HISTORICAL AND MATHEMATICAL DEVELOPMENT OF THE CHI-SQUARE DISTRIBUTION

A THESIS PRESENTED TO
THE FACULTY OF GRADUATE STUDIES
UNIVERSITY OF MANITOBA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

BY

ALLAN PHILLIP DONNER

MAY, 1967.

ABSTRACT

It was Karl Pearson who in 1900 discovered the goodness of fit criterion and showed that it followed the chi-square distribution. This significant discovery provided the impetus that has made the chi-square distribution one of the most useful in applied statistics. Its range of application is mainly accounted for by the non-parametric character of many chi-square tests.

In this thesis is given some indication of the variety of experimental problems to which chi-square may be applied. Supplementing this is an historical survey which traces the origins of the chi-square distribution back to the man who originally derived it in 1876, the German mathematician Helmert. In addition, several mathematical derivations of the curve are given, and its properties investigated.

ACKNOWLEDGEMENT:

The author would like to express his sincere appreciation to Professor G.I. Paul for his interest and encouragement in this work.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONT'D)

TABLE OF CONTENTS (CONT'D)

Page

INTRODUCTION

One of the most useful distributions in applied statistics is the chi-square distribution. It forms the underlying distribution for a multitude of statistical tests, including goodness of fit tests, and tests of independence.

Although the chi-square distribution was originally discovered in 1876 by the German mathematician H. Helmert, it was Karl Pearson in 1900 who introduced the goodness of fit criterion and showed that it asymptotically followed the chi-square distribution. Pearson's theoretical conclusions, however, were later proven to be entirely accurate only for the case in which the null hypothesis provided exact values for the expected frequencies. For the case in which population parameters must be estimated from sample data, Sir Ronald Fisher in 1924 showed that the distribution of the test criterion has one less degree of freedom for each such estimate made. His theoretical conclusions were subsequently backed up by various published sampling experiments.

Dr. F. Yates, who introduced the correction for continuity in 1934, H. Mann and A. Wald, who arrived at an expression for the "best" number of classes (1942), and George Barnard, who resolved problems of interpretation relating to the 2 x 2 contingency table (1947), were others who played significant roles in the development of the goodness of fit and other chi-square tests.

Because of its non-parametric nature, the chi-square test has found application in many statistical realms.

However, this by no means spells the extent of chi-square's contribution to statistical inference. In addition, it has valuable application in tests of independence, homogeneity tests, estimation of arbitrary parameters, experimental design and linear regression.

There are two basic forms of the chi-square statistic, the continuous form and the discrete form. Continuous chi-square is defined as the sum of squares of n independent, normally distributed variables, each with zero mean and unit variance. In notational form,

$\chi^2_{(n)} = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$, where the Xi are NID(μ,σ^2)

n is called the number of degrees of freedom.

The probability density function of the continuous chi-square statistic is given by

$$f(\chi^2) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} (\chi^2)^{\frac{n}{2}-1} e^{-\frac{\chi^2}{2}} \quad , 0 \leq \chi^2 \leq \infty$$

$\Gamma(\frac{n}{2})$ is the gamma function of $\frac{n}{2}$ where (m) is defined as

$$\Gamma(m) = \int_0^\infty X^{m-1} e^{-X} dX$$

Discrete chi-square is defined in terms of the observed and expected frequencies resulting from performing a random experiment a given number of times.

Formally, $\chi^2_{(n)} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

where k is the number of classes of frequencies

O_i is the observed frequency in the i^{th} class, $i = 1,2...k$

E_i is the expected frequency in the i^{th} class.

This is the goodness of fit criterion. The number of degrees of freedom n is interpreted as the number of independent expected frequencies. As n becomes infinitely large, the distribution of discrete chi-square approaches that of continuous chi-square.

In this thesis I will discuss the mathematical derivations of the chi-square distribution, describe its properties, explore its historical origins and background, and follow it through to its applications in modern statistics.

CHAPTER I

MATHEMATICAL DERIVATIONS OF THE CHI-SQUARE DISTRIBUTION

1. MATHEMATICAL INDUCTION

By definition, $\chi^2 = \sum_{i=1}^{n} X_i^2$, where the Xi are NID(0,1)

Let n = 1

Then $\chi^2 = X^2$, X is n (0,1)

$$f(X) = \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}} , -\infty < X < \infty$$

Put $Y = X^2$

$$X = \pm\sqrt{Y}$$

$$\frac{dX}{dY} = \frac{1}{2\sqrt{Y}}$$

$$\therefore g(y) = \frac{2}{\sqrt{2\pi}} e^{-\frac{Y}{2}} \frac{1}{2\sqrt{Y}} , 0 < Y < \infty$$

$$= \frac{1}{\sqrt{2\pi Y}} e^{-\frac{Y}{2}} , 0 < Y < \infty$$

But

$$\chi^2_{(1)} = \frac{1}{\Gamma(\frac{1}{2})2^{1/2}} e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{1}{2}-1} , 0 < \chi^2 < \infty$$

$$= \frac{1}{\sqrt{2\pi\chi^2}} e^{-\frac{\chi^2}{2}} , 0 < \chi^2 < \infty$$

X^2 follows a   dist. with 1 d.f.

Assume $\sum_{i=1}^{n} X_i^2$ is a $\chi^2_{(n)}$ and show under this assumption that

$$\sum_{i=1}^{n+1} X_i^2 \text{ is a } \chi^2_{(n+1)}$$

Let $U = X_1^2 + X_2^2 + \ldots X_n^2$

$V = X_{n+1}^2$

$$\therefore g(U) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} e^{-\frac{U}{2}} U^{\frac{n}{2}-1} , 0 < U < \infty$$

$$h(V) = \frac{1}{\sqrt{2\pi V}} e^{-\frac{V}{2}} , 0 < V < \infty$$

U and V are independent since the Xi's are independent.

$\therefore$ Joint density function of U and V is given by

$$\frac{1}{2^{m+1/2}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{1}{2}\right)} V^{-\frac{1}{2}} U^{\frac{m}{2}-1} e^{-\frac{U+V}{2}}, \quad \begin{array}{l} 0 < U < \infty \\ 0 < V < \infty \end{array}$$

Let Y = U + V $\qquad$ U = Y-Z

Z = V $\qquad$ V = Z

$$J = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1$$

$\therefore$ Joint density function of Y and Z is given by

$$\frac{1}{2^{m+1/2}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{1}{2}\right)} Z^{-\frac{1}{2}} (Y-Z)^{\frac{m}{2}-1} e^{-\frac{Y}{2}} \quad , \begin{array}{l} 0 \leq Y \leq \infty \\ 0 \leq Z \leq Y \end{array}$$

$\therefore$ Probability density function of Y is given by

$$K\int_0^Y Z^{-\frac{1}{2}}(Y-Z)^{\frac{m}{2}-1} dZ$$

Let Z = tY

dZ = Ydt

$\therefore$ Probability density function of Y becomes

$$K\int_0^1 (tY)^{-\frac{1}{2}}(Y-tY)^{\frac{m}{2}-1} Y dt = KY^{\frac{1}{2}}Y^{\frac{m}{2}-1}\int_0^1 t^{-\frac{1}{2}}(1-t)^{\frac{m}{2}-1}dt$$

$$= \frac{\beta\left(\frac{1}{2},\frac{m}{2}\right) Y^{\frac{m-1}{2}}}{2^{\frac{m+1}{2}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{1}{2}\right)} = \frac{Y^{\frac{m-1}{2}} e^{-\frac{Y}{2}}}{\Gamma\left(\frac{m+1}{2}\right) 2^{m+1/2}}$$

$\therefore$ Y = U + V follows the chi-square distribution with n + 1 degrees of freedom.

$\therefore$ By the principle of mathematical induction,

$\chi^2 = \sum_{i=1}^{m} X_i^2$, where the Xi are NID(0,1) follows the chi-square distribution with n degrees of freedom.

2. GAMMA DISTRIBUTION DERIVATION

Define $\Gamma(m) = \int_0^\infty e^{-X} X^{m-1} dX$

Then $f(x) = \dfrac{e^{-X} X^{m-1}}{\Gamma(m)}$, $0 < X < \infty$

is a probability density function called the gamma function with parameter n.

The moment generating function of the gamma distribution is $Mx(t) = E(e^{tx}) = \int_0^\infty \frac{1}{\Gamma(m)} e^{X(t-1)} X^{m-1} dX$

$$= \frac{1}{\Gamma(m)} \int_0^\infty e^{-X(1-t)} X^{m-1} dX$$

Let $U = X(1-t)$ $\quad X = U/1-t$

$dU = (1-t)\ dX \quad dX = dU/1-t$

$$Mx(t) = \frac{1}{\Gamma(m)} \int_0^\infty \frac{e^{-U} U^{m-1}}{(1-t)^{m-1}(1-t)} dU$$

$$= \frac{1}{\Gamma(m)} \frac{1}{(1-t)^m} \Gamma(m) = \frac{1}{(1-t)^m}, \quad t < 1$$

Let us find the distribution of the sum of two independent gamma variates, X_1 and X_2 with parameters ℓ and m respectively. Its moment generating function is

$M(x_1 + x_2)(t) = E(e^{t(x_1 + x_2)}) = E(e^{tx_1} e^{tx_2})$

$= E(e^{tx_1})E(e^{tx_2}) = Mx_1(t)Mx_2(t)$

$= \dfrac{1}{(1-t)^{\ell}} \dfrac{1}{(1-t)^m} = \dfrac{1}{(1-t)^{\ell+m}}$, which is the M.G.F.

of a gamma variate with parameter ℓ + m.

$\therefore$ Sum of 2 independent gamma variates with parameters l and m respectively is a gamma variate with parameter l + m.

Theorem 1

If X is $N(\mu,\sigma^2)$ then $\frac{\frac{1}{2}(X-\mu)^2}{\sigma^2}$ is a gamma variate with parameter $\frac{1}{2}$.

Proof

$$f(x)\,dX = \frac{1}{\sqrt{2\pi}}\sigma\, e^{-\frac{\frac{1}{2}(X-\mu)^2}{\sigma^2}}\,dX$$

Let $U = \frac{(X-\mu)^2}{2\sigma^2}$ $\quad dU = \frac{X-\mu}{\sigma^2}\,dX$

Also $X-\mu = \sigma\sqrt{2U}$; $X>\mu$

$X-\mu = \sigma\sqrt{2U}$; $X<\mu$

and $g(u)\,du = \frac{e^{-U}U^{-\frac{1}{2}}dU}{\Gamma(\frac{1}{2})}$

which shows that $U = \frac{1}{2}(X-\mu)^2$ is a gamma variate with parameter $\frac{1}{2}$.

Theorem 2

$\frac{1}{2}\chi^2$ is a gamma variate with parameter $\frac{n}{2}$ given that

$$\chi^2 = \sum_{i=1}^{m}\frac{(X_i-\mu)^2}{\sigma^2}$$

From the result in theorem 2 it follows that the probability density function of $\chi^2/2$ is

$$\frac{1}{\Gamma(\frac{m}{2})}\, e^{-\frac{\chi^2}{2}}\left(\frac{1}{2}\chi^2\right)^{\frac{m}{2}-1}$$

and the probability density function of χ^2 is

$$\frac{1}{\Gamma(\frac{m}{2})2^{\frac{m}{2}}}\, e^{-\frac{\chi^2}{2}}(\chi^2)^{\frac{m}{2}-1}\,d\chi^2$$

3. THE MOMENT GENERATING FUNCTION TECHNIQUE

Consider the probability density function

$$f(x) = \frac{1}{2^{m/2}\Gamma(\frac{m}{2})} \chi^{\frac{m}{2}-1} e^{-\chi/2}, \quad 0 < \chi < \infty$$

Where n is the number of "degrees of freedom".

The moment generating function of X is:

$$Mx(t) = E(e^{tX}) = \frac{1}{2^{m/2}\Gamma(\frac{m}{2})} \int_0^\infty e^{tX} X^{\frac{m}{2}-1} dX$$

$$= \frac{1}{2^{m/2}\Gamma(\frac{m}{2})} \int_0^\infty e^{-\frac{(1-2t)X}{2}} X^{\frac{m}{2}-1} dX$$

Let $U = \frac{(1-2t)X}{2}$ $\qquad X = \frac{2U}{1-2t}$

$dU = \frac{(1-2t)dX}{2}$ $\qquad dX = \frac{2dU}{1-2t}$

$$\therefore Mx(t) = \frac{1}{2^{m/2}\Gamma(\frac{m}{2})} \int_0^\infty \frac{e^{-U}(2U)^{m-2/2}}{(1-2t)^{m-2/2}} \cdot \frac{2}{1-2t} dU$$

$$= \frac{2^{m/2}}{2^{m/2}\Gamma(\frac{m}{2})(1-2t)^{m/2}} \int_0^\infty e^{-U} U^{\frac{m}{2}-1} dU$$

$$= \frac{1}{(1-2t)^{m/2}}$$

Consider $\chi^2 = \sum_{i=1}^{m} X_i^2$, where the Xi are NID(0,1)

$$M_{\chi^2}^{2}(t) = E(e^{t\sum X_i^2}) \qquad = \prod_{i=1}^{m} E(e^{tX_i^2}) = [M X_i^2(t)]^m$$

since all the Xi have the same distribution.

$$M_{xi}2\,(t) = E\left(e^{tX_i^2}\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tX_i^2} e^{-X_i^2/2}\, dX_i$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{X_i^2(1-2t)}{2}}\, dX_i$$

Let $U = \frac{Xi^2(1-2t)}{2}$, $X_i = \pm\sqrt{\frac{2U}{1-2t}}$

$dU = Xi(1-2t)dXi \quad dXi = \frac{\pm dU\sqrt{1-2t}}{\sqrt{2U}\,(1-2t)}$

$$\therefore\ M_{x_i^2}(t) = \frac{2}{\sqrt{2\pi}}\int_0^{\infty} e^{-U}\frac{1}{\sqrt{2U}\sqrt{1-2t}}$$

$$= \frac{1}{\sqrt{1-2t}\,\Gamma(\frac{1}{2})}\int_0^{\infty} e^{-U}U^{-\frac{1}{2}}\,dU$$

$$= \frac{1}{(1-2t)^{1/2}}\,,\quad 2t<1$$

$$\therefore\ M_{\chi}2(t) = \left[M\,X_i^2(t)\right]^m = \left[\frac{1}{(1-2t)^{1/2}}\right] = \frac{1}{(1-2t)^{m/2}}\,,\quad 2t<1$$

which is the M.G.F. of the previous distribution.

