Recall the Basics of Hypothesis Testing

The level of significance α, (*size of test*) is defined as the probability of X falling in w (rejecting H_0) when H_0 is true: $P(X \in w \,|\, H_0) = \alpha$.

		H_0 TRUE	H_1 TRUE
$X \notin w$	ACCEPT H_0	Acceptance good Prob $= 1 - \alpha$	Contamination Error of the second kind Prob $= \beta$
$X \in w$ (critical region)	REJECT H_0	Loss Error of the first kind Prob $= \alpha$	Rejection good Prob $= 1 - \beta$

Goodness-of-Fit Testing (GOF)

As in hypothesis testing, we are again concerned with the test of a null hypothesis H_0 with a test statistic T, in a critical region w_α, at a significance level α.

Unlike the previous situations, however, the alternative hypothesis, H_1 is now the set of all possible alternatives to H_0. Thus H_1 cannot be formulated, the risk of second kind, β, is unknown, and the power of the test is undefined.

Since it is in general impossible to know whether one test is more powerful than another, the theoretical basis for goodness-of-fit (GOF) testing is much less satisfactory than the basis for classical hypothesis testing.

Nevertheless, GOF testing is quantitatively the most successful area of statistics. In particular, Pearson's venerable Chi-square test is the most heavily used method in all of statistics.

GOF Testing: From the test statistic to the P-value.

Goodness-of-fit tests compare the experimental data with their p.d.f. under the null hypothesis H_0, leading to the statement:

> *If H_0 were true and the experiment were repeated many times, one would obtain data as far away (or further) from H_0 as the observed data with probability P.*

The quantity P is then called the P-value of the test for this data set and hypothesis. A small value of P is taken as evidence against H_0, which the physicist calls a bad fit.

From the test statistic to the P-value.

It is clear from the above that in order to construct a GOF test we need:

1. A test statistic, that is a function of the data and of H_0, which is a measure of the "distance" between the data and the hypothesis, and
2. A way to calculate the probability of exceeding the observed value of the test statistic for H_0. That is, a function to map the value of the test statistic into a P-value.

If the data X are discrete and our test statistic is $t = t(X)$ which takes on the value $t_0 = t(X_0)$ for the data X_0, the P-value would be given by:

$$P_X = \sum_{X:t \geq t_0} P(X|H_0)\,,$$

where the sum is taken over all values of X for which $t(X) \geq t_0$.

Example: Test of Poisson counting rate

Example of discrete counting data:
We have recorded 12 counts in one year, and we wish to know if this is compatible with the theory which predicts $\mu = 17.3$ counts per year.

The obvious test statistic is the absolute difference $|N - \mu|$, and assuming that the probability of n decays is given by the Poisson distribution, we can calculate the P-value by taking the sum in the previous slide.

