CAMBRIDGE

Cambridge University Press
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Excerpt
More information

Basics

© in this web service Cambridge University Press

CAMBRIDGE

Cambridge University Press
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Excerpt
More information

1

Statistical Distributions

The Normal Probability Distribution

The normal probability distribution has a long and rich history. Three names are always mentioned: Abraham de Moivre (b. 1667), Carl Friedrich Gauss (b. 1777), and Pierre-Simon Laplace (b. 1799). De Moivre is usually credited with the discovery of the normal distribution. Gauss introduced a number of important mathematical and statistical concepts derived from a normal distribution (1809). Adolphe Quetelet (b. 1796) suggested that the normal distribution was useful for describing social and biologic phenomena. In his study of the "average man" Quetelet characterized heights of army recruits with a normal distribution (1835). Twentieth-century statisticians Karl Pearson (b. 1857) and R. A. Fisher (b. 1890) added a few details, producing the modern normal distribution. Today's normal probability distribution has other less used names: Gaussian distribution, "bell-shaped curve," and Gauss-Laplacian distribution.

The algebraic expression of the normal probability distribution for a value denoted x is

$$f(x) = \frac{1}{\sigma_X\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma_X}\right]^2}.$$

The expression shows that the value of the function $f(x)$ is defined by two parameters: a mean value represented by μ and a variance represented by σ_X^2 (Chapter 27). The mean value μ determines location, and variance σ_X^2 determines spread or shape of the normal distribution (Figure 1.1). The usual estimates of these two parameters are the sample mean value (denoted $\bar{x}$) and the sample variance (denoted S_X^2). For a sample of n values $\{x_1, x_2, \ldots, x_n\}$,

$$\text{sample estimated mean value} = \bar{x} = \frac{1}{n}\sum x_i$$

and

$$\text{sample estimated variance} = S_X^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 \quad i = 1, 2, \ldots, n \ \ [13].$$

From the mathematical expression, the height $f(x - \mu)$ equals the height $f(\mu - x)$, making the normal distribution symmetric relative to the mean value μ. In addition, again seen from the expression $f(x)$, the normal distribution is always positive because e^{-z^2} is always positive for any value of z. The expression $f(x)$ dictates that a single maximum value occurs at the value $x = \mu$ and is $f(x_{\max}) = f(\mu) = 1/(\sigma_X\sqrt{2\pi})$ because e^{-z^2} is always less than 1 except

Note: the chapter number in parentheses indicates the chapter where the discussed statistical tool is further described and applied in a different context.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Excerpt
More information

Table 1.1 *Description: A Few Selected Critical Values and Their Cumulative Probabilities*[a] *from a Standard Normal Distribution ($\mu = 0$ and $\sigma = 1$)*

	Standard normal distribution								
Critical values (z)	−2.0	−1.5	−1.0	−0.5	0.0	0.5	1.0	1.5	2.0
Probabilities ($1 - \alpha$)	0.023	0.067	0.159	0.309	0.500	0.691	0.841	0.933	0.977
Probabilities (α)	0.977	0.933	0.841	0.691	0.500	0.309	0.159	0.067	0.023

[a] $P(Z \leq z) = 1 - \alpha$ making the value z a critical value (percentile/quantile) for a specific cumulative probability denoted $1 - \alpha$.

when $z = x - \mu = 0$. Therefore, the symmetric normal distribution has mean value, median value, and mode at the center of the distribution and "tails" that extend indefinitely for both positive and negative values of x (Figure 1.1). The total area enclosed by a normal probability distribution is 1.

A principal role of the normal distribution in statistics is the determination of probabilities associated with specific analytic results. In this context, the term *critical value* is frequently used to identify values from a normal distribution that are otherwise called quantiles or percentiles. These values and their associated probabilities typically arise as part of a statistical evaluation. For example, value $z = 1.645$ is the 95th percentile of a normal distribution with mean value $= \mu = 0$ and variance $= \sigma^2 = 1$ but is usually referred to as the critical value at the 95% significance level when applied to test statistics, confidence intervals, or other statistical summaries.