$\therefore$ The probability density function of $\chi^2 = \sum_{i=1}^{m} X_i^2$, X_i NID(0,1)

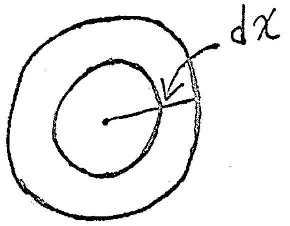
is $g(\chi^2) = \frac{1}{2^{m/2}\Gamma(\frac{m}{2})}(\chi^2)^{\frac{m}{2}-1}e^{-\chi^2/2}$, $0<\chi^2<\infty$

4. GEOMETRIC DERIVATION

By definition,

$$\chi^2 = \sum_{i=1}^{m} X_i^2 \quad ,\ \text{where the Xi are NID(0,1)}$$

This equation represents a hypersphere in n-dimensional space with radius χ . If we consider χ^2 as a parameter, then depending upon the value of χ, we have a family of concentric hyperspheres with centre the origin. In 2-space we can represent 2 of these hyperspheres (now circles) as



where $d\chi$ is an element of radius

For n-space, it can be shown that dv, the element of volume between 2 hyperspheres can be expressed in terms of $d\chi$, the element of radius, by the following formula

$dv = K\chi^{n-1}d\chi$, where K is a constant.

We also know that the joint density function of the Xi's is $k e^{-\frac{\Sigma X_i^2}{2}} = k e^{-\chi^2/2}$

Thus we can find the probability that a random value of χ lies in the interval $d\chi$. It is

$$k e^{-\frac{\Sigma X_i^2}{2}} dV = k e^{-\chi^2/2} K \chi^{n-1} d\chi$$

$$= c e^{-\chi^2/2} \chi^{n-1} d\chi = f(\chi) d\chi \, , \quad 0 < \chi < \infty$$

Since c is independent of χ, we may solve for it using the property that $\int_0^\infty f(\chi)\,d\chi = 1$

$$\therefore \int_0^\infty c e^{-\chi^2/2} \chi^{n-1} d\chi = 1$$

Let $u = \frac{\chi^2}{2}$ $\qquad = \sqrt{2u}$

$du = \chi\, d\chi \qquad d\chi = \frac{du}{\sqrt{2u}}$

$\therefore \left(c2^{\frac{n-2}{2}}\right) \int_0^\infty e{-u} \quad u^{\frac{n-2}{2}}$

$c = \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{n-2/2}}$

$\therefore$ The probability density function of χ is

$$f(\chi) = \frac{1}{2^{n-1/2}\Gamma\left(\frac{n}{2}\right)} e^{-\chi^2/2} \chi^{n-1}$$

Let $V = \chi^2 \quad \chi = \sqrt{V}$

$dV = 2\chi\, d\chi \quad d\chi = \frac{du}{2\sqrt{V}}$

Then, $f(V)\,dV = \frac{e^{-\frac{V}{2}} V^{n-1/2}}{2^{n-1/2}\Gamma\left(\frac{n}{2}\right) 2\sqrt{V}}\, dV$

$\therefore$ The probability density function of χ^2 is

$$\frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} e^{\frac{-\chi^2}{2}} \left(\chi^2\right)^{\frac{n}{2}-1} \quad , 0 < \chi^2 < \infty$$

n is called the number of degrees of freedom.

Suppose now we have P linear homogeneous constraints on the X_i's. Each of these constraints is represented by a hyperplane intersecting the n-dimensional hypersphere through the origin. This will result in a hypersphere of the same radius but of one dimension lower. Thus, with these P linear homogeneous constraints the distribution remains the same, except that the number of degrees of freedom changes from n to n P.

CHAPTER II

MATHEMATICAL DESCRIPTION OF THE CHI-SQUARE DISTRIBUTION AND ITS PROPERTIES

DESCRIPTION OF THE CHI-SQUARE FUNCTION

The probability density function of the chi-square distribution is given by $f(\chi^2) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})}(\chi^2)^{n/2-1}e^{-\chi^2/2}$

where n is the number of degrees of freedom.

As previously shown, this continuous distribution is that followed by the sum of squares of n independently distributed normal variates with zero means and unit variances (standard normal variates).

Note that

$$\lim_{\chi^2\to\infty}\frac{1}{2^{n/2}\Gamma(\frac{n}{2})}(\chi^2)^{n/2-1}e^{-\chi^2/2} = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})}\lim_{\chi^2\to\infty}\frac{(\chi^2)^{\frac{n}{2}-1}}{e^{\chi^2/2}}$$

An application of L'Hospital's Rule $\left(\frac{n-1}{2}\right)$ times shows this limit to be zero. Therefore the chi-square curve is asymptotic to the positive X-axis, and is skewed to the right. Also, for $n>2$

$$\lim_{\chi^2\to 0}(\chi^2)^{n/2-1}e^{-\chi^2/2} = 0$$

Therefore the left end-point of the chi-square curve for $n>2$ is the origin.

The extrema of the chi-square curve are found by setting the first derivative equal to 0.

Let X = χ^2

$$Y = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{n/2-1} e^{-x/2}$$

$$dY/dx = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} \left\{\left(\frac{n}{2}-1\right) x^{n/2-2} e^{-x/2} - \frac{1}{2} e^{-x/2} x^{n/2-1}\right\}$$

$$e^{-x/2} x^{n/2-1} \left\{\left(\frac{n}{2}-1\right)(1/x) - \frac{1}{2}\right\} = 0$$

$e^{-\frac{x}{2}} = 0$ implies $\chi = \infty$

$x^{\frac{n}{2}-1} = 0$ implies $\chi = 0$

$\left(\frac{n}{2} - 1\right) \frac{1}{X} - \frac{1}{2} = 0$ implies X = n-2

Using the second derivative test, it can be verified that the maximum of the chi-square curve occurs at χ^2 = n-2

PEARSONIAN DISTRIBUTIONS

Karl Pearson has shown that the standard frequency distributions may be represented as solutions of a certain differential equation. In particular, the chi-square curve can be characterized as a "Pearson type III curve", and as such can be written in the following form:

$$Y = c(X-\mu)^{\lambda-1} e^{-\alpha(X-\mu)} \quad , X > \mu \quad \alpha > 0 \quad \lambda > 0$$

where for the chi-square curve,

$\mu = 0$, $\lambda = \frac{n}{2}$, $\alpha = \frac{1}{2}$

MOMENTS AND CUMULANTS

It has been shown in Chapter I that the moment generating function of the chi-square distribution

is given by $\frac{1}{(1-2t)^{n/2}}$, $2t < 1$

From this, we can easily obtain the moments of χ^2:

$$\mu_1' = M'(t)\big|_{t=0} = -\frac{n}{2}(1-2t)^{-n/2-1}(-2)\big|_{t=0} = n = \mu$$

$$\mu_2' = M''(t)\big|_{t=0} = n(-n/2-1)(1-2t)^{-n/2-2}(-2)\big|_{t=0}$$

$$= \left(\frac{n^2}{2}+n\right)2 = n^2+2n$$

Therefore, $\mu_2 = \mu_2' - (\mu_1')^2 = n + 2n - n^2 = 2n$

Similarly, all other moments may be obtained.

However, we can also obtain the moments directly. For example, the mean is

$$\mu = E(\chi) = \int_0^\infty \chi f(\chi)\,d\chi$$

$$= \int_0^\infty \frac{\chi\,\chi^{n/2-1}}{\Gamma(\frac{n}{2})\,2^{n/2}} e^{-\chi/2}\,d\chi$$

$$= \frac{1}{\Gamma(\frac{n}{2})}\left(\frac{1}{2}\right)^{n/2}\int_0^\infty \chi^{n/2} e^{-\chi^2/2}\,d\chi$$

But $\Gamma(m) = \int_0^\infty \chi^{m-1} e^{-\chi}\,d\chi$

$$\therefore \Gamma\left(\frac{n}{2}+1\right) = \int_0^\infty (\chi/2)^{n/2} e^{-\chi/2}\left(\frac{1}{2}\right)d\chi$$

$$= \left(\frac{1}{2}\right)^{\frac{n}{2}+1}\int_0^\infty \chi^{\frac{n}{2}} e^{-\chi/2}\,d\chi$$

Since $\Gamma(m+1) = m\Gamma(m)$

$$\mu = \frac{1}{\Gamma(\frac{n}{2})\,2^{n/2}}\,2^{n/2+1}\left(\frac{n}{2}\right)\Gamma\left(\frac{n}{2}\right) = n$$

In the same manner, we can obtain the other moments of $X = \chi^2$.

A formula for the rth moment about the origin can be obtained from the gamma distribution. The r'th moment about the origin for the gamma distribution is given by

$$\mu_r' = \frac{1}{\Gamma(m)}\int_0^\infty e^{-\chi}\,\chi^{m-1+r}\,d\chi$$

$$= \frac{\Gamma(m+r)}{\Gamma(m)} = (m)(m+1)\ldots(m+r-1)$$

Since $\frac{1}{2}\chi^2$ is a gamma variate with parameter $\frac{n}{2}$, the r'th moment of the chi-square distribution is

$$\mu_r' = \frac{2^r\Gamma(r+\frac{n}{2})}{\Gamma(\frac{n}{2})} = (n)(n+2)(n+4)\ldots(n+2r-2)$$

The factor 2^r results from the fact that multiplying a variable by 2 multiplies its raw moments by 2^r.

From the moment generating function, it follows that the cumulant generating function

$$K(t) = \log M(t) = \log (1-2t)^{-\frac{n}{2}}$$

$$= -\frac{n}{2} \log (1-2t) = \frac{n}{2}\sum_{r=1}^{\infty}\frac{(2t)^r}{r}$$

From this, the rth cumulant is

$$K_r = (r-1)!\ 2^{r-1}n$$

This can also be obtained from the gamma distribution. Since the r'th cumulant for the gamma distribution with parameter m is $K_r = m(r-1)! = m\,\Gamma(r)$, the rth cumulant for chi-square is given by $K_r = 2^r\ \Gamma(r)\left(\frac{n}{2}\right) = 2^{r-1}(r-1)!\,n$

From the moments we find that:

(1) coefficient of skewness $= \gamma_1 = \sqrt{\mu_3^2/\mu_2^3} = \sqrt{8/n}$

(2) coefficient of kurtosis $= \gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{12}{n}$

Note that $\lim\limits_{n\to\infty}\gamma_1 = \lim\limits_{n\to\infty}\gamma_2 = 0$

This indicates that the chi-square distribution tends to normality for large n.

SPECIAL CASES: DEGREES OF FREEDOM ONE AND TWO

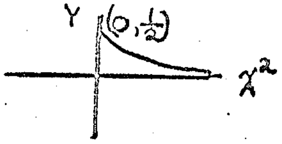
Let us now examine the special cases when the degrees of freedom, n, of the chi-square distribution are 1 and 2.

Case (1) n = 2

Then the equation of the curve reduces to

$$y = \frac{1}{2} e^{-\chi^2/2} \qquad 0 < \chi^2 < \infty$$

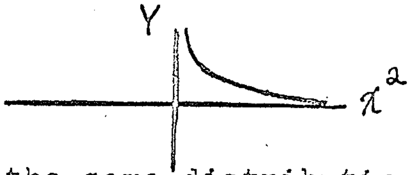
Its graph is as follows:



Case (2) n = 1

Then $y = \frac{1}{\sqrt{2\pi}} e^{-\frac{\chi^2}{2}} (\chi^2)^{-\frac{1}{2}} \qquad 0 < \chi^2 < \infty$

Its graph is as follows:



This is the same distribution as that followed by the square of a normal variate with 0 mean and unit variance.

ASYMPTOTIC DISTRIBUTION OF CHI-SQUARE

Asymptotically, chi-square follows a normal distribution. This was indicated previously when it was shown that in the limit both γ_1 and γ_2 approach 0. A formal proof now follows. The moment generating function of the chi-square distribution is given by $\frac{1}{(1-2t)^{\frac{n}{2}}}$, . However, the moment generating function of a distribution does not always exist. But the characteristic function, $E(e^{itx})$, $i = \sqrt{-1}$, always exists and, like the moment generating function, uniquely characterizes its distribution.

The characteristic function, (C.F.) for the chi-square distribution is $\phi_X(t) = \dfrac{1}{(1-2it)^{\frac{n}{2}}}$

Let $y = \frac{X-\mu}{\sigma}$, where $\mu = n$ and $\sigma^2 = 2n$

Then $\phi y(t) = e^{-\frac{it\mu}{\sigma}} \phi_X(\frac{t}{\sigma})$

$$\log_e \phi_y(t) = \frac{-itn}{\sqrt{2n}} - \frac{n}{2}\left[\log_e \left(1 - \frac{2it}{\sqrt{2n}}\right)\right]$$

Since $\log_e(1 + X) = X - \frac{X^2}{2} + \frac{X^3}{3} - + \cdots$

and letting $\frac{-2it}{\sqrt{2n}} = X$

$$\log_e \phi y(t) = \frac{-itn}{\sqrt{2n}} + \frac{n}{2}\left[\frac{-2it}{\sqrt{2n}} - \frac{(-2it)^2}{2(2n)} + \frac{(-2it)^3}{3(2n)^{3/2}} - + \cdots\right]$$

$\therefore \lim\limits_{n \to \infty} \log_e \phi_y(t) = -\frac{t^2}{2}$

$\therefore \lim\limits_{n \to \infty} \phi_y(t) = e^{-\frac{t^2}{2}}$ which is the C.F. of the standard normal distribution.

Thus, chi-square asymptotically follows a normal distribution.

Fisher has shown that for $n \geq 30$, $\sqrt{2X^2}$ is approximately normally distributed with mean $= \sqrt{2n}$ and variance $= 1$.

This can be proven quite readily:

$$P(\sqrt{2\chi^2}-\sqrt{2m} \leq X)$$

$$= P(\sqrt{2\chi^2} \leq X+\sqrt{2m} \quad)$$

$$= P(2\chi^2 \leq X^2+2X\sqrt{2m}+2m)$$

$$= P(\chi^2 \leq X^2/2+m+X\sqrt{2m} \quad)$$

$$= P\left(\frac{\chi^2-m}{\sqrt{2m}} \leq \frac{X^2}{2\sqrt{2m}}+X \quad \right)$$

Where Z is a random variable normally distributed with 0 mean and unit variance.

Thus, to test the significance of a value of based on more than 30 degrees of freedom, one can treat as a normal deviate with mean 0 and variance 1.

In addition, Wilson and Hilferty have shown that is approximately normally distributed with mean $1-\frac{2}{9m}$ and variance $\frac{2}{9n.}$. Garwood (1936) has shown that this approximation is better than that of Fisher's.

REPRODUCTIVE PROPERTY

The sum of k independently distributed chi-squares, with $n_1, n_2 \ldots n_k$ degrees of freedom respectively, is a chi-square with degrees of freedom.

This can be proved using the moment generating function technique:

$$\text{M.G.F. of } \sum^{k} \chi_i^2 = E(e^{t\sum^{k}\chi_i^2} \quad)$$

$$= \prod_i^k E(e^{t\chi_i^2}) \quad = \prod_i^k \frac{1}{(1-2t)^{m_i/2}}$$

$$= \frac{1}{(1-2t)^{\sum m_i/2}}$$

which is the M.G.F. of a chi-square with $\sum_{i=1}^{k} n_i$ degrees of freedom. As a corollary:

If the sum of two independent positive variates is a variate with $n_1 + n_2$ degrees of freedom, and one is a χ^2 with n_1 degrees of freedom, then the other is a χ^2 with n_2 d.f.