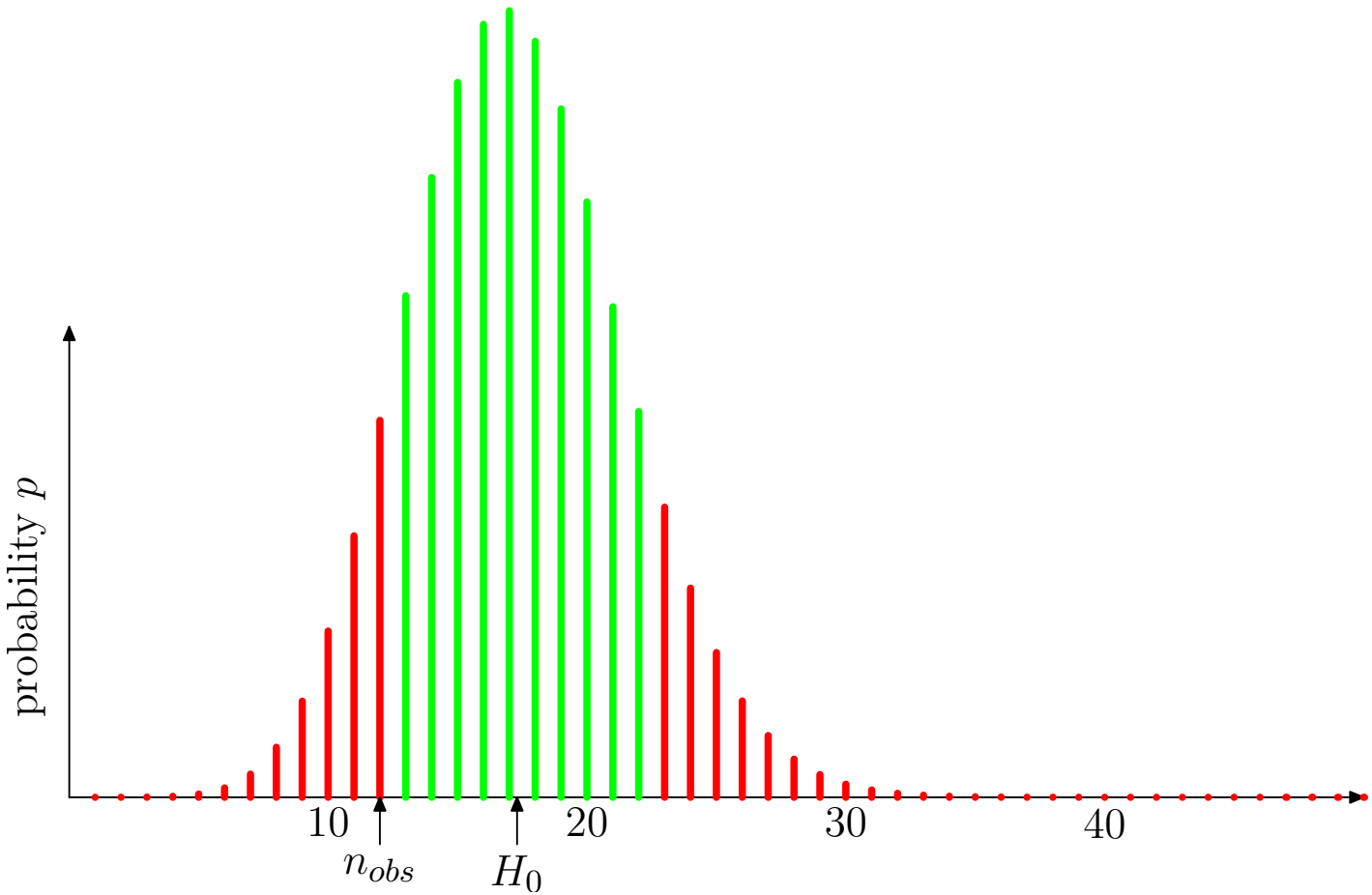
$$P_{12} = \sum_{n:|n-\mu|\geq 5.3} \frac{e^{-\mu}\mu^n}{n!} = \sum_{n=0}^{12} \frac{e^{-17.3}17.3^n}{n!} + \sum_{n=23}^{\infty} \frac{e^{-17.3}17.3^n}{n!}$$

Evaluating the above P-value, we get $P_{12} = 0.229$.

The interpretation is that the observation is not significantly different from the expected value, since one should observe a number of counts at least as far from the expected value about 23% of the time.

Poisson Test Example seen visually

The length of the vertical bars is proportional to the Poisson probability of n (the x-axis) for $H_0 : \mu = 17.3$.



Distribution-free Tests

When the data are continuous, the sum becomes an integral:

$$P_X = \int_{X:t>t_0} P(X|H_0)\,, \tag{1}$$

and this can become quite complicated to compute, so that one tries to avoid using this form. Instead, one looks for a test statistic such that the distribution of t is known independently of H_0.

Such a test is called a distribution-free test. We consider here mainly distribution-free tests, such that the P-value does not depend on the details of the hypothesis H_0, but only on the value of t, and possibly one or two integers such as the number of events, the number of bins in a histogram, or the number of constraints in a fit.

