TESTS OF SIGNIFICANCE

10.1 In mathematics, **mean** has several different definitions depending on the context.

In probability and statistics, **mean** and expected value are used synonymously to refer to one measure of the central tendency either of a probability distribution or of the random variablecharacterized by that distribution.[1] In the case of a discrete probability distribution of a random variable *X*, the mean is equal to the sum over every possible value weighted by the probability of that value; that is, it is computed by taking the product of each possible value *x* of *X* and its probability P(*x*), and then adding all these products together.

An analogous formula applies to the case of a continuous probability distribution. Not every probability distribution has a defined mean; see the Cauchy distribution for an example. Moreover, for some distributions the mean is infinite:

For a data set, the terms arithmetic mean, mathematical expectation, and sometimes average are used synonymously to refer to a central value of a discrete set of numbers: specifically, the sum of the values divided by the number of values. The arithmetic mean of a set of numbers x_1, x_2, ..., x_n is typically denoted by $\bar{x}$, pronounced "*x* bar". If the data set were based on a series of observations obtained by sampling from a statistical population, the arithmetic mean is termed the **sample mean** (denoted $\bar{x}$) to distinguish it from the **population mean**.

For a finite population, the **population mean** of a property is equal to the arithmetic mean of the given property while considering every member of the population. For example, the population mean height is equal to the sum of the heights of every individual divided by the total number of individuals. The sample mean may differ from the population mean, especially for small samples. The law of large numbers dictates that the larger the size of the sample, the more likely it is that the sample mean will be close to the population mean.

The **mean** may often be confused with the median, mode or the mid-range. The mean is the arithmetic average of a set of values, or distribution; however, for skewed distributions, the mean is not necessarily the same as the middle value (median), or the most likely (mode). For example, mean income is skewed upwards by a small number of people with very large incomes, so that the majority have an income lower than the mean. By contrast, the median income is the level at which half the population is below and half is above. The mode income is the most likely income, and favors the larger number of people with lower incomes. The median or mode are often more intuitive measures of such data. Nevertheless, many skewed distributions are best described by their mean – such as the exponential and Poisson distributions.

10.2 Goodness of fit: The **goodness of fit** of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. Such measures can be used in statistical hypothesis testing, e.g. to test for normality of residuals, to test whether two samples are drawn from identical distributions (see Kolmogorov–Smirnov test), or whether outcome frequencies follow a specified distribution (see Pearson's chi-squared test). In the analysis of variance, one of the components into which the variance is partitioned may be a lack-of-fit sum of squares.

In assessing whether a given distribution is suited to a data-set, the following tests and their underlying measures of fit can be used:

Kolmogorov–Smirnov test; In statistics, the **Kolmogorov–Smirnov test (K–S test)** is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test). The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. The null distribution of this statistic is calculated under the null hypothesis that the samples are drawn from the same distribution (in the two-sample case) or that the sample is drawn from the reference distribution (in the one-sample case). In each case, the distributions considered under the null hypothesis are continuous distributions but are otherwise unrestricted.

The two-sample K–S test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.

The Kolmogorov–Smirnov test can be modified to serve as a goodness of fit test. In the special case of testing for normality of the distribution, samples are standardized and compared with a standard normal distribution. This is equivalent to setting the mean and variance of the reference distribution equal to the sample estimates, and it is known that using these to define the specific reference distribution changes the null distribution of the test statistic: see below. Various studies have found that, even in this corrected form, the test is less powerful for testing normality than the Shapiro–Wilk test or Anderson–Darling test. However, other tests have their own disadvantages. For instance the Shapiro-Wilk test is known not to work well with many ties (many identical values).

Cramér–von Mises criterion; In statistics the **Cramér–von Mises criterion** is a criterion used for judging the goodness of fit of a cumulative distribution function F^* compared to a given empirical distribution function F_n, or for comparing two empirical distributions. It is also used as a part of other algorithms, such as minimum distance estimation.