This follows from the fact that, since the χ^2's are independent,

$$(1-2t)^{-\frac{n_1+n_2}{2}} = (1-2t)^{-\frac{n_1}{2}} \quad M_{\chi^2}(t)$$

$$\therefore \quad M_{\chi^2}(t) = (1-2t)^{\frac{-n_2}{2}}$$

A theorem which has useful application in homogeneity tests is the following one, due to Fisher.

FISHER'S THEOREM

Let A be a sum of squares of n independent normal standardized variates X_i, and suppose A = B + C, where B is a quadratic form in the X_i, distributed as χ^2 with h degrees of freedom. Then C is distributed as χ^2 with n-h degrees of freedom, and is independent of B.

The proof may be sketched as follows:

Since B is a quadratic form in the X_i and is distributed as with h degrees of freedom, B is a sum of squares of h orthogonal linear functions $Y_1, Y_2, \ldots, Y_h$ of the X_i. From the theory of linear transformations, we can find n-h further functions $Y_{h+1}, Y_{h+2}, \ldots Y_n$ which are mutually orthogonal and orthogonal to Y_1, $Y_2, \ldots, Y_h$ such that $\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i^2$

It can be shown that if the X_i are N1D(0,1) then the Y_i are N1D(0,1)

We have,

$$A = \sum_{i=1}^{m} X_i^2 = B+C = \chi^2_{(h)}+C = \sum_{j=1}^{h} Y_j^2 + \sum_{j=h+1}^{m} Y_j^2$$

$\therefore$ $C = \sum_{j=h+1}^{m} Y_j^2$ is a sum of squares of n-h independently normally distributed standard variables and therefore follows the chi-square distribution with n-h degrees of freedom.

This theorem can be extended in the following way:

If $A = B_1 + B_2 + \dots . B_k + C$,

where $A = \sum_{i=1}^{m} X_i^2$ and B_i, $i = 1, \dots k$ is a sum of squares of n_i variates Y_i which are independent linear functions of the X_i and $\sum_{i=1}^{k} m_i < m$, then C is distributed as χ^2 with $n-\sum_{i=1}^{k} m_i$ degrees of freedom, independently of the B_i.

THE PARTITION THEOREM

A converse of this theorem is the partition theorem. It gives the condition under which each member of a sum of k sums of squares will be independently distributed as χ^2. This theorem states:

Let the sum of squares of the n variables U_1, $U_2 \dots U_n$ be partitioned into a sum of k sums of squares, Q_1 $Q_2 \dots Q_k$ with f_1, $f_2, \dots f_k$ degrees of freedom, respectively:

The necessary and sufficient condition that $Q_1, Q_2 \cdots Q_k$ are stochastically independent and distributed as χ^2 with $f_1, f_2 \cdots\cdot f_k$ degrees of freedom, respectively, is that:

$f_1 + f_2 + f_k = n$

This theorem forms the theoretical basis for the analysis of variance.

DISTRIBUTION OF THE QUADRATIC FORM IN THE EXPONENT OF THE MULTIVARIATE NORMAL DISTRIBUTION

Another property which has very important applications involves the distribution of the quadratic form in the exponent of the multivariate normal distribution.

Let $X_1, X_2, \ldots X_n$ have the following joint distribution:

$$f(X_1 X_2 \ldots X_n) = \frac{1}{(2\pi)^{m/2}\sqrt{|V|}} \exp\left[\frac{(X-\mu)' V^{-1} (X-\mu)}{-2}\right]$$

where X is an nx1 column vector

μ is an nx1 column vector

V^{-1}, the inverse of the variance-covariance matrix, is a symmetric nxn matrix.

Let $Q = (X-\mu)^1 V^{-1} (X-\mu)$

The moment generating function of Q is $M_Q(t) = E(e^{tQ}) =$

$$\int\cdots\int \frac{1}{(2\pi)^{m/2}\sqrt{|V|}} \exp -\left[\frac{(X-\mu)' V^{-1} (X-\mu)}{2}\right] \exp(tQ)\, dx_1 dx_2 \ldots dx_m$$

$$= \int\cdots\int \frac{1}{(2\pi)^{m/2}\sqrt{|V|}} \exp -\left[\frac{(X-\mu)' V^{-1} (X-\mu)(1-2t)}{2}\right] dx_1 dx_2 \ldots dx_m$$

Note $\left|\frac{V}{1-2t}\right| = \frac{|V|}{(1-2t)^m}$

$$\therefore \int\cdots\int \frac{1}{(2\pi)^{m/2}\sqrt{|V/_{1-2t}|}} \frac{1}{(1-2t)^{m/2}} \exp\left[\frac{-(X-\mu)' V^{-1} (X-\mu)(1-2t)}{2}\right]$$

$$= \frac{1}{(1-2t)^{m/2}},\ t < 1/2$$

$\therefore$ Q follows a χ^2 distribution with n degrees of freedom.

RELATIONSHIP TO OTHER DISTRIBUTIONS

(a) NORMAL DISTRIBUTION

(1) By definition, chi-square is the sum of squares of normal, standardized variates.

(2) It has been shown that chi-square is asymptotically normally distributed with mean n and variance 2n. Fisher has shown that $\sqrt{2\chi^2}$ is approximately normally distributed with mean $\sqrt{2m}$ and variance 1. Wilson and Hilferty have shown that $\left(\frac{\chi^2}{m}\right)^{1/3}$ is approximately normal with mean $(1-\frac{2}{9n})$ and variance $\frac{2}{9n}$.

(b) GAMMA DISTRIBUTION

It has been shown that $\frac{\chi^2}{2} = \frac{\sum X_i^2}{2}, X_i\ NID(0,1)$ is a gamma variate with parameter $\frac{n}{2}$. This property simplifies the derivation of many other properties of χ^2.

(c) BETA DISTRIBUTION

There are two types of Beta variates, denoted by B_1 and B_2. The probability density function of B_1 is

$$B_1(X) = \frac{X^{l-1}(1-X)^{m-1}}{B(l,m)} \quad , 0 \leq X \leq 1$$

The probability density function of B_2 is $B_2(X)$

$$= \frac{X^{l-1}}{B(l,m)(1+X)^{l+m}} \quad , 0 \leq X \leq \infty$$

where $B(1,m) = \int_0^1 X^{l-1}(1-X)^{m-1}\,dX$

It can be shown that if X and Y are independent gamma variates with parameters ℓ amd m respectively, then $\frac{X}{X+Y}$ is a B_1 variate with parameters ℓ and m. Also, $\frac{X}{Y}$ is a B_2 variate with paramoters ℓ and m.

From this, and knowing the relation between the gamma and the chi-square distribution, the following theorem can be stated:

If the independent variates X and Y are distributed as chi-square with n_1 and n_2 degrees of freedom respectively, then $\frac{X}{X+Y}$ is a B_1 variate with parameters $\frac{n_1}{2}$ and $\frac{n_2}{2}$, and $\frac{X}{Y}$ is a B_2 variate with parameters $\frac{n_1}{2}$ and $\frac{n_2}{2}$.

(d) STUDENT'S t DISTRIBUTION

The random variable t is defined by $t = \frac{Z}{\sqrt{V/n}}$ where Z is n(0,1) and V is χ^2 with n degrees of freedom, and Z and V are independent.

(e) F DISTRIBUTION

The random variable F is defined as $F = \frac{V/n_1}{U/n_2}$ where V and U are independently distributed as chi-squares with n_1 and n_2 degrees of freedom respectively.

CHAPTER III

HISTORICAL DEVELOPMENT

The first known derivation of the chi-square distribution was obtained by Prof. W. Helmert of the Polytechnikum in Aachen, Germany in the year 1876. In a very general paper, entitled "On the Probability of the Sum of Powers in Errors of Observations and Some Related Questions", Helmert, working on a special case of this problem, derived by mathematical induction the distribution of n normally distributed standard variates. Twenty-five years later, Karl Pearson was to name it the chi-square distribution.

In Helmert's paper, he first sets up the multiple integral that must be evaluated in order to obtain the distribution of the sum of the m^{th} powers of n errors in observations, and proceeds to solve it for particular values of m and n, given that the errors follow a specified distribution. Having first dealt with the uniform distribution, Helmert then proceeds to the case in which the errors are normally distributed. Substituting the Gaussian function into his original formula, Helmert obtains the required distribution for the cases n = 1 and 2 with m = 1,2 and 3. He then concerns himself exclusively with the problem of what happens when m is fixed at 2, stating that for this value of the exponent, "the mathematical treatment is most convenient". Using the results he obtained for n = 1 and 2, he easily extends his findings to n = 3 and 4. These calculations now strongly suggest a general formula for the distribution of the sum of squares of n standard normal variates with n arbitrary. A formal proof by induction verifies this hypothesis, and the chi-square distribution has been derived for the first time.

KARL PEARSON

The chi-square distribution seems to have been neglected until Karl Pearson rediscovered it in 1900. In his classic paper in the Philosophical Magazine, he introduced the goodness of fit criterion, and established its distribution. Since all goodness of fit problems up to that time were solved by visual inspection, the χ^2 test has turned out to be one of the most useful in statistics.

Pearson starts off by considering the multivariate normal distribution, $= Z = Z_0 e^{-\frac{Q}{2}}$, of a system of variates with zero means, where Q is a quadratic form. He defines χ^2 to be Q and proceeds, by transforming the ellipsoid $\chi^2 = Q$ into a hypersphere, to develop its distribution. The result is that he expresses P, "the chances of a system of errors with as great or greater frequency than that denoted by ", as P =

$$\frac{\int_{\chi}^{\infty} e^{-\chi^2/2} \chi^{n-1} d\chi}{\int_{0}^{\infty} e^{-\chi^2/2} \chi^{n-1} d\chi}$$

He next expresses this probability in power series form, and from this develops the first probability table for the χ^2 distribution.

Pearson then applies these results to the goodness of fit problem. The problem is that of deciding whether a set of observed frequencies can be reasonably considered under random sampling to have arisen from a particular theoretical distribution. In other words, it must be decided whether the frequencies to be expected under this distribution are compatible with the observed frequencies. Pearson first assumes these expected values are known, and deals with an (n+1) fold grouping of frequencies under the sole restriction (which is crucial in order for the test to make any sense at all), that the sum of the observed frequencies equals the sum of the expected frequencies. Sincethis fixes one of the observations, we have left only n variables. These he assumes to have a multivariate normal distribution, and by a complicated transformation to polar coordinates he obtains the result that the quadratic form $Q = \chi^2 = \sum_{i=1}^{m} e_i^2/m_i$ where m_i is the i_{th} expected frequency (known a priori), and e_i is the i_{th} deviation of the observed frequency m_i' from m_i. This result, the goodness of fit criterion, Pearson aptly refers to as being "of very great simplicity and very easily applicable".

One calculates this result from the data and finds the corresponding P, the probability that a random sample of observed frequencies from the hypothetical distribution will give a more extreme value of χ^2. If P is between .1 and .9, the hypothesis of a "good" fit is almost always accepted. This value of P, Pearson notes, can be calculated directly from the power series form, or for $n < 13$ can be obtained from his tables. However, these tables, which give P as a function of χ^2 and $n^1 = n + 1$, are not completely accurate. This is because Pearson assumed, as long as the only linear constraint among the observations was that imposed by fixing the total sample size, that n^1 could always be taken as one more than the number of cells. It was Sir Ronald Fisher who showed more than twenty years later that the number of degrees of freedom used by Pearson was not correct.

Pearson concluded in his own paper that if the expected frequencies had to be estimated from the sample, the distribution of the test criterion would remain unchanged. He compares χ^2 based upon the true theoretical values with χ_s^2 based upon estimates from the sample data, and says the following about the differences between them:

> "I think we may conclude that χ only differs from χ_s by terms of the order of the squares of the probable errors of the constants of the sample distribution."

Thus he maintains (although as Cochran notes, "with some sign of hesitation",) that estimation of parameters from the sample data does not modify the distribution of χ^2, and in particular, its degrees of freedom. Pearson was to stick adamantly to this conclusion until 1922.

Although Pearson granted, and indeed himself showed (1915) that the degrees of freedom had to be reduced by one for every linear restriction imposed upon the observations, he drew a distinction between this and the estimation of population parameters. This is clear when he says,

> "If the sampled population is not known we can only put for the marginal totals the values given by the sample itself and test from this substitution the degree of divergence from independence."

As Fisher points out, however, Pearson never suggests any correction for this. Pearson states (1922):

> "We certainly do not by using sample constants reduce in any way the random sampling degrees of freedom. What we actually do is replace the accurate value of χ^2, unknown to us and cannot be found, by an approximate value, with the same justification as the astronomer claims, when he calculates his probable error on his observations, and not on the mean square error from an infinite population of errors unknown to him."

SIR RONALD FISHER

It was Fisher, who in papers in 1922 and 1924, finally cleared the issue up, at least theoretically. In his 1922 article, Fisher dealt principally with the 2x2 contingency table. He showed that n^1, which is entered into Pearson's table along with the calculated χ^2 to obtain the corresponding P, should be taken as one more than the degrees of freedom in the distribution. Since in a test of independence of two attributes, the marginal totals are used to estimate the expected frequencies, they are considered fixed. This leaves only one degree of freedom by which the expected values might differ from the observed values, not three, as Pearson would claim. Fisher also points out that the same problem can be examined by testing the difference between two proportions, using the normal approximation to the binomial distribution. This procedure should be equivalent to the χ^2 test, but only if the degrees of freedom are taken as one, will the two tests turn out to be identical.

Fisher's theoretical conclusions were backed up by sampling experiments on the 2x2 table conducted by Yule (1922). By comparing values of observed χ^2's calculated from 350 tables with 100 observations each, against the v alues to be expected from the theoretical distribution, he found that the two sets of χ^2's were compatible only if each table was assumed to have a single degree of freedom.

Fisher's 1924 paper deals with the general case in which population parameters must be estimated in order to obtain the expected frequencies. First he points out that the distribution of χ^2 will depend upon the method of estimation. Reasonably enough, he chooses to estimate the parameters in such a way that χ^2 is a minimum. He then shows that if the number of observations is large, this method is equivalent to that of maximum likelihood.

Fisher's main accomplishment in this paper is showing that if a single estimated parameter is used to calculate the expected frequencies, the resulting χ^2 differs from the true χ^2 by the square of a quantity normally distributed with unit variance. Since this quantity has a single degree of freedom attached ot it, it is evident that for each estimated parameter, one degree of freedom is lost from the distribution of χ^2.

Thus Fisher cleared up the issue theoretically. Pearson, however, taking up a new line of defense, still maintained the degrees of freedom for a 2x2 contingency table should be three. He claimed now, and as late as 1932, that fixing the marginal totals created a "spurious" contingency table and that the data should not be arranged in a 2x2 table at all. Many others also questioned the logic of having contingency tables with "fixed" marginal totals. One would expect that in many experiments, the marginal totals would change from sample to sample.