Then the mapping from t to P-value can be calculated once for all and published in tables, of which the well-known χ^2 tables are an example.

Pearson's Chi-square Test

An obvious way to measure the distance between the data and the hypothesis H_0 is to

1. Determine the expectation of the data under the hypothesis H_0 .
2. Find a metric in the space of the data to measure the distance of the observed data from its expectation under H_0 .

When the data consists of measurements $\mathbf{Y} = Y_1, Y_2, \ldots, Y_k$ of quantities which, under H_0 are equal to $\mathbf{f} = f_1, f_2, \ldots, f_k$ with covariance matrix $\underset{\sim}{V}$, the distance between the data and H_0 is clearly:

$$T = (\mathbf{Y} - \mathbf{f})^T \, \underset{\sim}{V}^{-1} \, (\mathbf{Y} - \mathbf{f})$$

This is just the Pearson test statistic usually called chi-square, because it is distributed as $\chi^2(k)$ under H_0 if the measurements $\mathbf{Y}$ are Gaussian-distributed. That means the P-value may be found from a table of $\chi^2(k)$, or by calling PROB(T,k).

Pearson's Chi-square test for histograms

Karl Pearson made use of the asymptotic Normality of a multinomial p.d.f. in order to find the (asymptotic) distribution of:

$$T = (\mathbf{n} - N\mathbf{p})^T \, \underset{\sim}{V}^{-1} \, (\mathbf{n} - N\mathbf{p})$$

where $\underset{\sim}{V}$ is the covariance matrix of the observations (bin contents) $\mathbf{n}$ and N is the total number of events in the histogram.

In the usual case where the bins are independent, we have

$$T = \frac{1}{N} \sum_{i=1}^{k} \frac{(n_i - N p_i)^2}{p_i} = \frac{1}{N} \sum_{i=1}^{k} \frac{n_i^2}{p_i} - N \, .$$

This is the usual χ^2 goodness-of-fit test for histograms. The distribution of T is generally accepted as close enough to $\chi^2(k-1)$ when all the expected numbers of events per bin (Np_i) are greater than 5. Cochran relaxes this restriction, claiming the approximation to be good if not more than 20% of the bins have expectations between 1 and 5.

Chi-square test with estimation of parameters

If the parent distribution depends on a vector of parameters $\boldsymbol{\theta} = \theta_1, \theta_2, \ldots, \theta_r$, to be estimated from the data, one does not expect the T statistic to behave as a $\chi^2(k-1)$, except in the limiting case where the r parameters do not actually affect the goodness of fit.

In the more general and usual case, one can show that when r parameters are estimated from the same data, the cumulative distribution of T is intermediate between

$\chi^2(k-1)$ (which holds when θ is fixed) and
$\chi^2(k-r-1)$ (which holds for an optimal estimation method),
always assuming the null hypothesis.

The test is no longer distribution-free, but when k is large and r small, it is often nearly distribution-free.

In practice, the r parameters will be well chosen, and T will usually behave as $\chi^2(k-r-1)$. This is called Wilks' Theorem.

Binned Likelihood (Likelihood Chi-square)

Pearson's Chi-square is a good test statistic when fitting a curve to points with Gaussian errors,

but for fitting histograms,
we can make use of the known distribution of events in a bin,
which is not exactly Gaussian:

- It is Poisson-distributed if the bin contents are independent (no constraint on the total number of events).
- Or it is Multinomial-distributed if the total number of events in the histogram is fixed.

reference: Baker and Cousins, *Clarification of the Use of Chi-square and Likelihood Functions in Fits to Histograms* NIM 221 (1984) 437

Our test statistic will be the binned likelihood, which Baker and Cousins called the likelihood chi-square because it behaves as a χ^2, although it is derived as a likelihood ratio.