An essential property of the normal probability distribution is that a standard distribution exists; that is, a single normal distribution can be used to calculate probabilities for values from any normal distribution using the parameters μ and σ_X^2. This standard normal distribution has mean value $\mu = 0$ and variance $\sigma^2 = 1$. Table 1.1 gives a sense of the relationship between the critical values (quantiles/percentiles) and their associated probabilities from this standard normal distribution. Of course, more extensive tables exist, and typically these probabilities are computer calculated as part of a statistical analysis.

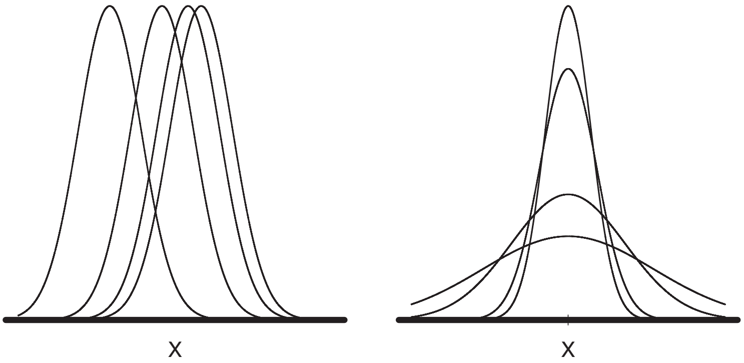

Figure 1.1 Four Normal Probability Distributions with Different Mean Values and the Same Variance (Left) and Four Normal Distributions with the Same Mean Value and Different Variances (Right)

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Excerpt
More information

For example, the probability that a random value from the standard normal distribution is less than a critical value 1.5 is 0.933, called a *cumulative normal probability*. In symbols, for a random value represented by Z from a standard normal distribution, then

$$\text{cumulative normal probability} = P(Z \leq z) = P(Z \leq 1.5) = 0.933.$$

Geometrically, the cumulative probability $P(Z \leq z)$ is the area enclosed by the normal distribution to the left of the value z.

A miraculous property of the normal distribution allows this standard normal distribution to be used to calculate probabilities for any normal distribution. The process of obtaining a probability for a specific value is simply a change in measurement units. When a normally distributed value is measured in terms of standard deviations (square root of the variance, denoted σ_X) relative to above or below the mean value, the associated cumulative probabilities from all normal distributions are the same. For example, the probability that a value is more than two standard deviations below a mean value μ is 0.023 for all normal distributions. Thus, from any normal distribution $P(X \leq \mu - 2\sigma_X) = 0.023$. Therefore, to find a probability for a specific value, the units of the original value are converted to units of standard deviations. For example, an observation of $x = 25$ feet sampled from a normal probability distribution with mean value $\mu = 10$ and variance $\sigma_X^2 = 100$ is $Z = (x - \mu)/\sigma_X$ standard deviations above the mean, and Z has a normal probability distribution with mean $\mu = 0$ and standard deviation $\sigma_X = 1$ (Table 1.1). Therefore, the probabilities associated with Z are the same as the probabilities associated with X when the value x is converted to standard deviations. Specifically, the value $x = 25$ feet is $z = (25 - 10)/10 = 1.5$ standard deviations above the mean, making $P(X \leq 25.0) = P(Z \leq 1.5) = 0.933$ because $P(X \leq x$ natural units$) = P(Z \leq z$ standard deviation units$)$ from a standard normal distribution (Table 1.1).

To determine a value in natural units associated with a specific probability, the process is reversed. The value Z from the standard normal distribution is converted to original units. The critical value $z_{1-\alpha}$ associated with the probability $1 - \alpha$ is used to calculate $X = \mu + z_{1-\alpha}\sigma_X$ where $z_{1-\alpha}$ represents a value from the standard normal distribution. The symbol $1 - \alpha$ traditionally represents the cumulative probability $P(X \leq x) = 1 - \alpha$ making x the $(1 - \alpha)$ – level quantile or the $(1 - \alpha) \times 100$ percentile of the normal probability distribution of a variable X. For the example, when $1 - \alpha = 0.933$, then $z_{0.933} = 1.5$ (Table 1.1) and $X = 10 + 1.5(10) = 25$ feet for a normal probability distribution with mean value $\mu = 10$ and variance ${\sigma^2}_X = 100$. As required, the associated probability is again $P(Z \leq 1.5) = P(X \leq 25.0) = 0.933$. Figure 1.2 schematically displays the relationship between cumulative probabilities, observed values, and standard deviations for all normal distributions.