GEORGE BARNARD

George Barnard, examining the philosophy of the situation in 1947, resolved the problem with great thoroughness. He verified that although logically three different kinds of tests of independence can be executed on the data in a 2x2 table, each involves only a single degree of freedom. The appropriate test depends upon the abstract picture the experimenter has in mind. However, the adoption of the correct picture depends upon information not always supplied in a real situation. It is this information that determines the differences between the three tests.

What Barnard refers to as "independence trials" corresponds logically to the case in which the marginal totals are allowed to vary from sample to sample. For example, in testing the hypothesis of independence between marriage adjustment and level of education, one would not expect the numbers in each category to remain fixed from trial to trial. However, since two parameters must be estimated in order to test the hypothesis of independence, and the total sample size is fixed, we have left only 4-2-1 = 1 degree of freedom.

Barnard's "comparative trials" correspond logically to the situation in which one set of marginal totals is fixed. Here the null hypothesis is that the proportion in the two categories are equal. For example, one may have two urns, A and B, each containing balls labelled either I or II. The null hypothesis is that the proportion of balls marked I in urn A equals the proportion of balls marked I in urn B is equal to P, say. It is assumed that the number of balls in each urn is fixed (although these two totals are independent). However, this means one parameter (P) must be estimated. Furthermore, fixing a set of marginal totals imposes an additional linear restriction on the observations. Thus again we are left with a single degree of freedom.

"Double dichotomy" trial is the name given by Barnard to the situation which corresponds logically to fixing both sets of marginal totals. An example, as pointed out by Cochran (1952), is Fisher's well-known tea tasting experiment in which a lady attemptsto guess whether milk or tea was added first in her cup. In the test situation, one classification corresponds to what quantity is added first, and the other classification, to what the lady guessed was added first. Since she knew how many cups were of each kind, it is assumed she matched her guesses to these numbers. Again the null hypothesis is that of independence between the two classifications, but now there are three linear restrictions, corresponding to fixing the two sets of marginal totals and the total sample size. However, there are no estimated parameters since fixing all the sub-totals and the grand total permits the exact calculation of the various probabilities. Again we are left with one degree of freedom.

Barnard's paper was a valuable contribution. He showed whether the marginal totals in 2x2 independence table are interpreted as being fixed or not from trial to trial, the resulting χ^2 has in either case one degree of freedom.

Besides the "degrees of freedom" battle, the goodness of fit test provoked discussion on other points. Some of these points are concerned with the minimum allowable number of observations per cell, the optimum number of classes in the observations, and the need for a correction for continuity.

MINIMUM NUMBER OF OBSERVATIONS AND THE CONTINUITY CORRECTION

There are a variety of proofs for the limiting distribution of the chi-square criterion. These include Pearson's original derivation, and Cramer's method, which makes use of the characteristic function. Morgenstern (1958) has given a proof based on mathematical induction. Still another is Rybarz's proof (1959) using the asymptotic normality of the Poisson distribution (an idea originally stated by Fisher).

All these proofs, however, have in common one assumption, and this is that the total sample size is large, and that the theoretical numbers in each class are not too small. There are two types of approximations made by these authors that make these assumptions necessary:

(1) Rough approximations which include Stirling's approximation to the factorial and the omission of all but the first two terms of the Taylor series expansion of $\log (1 + X) = X - \frac{X^2}{2} + \frac{X^3}{3} + \ldots.$

(2) The approximation of a discrete distribution by a continuous one.

In Pearson's original proof, he took the observed multinomially distributed X_i's as each having a normal distribution about the corresponding expected value, m_i. As Cochran points out, this immediately committed him to a large number of observations per cell. Camp (1931) has set bounds on the maximum error involved in this.

However, it is not necessary to rigorously make the normality assumption. It may be alternatively assumed that the discrete distribution of the test criterion can be approximated by the continuous chi-square distribution. This assumption arises from the fact that, in order to facilitate the calculation of the tabular probabilities corresponding to different values of χ^2, one must replace a discrete sum by an integration over the appropriate region. The error involved in both types of approximation increases as the sample size decreases; however, it is impossible to set an absolute lower bound on the pemissible number of observations.

What is adequate for one application may be unsuitable for another. Although many statisticians disagree on this point, the generally accepted rule is that the total sample size should be at least 50, and that the minimum allowable cell frequency should be 5. A number of studies have been made on this, some of which I shall now refer to.

Hoel (1938) performed a study in which he concluded about the first type of error (rough approximations) that "the actual error committed by using the customary first approximation is much smaller than the order of the neglected terms would indicate, and therefore the range of applicability of P is wider than has been supposed".

Neyman and Pearson have studied the problem when the sample size is ten. After performing a sampling experiment they concluded that the exact χ^2 differs from the tabular χ^2 by very little. In the region of significance from .01 to .10, the greatest difference between the two P's was .061, and in most cases was far lower.

In the case in which all expectations are small, but there is a large number of degrees of freedom(Cochran suggests no less than 60), the normal approximation to the chi-square distribution may be used. Haldane (1937) has worked out exact expressions for the mean and variance of χ^2, that can be used for this purpose. For an rxc contingency table, his result for the mean of χ^2 is $\frac{(r-1)(c-1)}{N-1}\cdot N$, where N is the sample size. His expression for the variance is very cumbersome. Dawson (1957) has substituted a simpler but equivalent expression:

$$\mathrm{Var}\,(\chi^2) = \frac{2N}{N-3} \cdot (n_1-U_1)(n_2-U_2) + \frac{N^2}{N-1}(U_1U_2)$$

$$\text{where } n_1 = \frac{(r-1)(N-r)}{N-1}$$

$$n_2 = \frac{(c-1)(N-c)}{N-1}$$

$$U_1 = \frac{N\Sigma R_i^{-1} - r}{N-2}$$

$$U_2 = \frac{N\Sigma C_i^{-1} - c^2}{N-2}$$

where r_i is the sum of the entries in the ith row, and c_j the sum of the entries in the jth column.

To correct for the bias introduced in assuming that the discrete frequencies can be approximated by a continuous distribution, Yates (1934) suggested subtracting .5 from the absolute values of each of the deviations, before squaring them. This adjustment, called Yates' continuity correction, will lower the value of χ^2 and yield more exact probabilities. Yates himself stated that "the worker will not be led badly astray if he applies the ordinary chi-square test (after correcting for continuity) to tables giving expectations as low as 10, so long as the corresponding distributions are reasonably symmetrical".

In the same paper, Yates offers an example in which two of the expected frequencies are between 2 and 3. He concludes that the exact results are very well approximated by using χ^2 along with the continuity correction.

Fisher stated his belief that while it is permissible to deal with frequencies like 2 or 3 (along with the continuity correction) in a 2x2 contingency table, for all other cases the minimum allowable expected frequency should be 5. There seems to be some confusion here, as well as in other parts of the literature, about the distinction between observed frequencies around 5 and expected frequencies around 5.

As was stated before, the commonly accepted practice is to allow a minimum of five observations per cell. If this condition is not satisfied, adjacent cells are to be pooled until it is, with a resulting loss in degrees of freedom (number of additional degrees of freedom lost = number of pooled cells). Cochran (1952) points out that this can be a dangerous practice, especially in fitting bell-shaped curves such as the normal. This is because discrepancies are often likely to appear at the tails, where observations are most scarce. It is Cochran's belief that the recommended minimum of 5 is too conservative, and that too rigid an application of the rule may result in a substantial loss of power. However, he adds that this is just an opinion since not enough research has been done to make the test quite clear.

One study, however, has beendone by Neyman and Pearson (1931). They concluded after performing a sampling experiment that the error will not be large if the groups which are pooled contain only a small portion of the total frequency. In any case, they point out that if the χ^2 based on pooling is insignificant, there is no problem, since the true minimum χ^2 would show a still better fit.

For reference, I have constructed the following table which shows how various statisticans feel about the issue as a whole.

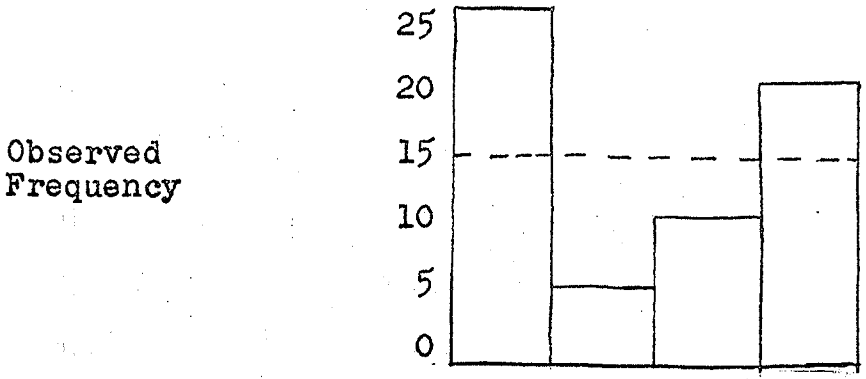
Author	Recommendations	Source
Pearson, Karl	Makes no specific recommenations but states that "no cell (should) be taken so small that its contents are very small compared with the size of the sample".	"On the General Theory of Multiple Contingency" BKA, VXI, 1916, pp 147-158.
Fisher, R.	Minimum expectation should be 5, except in the case of the 2x2 table, where frequencies of 2 and 3 may be dealt with using Yates' correction for continuity.	"Statistical Methods for Research Workers" 11th edition, Hafner Publ. Co., N.Y., 1950 pp 931
Yates, F.	"The worker will not be led badly astray if he applies the ordinary chi-square test (after correcting for continuity) to tables giving expectations as low as 10, as long as the corresponding distributions are reasonably symmetrical". No minimum sample size is demanded, but if smallest expectation is less than 500, continuity correction should be used.	"Contingency Tables Involving Small Numbers and the χ^2 test". JRSS Supl. Vol. 1, 1934, pp 229

Author	Recommendations	Source
Snedecor, G.	"No expected number should be smaller than 5, and when possible, all should be 10 or more. (The correction for continuity) is well worth making if any critical decision is involved, especially if the numbers in any class fall below 50."	"Statistical Methods" Iowa State College Press, Ames, Iowa, 1940, pp 168-170
Yule, G. and Kendall, M.	The sample size should be "reasonably large ... It is difficult to say exactly what constitutes largeness, but as an arbitrary figure, we may say N should be at least 50. No theoretical cell frequency should be small. Here again it is hard to say what constitutes smallness but 5 should be regarded as the very minimum, and 10 is better".	"An Introduction to the Theory of Statistics" Charles Griffin and Co. Ltd., London, 1937, p. 422.
Cramer, H.	The minimum expected frequency should be 10.	"Mathematical Methods of Statistics" Princeton University Press, Princeton 1946, p 420 f.
Hoel, P.	"The approximation is usually satisfactory provided that the (expected values are) >5."	"Introduction to Mathematical Statistics", J. Wylie & Sons Inc., N.Y., 1947, p. 191.
Cochran W.	For the 2x2 table, Fisher's exact test should be used if the sample size $n<20$, or if $20<n<40$, and the smallest expectation is less than 5. For $n>40$, χ^2 should be corrected for continuity, if the smallest expectation is less than 500. However, for tables with more than 1 degree of freedom and some expectations under 5, χ^2 should be used, uncorrected. If all expectations	"The χ^2 Test of Goodness of Fit" Annals of Mathematical Statistics, Vol. 23, 1952, p. 334

Author	Recommendations	Source
	are less than 5, but there are more than 60 degrees of freedom, use the normal approximation to the exact distribution, using the exact mean and variance of χ^2 .	

NUMBER AND WIDTH OF CLASSES

The choice of the proper number and width of the classes for the goodness of fit test has also been subjected to scrutiny. These are important decisions, since different choices may change the result of the test. Williams (1950) gives a simple example of this. Consider the following frequency distribution, where the dashed line represents the distribution under the null hypothesis, the numerals I, II, III and IV, one set of class frequency divisions, and I^1 and II^1, another set.



Then χ^2 under the four-fold division yields $\chi^2 =$

$$\frac{(25-15)^2}{15} + \frac{(5-15)^2}{15} + \frac{(10-15)^2}{15} + \frac{(20-15)^2}{15}$$

= 16.7, significant at the 5% level.

But χ^2 under the two-fold division yields $\chi^2 = \frac{(30-30)^2}{30} + \frac{(30-30)^2}{30} = 0$, which is non-significant.

The most thorough study in this area has been done by Mann and Wald (1942). By maximizing the power function of the test at about the point where the power is ½ (according to Cochran an "arbitrary" but "reasonable" choice), the theoretically "best" number of classes k emerges as

$$k = 4\sqrt[5]{\frac{2(n-1)^2}{c}}$$

where $$\int_c^\infty \frac{1}{\sqrt{2\pi}}\, e^{-\chi^2/2}\, d\chi = \alpha$$

and α is the desired level of significance. At the 5% level, c is 1.64.

The class limits are then chosen so that the number of theoretical frequencies in each class is equal to $\frac{N}{k}$. Choosing the class limits in this way has the obvious advantage that it removes the subjective element, although it is a fairly complicated procedure, and requires that the data be ungrouped.

Although Mann and Wald rigorously proved their findings only for very large N ($N \geq 450$ at the 5% level of significance), they maintain that their results hold for sample sizes as low as 200. For sample sizes even smaller, it is their belief that their results still hold approximately.

CHAPTER IV - APPLICATIONS

GOODNESS OF FIT TESTS

(a) GENERAL GOODNESS OF FIT

It is often wished to test whether a set of observed frequencies resulting from some experiment are compatible with the frequencies to be expected if the data came from a particular theoretical distribution.

Let the total sample size be N, the number of classes of frequencies (cells) k, the observed frequencies m_1, m_2 ....mk, and the expected frequencies m_1^1, m_2^1...m_k^1. Then the criterion $\chi^2 = \sum_{i=1}^{k} \frac{(m_i - m_i')^2}{m_i'}$ where $\sum^{k} m_i = \sum^{k} m_i' = N$ gives a measure of the "goodness of fit". Associated with each such calculated χ^2 is its degrees of freedom, the number of independent observations in the sample. For the goodness of fit application it can be interpreted as the number of classes k less the number of independent restrictions imposed upon the theoretical and observed frequencies. Owing to certain assumptions made in the derivation of the limiting distribution of , each such restriction must be linear and homogeneous. The one universal restriction (without which it becomes very difficult to interpret the test) is that $\sum m_i = \sum m_i'$.

If this is the only restriction, the degrees of freedom become k-1. Often, however, other restrictions are imposed, such as requiring that the arithmetic means of the observed and expected frequencies are equal. Sometimes, also, the expected frequencies are not known a priori, and they must be estimated from the sample data. That is, parameters of the hypothetical distribution are estimated by the information at hand. Sir Ronald Fisher has shown that each such estimated parameter has the same effect as a linear constraint and therefore causes the degrees of freedom to be reduced by one. For example, in fitting a normal distribution to observed data, usually both the mean and the variance of the distribution are not known, and consequently must be estimated from the sample. This involves the loss of two additional degrees of freedom.