Binned Likelihood for Poisson bins

For bin contents n_i that are Poisson-distributed with mean μ_i, we have:

$$L = \prod_{bins} e^{-\mu_i} \mu_i^{n_i} / n_i!$$

$$-2 \ln L = 2 \sum_i [\mu_i - n_i \ln \mu_i + \ln n_i!]$$

Now define L_0 as $L(n_i = \mu_i)$, the likelihood for a perfect fit:

$$-2 \ln L_0 = 2 \sum_i [n_i - n_i \ln n_i + \ln n_i!]$$

Now we subtract the last two equations above to get rid of the nasty term in $n!$ and this gives the log of the likelihood ratio L/L_0 :

$$\text{Poisson} \quad \chi^2_\lambda = -2 \ln(L/L_0) = 2 \sum_i [\mu_i(\theta) - n_i + n_i \ln(n_i/\mu_i(\theta))]$$

where the last term is defined as =0 when $n_i = 0$.

Binned Likelihood for Multinomial bins

For bin contents n_i that are Multinomial-distributed with mean μ_i, we have:

$$L = N! \prod_{bins} (\mu_i/N)^{n_i}/n_i!$$

$$-2 \ln L = -2[\ln N! - k \ln N + \sum_i (n_i \ln \mu_i - \ln n_i!)]$$

As before, we define L_0 as $L(n_i = \mu_i)$, the likelihood for a perfect fit, and taking the ratio L/L_0 only one term remains, so we obtain:

$$\text{multinomial} \quad \chi^2_\lambda = -2 \ln(L/L_0) = 2 \sum_i [n_i \ln(n_i/\mu_i(\theta))]$$

where the terms in the sum are defined as $=0$ when $n_i = 0$.

Note that the multinomial form assumes that the μ_i obey the constraint $\sum_i \mu_i = \sum_i n_i = N$, whereas with the Poisson form the fitted values of μ_i will automatically satisfy this constraint.

Binned Likelihood as a GOF Test

Using the more common Poisson form,

$$t = -2 \ln \lambda = 2 \sum_i [\mu_i(\theta) - n_i + n_i \ln(n_i/\mu_i(\theta))]$$

asymptotically obeys a Chi-square distribution, with number of degrees of freedom equal to the number of bins minus one. It is therefore a GOF test.

Faced with the enormous popularity of Pearson's T-statistic, binned likelihood is not used as much as it should be, so practical experience is somewhat limited, but all indications are that it is superior to Pearson's T for histograms, both for parameter fitting and for GOF testing.

If we make the bins more numerous and narrower, the efficiency as an estimator improves, but the power of the GOF test at some point degrades. In the limit of infinitely narrow bins, all n_i are either zero or one, and the binned likelihood tends to the unbinned likelihood (not good for GOF).

Runs test

A drawback of the T statistic is that the signs of the deviations $(n_i - Np_i)$ are lost. The runs test is based on the signs of the deviations. The main interest of this test is that, *for simple hypotheses*, it is independent of the χ^2 test on the same bins and thus brings in new information.
Under hypothesis H_0, all patterns of signs are equally probable. This simple fact allows us to write the following results [Wilks, p. 154]. Let M be the number of positive deviations, N the number of negative deviations, and R the total number of runs, where a *run* is a sequence of deviations of the same sign, preceded and followed by a deviation of opposite sign (unless at the end of the range of the variable studied). Then

$$P(R = 2s) = \frac{2\binom{M-1}{s-1}\binom{N-1}{s-1}}{\binom{M+N}{M}}$$

$$P(R = 2s-1) = \frac{\binom{M-1}{s-2}\binom{N-1}{s-1} + \binom{M-1}{s-1}\binom{N-1}{s-2}}{\binom{M+N}{M}}$$

Runs test

The critical region is defined as improbably low values of $R : R \leq R_{\text{min}}$. Given the probability of R, one can compute R_{min} corresponding to the significance level required. The expectation and variance of R are

$$E(R) = 1 + \frac{2MN}{M+N}$$

$$V(R) = \frac{2MN(2MN - M - N)}{(M+N)^2(M+N-1)}.$$

Although the runs test is usually not as powerful as Pearson's χ^2 test, it is (asymptotically) independent of it and hence the two can be combined to produce an especially important test (see below, Combining Independent Tests).

