = \sum_j r_{1j}$ and $r_2 = \sum_j r_{2j}$ denote the total number of runs of a's and b's respectively. For a given set of numbers $r_{11}, r_{12}, \ldots, r_{1n_1}$ there are $\frac{r_1!}{r_{11}!\, r_{12}! \cdots r_{1n_1}!}$ ways of arranging the r_1 runs of a's. Similarly there are $\frac{r_2!}{r_{21}!\, r_{22}! \cdots r_{2n_2}!}$ ways of arranging the r_2 runs of b's.

Since the runs of a's and b's alternate, either $r_1 = r_2$, $r_1 = r_2 - 1$, or $r_1 = r_2 + 1$. If $r_1 = r_2 + 1$, the sequence must begin and end with an a. If $r_1 = r_2 - 1$, the sequence must begin and end with a b. However, if $r_1 = r_2$ the sequence can begin with either letter. For the first two cases there is no choice of beginning letter. However, for the third case ($r_1 = r_2$) a given arrangement of runs of a's can be fitted into a given arrangement of runs of b's in two ways, either with a run of a's first or with a run of b's first. Therefore, for the third case the number of arrangements is twice as large. In every case the total number of ways of getting the set r_{ij} ($i = 1,2$; $j = 1,2,\ldots,n_i$) is

$$N(r_{ij}) = \frac{r_1!}{r_{11}!\, r_{12}! \cdots r_{1n_1}!} \cdot \frac{r_2!}{r_{21}!\, r_{22}! \cdots r_{2n_2}!} \cdot F(r_1, r_2) \qquad (1)$$

where $F(r_1, r_2) = \begin{cases} 2 \text{ if } r_1 = r_2 \\ 1 \text{ if } r_1 \neq r_2 \end{cases}$

Since there are $\frac{n!}{n_1!\, n_2!}$ possible arrangements of a's and b's, each of which is equally likely, the joint frequency function of the given set r_{ij} is

$$P(r_{ij}) = \frac{r_1!}{r_{11}!\, r_{12}! \cdots r_{1n_1}!} \cdot \frac{r_2!}{r_{21}!\, r_{22}! \cdots r_{2n_2}!} \cdot \frac{F(r_1, r_2)}{\frac{n!}{n_1!\, n_2!}} \qquad (2)$$

Now let us determine the joint frequency function of the r_{ij}. It is easier at first to find the joint frequency function of r_{ij} and r_2. To do this we sum (2), for fixed r_{ij} and r_2, with respect to all r_{2j}. We wish to sum the multinomial coefficient $\frac{r_2!}{r_{21}!\, r_{22}! \cdots r_{2n_2}!}$ for all r_{2j} such that $\sum_{j=1}^{n_2} j\, r_{2j} = n_2$ and $\sum_{j=1}^{n_2} r_{2j} = r_2$.

In order to do this, consider the multinomial expression

$$\left(x + x^2 + \cdots + x^{n_2}\right)^{r_2} = \sum \frac{r_2!}{r_{21}!\, r_{22}! \cdots r_{2n_2}!}\, x^{r_{21} + 2r_{22} + \cdots + n_2 r_{2n_2}}$$

It is obvious that the coefficient of x^{n_2} in this expansion is $\sum \frac{r_2!}{r_{21}!\, r_{22}! \cdots r_{2n_2}!}$ under the condition $\sum_{j=1}^{n_2} j\, r_{2j} = n_2$, $\sum_{j=1}^{n_2} r_{2j} = r_2$.

However, it is clear that this is also the coefficient of x^{n_2} in the infinite expression $(x + x^2 \ldots)^{r_2}$. Now

$$\left(x + x^2 + \cdots\right)^{r_2} = \left(\frac{x}{1-x}\right)^{r_2} = \frac{x^{r_2}}{(1-x)^{r_2}}$$

$$= x^{r_2} \sum_{\tau=0}^{\infty} \frac{(r_2 - 1 + \tau)!}{(r_2 - 1)!\, \tau!}\, x^{\tau}$$