The calculated value of χ^2, together with its degrees of freedom, can be compared to a tabular χ^2. If the calculated χ^2 exceeds the tabular χ^2, at the chosen level of significance α, we conclude that a system of deviations such as that observed will occur less than 100α % of the time under random sampling, and therefore we reject the hypothesis that the observed sample follows the theoretical distribution that yielded the expected frequencies.

This test was discovered by Karl Pearson, who showed that the χ^2 criterion asymptotically followed the continuous chi-square distribution. This would indicate that the sample size should be large and that the frequency in any one cell should not be too low. As a general rule, the minimum expected frequency in any one cell is taken to be 5; if this condition is not met, adjacent cell frequencies are grouped together until it is. This procedure has the effect of reducing the degrees of freedom by the number of pooled cells. Since the sensitivity of the test is then diminished, the pooling technique is not to be greatly encouraged.

The fact that we are using a continuous distribution to approximate a discrete set of data would also indicate that a correction for continuity is desirable. The commonly accepted one, that proposed by Yates, involves the subtraction of .5 from the absolute values of each of the deviations, before they are squared. It is of practical value only in situations where there is a single degree of freedom. In this case, since there is a very limited number of possible values for χ^2, it is obvious that the distribution is discrete. By reducing the calculated χ^2, Yates' correction factor partially compensates for this.

The χ^2 test for goodness of fit is treated as a one-sided test in most cases and the hypothesis of a good fit is rejected whenever a calculated statistic exceeds a critical value that corresponds with a particular level of significance. To reject the hypothesis under test whenever a calculated χ^2 statistic is so small that the probability of its occurrence under the given hypothesis is less than α, is to admit that there is the possibility in fitting frequencies or curves that the fit can be too good. In some areas of research a two-sided test of this type is useful in detecting some discrepancies in the conduct of the experiment such as non-randomness in the observed sample.

(b) PARTITION OF DEGREES OF FREEDOM

If the goodness of fit test is significant, it is often of interest to investigate where the discrepancy lies. We can then test the significance of any linear function of the deviations, $L = \sum_i g_i(m_i - m_i^1)$, where the g_i are numbers chosen in advance. The test criterion is $\chi^2 = \frac{L^2}{\text{Var}(L)}$, and is approximately distributed as chi-square with one degree of freedom. For tests involving the Poisson, binomial or normal distributions, Cochran (1954) has given special formulae for var (L). As an example, consider the fitting of a Poisson distribution to the data shown in the table below.

i	m_i	m_i^1	$\frac{(m_i - m_i^1)^2}{m_i^1}$
0	52	47.65	0.40
1	67	77.04	1.31
2	58	62.28	0.29
3	52	33.56	10.13
4	7	13.56	3.17
5	3	4.39	0.44
6	1	1.52	0.18
Total	240	240.00	15.92

The total $\chi^2 = \sum_{i=0}^{6} \frac{(m_i - m_i')^2}{m_i'} = 15.92$ with 5 d.f. is significant at the 1% level. The large contribution from $i = 3$ attracts attention, and therefore we test the deviation $L = m_3 - m_3^1 = 52 - 33.56 = 18.44$. For the situation in which the Poisson mean m must be estimated from the data, Cochran's formula for var (L) is

$$\sum_i g_i^2 m_i - \frac{\left(\sum_i g_i m_i\right)^2}{N} - \frac{\left[\sum g_i m_i (i-\hat{m})\right]^2}{N\hat{m}}$$

where N is the total sample size, and $\hat{m}$ is the estimate of m. In our example,

$$\hat{m} = \frac{\sum_i i m_i}{N} = \frac{388}{240} = 1.6167$$

$$\text{var}\ (L) = 33.56 - \frac{(33.56)^2}{240} \left[1 + \frac{(3-1.6167)^2}{1.6167}\right] = 23.31$$

and $\chi^2_{(1)} = \frac{(18.44)^2}{23.31} = 14.59$, significant at the 1% level.

It is thus seen that the deviation corresponding to i = 3 constitutes the major part of the total χ^2. Cochran points out, however, that for the test to be strictly valid, the deviation should be picked out before seeing the data. If the test is applied, as here, to a deviation that looks unusually large, the calculated χ^2 will be too high.

A similar investigation may be performed when fitting the binomial or the normal distributions. In these cases Cochran's formulae for var (L) are:

(1) <u>Binomial distribution, parameters n and p</u>

$$\text{Var}\ (L) = \sum_i g_i{}^2 m_i - \frac{(\sum g_i m_i)^2}{N} - \frac{\left[\sum_i g_i m_i (i-np\right]^2}{Nn\hat{p}\hat{q}}$$

where $\hat{p}$ = sample estimate of p

$\hat{q} = 1 - \hat{p}$

N = total number of observations

(2) Normal distribution

$$\mathrm{Var}(L) = \sum g_i^2 m_i - \frac{(\sum g_i m_i)^2}{N} - \frac{(\sum g_i d_i m_i)^2}{NS^2} - \frac{(\sum g_i m_i (d_i^2 - S^2))^2}{2NS^4}$$

where

d_i = (midpoint of ith class) - (sample mean)

S^2 = sample estimate of variance

TESTS OF INDEPENDENCE

(a) r x s Contingency Tables

By arranging the data into an r x s contingency table, it is possible to test the independence of 2 attributes A and B, divided into s and r classes respectively. The contingency table may appear as follows:

O_{11}	O_{12}				O_{1s}	$T_{1.}$
O_{21}	O_{22}					$T_{2.}$
			O_{ij}		O_{is}	$T_{i.}$
O_{r1}	O_{r2}		O_{rj}		O_{rs}	$T_{r.}$
$T_{.1}$	$T_{.2}$		$T_{.j}$		$T_{.s}$	$T_{..}$

If the classification of individuals into columns is independent of the classification of individuals into rows, then the joint relative frequencies in the body of the table differ from the product of the marginal frequencies by amounts equal to chance fluctuations. The expected frequencies E_{ij} for the body of the table can be calculated, under the hypothesis of independence between the row and column classifications, as products of the ith and jth marginal totals divided by the total number of frequencies in the table. Given that $T_{i.}$, $T_{.j}$, and $T_{..}$ represent the ith row total, the jth column total, and the grand total respectively,

$$E_{ij} = \frac{T_{i.}T_{.j}}{T_{..}}$$

The use of the marginal totals and the grand total in the calculation of the expected frequencies in the table fixes these totals in the sense that the sample space for the purpose of calculating the probability of the observed frequencies consists of all possible tables with these marginal totals. The chi-square statistic is calculated as

$$\chi^2 = \sum_{i,j}^{r,s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

with degrees of freedom equal to (r-1)(s-1). A sufficiently large χ^2 causes rejection of the hypothesis of independence, and one may wish to partition χ^2 with the object of discovering a more specific reason for rejecting the hypothesis of independence.

Lancaster (1949) showed that every r x s contingency table can be reduced to (r-1)(s-1) two by two contingency tables in a unique way and such that the sum of the single degree of freedom chi-squares is equal to the original χ^2 obtained from the r x s table.

The component single degree of freedom tables can be constructed as follows: For each frequency in the last r-1 rows and s-1 columns, there is a 2x2 table having this frequency in its lower right hand cell. The remaining three cells are composed of the frequencies above and to the left of the frequency chosen for the lower right hand cell.

For example, for a 3 x 4 table, the 6 component 2 x 2 tables are

O_{11}	O_{12}
O_{21}	O_{22}

$O_{11}+O_{12}$	O_{13}
$O_{21}+O_{22}$	O_{23}

$O_{11}+O_{12}+O_{13}$	O_{14}
$O_{21}+O_{22}+O_{23}$	O_{24}

$O_{11}+O_{21}$	$O_{12}+O_{22}$
O_{31}	O_{32}

$O_{11}+O_{12}+O_{21}+O_{22}$	$O_{13}+O_{23}$
$O_{31}+O_{32}$	O_{33}

$O_{11}+O_{12}+O_{13}+O_{21}+O_{22}+O_{23}$	$O_{14}+O_{24}$
$O_{31}+O_{32}+O_{34}$	O_{34}

Kimball (1950) has given formulae for the χ^2's of each of the component tables that do not require knowledge of the expected frequencies. Each of these formulae are of the same form as that for an ordinary 2x2 table. As a result, computation of the single degree of freedom chi-squares requires no more effort than the computation of chi-squares from (r-1)(s-1) 2x2 tables.

As an example consider the 3x3 table. If the observed frequencies and marginal tables are

O_{11}	O_{12}	O_{13}	$T_{1.}$
O_{21}	O_{22}	O_{23}	$T_{2.}$
O_{31}	O_{32}	O_{33}	$T_{3.}$
$T_{.1}$	$T_{.2}$	$T_{.3}$	$T_{..}$

Kimball's formulae are

$$\chi_1^2 = \frac{T_{..}[(T_{21}(T_{.2}O_{11}-T_{.1}O_{12}) - T1.(T_{.2}O_{21}-T_{.1}O_{22})]^2}{T_{1.}T_{2.}T_{..}T_{12}(T_{1.}+T_{2.})(T_{..}+T_{12})}$$

$$\chi_2^2 = \frac{T_{..}^2[O_{23}(O_{11}+O_{12}) - O_{13}(O_{21}+O_{22})]^2}{T_{1.}T_{2.}T_{.3}(T_{1.}+T_{2.})(T_{.1}+T_{.2})}$$

$$\chi_3^2 = \frac{T_{..}^2[(O_{32}(O_{11}+O_{21}) - O_{31}(O_{12}+O_{22})]^2}{T_{3.}T_{2.}T_{.2}(T_{1.}+T_{2.})(T_{.1}+T_{.2})}$$

$$\chi_4^2 = \frac{T_{..}[(O_{33}(O_{11}+O_{12}+O_{21}+O_{22}) - (O_{13}+O_{23})(O_{32}+O_{31})]^2}{T_{3.}T_{.3}(T_{1.}+T_{2.})(T_{.1}+T_{.2})}$$

(b) r x 2 tables

The r x 2 table is a special case of the r x s table. It can be represented as follows:

	Category A	Category B	Total	Proportion in Cat.
1	X_1	$n_1 - X_1$	n_1	$P_1 = \frac{X_1}{n_1}$
2	X_2	$n_2 - X_2$	n_2	$P_2 = \frac{X_2}{n_2}$
. . .				
r	X_r	$n_r - X_r$	n_r	$P_r = \frac{X_r}{n_r}$
Totals	T_x	$T - T_x$	T	$\hat{P} = \frac{T_x}{T}$

The X_i refer to the observed values. A short-cut method of calculating χ^2 is given by $\chi^2 = \frac{\sum_{i=1}^{r} X_i P_i - \hat{P} T_x}{\hat{p}\hat{q}}$

with r-1 degrees of freedom.

As a further step, it may be desirable to test whether the value of P in the first r_1 rows differs from the value of P in the next r_2 rows, where $r_1 + r_2 = r$.

χ^2 may be divided into 3 components as follows:

	Degrees of Freedom
Difference between P's in first r_1 and last r_2 rows	1
Variation among P_i's within first r_1 rows	$r_1 - 1$
Variation among P_i's within last r_2 rows	$r_2 - 1$
	$r - 1$

A separate χ^2 can then be calculated for each comparison. This is done by subdividing the sum of squares $\left[\sum_{j=1}^{n} X_j P_j - \hat{p} T_x\right]$ into the three relevant components and then dividing each component by Pq.

As an example, consider the following 4 x 2 table

	$\mathcal{A}$ X_j	B	Total n_j	Proportion in Cat. P_j
1	86	814	900	0.095556
2	117	1038	1155	0.101299
3	49	1475	1524	0.032152
4	61	1580	1641	0.037172

$$\hat{P} = 0.059962$$
$$\hat{q} = 0.940038$$
$$\hat{P}\hat{q} = 0.056367$$

$$\text{Total } \chi^2_{(3)} = \frac{\sum_{j=1}^{4} X_j P_j - \hat{p} T_x}{\hat{p}\hat{q}}$$

$$= \frac{(86)(0.095556) + \ldots + (61)(0.037172) - (313)(0.059962)}{0.056367}$$

= 91.27, significant at the 1% level.

There is evidence in the data that the value of P in the first two rows differs from the value of P in the last two rows. To test this hypothesis, we first combine the data into the following 2 x 2 table:

	α X_j	B	Total n_j	Proportion in Cat. P_j
1 and 2	203	1852	2055	0.098783
3 and 4	110	3055	3165	0.034755

We can now obtain the following χ^2's

Rows 1 and 2 vs. Rows 3 and 4

$$= \frac{(203)(0.098783) + (110)(0.034755) - 313(0.055962)}{0.056367}$$

= 90.62, significant at the 1% level.

Row 1 vs. Row 2

$$= \frac{(86)(0.095556) + 117(0.101299) - 203(0.098783)}{0.056367}$$

= 0.30, not significant.

Row 3 vs. Row 4

$$= \frac{(49)(0.032152) +(61)(0.037172) - 110(0.034755)}{0.056367}$$

= 0.35, not significant.

From this we can conclude that the significance of the original χ^2 with 3 d.f. was due to the difference in the value of P in the first two rows from that in the last two rows.

In general, one may divide the rows into any number of groups, based upon the experimenter's discretion. Then the variation in P among the groups, and also within each group may be tested for significance.

(c) 2 x 2 tables

The 2x2 contingency table may be represented as follows:

			Totals
	X_1	X_2	T_x
	Y_1	Y_2	T_y
Totals	T_1	T_2	T

where X_i's and Y_i's are the observed values

A short-cut method of calculating   is given by

$$\chi^2 = \frac{(|X_1Y_2 - X_2Y_1| - T/2)^2 T}{T_1T_2T_xT_y}$$

The subtraction of $\frac{T}{2}$ represents Yates' correction factor, which is advisable for tables with only a single degree of freedom.

Fisher has developed an exact test for the 2 x 2 case. He has shown that the probability of the above table is given by

$$P = \frac{(T_1!\ T_2!\ \ T_x!\ \ T_r!)}{T\ !} \quad \frac{1}{X_1!\ X_2!\ Y_1!\ Y_2!}$$

His method is to calculate the probability of 2 x 2 tables as extreme or more extreme than the one in question, with the marginal totals assumed fixed. If the sum of these probabilities is less than the significance level, the null hypothesis of independence is rejected. For example, consider the table

19	10
12	2

The following tables would be considered more extreme

18	11
13	1

17	12
14	0

The three probabilities for these tables would be summed and if the total fell short of .05, the null hypothesis would be rejected at the 5% level of significance.

Cochran (1959) recommends that this test be used when the sample size is less than 20, or when the sample size is less than 40 and one of the expected frequencies is less than 5.

(d) Combining of Contingency Tables

It often happens that we obtain a number of tables with similar data, but from different testing areas. Some or all of the individual χ^2's may be insignificant. However, we can gain an over-all picture of the significance by pooling both the separate values of χ^2 and their degrees of freedom to obtain a single value for each. The χ^2 test may then be applied as if the data came from a single table.

This method has the disadvantage that it takes no account of the signs of the differences in the different samples. Hence, as pointed out by Cochran (1954), it lacks power in detecting a difference that shows up consistently in the same direction in all or most of the individual tables.