Binned and Unbinned Data

Binned Data

By combining events into histogram bins (called data classes in the statistical literature), some information is lost: the position of each event inside the bin. The loss of information may be negligible if the bin width is small compared with the experimental resolution, but in general one must expect tests on binned data to be inferior to tests on individual events.

Tests on Unbinned Data

Unfortunately, the requirement of distribution-free tests restricts the choice of tests for unbinned data, and we will consider only those based on the order statistics (or empirical distribution function).

Since order statistics can be defined only in one dimension, this limits us to data depending on only one random variable, and to simple hypotheses H_0 (no free parameters $\boldsymbol{\theta}$).

Order statistics

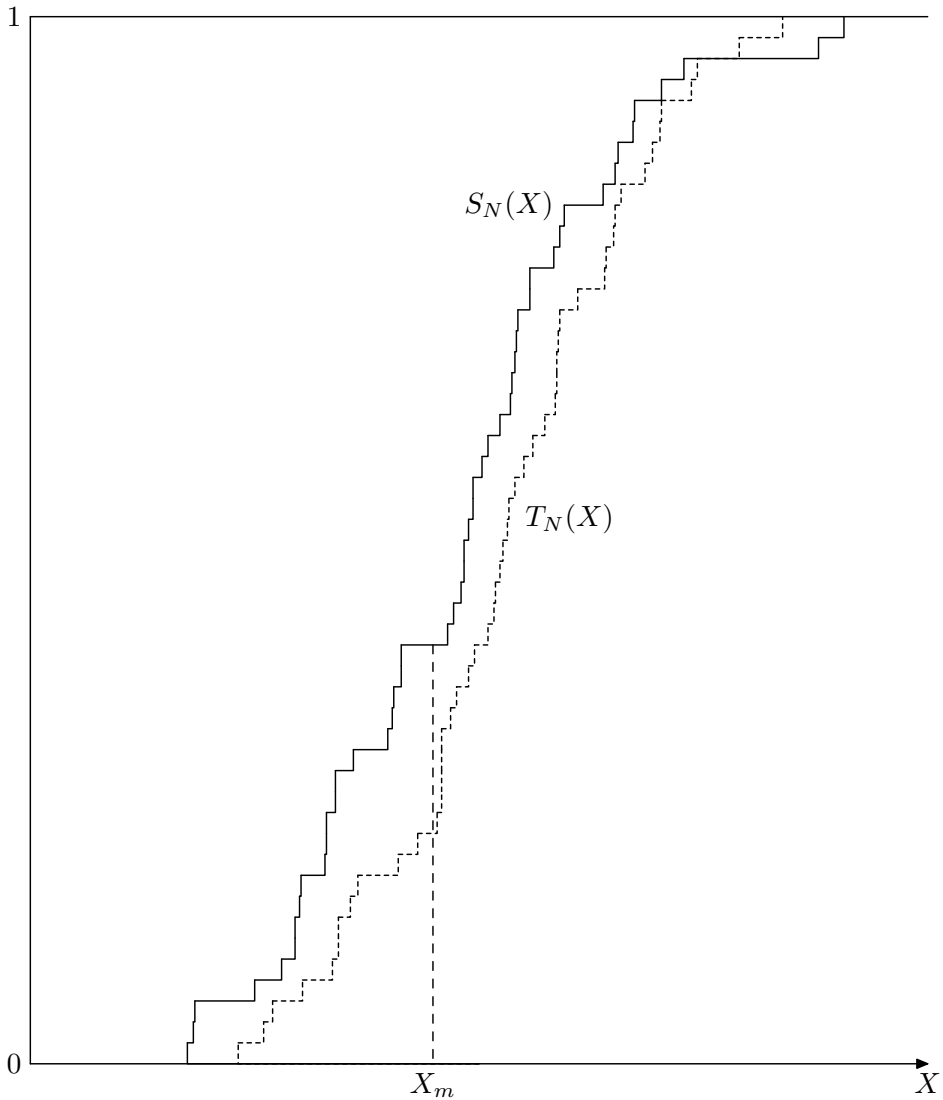
Given N independent observations $X_1, \ldots, X_N$ of the random variable X, let us reorder the observations in ascending order, so that $X_{(1)} \leq X_{(2)} \leq, \ldots, \leq X_{(N)}$ (this is always permissible since the observations are independent).