The coefficient of x^{n_2} in the expression on the right hand side is the coefficient of the term for which $r_2 + t = n_2$, that is $t = n_2 - r_2$. Therefore

$$\sum \frac{r_2!}{r_{21}!\, r_{22}! \cdots r_{2n_2}!} = \frac{(r_2 - 1 + n_2 - r_2)!}{(r_2 - 1)!\, (n_2 - r_2)!}$$

$$= \frac{(n_2 - 1)!}{(r_2 - 1)!\, (n_2 - r_2)!}$$

Therefore the joint frequency function of the r_{ij} and r_2 is

$$P(r_{ij}, r_2) = \frac{r_1!}{r_{11}!\, r_{12}! \cdots r_{1n_1}!} \cdot \frac{(n_2 - 1)!}{(r_2 - 1)!\, (n_2 - r_2)!} \cdot F(r_1, r_2) \Big/ \frac{n!}{n_1!\, n_2!} \qquad (3)$$

Now we sum (3) with respect to r_2

$$\sum_{r_2} \frac{(n_2-1)!}{(r_2-1)!\,(n_2-r_2)!} \cdot F(r_1,r_2) = \frac{(n_2-1)!}{(r_1-2)!\,(n_2-r_1+1)!} \cdot 1 + \frac{(n_2-1)!}{(r_1-1)!\,(n_2-r_1)!} \cdot 2$$

$$+ \frac{(n_2-1)!}{r_1!\,(n_2-r_1-1)!} \cdot 1 = \frac{(n_2+1)!}{r_1!\,(n_2-r_1+1)!}$$

This gives us the joint frequency function of the r_{1j}

$$P(r_{1j}) = \frac{r_1!}{r_{11}!\,r_{12}!\cdots r_{1n_1}!} \cdot \frac{(n_2+1)!}{r_1!\,(n_2-r_1+1)!} \Big/ \frac{n!}{n_1!\,n_2!} \qquad (4)$$

with a similar expression holding for the joint frequency function of the r_{2j}.

Another important distribution is the joint frequency function of r_1 and r_2. We get this by summing (3), for fixed r_1 and r_2, with respect to all r_{1j}. This is the same method that we used to obtain (3) by summing (2) with respect to all r_{2j}. The result is

$$P(r_1,r_2) = \frac{(n_1-1)!}{(r_1-1)!\,(n_1-r_1)!} \cdot \frac{(n_2-1)!}{(r_2-1)!\,(n_2-r_2)!} \cdot F(r_1,r_2) \Big/ \frac{n!}{n_1!\,n_2!}$$

$$= \binom{n_1-1}{r_1-1} \cdot \binom{n_2-1}{r_2-1} \cdot F(r_1,r_2) \Big/ \binom{n_1+n_2}{n_1} \qquad (5)$$

We find the frequency function of r_1 by summing (5) with respect to r_2, obtaining

$$P(r_1) = \frac{(n_1-1)!}{(r_1-1)!\,(n_1-r_1)!} \cdot \frac{(n_2+1)!}{r_1!\,(n_2+1-r_1)!} \Big/ \frac{n!}{n_1!\,n_2!}$$

$$= \binom{n_1-1}{r_1-1} \cdot \binom{n_2+1}{r_1} \Big/ \binom{n_1+n_2}{n_1} \qquad (6)$$

We use the notation $nC_r = \binom{n}{r}$

The frequency function of the total number of runs of a's and b's is of considerable interest in the application of run theory. Let $u = r_1 + r_2$, the total number of runs. To find the frequency function of u we must sum (5) over all values of r_1 and r_2 such that $r_1 + r_2 = u$. We have two cases, (1) $u = 2k$ (even) and (2) $u = 2k-1$ (odd). If $u = 2k$ (even), there is one pair of values to consider, $r_1 = r_2 = k$. If $u = 2k-1$ (odd), there are two pairs of values to consider, $r_1 = k$, $r_2 = k-1$ or $r_1 = k-1$, $r_2 = k$. Hence, from (5) we have

$$P(u = 2k) = \binom{n_1 - 1}{k-1} \cdot \binom{n_2 - 1}{k-1} \cdot 2 \Big/ \binom{n_1 + n_2}{n_1}$$

$$P(u = 2k-1) = \frac{\binom{n_1 - 1}{k-1} \cdot \binom{n_2 - 1}{k-2} + \binom{n_1 - 1}{k-2} \cdot \binom{n_2 - 1}{k-1}}{\binom{n_1 + n_2}{n_1}} \qquad (7)$$

The function $P(u \leq u') = \sum_{u=2}^{u'} P(u)$ has been tabulated for various values of n_1 and n_2 and u'. Such tables enable us to test whether a sample value of u is unusually large or small as compared to what would be expected if the sequence of values constituted a random sequence.