An alternative procedure is to immediately pool the data into one table, and compute χ^2 in the usual way. However, this is possible only if the theoretical probabilities are assumed to be the same in each of the tables.

A third procedure, applicable only to 2 x 2 tables, is suggested by Cochran (1954). One computes the separate χ values, and adds them, taking into account for each table the sign of the difference between the observed proportions. Since χ is approximately normally distributed with mean 0 and unit variance, the sum of g independent χ values is approximately normally distributed with mean 0 and variance g. Therefore the test criterion $Z = \frac{\Sigma \chi}{\sqrt{g}}$ is approximately a standard normal variate.

Cochran particularly recommends this test if the totals of the individual tables do not differ greatly, and if the proportions are all in the range .2 to .8. If these conditions are not satisfied, addition of the χ's tends to lose power.

(e) Coefficients of Contingency

A measure of association between qualitative data arranged in some meaningful way in a contingency table is provided by Pearson's coefficient of contingency. It is given by $c = \sqrt{\chi^2/N+\chi^2}$, where N is the total number of observations.

There are other such criteria such as $c^1 = \frac{\chi^2}{N(t-1)}$, applicable to an rxc table, where t is the smaller of r and c. This criterion lies between 0 and 1.

(f) Analysis of Variance

It should be noted that the χ^2 test of independence between two groups can be replaced by a one-way analysis of variance on the means of these groups. Either method is satisfactory.

For an r x c contingency table, the F-value from the equivalent analysis of variance is related to the χ^2 with $(r-1)(c-1) = k$ degrees of freedom as follows:

$$F = \frac{\chi^2(T-C)}{k\ (T-\chi^2)},$$

where T is the total number of observations

TESTS OF HOMOGENEITY

(a) mx n contingency tables

Consider the following problem. We are given m samples, each sample divided into n classes. We wish to test the hypothesis that the frequencies of the classes for each sample fall into some ratio, say 9:3:3:1 for n = 4. The problem may be presented in the following contingency table:

	1	2		n	
1	11	12			T_1
2	21				T_2
• • • m					
	B_1	B_2		B_n	T

In this type of experiment, there are three useful χ^2's that can be calculated. For each, the null hypothesis is the same: that in each of the m samples, the ratio of the frequencies of the n classes is D_1: D_2... D_n , for some integers D_1, D_2....D_n . The alternative hypotheses for each, however, differ. Following Snedecor's notation, these χ^2's may be described as follows:

(1) Sum χ^2

For each of the n samples a χ^2 with (n-1) degrees of freedom is calculated. The expected frequencies for the cell in the i'th sample and the jth class will be given by

$$\frac{D_j T_i}{\sum^m D_j}, \ i = 1, 2 \dots m$$

These mχ^2's will themselves each give information on whether the hypothesized ratio is being followed in the corresponding sample. We now add the m values of χ^2, each with n-1 degrees of freedom, to obtain a "sum-χ^2" with m(n-1) degrees of freedom. This sum-χ^2 may be significant even if all its components are insignificant. The alternative hypothesis is that the population ratios differ from those hypothesized, with no distinction between excess and deficit, that is, no distinction of sign. This distinction is removed by the squaring of the deviations in the component χ^2's.

(2) Pooled χ^2

In this χ^2, the m samples are treated as one large sample. To calculate it, we use the observed totals B_1, B_2, ...B_n together with the corresponding expected totals

$$\frac{D_1 T}{\sum D_i}, \frac{D_2 T}{\sum D_i}, \ldots \frac{D_m T}{\sum D_i}$$

This "pooled χ^2 " has n-1 degrees of freedom. The alternative hypothesis is that there is a predominating tendency toward deviations with a common sign; that is, the samples deviate in the same direction from the hypothetical values.

(3) Heterogeneity χ^2

This is obtained by subtracting the "pooled χ^2" from the "sum χ^2." It has $m(n-1) - (n-1) = (m-1)(n-1)$ degrees of freedom. It measures the inconsistency of the deviations of the sample ratios from the hypothetical. That is, the alternative hypothesis states that there are deviations, signs considered.

(b) HOMOGENEITY OF VARIANCES

Bartlett's Test of Homogeneity is used to test the hypothesis that $\sigma_1^2 = \sigma_2^2 = \ldots\ldots = \sigma_k^2$ where the σ_i^2 are the population variances corresponding to the k independent sample variances, S_1^2, $S_2^2, \ldots S_k^2$, coming from samples of sizes $n_1, n_2, \ldots n_k$, respectively.

Bartlett has shown that

$$M = \frac{(\sum_{i}^{k}[n_i - 1)] \log \bar{S}^2 - \sum_{i=1}^{k}[(n_i-1)(\log S_i^2)]}{C}$$

where $\bar{S}^2 = \sum_{i}^{k} \frac{S_i^2}{k}$

and C, the correction factor,

$$= 1 + \frac{1}{3(k-1)} \sum_{i}^{k} \left[\frac{1}{(n_i - 1)} - \frac{1}{\sum_{i}(n_i-1)}\right]$$

is distributed approximately as χ^2 with k-1 degrees of freedom. M is calculated from the sample and compared with the appropriate tabular χ^2.

(c) HOMOGENEITY OF CORRELATION COEFFICIENTS

Consider the problem of testing whether k correlation coefficients $r_1, r_2 \ldots r_k$ can be regarded as homogeneous. R.A. Fisher has shown that $Z_i = .5 \ln\left(\frac{1+r_i}{1-r_i}\right)$ is approximately normally distributed with mean $= .5 \ln\left(\frac{1+\rho_i}{1-\rho_i}\right)$, where ρ_i is the population correlation coefficient. $\frac{1}{n_i-3}$ is the variance of r_i calculated from the i^{th} sample of size n_i. Tables have been published that allow for immediate conversion between r_i and Z_i.

Let $\bar{Z}w = \sum_{i=1}^{k} \frac{(n_i-3)Z_i}{\Sigma(n_i-3)}$, the weighted mean of the Z_i

Then $N = \sum_{i}^{k} \frac{(Z_i - \bar{Z}w)}{1/\sqrt{n_i-3}} = \sum_{i=1}^{k} (n_i-3)(Z_i - \bar{Z}w)^2$

$$= \Sigma(n_i-3)Z_i^2 - \frac{\Sigma(n_i-3)Z_i}{[\Sigma n_i - 3]}$$

is approximately a χ^2 with k-1 degrees of freedom. If the null hypothesis that the correlation coefficients are homogeneous is accepted, then a pooled estimate of the true may be obtained by converting $\bar{Z}w$ back to r.

(d) CONFIDENCE INTERVALS AND TESTS OF HYPOTHESES FOR σ^2.

χ^2 is defined as the sum of squares of independent normally distributed variables with zero means and unit variances. That is, $\chi^2_{(n)} = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$, where the X_i are normally distributed with mean μ and variance σ^2.

It follows then that $\frac{\Sigma(X_i - \bar{X})^2}{\sigma^2}$ which is equal to $\frac{(n-1)s^2}{\sigma^2}$ is a χ^2 with n-1 degrees of freedom, since we lose one degree of freedom in estimating μ.

Therefore we can find $\chi^2_{.025}$ and $\chi^2_{.975}$ such that

$$\text{Prob.} \quad \left(\chi^2_{.975} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{.026} \right) = .95$$

$$\text{or Prob.} \quad \frac{(n-1)s^2}{\chi^2_{.025}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{.975}} = .95$$

where $\chi^2_{.025}$ is that value of χ^2 such that $P(\chi^2 \geq \chi^2_{.025})$ $= .025$ and $^2_{.975}$ is such that $P(\chi^2 \geq \chi^2_{.975}) = .975$.

Consequently,

$$\left(\frac{(n-1)s^2}{\chi^2_{.025}}, \frac{(n-1)s^2}{\chi^2_{.975}} \right)$$

constitutes a 95% confidence interval for σ^2.

Similarly, the hypothesis that σ^2 equals a certain constant σ_0^2 can be tested by compting the statistic,

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

and comparing it with the tabular χ^2 for (n-1) degrees of freedom.

(e) GENETICAL APPLICATIONS

(1) Mendelian Ratios

The chi-square statistic is well suited for testing the goodness of fit of observations to genetical ratios. Breeding experiments, that are designed to provide information about the mode of inheritance of certain genetical characteristics, consist mainly of crosses betweenselected individuals. The progeny that result from a cross are then identified and classified according to the presence or absence of a particular characteristic; as exhibited by the appearance of the individual. Genetical theory is responsible for the ratio of kinds of individuals that one expects given a certain hypothesized mode of inheritance. For example, if one makes a cross between two filial generation diploid individuals in which two loci are involved, and if the loci are on separate chromosomes with a condition of simple dominance at each, then one should expect a ratio of 9AB : 3 Ab: 3 aB: lab where A and B are characteristics at the first and second loci respectively and a and b are their respective alleles (Sinnott et al 1950).

The genetical hypothesis that gives rise to the 9:3:3:1 ratio can be tested by means of a chi-square statistic that is calculated from a random sample of observations drawn from the population in question. This chi-square has three degrees of freedom and if it is found significant one concludes that the 9:3:3:1 ratio is not the correct model. The deficiency in the model may be that the loci are not independent and/or that dominance at one or both of the loci is not the case. Chi-square with three degrees of freedom can be partitioned into three parts, each with one degree of freedom, in order to find where the discrepancy lies.

To do this, we construct three orthogonal contrasts, with the following coefficients:

	AB	Ab	aB	ab
(1)	1	1	-3	-3
(2)	1	-3	1	-3
(3)	1	-3	-3	9
P	9/16	3/16	3/16	1/16

Each contrast results in a χ^2 with one degree of freedom. Contrast (1) measures the deviation from a 3:1 ratio at the A locus and $\chi^2 = \frac{[AB + Ab - 3(aB + ab)]^2}{m\sum P_i k^2_i}$ where $P_1 = 9/16$, $P_2 = 3/16$, $P_3 = 3/16$, $P_4 = 1/16$

$k_1 = 1$, $k_2 = 1$, $k_3 = -3$, $k_4 = -3$

n = total sample size.

Contrast (2) leads to a test of the hypothesis that the B locus segregates in a 3:1 ratio and $\chi^2 = \frac{[AB+aB-3(Ab+ab)]^2}{m\sum_i P_i k_i^2}$

where the P_i's are the same as in the previous test, and $k_1 = 1$, $k_2 = -3$, $k_3 = -1$, $k_4 = -3$

Contrast (3) leads to a test for linkage or for independence between the two loci A and B.

$$\chi^2 = \frac{[AB + 9(ab) - 3(Ab + ab)]^2}{m\sum P_i k_i^2}$$

where the P_i's are as before and $k_1 = 1$, $k_2 = -3$, $k_3 = -3$ $k_4 = 9$.

Partitioning χ^2 in this way enables one to locate the discrepancy in the null hypothesis more specifically.

(2) BLOOD GROUP ESTIMATION

Stevens (1950) used the chi-square distribution in developing a goodness of fit test for the special situation in which observed phenotypic frequencies were compared to those expected under a genetical hypothesis.

In particular he was concerned with the estimation of gene frequencies in genetical populations that were classified according to the A-B-O blood group system. He used a method developed by Bernstein (1930), to obtain efficient estimates of the three proportions. Denoting the phenotypes and their numbers by O, A, B, and AB, and their total number by n, the first estimates of the frequencies of the genes A, B, and O, are given by

$$p^1 = 1 - \sqrt{\frac{O+B}{m}}$$

$$q^1 = 1 - \sqrt{\frac{O+A}{m}}$$

$$r^1 = \sqrt{\frac{O}{m}}$$

These preliminary estimates, which are inefficient, have a sum which falls short of 1 by the quantity $D = 1 - (p^1 + q^1 + r^1)$.

Efficient estimates are found by transforming the preliminary estimates as follows:

$$p = p^1 \ (1 + D/2)$$

$$q = q^1 \ (1 + D/2)$$

$$r = (r^1 + D/2) \ (1 + D/2)$$

The preliminary estimates p^1, q^1 and r^1, added to 1-D. It is easily shown that the transformed estimates p, q, and r add to $1-D^2/4$; and since D is the difference between unity and a number between 1 and 0, fairly close to 1, D is a quantity that is very small in absolute value, and $D^2/4$ is considerably smaller. If one applies the transformation to the gene frequencies a second time the discrepancy of the sum of the revised frequencies from unity becomes $D^4/64$ and one sees that this can be made arbitrarily close to zero with repeated application of the transformation. One concludes that $E(D) = 0$.

It can be shown also that the variance of D is $\frac{1}{2n(1+r/pq)}$, and with the assumption that D is normally distributed $2nD^2(1+r/pq)$ is distributed as chi-square with one degree of freedom. This is the criterion that can be used to test the genetic hypothesis that the relative freq uencies are in equilibrium.

GENERAL ESTIMATION OF PARAMETERS AND TESTS OF HYPOTHESIS

(a) Comparison of Estimates and Parametric Values

Let X be the estimate of a parameter E. Then when the number of observations n is large, by virtue of the central limit theorem, X will tend to be normally distributed around E with variance inversely proportional to n, say $V = v/n$, where v is a function of E and not of n. For example, if E is the population mean, then v is σ^a, the population variance.

To test the hypothesis that X comes from a population having parametric value E_o, that is H_o: $E = E_o$, we can use the usual normal deviate test to compare the difference $X - E_o$ to its standard error. However, this is equivalent to writing $\chi^2 = \frac{(X-E_o)^2}{v/n}$, where χ^2 is the square of a standard normal variate and has one degree of freedom.

This test can be generalized quite easily to the case of 2 parameters. Suppose that X and Y are the sample estimates of the parameters E and N, respectively. In the limit, X and Y will have a bivariate normal distribution around E and N respectively. Let $V_1 = \frac{v_1}{n_1}$ and $V_2 = \frac{v_2}{n_2}$ be the variance of X and Y, respectively. Then the hypothesis

$$H_o : E = E_o \qquad N = N_o$$

can be tested by

$$\chi^2 = \left[\frac{(X-E_o)^2}{V_1/n_1} - \frac{2\rho\,(X-E_o)(Y-N_o)}{V_1/n_1\ V_2/n_2} + \frac{(Y-N_o)^2}{V_2/N_2}\right]\frac{1}{1-\rho^2}$$

where ρ is the correlation coefficient between X and Y and has two degrees of freedom.

(b) Probability Ellipses

A confidence interval for the single parameter E will be given by $X \pm Z_o\ V(X)$, assuming that χ is normally distributed about E. Z_o is the value associated with the unit normal curve such that the area under the curve between $-Z_o$ and Z_o is 1-P.