From the symmetry of the normal probability distribution $f(x)$, it follows that the probabilities are symmetric. For a normally distributed value X, then $P(X \geq c) = P(X \leq -c)$. For example, for a standard normal distribution $P(Z \geq 1.0) = P(Z \leq -1.0) = 0.159$ because $P(Z \leq -1.0) = 0.159$ (Table 1.1 and Figure 1.2). In symbols, for example,

$$P(X \leq \mu - 2\sigma) = P(X \geq \mu + 2\sigma).$$

The ubiquitous role of the normal distribution in statistical analysis stems from the *central limit theorem*. Probability theory is not simple, and neither is the central limit theorem. Leaving out considerable detail, a statement of this indispensable theorem is as follows:

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Excerpt
More information

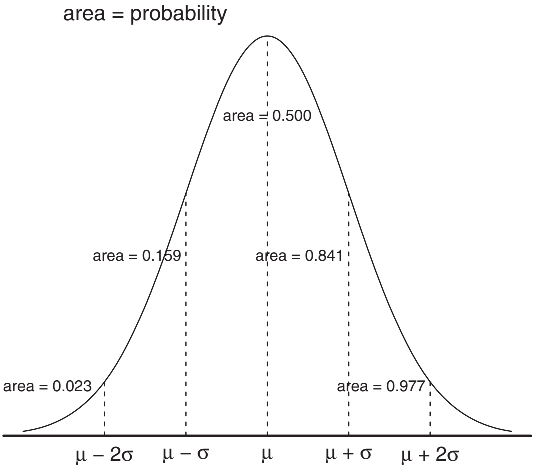

Figure 1.2 Normal Distribution Probabilities (Area to the Left = Cumulative Probability = $1 - \alpha$) and Their Associated Critical Values (Percentiles/Quantities)

> When a sample mean value represented by $\bar{X}$ is estimated from n independent random observations $\{x_1, x_2, \ldots, x_n\}$ sampled from the same distribution with mean value μ and variance σ_X^2, the distribution of
>
> $$Z = \sqrt{n}\left[\frac{\bar{X} - \mu}{\sigma_X}\right]$$
>
> converges to a normal distribution with mean = 0 and variance = 1 as the sample size n increases.

The most important feature of the central limit thereon is what is missing. No mention is made of the properties of the sampled population. The extraordinary usefulness of this theorem comes from the fact that it applies to many kinds of mean values from many situations. Thus, for a sample size as few as 20 or 30 observations, a normal distribution is frequently an accurate description of the distribution of a large variety of summary statistics. Therefore, different kinds of statistical summaries can be evaluated with approximate normal distribution probabilities for moderate sample sizes. In fact, data sampled from symmetric distributions likely produce mean values with close to symmetric distributions and, therefore, are often accurately approximated by a normal distribution for a sample sizes as small as 10. Ratio summaries are similarly evaluated because the logarithm of a ratio frequently produces values with an approximate symmetric distribution. Of note, when values are sampled from a normal distribution, sums, and means of these values have exactly normal distributions for any sample size.

To illustrate the central limit theorem, consider a sample of $n = 10$ independent and randomly sampled values distributed between 0 and 1. One such a sample is

$$\{0.593, 0.726, 0.370, 0.515, 0.378, 0.418, 0.011, 0.532, 0.432 \text{ and } 0.094\}$$