Anderson–Darling test; The **Anderson–Darling test** is a statistical test of whether a given sample of data is drawn from a given probability distribution. In its basic form, the test assumes that there are no parameters to be estimated in the distribution being tested, in which case the test and its set of critical values is distribution-free. However, the test is most often used in contexts where a family of distributions is being tested, in which case the parameters of that family need to be estimated and account must be taken of this in adjusting either the test-statistic or its critical values. When applied to testing if a normal distribution adequately describes a set of data, it is one of the most powerful statistical tools for detecting most departures from normality.[1][2] ***K*-sample Anderson–Darling tests** are available for testing whether several collections of observations can be modelled as coming from a single population, where the distribution functiondoes not have to be specified.

In addition to its use as a test of fit for distributions, it can be used in parameter estimation as the basis for a form of minimum distance estimation procedure.

The test is named after Theodore Wilbur Anderson (born 1918) and Donald A. Darling (born 1915), who invented it in 1952.

Shapiro–Wilk test; The **Shapiro–Wilk test** is a test of normality in frequentist statistics. It was published in 1965 by Samuel Sanford Shapiro and Martin Wilk.

Chi Square test; A **chi-squared test**, also referred to as **chi-square test** or χ^2 **test**, is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Also considered a chi-squared test is a test in which this is *asymptotically* true, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a chi-squared distribution as closely as desired by making the sample size large enough. The chi-square (I) test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. Do the number of individuals or objects that fall in each category differ significantly from the number you would expect? Is this difference between the expected and observed due to sampling error, or is it a real difference?

Examples of chi-squared tests

The following are examples of chi-squared tests where the chi-squared distribution is approximately valid:

Pearson's chi-squared test

Pearson's chi-squared test, also known as the chi-squared goodness-of-fit test or chi-squared test for independence. When the chi-squared test is mentioned without any modifiers or without other precluding context, this test is usually meant (for an exact test used in place of χ^2, see Fisher's exact test).

Yates's correction for continuity

Using the chi-squared distribution to interpret Pearson's chi-squared statistic requires one to assume that the discrete probability of observed binomial frequencies in the table can be approximated by the continuous chi-squared distribution. This assumption is not quite correct, and introduces some error.

To reduce the error in approximation, Frank Yates, an English statistician, suggested a correction for continuity that adjusts the formula for Pearson's chi-squared test by subtracting 0.5 from the difference between each observed value and its expected value in a 2 × 2 contingency table.[1] This reduces the chi-squared value obtained and thus increases its p-value.

Other tests

- Cochran–Mantel–Haenszel chi-squared test.
- McNemar's test, used in certain 2 × 2 tables with pairing
- Tukey's test of additivity
- The portmanteau test in time-series analysis, testing for the presence of autocorrelation
- Likelihood-ratio tests in general statistical modelling, for testing whether there is evidence of the need to move from a simple model to a more complicated one (where the simple model is nested within the complicated one).

Exact chi-squared distribution

One case where the distribution of the test statistic is an exact chi-squared distribution is the test that the variance of a normally distributed population has a given value based on a sample variance. Such a test is uncommon in practice because values of variances to test against are seldom known exactly.

Chi-Square Test Requirements

1. Quantitative data. 2. One or more categories. 3. Independent observations. 4. Adequate sample size (at least 10). 5. Simple random sample. 6. Data in frequency form. 7. All observations must be used.

Chi-squared test for variance in a normal population

If a sample of size *n* is taken from a population having a normal distribution, then there is a result (see distribution of the sample variance) which allows a test to be made of whether the variance of the population has a pre-determined value. For example, a manufacturing process might have been in stable condition for a long period, allowing a value for the variance to be determined essentially without error. Suppose that a variant of the process is being tested, giving rise to a small sample of *n* product items whose variation is to be tested. The test statistic *T*in this instance could be set to be the sum of squares about the sample mean, divided by the nominal value for the variance (i.e. the value to be tested as holding).
Then *T* has a chi-squared distribution with $n - 1$ degrees of freedom. For example if the sample size is 21, the acceptance region for *T* for a significance level of 5% is the interval 9.59 to 34.17.