In the case of two parameters we choose a value for χ^2 with two degrees of freedom such that the probability is

1-P that the point representing the true values of the parameters lies within the ellipse having the following equation:

$$= \left[\frac{(X-E_0)^2}{V_1/N_1} - \frac{2\rho(X-E_0)(Y-N_0) \quad +(Y-N_0)}{V_1/N_1 \quad V_1/N_1 \qquad V_2/N_2} \right] \frac{1}{1-\rho^2}$$

As in the case of a confidence interval for one parameter, the value of P is somewhat arbitrary. Stevens (1950) suggests that P = .20 is a good choice. The probability is .80 that the point representing the true values lies within the ellipse. The "standard ellipse" is that obtained by putting $\chi^2 = 1$ in this equation. The probability is .39 that the true values will lie within the standard ellipse. Since the median of χ^2 with 2 degrees of freedom is 1.386, putting $\chi^2 = 1.386$ into the equation gives us an ellipse which has a probability of .50 of containing the point representing the true parametric values. This ellipse is called the probable ellipse.

(c) Comparison of 2 sets of data

Suppose we have 2 samples of sizes n_1 and n_2 drawn from the same population and yielding estimates X_1 and X_2, respectively, of a parameter E. The variance of the difference $X_1 - X_2$ is given by $\frac{v}{n_1} + \frac{v}{n_2} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)v$

where v is a function of E alone. Therefore, to test the hypothesis that the two samples come from the same population with respect to E, that is $H_0 : E_1 = E_2$, we can use the criterion $\chi^2 = \frac{(X_1 - X_2)^2}{v(\frac{1}{n_1} + \frac{1}{n_2})}$

with one degree of freedom.

The generalization to the case of two parameters follows directly. Let X_1 and Y_1 be the estimates of the parameters E and N respectively from the first sample, and X_2 and Y_2 be their estimates from the second. Let $v_1 (\frac{1}{n_1} + \frac{1}{n_2})$ be the variance of $X_1 - X_2$and $v_2 (\frac{1}{n_1} + \frac{1}{n_2})$, the variance of $Y_1 - Y_2$. Then the hypothesis that the samples come from the same population, i.e. H_o: $\begin{matrix} E_1 = E_2 \\ N_1 = N_2 \end{matrix}$

is tested by χ^2 with two degrees of freedom, where

$$\chi^2 = \frac{1}{1-p^2} \left[\frac{(X_1-X_2)^2}{V_1(\frac{1}{n_1} + \frac{1}{n_2})} - \frac{2P(\ _1 - \ _2)(Y_1 - Y_2)}{V_1V_2(\frac{1}{n_1} + \frac{1}{n_2})} + \frac{(Y_1 - Y_2)^2}{V_2(\frac{1}{n_1} + \frac{1}{n_2})} \right]$$

(d) <u>Heterogeneity of more than two sets of data</u>

Suppose we have $t > 2$ samples and we wish to test whether they come from the same population with respect to a certain parameter, E. Let the sample sizes and the estimates of E be respectively, n_1, n_2,...n_t and X_1, X_2, ...X_t. Let $\frac{v}{n_1}$ be the variance of X_i, and let $\bar{X} = \frac{\Sigma n_i X_i}{\Sigma n_i}$.

be the weighted mean of the estimates. Then we can test the hypothesis: $H_o : E_1 = E_2 = \ldots\ldots = E_t$ by the criterion

$\chi^2 = \frac{\sum_{i=1}^{t}(X_i - \bar{X})^2}{V/n_i}$ with t-1 degrees of freedom.

Now let us consider the generalization to two parameters, E and N. Let Y_1, Y_2 ....Y_t be the estimates of N from the t samples, and let $\overline{Y} = \frac{\sum m_i Y_i}{\sum m_i}$ be their weighted mean.

Then the hypothesis that the t samples come from the same population with respect to the parameters E and N, that is,

$$H_o : E_1 = E_2 = \ldots\ldots E_t$$
$$N_1 = N_2 = \ldots\ldots N_t$$

is tested by

$$\chi^2 = \frac{1}{1-\rho^2}\left[\sum_{i=1}^{t}\frac{(X_i-\overline{X})^2}{V_X/m_i} - 2\rho\sum_{i=1}^{t}\frac{(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{V_X/m_i \cdot V_Y/m_i}} + \sum_{i=1}^{t}\frac{(Y_i-\overline{Y})^2}{V_Y/m_i}\right]$$

with 2t-2 degrees of freedom.

(e) Approximations Made And Their Improvement

The tests described in this section are not exact. In the first place the distributions of the estimates are not exactly normal, and secondly variances of the estimates are not known exactly. Stevens (1950) suggests that the reliability of the tests increases with a suitable transformation of the data. It is well known that transformations that stabilize the variance also make the distributions more normal.

INDICES OF DISPERSION

χ^2 can be used in the situation where a goodness of fit test to the binomial or Poisson distribution is required, but in cases where the observed values are so few that it is not worthwhile to obtain expected values. In this case criteria are available that will test whether the sample variance is compatible with the theoretical variance.

(a) Binomial Index

For the binomial distribution, the index of dispersion is $\chi^2_{k-1} = \sum_{i}^{k} \frac{(X_i - X)^2}{\bar{X}} + \sum_{i=1}^{k} \frac{(X_i - \bar{X})^2}{m - \bar{X}}$

where X_1, X_2,... X_k are the number of successes for each of k samples of n trials.

(b) Poissonian Index

For situations in which p is very small and n very large, the binomial index of dispersion reduces to the Poisson index of dispersion, $\chi^2_{k-1} = \sum_{i=1}^{k} \frac{(X_i - \bar{X})^2}{\bar{X}}$

This criterion can be used for testing the hypothesis that the data come from a Poisson distribution.

LINEAR REGRESSION IN N x 2 TABLES

Suppose that we have a contingency table with N rows and 2 columns, one column containing observations denoted by X_j, j= 1....N, the other column containing observations denoted by Y_j, j = 1...N. Let $N_j = X_j + Y_j$ denote the marginal totals. Let $P_j = \frac{X_j}{N_j}$ be interpreted as the observed proportion of successes in the j^{th} member of the sample. The situation may arise where there is a variate Z_j, assigned to each of the rows, which is linearly related to P_j. Alternatively it may be possible to assign such a variate to the rows, assuming that they fall into some natural order. Then a continuous scale may be created which will be linearly related to the P_j's. The problem may be represented as follows:

			A		
	X_j	Y_j	n_j	P_j	Z_j
	X_1	Y_1	n_1	P_1	Z_1
	X_2	Y_2			
	$\cdot$				
B					
	$\cdot$				
	X_N	Y_N	n_N	P_N	Z_N
	T_x	T_y	T		

$$n_j = X_j + Y_j$$

$$P_j = \frac{X_j}{n_j}$$

Let $\hat{P}$ be an estimate of P from the total sample:

$$\hat{P} = \frac{\sum X_i}{\sum m_i}.$$

Then the regression coefficient b of P_j on Z_j is

$$b = \frac{\sum m_i(P_i - \hat{P})(Z_i - \bar{Z}_w)}{\sum m_i(Z_i - \bar{Z}_w)^2}$$

where $Z\bar{w}$ is the weighted mean of the Z_j.

The χ^2 for regression, with one degree of freedom, is

$$\chi^2 = \frac{\left[\sum_{i=1}^{N} X_i Z_i - \frac{T_x}{T}\left(\sum_{i=1}^{N} m_i Z_i\right)\right]}{\hat{P}(1-\hat{P})\left[\sum_{i=1}^{N} m_i Z_i^2 - \frac{\left(\sum m_i Z_i\right)^2}{T}\right]}$$

The total χ^2 from the N x 2 table can be partitioned as follows:

	Degrees of Freedom
Regression of P_j on Z_j	1
Deviations from regression	N-2
Total	N-1

As an example, consider the following data. One hundred and ninety six hospital patients were classified as to the degree of infiltration (a measure of a certain type of skin damage) and change in condition after 48 weeks of treatment. The total χ^2, 6.88 with 4 d.f., is insignificant. However it is noticed that the P_j (the proportions of patients with severe infiltration) decline steadily from the "markedly improved" class to the "worse" class. This suggests that a regression of the P_j on the clinical change might provide a more sensitive test.

In order to compute the regression, we assign the scores 3, 2, 1, 0 and -1 respectively to the five classes of clinical change:

		Degree of Infiltration		Total		
Clinical Change	Z_j	0-7	8-15	N_j	$P_j=X_j/n_j$	N_jZ_j
Marked Improvement	3	11	7	18	.39	54
Moderate Improvement	2	27	15	42	.36	84
Slight Improvement	1	42	16	58	.28	58
Stationary	0	53	13	66	.20	0
Worse	-1	11	1	12	.08	-12
Total		144		196		184

$$\chi^2_{(1)} = \frac{\left(\sum X_i Z_i - \frac{T_x}{T}\sum m_i Z_i\right)^2}{\hat{P}(1-\hat{P})\left[\sum m_i Z_i^2 - \frac{(\sum m_i Z_i)^2}{T}\right]}$$

$$= \frac{(7)(3) + (15)(2) + \ldots + (1)(-1) - \frac{(52)(184)}{196}\quad{}^2}{(.26531)(1-.26531)\ (54)(3) + (84)(2) + \ldots + (-12)(-1) - \frac{(184)^2}{196}}$$

= 6.666, significant at the 1% level.

The total χ^2 has now been subdivided as follows:

	d.f.	χ^2
Regression of P_j on Z_j	1	6.67
Deviations from Regression	3	0.21
	4	6.88

TESTS OF SECOND-ORDER INTERACTION

2 x 2 x 2 Factorial Experiments

In a factorial experiment with three factors each at two levels, the main effects and the interactions can be tested by an F statistic, as is well known. However, it is also possible to use the goodness of fit criterion to test the second order interaction.

If the three factors are denoted by A, B, and C, and the levels denoted by 1 and 2, we can set up the following contingency table.

A_1				A_2				
B_1		B_2		B_1		B_2		
C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2	Total
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	n
M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	n

In this table, the X_i are the observed values, and the M_i are the expected values, where $\sum X_i = \sum M_i = n$. An example using this method can be found in an article by M. Kastenbaum and D. Lamphiear (1959).

To test the second order or ABC interaction we test whether the BC interaction is the same for A_1 and A_2. The null hypothesis is that

$$\frac{M_1/M_2}{M_3/M_4} = \frac{M_5/M_6}{M_7/M_8}$$

or $M_1\ M_4\ M_6\ M_7 = M_2\ M_3\ M_5\ M_8$

Because of this relation, and since $\sum M_i = n$, which is fixed, we have only to estimate six of the M_i's. Therefore we have left 8-6-1 = 1 degree of freedom.

By the method of maximum likelihood, the best estimates of the M_i are:

$$\begin{aligned} M_1 &= X_1 + k & M_5 &= X_5 - k \\ M_2 &= X_2 - k & M_6 &= X_6 + k \\ M_3 &= X_3 - k & M_7 &= X_7 + k \\ M_4 &= X_4 + k & M_8 &= X_8 - k \end{aligned}$$

where k is found from the equation

$$(X_1 + k)(X_4 + k)(X_6 + k)(X_7 + k) = (X_2 - k)(X_3 - k)(X_5 - k)(X_8 - k)$$

The criterion $\chi^2 = \dfrac{\sum(X_i - M_i)^2}{M_i}$ with one degree of freedom tests the null hypothesis of no 3-factor interaction.

r x s x t FACTORIAL EXPERIMENTS

The foregoing test can be extended to the general (r x s x t) contingency table. In this case, estimation of the parameters will involve the solution of (r-1)(s-1)(t-1) third degree equations, and the resulting χ^2 will have (r-1)(s-1) (t-1) degrees of freedom. In general, the null hypothesis of no second-order interaction for an (r x s x t) contingency table is given by

$$\frac{P_{rst}\,P_{ijt}}{P_{ist}\,P_{rjt}} = \frac{P_{rsk}\,P_{ijk}}{P_{isk}\,P_{rjk}}$$

where i = 1, 2, ...r-1

j = 1, 2, ...s-1

k = 1, 2, ...t-1

and the P_{ijk}'s are the parameters of the multinomial distribution

$$\phi = \frac{N!}{\prod_i \prod_j \prod_k m_{ijk}} \prod_i \prod_j \prod_k P_{ijk}^{m_{ijk}}$$

for $\sum_i \sum_j \sum_k m_{ijk} = N$

and $\sum_i \sum_j \sum_k P_{ijk} = 1$

Since solving the (r-1)(s-1) (t-1) simultaneous equations is extremely laborious, Kastenbaum and Lamphiear (1959) have given a computational procedure well-suited for a desk calculator.

NON-PARAMETRIC TESTS

Non-parametric tests are those which do not depend upon any particular frequency distribution, nor any particular parameter. The first non-parametric test developed was the goodness of fit test itself. Two others which will be discussed now are the test for medians and Friedman's two-way analysis of variance test.

MEDIAN TEST

Consider the problem of testing the hypothesis that two populations, denoted by X and Y, have the same median. Let X_1, X_2 .... X_{N1}, and Y_1, Y_2,...Y_{N2} be the ordered samples from the X and Y populations, respectively, and let Z_1, Z_2,...Z_{N1+N2} be the order statistics from the combined sample. Denote the median of the combined sample by $\bar{Z}$. If the null hypothesis is true, we should expect the number of X's that exceed $\bar{Z}$, say M_1, to equal $\frac{N_1}{2}$, and the number of Y's that exceed $\bar{Z}$, say M_2, to equal $\frac{N_2}{2}$.

It then follows that M_1 and M_2 have the following hypergeometric probability distribution:

$$g\,(M_1, M_2) = \frac{\binom{N_1}{M_1}\binom{N_2}{M_2}}{\binom{N_1 + N_2}{M_1 + M_2}}$$

This is the same distribution as that followed by the frequencies in a 2x2 contingency table, under the hypothesis of independence. The corresponding contingency table is

	<Median	>Median	Total
Sample I	$N_1 - M_1$	M_1	N_1
Sample II	$N_2 - M_2$	M_2	N_2
	$(N_1 + N_2) - (M_1 + M_2)$	$M_1 + M_2$	$N_1 + N_2$

The null hypothesis that the two populations have a common median can be tested by the χ^2 criterion with one degree of freedom. However, if either N_1 or N_2 is less than 10, Fisher's exact test should be used.

This test is easily extended to more than two samples. Suppose we wish to test whether k populations have the same median. Let $\overline{Z}$ be the median of the k combined samples. If M_i, $i = 1, 2, \ldots k$ is the number of observations from the i^{th} sample that exceed $\overline{Z}$, then the distribution of $M_1, M_2, \ldots . M_k$. is

$$g\ (M_1, M_2, \ldots M_k) = \prod_{i=1}^{k} \binom{N_i}{M_i} \Big/ \binom{N}{M}$$

This is the same distribution as that followed by the frequencies in the following k x 2 contingency table, under the hypothesis of independence:

	Median	Median	
Sample 1	$N_1 - M_1$	M_1	N_1
Sample 2	$N_2 - M_2$	M_2	N_2
⋮			
k	$N_k - M_k$	M_k	N_k
		M	N

The null hypothesis may be tested by the χ^2 criterion with k-1 d.f. For this problem χ^2 may be put in the following form:

$$\chi^2 = N(N-1)\sum_i^k \frac{1}{m_i}\left(m_i - n_i \sum^k \frac{m_i}{N}\right)^2 \frac{1}{\sum_i^n m_i \left(N - \sum_i^k m_i\right)}$$

FRIEDMAN'S 2-WAY ANALYSIS OF VARIANCE TEST

Friedman (1937) proposed a non-parametric procedure for use as a test for differences among treatment means in a randomized-block design. His test involved ranking the treatments within each block from lowest to highest, and then determining the probability that the different columns of ranks came from the same population. After ranking, the procedure requires the sum of the ranks for each treatment, and the following criterion as a test of the null hypothesis of no differences among treatment means:

$$\chi_M^2 = \frac{12}{bt(t+1)} \sum r_i^2 - 3b(t+1)$$

where t is the number of treatments, b is the number of blocks, and r_i the sum of the ranks for the ith treatment. This criterion is approximately distributed as chi-square with t-1 degrees of freedom. The approximation is poorest for small sums of t and b, but in this case exact probability tables can be used.