The ordered observations $X_{(i)}$ are called the order statistics. Their cumulative distribution is called the empirical distribution function or EDF.

$$S_N(X) = \begin{cases} 0 & X < X_{(1)} \\ i/N \quad \text{for} & X_{(i)} \leq X < X_{(i+1)}, \quad i = 1, \ldots, \quad N-1\,. \\ 1 & X_{(N)} \leq X \end{cases}$$

Note that $S_N(X)$ always increases in steps of equal height, N^{-1}.

Order statistics



Example of two cumulative distributions, $S_N(X)$ and $T_N(X)$

For these two data sets, the maximum distance $S_N - T_N$ occurs at $X = X_m$.

We shall consider different norms on the difference $S_N(X) - F(X)$ as test statistics.

Smirnov - Cramér - von Mises test

Consider the statistic

$$W^2 = \int_{-\infty}^{\infty} [S_N(X) - F(X)]^2 f(X) dX \,,$$

where $f(X)$ is the p.d.f. corresponding to the hypothesis H_0, $F(X)$ is the cumulative distribution, and $S_N(X)$ is defined as above, which gives

$$\begin{aligned} W^2 &= \int_{-\infty}^{X_1} F^2(X) dF(X) + \sum_{i=1}^{N-1} \int_{X_i}^{X_{i+1}} \left[\frac{i}{N} - F(X)\right]^2 dF(X) \\ &\quad + \int_{X_N}^{\infty} [1 - F(X)]^2 dF(X) \\ &= \frac{1}{N}\left\{\frac{1}{12N} + \sum_{i=1}^{N}\left[F(X_i) - \frac{2i-1}{2N}\right]^2\right\} , \end{aligned}$$

using the properties $F(-\infty) \equiv 0$, $F(+\infty) \equiv 1$.

Smirnov - Cramér - von Mises test

The Smirnov-Cramér-von Mises test statistic W^2 has mean and variance

$$E(W^2) = \frac{1}{N}\int_0^1 F(1-F)dF = \frac{1}{6N}$$
$$V(W^2) = E(W^4) - [E(W^2)]^2 = \frac{4N-3}{180N^3}\,.$$

Smirnov has calculated the critical values of NW^2

Test size α	Critical value of NW^2
0.10	0.347
0.05	0.461
0.01	0.743
0.001	1.168

It has been shown that, to the accuracy of this table, the asymptotic limit is reached when $N \geq 3$.

Smirnov - Cramér - von Mises test

When H_0 is composite, W^2 is not in general distribution-free. When X is many-dimensional, the test also fails, unless the components are independent. However, one can form a test to compare two distributions, $F(X)$ and $G(X)$. Let the number of observations be N and M, respectively, and let the hypothesis be H_0: $F(X) = G(X)$. Then the test statistic is

$$W^2 = \int_{-\infty}^{\infty} [S_N(X) - S_M(X)]^2 d\left[\frac{NF(X) + MG(X)}{N + M}\right].$$

Then the quantity

$$\frac{MN}{M + N} W^2$$

has the critical values shown in the table above.