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Excerpt
More information

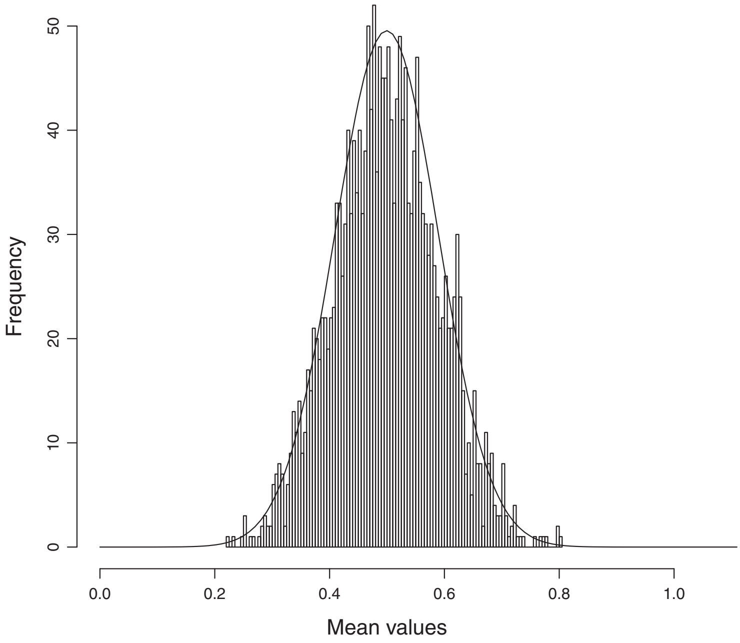

Figure 1.3 An Illustration of the Central Limit Theorem Consisting of 1000 Mean Values Each Estimated from 10 Randomly Sampled Values between 0 and 1 ($n = 10$)

with mean value $\bar{x} = 0.406$. Thus, a distribution of 1000 such mean values each calculated from 10 observations randomly sampled from the same distribution is accurately approximated by a normal distribution with a mean value $\mu = 0.5$ and variance $\sigma^2_{\bar{x}} = \sigma^2_x/n = (1/12)/n = 0.0083$ (Figure 1.3, solid line).

The convergence described by the central limit theorem becomes slower and less accurate as the population sampled becomes less symmetric. Transformations such as the square root or logarithm or other specialized functions of observations produce values with an approximate normal distribution (Chapter 6). These transformations typically make large reductions in extreme values and relatively smaller reductions in small values tending to create a more symmetric distribution. When the distribution of data is extremely asymmetric, the mean value is not a useful summary, and alternatives such as the median value or other statistical measures are more meaningful. Thus, when a mean value is a worthwhile summary, even for modest sample sizes of 20 or 30 observations, it is likely to have at least an approximate normal distribution.

The *t*-Distribution

A probability from a *t*-distribution describes the properties of a test statistic designed to evaluate the sample mean value $\bar{x}$ consisting of n independently sampled observations from a normal distribution with mean μ. Thus, the expression

$$t \text{ statistic} = T = \frac{\bar{x} - \mu}{S_{\bar{x}}} = \sqrt{n}\left[\frac{\bar{x} - \mu}{S_X}\right]$$

CAMBRIDGE

Cambridge University Press
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Excerpt
More information

Table 1.2 *Description: Standard Normal and t-Distribution Critical Values for Five Selected Probabilities (Degrees of Freedom = df =* {2, 10, 20, 40*, and* 60}*)*

Normal distribution		*t*-Distributions				
Probabilities $(1 - \alpha)$	$z_{1-\alpha}$	*df* = 2	*df* = 10	*df* = 20	*df* = 40	*df* = 60
0.990	2.326	6.965	2.764	2.528	2.423	2.390
0.975	1.960	4.303	2.228	2.086	2.021	2.000
0.950	1.645	2.920	1.812	1.725	1.684	1.671
0.900	1.282	1.886	1.372	1.325	1.303	1.296
0.800	0.842	1.061	0.879	0.860	0.851	0.848

has a *t*-distribution (Table 1.2) with a mean value of 0.0. The evaluation of the sample mean value $\bar{x}$ is much like the *Z*-value based on the central limit theorem. The difference is that, unlike the normal distribution test statistic *Z* where the variance is a known value, the variance used in the calculation of the *t* statistic is estimated from the sampled data that generated the mean value (in symbols, $S_{\bar{X}}^2 = S_X^2/n$) (Chapter 27).