Akaike information criterion; The **Akaike information criterion** (**AIC**) is a measure of the relative quality of a statistical model for a given set of data. As such, AIC provides a means for model selection.

AIC deals with the trade-off between the goodness of fit of the model and the complexity of the model. It is founded on information theory: it offers a relative estimate of the information lost when a given model is used to represent the process that generates the data.

AIC does not provide a test of a model in the sense of testing a null hypothesis; i.e. AIC can tell nothing about the quality of the model in an absolute sense. If all the candidate models fit poorly, AIC will not give any warning of that.

Definition

For any statistical model, the AIC value is

$$AIC = 2k - 2\ln(L)$$

where *k* is the number of parameters in the model, and *L* is the maximized value of the likelihood function for the model.

Given a set of candidate models for the data, *the preferred model is the one with the minimum AIC value.* Hence AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting (increasing the number of parameters in the model almost always improves the goodness of the fit).

AIC is founded in information theory. Suppose that the data is generated by some unknown process f. We consider two candidate models to represent f: g_1 and g_2. If we knew f, then we could find the information lost from using g_1 to represent f by calculating the Kullback–Leibler divergence, $D_{KL}(f \parallel g_1)$; similarly, the information lost from using g_2 to represent f could be found by calculating $D_{KL}(f \parallel g_2)$. We would then choose the candidate model that minimized the information loss.

We cannot choose with certainty, because we do not know f. Akaike (1974) showed, however, that we can estimate, via AIC, how much more (or less) information is lost by g_1 than by g_2. It is remarkable that such a simple formula for AIC results. The estimate, though, is only valid asymptotically; if the number of data points is small, then some correction is often necessary (see AICc, below).

How to apply AIC in practice

To apply AIC in practice, we start with a set of candidate models, and then find the models' corresponding AIC values. There will almost always be information lost due to using one of the candidate models to represent the "true" model (i.e. the process that generates the data). We wish to select, from among R candidate models, the model that minimizes the information loss. We cannot choose with certainty, but we can minimize the estimated information loss.

Denote the AIC values of the candidate models by AIC_1, AIC_2, AIC_3, …, AIC_R. Let $\mathrm{AIC}_{\min}$ be the minimum of those values. Then $\exp((\mathrm{AIC}_{\min}-\mathrm{AIC}_i)/2)$ can be interpreted as the relative probability that the ith model minimizes the (estimated) information loss.[1]

As an example, suppose that there were three models in the candidate set, with AIC values 100, 102, and 110. Then the second model is $\exp((100-102)/2) = 0.368$ times as probable as the first model to minimize the information loss; similarly, the third model is $\exp((100-110)/2) = 0.007$ times as probable as the first model to minimize the information loss.

In this example, we would omit the third model from further consideration. We then have three options: (1) gather more data, in the hope that this will allow clearly distinguishing between the first two models; (2) simply conclude that the data is insufficient to support selecting one model from among the first two; (3) take a weighted average of the first two models, with weights 1 and 0.368, respectively, and then do statistical inference based on the weighted multimodel.[2]

The quantity $\exp((\mathrm{AIC}_{\min}-\mathrm{AIC}_i)/2)$ is the *relative likelihood* of model *i*.

If all the models in the candidate set have the same number of parameters, then using AIC might at first appear to be very similar to using the likelihood-ratio test. There are, however, important distinctions. In particular, the likelihood-ratio test is valid only for nested models whereas AIC (and AICc) has no such restriction.

Hosmer–Lemeshow test; The **Hosmer–Lemeshow test** is a statistical test for goodness of fit for logistic regression models. It is used frequently in risk prediction models. The test assesses whether or not the observed event rates match expected event rates in subgroups of the model population. The Hosmer–Lemeshow test specifically identifies subgroups as the deciles of fitted risk values. Models for which expected and observed event rates in subgroups are similar are called well calibrated.