Kolmogorov test

The test statistic is now the maximum deviation of the observed distribution $S_N(X)$ from the distribution $F(X)$ expected under H_0. This is defined either as

$$D_N = \max \ |S_N(X) - F(X)| \qquad \text{for all } X$$

or as

$$D_N^{\pm} = \max \{\pm[S_N(X) - F(X)]\} \quad \text{for all } X\,,$$

when one is considering only one-sided tests. It can be shown that the limiting distribution of $\sqrt{N}D_N$ is

$$\lim_{N\to\infty} P(\sqrt{N}D_N > z) = 2\sum_{r=1}^{\infty}(-1)^{r-1}\exp(-2r^2z^2)$$

and that of $\sqrt{N}D_N^{\pm}$ is

$$\lim_{N\to\infty} P(\sqrt{N}D_N^{\pm} > z) = \exp(-2z^2)\,.$$

Kolmogorov Test

Alternatively, the probability statement above can be restated as

$$\lim_{N\to\infty} P[2N(D_N^{\pm})^2 \leq 2z] = 1 - e^{-2z^2}.$$

Thus $4N(D_N^{\pm})^2$ have a $\chi^2(2)$ distribution.
The limiting distributions are considered valid for $N \approx 80$.

We give some critical values of $\sqrt{N}D_N$.

Test size α	Critical value of $\sqrt{N}D_N$
0.01	1.63
0.05	1.36
0.10	1.22
0.20	1.07

Two-Sample Kolmogorov Test

The equivalent statistic for comparing two distributions $S_N(X)$ and $S_M(X)$ is

$$D_{MN} = \max |S_N(X) - S_M(X)| \qquad \text{for all } X$$

or, for one-sided tests

$$D^{\pm}_{MN} = \max \{\pm[S_N(X) - S_M(X)]\} \qquad \text{for all } X\,.$$

Then $\sqrt{MN/(M+N)}D_{MN}$ has the limiting distribution of $\sqrt{N}D_N$ and $\sqrt{MN/(M+N)}D^{\pm}_{MN}$ have the limiting distribution of $\sqrt{N}D^{\pm}_N$.

Kolmogorov Test

Finally, one may invert the probability statement about D_N to set up a confidence belt for $F(X)$. The statement

$$P\{D_N = \max|S_N(X) - F(X)| > d_\alpha\} = \alpha$$

defines d_α as the α-point of D_N. If follows that

$$P\{S_N(X) - d_\alpha \leq F(X) \leq S_N(X) + d_\alpha\} = 1 - \alpha\,.$$

Therefore, setting up a belt $\pm d_\alpha$ about $(S_N(X)$, the probability that $F(X)$ is entirely within the belt is $1 - \alpha$ (similarly d_α can be used to set up one-sided bounds). One can thus compute the number of observations necessary to obtain $F(X)$ to any accuracy. Suppose for example that one wants $F(X)$ to precision 0.05 with probability 0.99, then one needs $N = (1.628/0.05)^2 \sim 1000$ observations.

Better tests using the EDF

Users of the Kolmogorov test will probably notice that the maximum difference D_N or D_{MN} almost always occurs near the middle of the range of X. And it is constrained to be zero at the two end points.

This has led Anderson and Darling to propose an improved test which gives more weight to the ends of the range. The Anderson–Darling test is an example of a definite improvement on the Kolmogorov test, but one which comes at the expense of losing the distribution-free property.

This means that the P-value must be computed differently depending on whether one is testing for Normally distributed data or uniformly distributed data, for example.

Tests of this kind are outside the scope of this course; the reader is referred to the book of D'Agostino and Stephens (eds.).

The likelihood function is not a good test statistic (1)

Suppose that the N observations $\mathbf{X}$ have p.d.f. $f(\mathbf{X})$ and log-likelihood function

$$\lambda = \sum_{i=1}^{N} \ln f(X_i)\,.$$

One can, in principle, compute the expectation and the variance of λ,

$$E_{\mathbf{X}}(\lambda) = N \int \ln f(\mathbf{X}) \cdot f(\mathbf{X}) d\mathbf{X}\,,$$

$$V_{\mathbf{X}}(\lambda) = N \int [\ln f(\mathbf{X}) - N^{-1} E_{\mathbf{X}}(\lambda)]^2 f(\mathbf{X}) d\mathbf{X}\,,$$

and even higher moments, if one feels that the Normality assumption is not good enough. We can therefore convert the value of λ into a P-value.