Another probability function of considerable interest in the application of the theory of runs is the probability of getting at least one run of a's of length s or greater, or in other words the probability that at least one of the variables r_{1s}, r_{1s+1}, r_{1s+2}, ... in the distribution (4) is ≥ 1. Mosteller has solved this problem for the case $n_1 = n_2 = n$. To obtain this probability we put $n_1 = n_2 = n$ in (4), thus obtaining

$$P(r_{1j}) = \frac{r_1!}{r_{11}!\, r_{12}! \cdots r_{1n}!} \cdot \frac{(n+1)!}{r_1!\,(n - r_1 + 1)!} \Big/ \frac{(2n)!}{(n!)^2} \qquad (8)$$

and sum over all terms such that at least one of the variables $r_{is}, r_{i,s+1}, \ldots \geq 1$. We can accomplish the same thing by summing over all terms such that all of these variables are zero, and subtracting the result from unity. To do this we must sum the multinomial coefficient $\frac{r_i!}{r_{i1}!\, r_{i2}! \cdots r_{in}!}$ in (8) over all values of $r_{i1}, r_{i2}, \ldots, r_{in}$ such that $r_{is} = r_{i,s+1} = \cdots = r_{in} = 0$, $\sum_{j=1}^{n} j\, r_{ij} = n$, $\sum_{j=1}^{n} r_{ij} = r_i$, and then sum with respect to r_i.

By the same argument used to derive (3), we see that the sum of the multinomial coefficient is given by the coefficient of x^n in the expansion of

$$\begin{aligned}(x + x^2 + \cdots + x^{s-1})^{r_i} &= x^{r_i}(1 - x^{s-1})^{r_i}(1-x)^{-r_i} \\ &= x^{r_i}(1 - x^{s-1})^{r_i} \sum_{\tau=0}^{\infty} \binom{r_i - 1 + \tau}{r_i - 1} x^{\tau} \\ &= x^{r_i} \sum_{j=0}^{r_i} (-1)^j \binom{r_i}{j} x^{j(s-1)} \sum_{\tau=0}^{\infty} \binom{r_i - 1 + \tau}{r_i - 1} x^{\tau} \\ &= \sum_{\tau=0}^{\infty} \sum_{j=0}^{r_i} (-1)^j \binom{r_i}{j} \binom{r_i - 1 + \tau}{r_i - 1} x^{\tau + r_i + j(s-1)}\end{aligned}$$

Therefore, the sum of the multinomial coefficient is given by

$$\sum_{j=0}^{r_i} (-1)^j \binom{r_i}{j} \binom{n - j(s-1) - 1}{r_i - 1}$$

Thus the desired probability of at least one run of length s or greater is

$$P(\text{at least one } r_{ij} > 1,\ j \geq s) = 1 - \frac{\sum_{r_i} \sum_{j=0}^{r_i} (-1)^j \binom{r_i}{j} \binom{n - j(s-1) - 1}{r_i - 1} \binom{n+1}{r_i}}{\binom{2n}{n}} \qquad (9)$$

Applying similar methods to each of the multinomial coefficients in (2), Mosteller has obtained the probability of getting at least one run of a's or b's of length s or greater.

In order to indicate how to find moments of run variables, let us consider the case of r_1. We shall first find the factorial moments $E\left[X^{(a)}\right]$ where $x^{(a)} = x(x-1)(x-2)\cdots\cdots(x-a+1) = \frac{x!}{(x-a)!}$ for they are easier to find than ordinary moments in the present problem. From them the ordinary moments may be found since $E\left[X^{(i)}\right]$ is a linear function of the first i ordinary moments. Letting $i = 1,2,3,\ldots,a$ we obtain a system of a linear equations which may be solved to obtain the ordinary moments as linear functions of the factorial moments. We have

$$E\left[r_1^{(a)}\right] = \sum_{r_1=1}^{n_1} r_1^{(a)} P(r_1) = \sum_{r_1=1}^{n_1} \frac{r_1!}{(r_1-a)!} P(r_1) \qquad (10)$$