The *t*-distribution is defined by a single parameter called the degrees of freedom (denoted *df*) determined by the number of observations used to estimate the mean value and its variance. Thus, a different *t*-distribution exists for every sample size. Figure 1.4 displays the

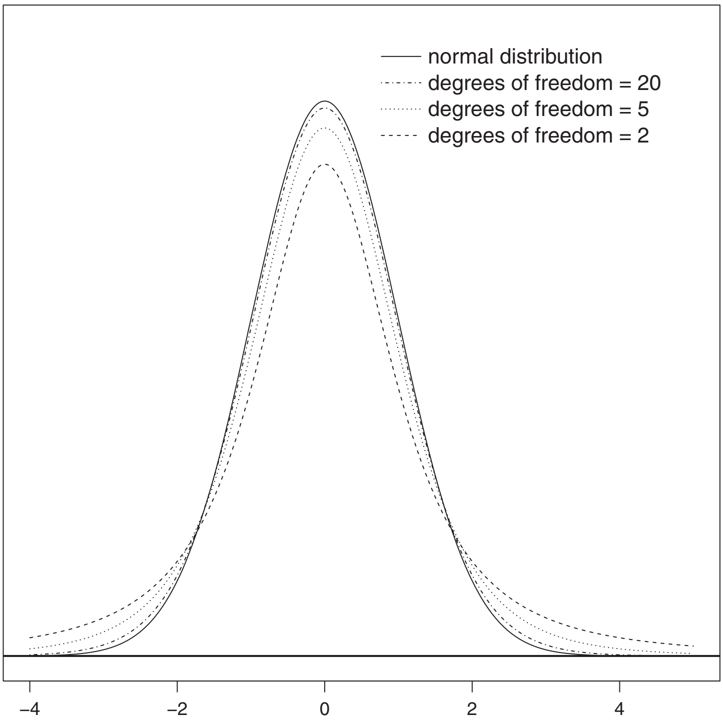

Figure 1.4 Three Representative *t*-Distributions with Degrees of Freedom of 2, 5, and 20 and a Standard Normal Distribution ($\mu = 0$ and $\sigma^2 = 1$)

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Excerpt
More information

standard normal distribution and three t-distributions. Because this statistical distribution accounts for the added variability due to an estimated variance, the cumulative probabilities from the t-distribution are larger than the corresponding cumulative probabilities from a standard normal distribution. For example, this increased variability, for degrees of freedom = 30, causes the modest difference $P(T \geq 2.0) = 0.027$ and $P(Z \geq 2.0) = 0.023$.

The critical values from a standard normal distribution $Z_{1-\alpha}$ are approximately equal to critical values from a t-distribution (denoted $t_{1-\alpha,df}$) for degrees of freedom (df) greater than 30 or so. As the sample size increases, the difference decreases (Figure 1.4). In symbols,

$$Z_{1-\alpha} \approx t_{1-\alpha,df} \quad \text{for } df = \text{degrees of freedom} > 30.$$

Table 1.2 illustrates a few selected t-distribution cumulative probabilities. The mean of the symmetric t-distribution is 0, and the variance is $df/(df-2)$ for $df > 2$. Thus, as the degrees for freedom (sample size) increases, these two parameters more closely correspond to the mean value of 0.0 and variance of 1.0 of the standard normal distribution.

Table 1.2 further indicates that differences between a t-distribution and a normal distribution are small and generally negligible for sample sizes greater than 30 or 40 observations. For a sample size where the degrees of freedom are 60, then $t_{0.95,60} = 1.671$ and $z_{0.95} = 1.645$. Thus, for large sample sizes, ignoring the difference between a t-distribution and normal distribution has negligible effects, as noted in Figure 1.4. From a practical point of view, this similarity indicates that the difference between the estimated variance (S_X^2), and the variance estimated (σ_X^2) becomes unimportant unless the sample is small ($n < 30$).

For small sample sizes, the t-distribution provides an analysis of a mean value when the data are sampled from a normal distribution. The central issue, however, for a small sample size is bias. An error in measuring a single observation or a loss of a single observation, for example, can have considerable influence when the sample size is small. If 10 observations are collected, a single biased or missing value represents 10% of the data. When a student described an experiment based on six observations to statistician R. A. Fisher, he is said to have replied, “You do not have an experiment, you have an experience.” Thus, for small sample sizes, the accuracy of the usually unsupported assumption that the data consist of independent and randomly sampled unbiased observations from a normal distribution becomes critically important. Furthermore, exact statistical analyses exist for small samples sizes that do not depend on the properties of the population sampled (Chapters 8 and 15).