The likelihood function is not a good test statistic (2)

At first glance, we might expect the value of the likelihood to be a good measure of GOF, since we know the maximum of the likelihood gives the best estimates of parameters.

But in m.l. estimation, we are using the likelihood for fixed data as a function of the parameters in the hypothesis, whereas in GOF testing we use the likelihood for a fixed hypothesis as a function of the data, which is very different.

And indeed, it does not make a good GOF test statistic. This can be seen in different ways, but the first clue should come when we see that the likelihood does not measure the "distance" between the data and the hypothesis.

In addition, the value of the likelihood function is most certainly NOT distribution-free.

The likelihood function is not a good test statistic (3)

Suppose for example that the hypothesis under test H_0 is the uniform distribution (which can always be arranged by a simple coordinate transformation). For this hypothesis, it is easily seen that the likelihood has no power at all as a GOF test statistic, since all data sets (with the same number of events) have exactly the same value of the likelihood function, no matter how well they fit the hypothesis of uniformity.

More extensive studies show a variety of examples where the value of the likelihood function has no power as a GOF statistic and no examples where it can be recommended.
Joel Heinrich (2001) has written a report on that: CDF memo 5639.

Combining Independent Tests

It may happen that two or more different tests may be applied to the same data, or the same test applied to different sets of data, in such a way that although no one test is significant by itself, the combination of tests becomes significant.

When using a combined test, one must of course know the properties of the individual tests involved, and in addition two new problems arise:

(a) establishing that the individual tests are independent

(b) finding the best test statistic for the combined test

(c) finding the significance level of the combined test.

Combining Independent Tests, continued

1. Different tests on the same data set

Of all the usual distribution-free tests, only the Pearson χ^2 test and the runs test are (asymptotically) independent [Kendall, II, p. 442]. Intuitively this is clear, since Pearson's test does not depend on the ordering of the bins or on the signs of the deviations in each bin, while this is the only information used in the runs test. In fact, Pearson's test, although probably the most generally used test of fit, has been criticized for its lack of power precisely because it does not take account of this information.

2. The same test on different data sets

Different data sets, even from the same experiment, are in general independent, so the same test can be applied to all the data sets, and the tests will be independent.

Test statistic for the combined test

Suppose that two independent tests yield individual p-values p_1 and p_2. The obvious test statistic for the combined test is $t = p_1 p_2$.

It turns out that if nothing more is known (only the two p-values) it is not possible to find the optimal test statistic, but $t = p_1 p_2$ is a reasonable choice giving equal weight to the two component tests. more on this later
It might be supposed that the p-value of the combined test would be $p_{12} = p_1 p_2$, but it is easily seen that if p_1 and p_2 are both uniformly distributed between zero and one, then the product $p_1 p_2$ would not be uniform on the same interval.

It can be shown that, with the test statistic $t = p_1 p_2$, the p-value is

$$p_{12} = t[1 - \ln(t)]$$

which is larger than t.

P-value of the combined test

A better way to combine tests can be seen by considering how to combine the results of two Chi-square tests.
Suppose we have applied Pearson's test to two independent data sets, with the results

$$\chi_1^2(n_1) \to p_1 \qquad \text{and} \quad \chi_2^2(n_2) \to p_2$$

Clearly the combined test consists of adding the two χ^2 and finding the p-value of that sum for $n_1 + n_2$ degrees of freedom.

This suggests that the general way to combine tests with p-values p_1 and p_2 is to convert the two p-values to χ^2 and add the values of χ^2.

But for how many degrees of freedom?
That depends on the relative "weight" of the two tests, which may not be specified, in which case the solution to the problem is not unique.