In order to evaluate (10) we use the following identity:

$$\sum_{i=0}^{B} \frac{A!}{(c+i)!(A-c-i)!} \cdot \frac{B!}{i!(B-i)!} = \frac{(A+B)!}{(c+B)!(A-c)!} \qquad (11)$$

which follows at once by equating coefficients of x^c in the expansion of

$$(1+x)^A \left(1+\tfrac{1}{x}\right)^B = \frac{(1+x)^{A+B}}{x^B} \qquad (12)$$

Substituting $P(r_1)$ from (6) into (10), simplifying, and using (11), we have

$$E\left[r_1^{(a)}\right] = (n_2+1)^{(a)} \sum_{r_1} \frac{(n_1-1)!}{(r_1-1)!(n_1-r_1)!} \cdot \frac{(n_2+1-a)!}{(r_1-a)!(n_2+1-r_1)!} \Big/ \frac{n!}{n_1!n_2!}$$

$$= (n_2+1)^{(a)} \frac{(n-a)!}{(n_1-a)!\,n_2!} \Big/ \frac{n!}{n_1!n_2!} \qquad (13)$$

From this result we find

$$E(r_1) = \frac{(n_2+1)n_1}{n}$$

$$\sigma^2_{r_1} = \frac{(n_2+1)^{(2)} n_1^{(2)}}{n \cdot n^{(2)}}$$

a similar expression holds for $E[r_2^{(a)}]$.

K Kinds of Elements

The theory of runs has been extended to the case of several kinds of elements by Mood. If there are k kinds of elements, say $a_1, a_2, \ldots, a_k$, denote by r_{ij} the number of runs of a_i of length j. Let r_i be the total number of runs of a_i. Mood has shown that the joint frequency function of the r_{ij} is given by

$$P(r_{ij}) = \prod_{i=1}^{k} \frac{r_i!}{r_{i1}!\, r_{i2}! \cdots r_{in_i}!} \cdot F(r_1, r_2, \ldots, r_k) \Big/ \frac{n!}{n_1!\, n_2! \cdots n_k!} \qquad (14)$$

where $F(r_1, r_2, \ldots, r_k)$ is the number of ways r_1 objects of one kind, r_2 objects of a second kind, and so on, can be arranged so that no two adjacent objects are of the same kind. The argument for establishing (14) is very similar to that for the case of $k = 2$. The joint frequency function of $r_1, r_2, \ldots, r_k$ is given by

$$P(r_1, r_2, \ldots, r_k) = \prod_{i=1}^{k} \binom{n_i - 1}{r_i - 1} \cdot F(r_1, r_2, \ldots, r_k) \Big/ \frac{n!}{n_1!\, n_2! \cdots n_k!} \qquad (15)$$

which we state without proof.

Application of Run Theory

Let $x_1', x_2', \ldots, x_n'$ denote a random sample of size n from a population with a continuous frequency function. Let $\tilde{x}$ be the median value of the sample. Each sample value greater than $\tilde{x}$ will

be called a and each sample value less than $\tilde{x}$ will be called b. In what follows we will assume that n is an even number, say $n = 2k$, so that there will be the same number of values (k) above and below the median. In this case we choose the median, $\tilde{x}$, to be the arithmetic mean of the two middle values in the sample. If n is an odd number, say $n = 2k+1$, then we ignore the median and the same arguments apply. Thus we have k a's and k b's in the sample.

Denote by $x_1, x_2, \ldots, x_n$ the ordered set of values corresponding to the above sample. Now any particular random variable in the sample has the same probability of occurring in a specified order position in the ordered set as any other variable. For example, the first sample value to be drawn, x_1', has the same probability of being the largest value, x_n, in the ordered set as the second sample value, x_2', has. Thus each of the n! possible permutations of the variables $x_1', x_2', \ldots, x_n'$ has the same probability of being the ordered set of values denoted by $x_1, x_2, \ldots, x_n$.

Let babbaa....a denote any permutation of the k a's and the k b's. Consider the number of the n! permutations of the x''s that will yield this particular permutation of a's and b's. The first b in this permutation means that the first sample value x_1' was smaller than the median. Thus x_1' could have occupied any one of the first k order positions in the ordered set. The first a in this permutation means that the second sample value x_2' was larger than the median. Thus x_2' could have occupied any one of the last k order positions in the ordered set. Hence, there are k choices of order positions for the first b and also for the first a. In a similar manner there are k-1 remaining choices of order positions for the second b and the second a. Filling order positions in this

manner, there are k! k! choices of order positions for the x''s to yield the above arrangement of a's and b's. This number of choices does not depend upon the particular permutation of a's and b's. Furthermore, all order permutations of the x''s have the same probability of occurring. Thus all distinct permutations of the a's and b's have the same probability of occurring.