The Chi-Square Probability Distribution

Karl Pearson (circa 1900) introduced the chi-square probability distribution as a way to evaluate a test statistic that combines estimates of variability from different sources into a single statistical summary. His chi-square distribution is defined by the following theorem:

> If $Z_1, Z_2, \ldots, Z_m$ are m independent and normally distributed random variables each with mean value = 0 and variance = 1, then the sum of squared z-values,
>
> $$X^2 = Z_1^2 + Z_2^2 + \cdots + Z_m^2,$$
>
> has a chi-square distribution with m degrees of freedom.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Excerpt
More information

Table 1.3 *Description: A Few Selected Critical Values and Their Cumulative Probabilities from Chi-Square Probability Distributions ($df = \{1, 2, 10, 30, 50,$ and $100\}$)*

	Degrees of freedom (df)					
Probabilities ($1 - \alpha$)	1	2	10	30	50	100
0.990	6.635	9.210	23.209	50.892	76.154	135.807
0.975	5.024	7.378	20.483	46.979	71.420	129.561
0.950	3.841	5.991	18.307	43.773	67.505	124.342
0.900	2.706	4.605	15.987	40.256	63.167	118.498
0.800	1.642	3.219	13.442	36.250	58.164	111.667

Each chi-square distribution is a member of a family of probability distributions identified by an associated degree of freedom (again denoted *df*). The degrees of freedom of a chi-square distribution completely define its location and shape and, therefore, its properties. The degrees of freedom associated with a specific chi-square distribution depend on the number of independent *z*-values in the sum that makes up the chi-square statistic denoted X^2. For most chi-square statistics the values $\{Z_1^2, Z_2^2, \ldots, Z_m^2\}$ are not independent ($df < m$). This lack of independence is dealt with by adjusting the degrees of freedom. Thus, the degrees of freedom are occasionally difficult to determine but are usually part of the description of a specific statistical application or are computer generated with statistical software. Table 1.3 contains a few chi-square critical values (denoted $X^2_{1-\alpha,df}$) and their corresponding cumulative probabilities giving a sense of this probability distribution. For example, the chi-square distribution 90th percentile ($1 - \alpha = 0.90$) value $X^2_{0.90,2}$ is 4.605 when the degrees of freedom are two ($df = 2$).

Thus, for the chi-square variable represented by X^2, the associated cumulative probability is $P(X^2 \leq 4.605) = 0.90$. Figure 1.5 displays four representative chi-square distributions.

The mean value of a chi-square distribution equals the degrees of freedom (*df*), and the variance is 2*df*. To evaluate a chi-square value directly, it is handy to know that a chi-square distributed value less than its mean value (*df*) has a cumulative probability always greater than 0.3 for all chi-square distributions or always $P(X^2 \leq df) > 0.3$.

The essence of a chi-square statistic is that it combines a number of summary values each with a standard normal distribution into a single measure of variability. For example, consider four independent sample mean values $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, and $\bar{x}_4$ each estimated from four samples consisting of n_j normally distributed observations. In addition, the mean values and the variances of each sampled source are the same, represented by μ and σ_X^2. A chi-square comparison of these sample mean values addresses the statistical question: Are the differences among the four sample mean values likely to have occurred by chance alone? If the answer is yes, the test statistic has a chi-square distribution with four degrees of freedom. If the answers is no, it is likely that a larger and less probable test statistic X^2 occurs (Chapter 13). Typically, a significance probability (*p*-value; to be discussed) is useful in choosing between these two alternatives. The probability calculated from a test statistic X^2 with a chi-square probability distribution indicates the likelihood that the observed variability among the mean values occurred by chance alone. In other words, the chi-square distribution summarizes the observed variation relative to a known and fixed population variance (Chapter 13). Like many test statistics, it is a comparison of data to theoretical

© in this web service Cambridge University Press www.cambridge.org