TOLERANCE LIMITS

In the manufacturing of industrial products, limits beyond which only a small fraction of items is expected to fall are constructed. These limits L_1 and L_2 are called tolerance limits.

Let us assume that we have a population which is normally distributed with mean μ and variance σ^2. To find two symmetric tolerance limits such that the probability is P that 100(1-2E)% of the population lies between these two limits, we solve the following equation for Z_{1-E}:

$$\int_{\mu-\sigma Z_{1-E}}^{\mu+\sigma Z_{1-E}} \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(X-\mu)^2}{2\sigma^2}}\, dX = 1-2E$$

For example, 1-2E = .95 gives the tolerance limits as $\mu \pm 1.960$.

Suppose now that μ is known but that σ is not. Let S be its sample estimate. Then the condition that $\mu \pm S\ell$ will include at least (1-2E)% of the population is that

$$S\ell > \sigma Z_{1-E} \quad \text{or} \quad S/\sigma > Z_{1-E}/\ell$$

But $\chi^2 = \underline{(n-1)S^2}$

$$\therefore \frac{S}{\sigma} = \sqrt{\frac{\chi^2}{n-1}}$$

If P is the probability that at least 100(1-2E)% of the population is included between these limits in repeated random sampling, then $\Pr\left(S/\sigma > Z_{1-E}/\ell\right) = P$ has the solution

$$\sqrt{\chi^2_{1-P}/_{n-1}} = Z_{1-E}/\ell$$

$$\ell = Z_{1-E}\sqrt{n-1/\chi^2_{1-P}}$$

Therefore, the sample tolerance limits are

$$\left(\mu - S Z_{1-E}\sqrt{n-1/\chi^2_{1-P}}\ ,\ \mu + S Z_{1-E}\sqrt{n-1/\chi^2_{1-P}}\ \right)$$

In the general case, where both μ and σ are unknown, it may be shown that the value of ℓ is approximately given by $\ell = Z_{1-E}\sqrt{\frac{n-1}{\chi^2_{1-P}}}\ (1 + \frac{1}{2n})$

CHAPTER V - NON-CENTRAL ^{2}

DEFINITION AND DESCRIPTION

χ^2 is defined as the sum of squares of n independent random variables that are normally distributed with zero means and unit standard deviations.

Consider now n independent and normally distributed variables, E_1, E_2,...E_n with means A_1, A_2, A_n and each with the same standard deviation σ. Let $X_i = E_i - A_i$.

Then non-central χ^2 is defined by

$$\chi'^2 = \sum_{i=1}^{n} \frac{(X_i + A_i)^2}{\sigma^2} = \sum_{i=1}^{n} \varepsilon_i^2$$

χ'^2 is a generalized form of χ^2.

The distribution of χ'^2 can be shown to be

$$f(\chi'^2) = \exp \frac{\left[-\chi'^2/2\, e^{-\lambda/2}(\chi'^2)^{n/2-1}\right]}{2^{n/2}\,\Gamma(n/2)} \left[1 + \frac{\chi'^2\lambda}{2} + \frac{(\chi'^2\lambda)^2}{(n)(n+2)2!4} + \cdots\right]$$

where λ is the parameter of non-centrality, and n is called the number of degrees of freedom.

The moment generating function of χ'^2 is

$$M(t) = \frac{e^{(\lambda t/1-2t)}}{(1-2t)^{\frac{n}{2}}}, \quad t < 1/2$$

The r^{th} cumulant can then be shown to be

$K_r = 2^{r-1}(r-1)!\,(n+r)$. In particular,

$K_1 = n + \lambda = \mu_{\chi'^2}$

$K_2 = 2(n + 2\lambda) = \sigma^2_{\chi'^2}$

$K_3 = 8(n + 3\lambda) =$

$K_4 = 48\ (n + 4)$

APPROXIMATIONS TO THE χ'^2 DISTRIBUTION

(a) The χ^2 approximation

The distribution of $\dfrac{\chi'^2}{P}$, where

$P = \dfrac{n + 2\lambda}{n + \lambda} = 1 + \dfrac{\lambda}{n+\lambda}$, can be

approximated by that of χ^2 with r degrees of freedom, where $r = \dfrac{(n + \lambda)^2}{n + 2\lambda} = n + \dfrac{\lambda^2}{n+2\lambda}$

r is in general a fraction.

(b) The normal approximation

Fisher has shown that $\sqrt{2\chi^2}$ is approximately normally distributed with mean $\sqrt{2n}$ and variance 1. There is a similar approximation to the distribution of χ'^2. Furthermore, χ'^2 approaches normality faster than χ^2.

Patnaik (1949) has shown that $\sqrt{\dfrac{2\chi'^2(n+\lambda)}{n+2\lambda}}$ is approximately normally distributed with mean $\sqrt{\dfrac{2(n+\lambda)^2}{n+2\lambda} - 1}$ and variance 1.

APPLICATIONS OF THE NON-CENTRAL DISTRIBUTION TO THE POWER FUNCTION OF TESTS

Suppose we wish to test the null hypothesis that a random sample of observations, E_1, E_2, $\ldots E_n$, comes from a population normally distributed with mean 0 and variance 1. Then the test criterion is $\chi^2 = \sum E_i^2$. If we wish to compute the power of the test, we must find the probability that this criterion exceeds the critical value χ_0^2 under some alternative hypothesis. Let this alternative hypothesis be that the E_i come from populations with unit variances but different means, A_1 , $A_2, \ldots A_n$. Then the distribution of non-central χ^2 can be used to supply the power function, given by

$$\int_{\chi_0^2}^{\infty} f(\chi'^2/\lambda)\, d\chi'^2$$

where the null hypothesis is that $\lambda = \sum A_i^2 = 0$, and the alternative hypothesis is the composite hypothesis, including the family of alternatives for which $\sum A_i^2 = \lambda$.

The non-central χ^2 distribution can also be used to determine the power of the goodness of fit test. The crux of this test is that if n_i denotes the observed frequencies and $N\pi_i$ the expected frequencies, then $\sum_{i=1}^{k} \frac{(m_i - N\pi_i)^2}{N\pi_i}$, where k is the number of classes, is approximately distributed as χ^2 with k-1 degrees of freedom.

Patnaik (1949) has shown that if the true expectations are NP_i rather than $N\pi_i$, where $\sum P_i = \sum \pi_i = 1$, then $\sum_{i=1}^{k} \frac{(m_i - N\pi_i)^2}{NP_i}$ is approximately distributed as non-central χ^2 with k-1 degrees of freedom, With this result, one can determine the power of the goodness of fit test of any simple hypothesis (specifying probabilities π_i) with respect to simple alternative hypotheses (specifying probabilities P_i). Knowing the power function, several important problems can be solved. As examples, Patnaik cites the following:

"(1) For a given sample size N and number of groups k, what is the chance of establishing the inadequacy of the null hypothesis, using a given significance level?

(2) For a given k, how many observations are necessary to give a chance, say of 90%, of establishing significance at the 5% level?

(3) For a given k and N, how large a departure of H_1 and H_o will be detected with a given chance?"

Illustrations of these applications are given by Patnaik (1949).

SUMMARY AND DISCUSSION

The distribution of continuous chi-square, defined as the sum of squares of standard normal variates, was originally derived by F.R. Helmert in 1876 by mathematical induction. Chapter I gives this derivation as well as three others.

Only the normal distribution, to which it is intimately related, outranks the chi-square distribution in general usefulness. Accounting for this are its general yet diversified properties, and it is with application in mind that these properties are discussed in Chapter II. Of particular importance are the reproductive property of chi-square and its identification with the exponent of the multivariate normal distribution. The former property facilitates the location of underlying areas of significance with pin-point precision, while the latter extendsthe breadth of chi-square's inferential powers to parameters of any distribution. Both form the theoretical basis for many of the tests described in Chapter IV.

In need of some criterion to measure the quality of curve-fitting,Karl Pearson introduced the goodness of fit statistic, or discrete chi-square, and established its asymptotic distribution as that of continuous chi-square. As the oldest of the non-trivial significance tests, as well as one of the most widely used, chi-square has an absorbing, controversial history. In particular the "degrees of freedom battle", waged for over twenty years by Karl Pearson and Sir Ronald Fisher, has theoretical and philosophical implications worthy of careful study. The flavour of this and other historical highlights are discussed in Chapter III.

Over the years chi-square has turned into an enormously useful device with a range of applications far greater than the specific problem to which it was initially applied. Although by no means complete, the selection of tests in Chapter IV has been chosen with this range in mind. As well as covering the standard applications, including goodness of fit tests, tests of independence, and tests of homogeneity, less well-known procedures are also investigated. These include tests of second-order interactions in factorial experiments, bivariate confidence ellipses, and the estimation of gene frequencies in genetical populations.

Chapter V contains a discussion of the non-central chi-square distribution, discovered by Patnaik in 1949. Particularly emphasized is its relationship to the central chi-square distribution.

BIBLIOGRAPHY

A. BOOKS

Anderson, R.L., and Bancroft, T.A.,
Statistical Theory in Research
McGraw-Hill Co. Inc.,
New York, Toronto and London 1952.

Cramer, H.,
Mathematical Methods of Statistics
Princeton University Press
Princeton, N.J., 1946.

Fisher, R.A.,
Statistical Methods for Research Workers
Oliver and Boyd, Ltd.,
Edinburgh and London, 1932.

Goulden, C.H.,
Methods of Statistical Analysis
John Wiley and Sons Inc., N.Y., 1939.

Greenhood, E.,
A Detailed Proof of the Chi-Square Test of Goodness
of Fit, Harvard University Press, 1939.

Hald, A.,
Statistical Theory with Engineering Applications
New York, Wiley, 1952.

Helmert, F.R.,
Ueber Die Wahrscheinlichkeit Der Potenzsummen Der
Beobachtungsfehler Und Ueber Einige Damit Im
Zusammenhange Stehende Fragen
Zeitschrift Fur Mathematik Und Physik, 21, 1876,
192-218.

Hoel, P.G.,
Introduction to Mathematical Statistics
John Wiley and Sons, Inc., N.Y., 1947.

Hogg, R.L., and Craig, A.T.,
Introduction to Mathematical Statistics
The Macmillan Company, N.Y., 1965.

Kenney, J.F., and Keeping, E.S.,
Mathematics of Statistics
D. Van Nostrand Co. Inc.,N.Y.,

Maxwell, A.E.,
Analyzing Qualitative Data,
John Wiley and Sons, N.Y., 1961.

BIBLIOGRAPHY

A. BOOKS (CONT'D)

Mood, A.M.,
Introduction to the Theory of Statistics
McGraw-Hill Book Co. Inc., N.Y., 1950.

Peters, C.C., and Van Voorhis, W.R.,
Statistical Procedures and Their Mathematical Bases,
McGraw-Hill Co., N.Y., 1940.

Snedecor, G.W.,
Statistical Methods,
Iowa State Press, Ames, Iowa, 1956.

Steele, R., and Torrie, J.,
Principles and Procedures of Statistics
McGraw-Hill Co. Inc., 1960.

Weatherburn, C.E.,
Mathematical Statistics,
Cambridge University Press, 1947.

Yule, G.U., and Kendall, M.,
Introduction to Theory of Statistics,
C. Griffin and Co., 1937.

B. JOURNAL ARTICLES

Adler, F.,
"Yates' Correction and the Statisticians"
Jour. Am. Stat. Assoc.,
Vol. 46, 1951, pp. 490-501.

Camp, B.H.,
"The Multinomial Solid and the Chi-Test",
Transactionsof the American Mathematical Society,
Vol. 31, 1929, pp. 133-144.

Cochran, W.G.,
"Some Methods for Strengthening the Common χ^2 Test",
Biometrics, Vol. 10, 1954, pp. 417-451.

Cochran, W.G.,
"The χ^2 Test of Goodness of Fit",
Annals of Math. Stat.,
Vol. 23, 1952, pp. 315-345.

Fisher, R.A.,
"On the Interpretation of Chi-Square from Contingency
Tables, and the Calculation of P".
J.R.S.S., Vol. 85, 1922, pp. 87-94

BIBLIOGRAPHY

B. JOURNAL ARTICLES (CONT'D)

Fisher, R.A.,
"The conditions under which chi-square measures the discrepancy between observation and hypothesis."
J.R.S.S., Vol. 87, 1924, pp. 442-450.

Haldane, J.B.S.,
"The Exact Value of the Moments of the Distribution of χ^2 as a Test of Goodness of Fit when Expectations are Large", Biometrika, Vol. 29, 1937, pp. 133-143.

Kastenbaum, M.A., and Lamphiear, D.E.,
"Calculation of Chi-Square to Test the No. 3 Factor Interaction Hypothesis".
Biometrics 1959, Vol. 15, pp. 107-115.

Kimball, A.W.,
"Short-Cut Formulas for the Exact Partition of χ^2 in Contingency Tables".
Biometrics, Vol. 10, 1954, pp. 452-458.

Lancaster, H.O.,
"The Derivation and Partition of   in Certain Discrete Distributions".
Biometrika, Vol. 36, 1949, pp. 117-129.

Mann, H.B., and Wald, A.,
"On the Choice of the Number of Class Intervals in the Application of the Chi-Square Test."
Annals of Math. Stat.
Vol. 13, 1942, pp. 306-317.

Patnaik, P.B.,
"The Non-Central χ^2 - and F- Distributions and Their Applications", Biometrika, Vol. 36, 1949, pp. 202-232.

Pearson, Karl,
"On the Criterion that a Given System of Deviations From the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling".
Phil. Mag. Series 5, Vol. 150, 1900, pp. 157-172.

Pearson, Karl,
"Experimental Discussion of the (χ^2, P) Test for Goodness of Fit",
Biometrika, 1932, Vol. 27, pp. 351-381.

BIBLIOGRAPHY

B. JOURNAL ARTICLES (CONT'D)

Pearson, Karl,
"On the Theories of Multiple and Partial Contingency",
Biometrika, Vol. XI, pp. 145-158.

Yule, G.U.,
"On the Application of the χ^2 Method to Association and Contingency Tables with Experimental Illustrations."
J.R.S.S., Vol. 87, 1922, pp. 76-82.