It follows from the above discussion that all of the run frequency functions (2), (3), (4), (5), (6), (7) are applicable, for $n_1 = n_2 = n$, to all possible arrangements of the a's and b's. Thus we see that by classifying sample values into a's and b's and using the theory of runs we have a method for testing randomness in a sample as far as order is concerned.

The more commonly used tests of randomness based on run theory are:

(1) Number of runs of a's, for which the distribution is (6). For given values of n_1 and n_2, the test consists of finding the largest value of r_1 (the number of runs of a's), say r_1^o, for which $P[r_1 \geq r_1^o] \leq \epsilon$, e.g. for $\epsilon = .05$. A similar statement may be made concerning runs of b's.

(2) Total number of runs of a's and b's having distribution (7). Again, the test consists of finding the largest value of u, say u^o, for which $P[u \geq u^o] \leq \epsilon$, for given values of n_1 and n_2.

(3) At least one run of a's (or b's) of at least length s, for $n_1 = n_2 = n$, based on the distribution (9). The test consists of finding the smallest value of s for which the probability (9) is $\leq \epsilon$.

As an illustration of the application of the preceding theory consider the following example. Samples are taken every morning and afternoon from a production line to check the diameter of a

part. Suppose the following diameters are obtained.

.220, .213, .221, .214, .219, .214, .222, .216, .212, .221, .223, .214, .221, .216, .217, .215.

The median value of this sample is .218. If each value above the median is assigned the letter a while each value below the median is assigned the letter b, we obtain the following sequence of letters.

abaaababbaababbb

We will now use the total number of runs, u, in order to test the hypothesis of randomness in the above sequence. Here $u = 10$. We wish to see whether this value of u is large or small as compared to the number of runs expected in a randomly selected sequence of the same length. We stated earlier in the chapter that the function $P(u \leq u') = \sum_{u=2}^{u'} P(u)$ has been tabulated for various values of n_1 and n_2 and u'. Below we have a few entries from one of these tables.

TABLE 1

$n_1 = n_2$	5	10	15	20	25	30	40	50	60	70	80	90	100
$u_{.05}$	3	6	11	15	19	24	33	42	51	60	70	79	88
$u_{.95}$	8	15	20	26	32	37	48	59	70	81	91	102	113

In this table $u_{.05}$ is the value of u such that $P(u \leq u_{.05}) \leq .05$ and $u_{.95}$ is the value of u such that $P(u \leq u_{.95}) \geq .95$. These values may, therefore, be used as 5 per cent critical values for testing randomness against the alternative of too few or too many runs. In our example $u = 10$, $n_1 = 8$, and $n_2 = 8$. By interpolation we find that $u_{.05}$ is approximately 5 and $u_{.95}$ is approximately 12. The sample value of 10 is not critical. Thus we have no reason to doubt the randomness of the sample.

Too many runs may occur if the machines have a tendency to turn out larger parts in the morning and smaller parts in the afternoon. In this case we would get a sequence like the following. abababa..... Too few runs may occur if the machines gradually produce larger parts from day to day. In this case the earlier observations would be mostly below the median and the later ones above. We would then get a sequence like the following. bbbbaaaaaaa.....

BIBLIOGRAPHY

1. I. W. Burr — Engineering Statistics and Quality Control, McGraw - Hill, 1953

2. W. E. Deming — Some Theory of Sampling, John Wiley, 1950

3. H. F. Dodge & H. G. Romig — Sampling Inspection Tables, John Wiley, 1944

4. T. C. Fry — Probability and its Engineering Uses, Van Nostrand, 1928

5. P. G. Hoel — Introduction to Mathematical Statistics, John Wiley, 1954

6. A. M. Mood — Introduction to the Theory of Statistics, McGraw - Hill, 1950

7. W. A. Shewhart — Economic Control of Quality of Manufactured Product, Van Nostrand, 1931

8. S. S. Wilks — Mathematical Statistics, Princeton University Press 1946