Goodness-of-Fit – Pitfalls and Power
Example: Testing Consistency of Two Histograms

Sometimes we have two histograms and are faced with the question: “Are they consistent?”

That is, are our two histograms consistent with having been sampled from the same parent distribution?

Each histogram represents a sampling from a multivariate Poisson distribution.
Testing Consistency of Two Histograms

There are two variants of interest to this question:

1. We wish to test the hypothesis (absolute equality):

   \[ H_0: \text{The means of the two histograms are bin-by-bin equal}, \text{ against} \]

   \[ H_1: \text{The means of the two histograms are not bin-by-bin equal}. \]

2. We wish to test the hypothesis (shape equality):

   \[ H'_0: \text{The densities of the two histograms are bin-by-bin equal}, \text{ against} \]

   \[ H'_1: \text{The densities of the two histograms are not bin-by-bin equal}. \]
Testing Consistency of Two Histograms – Some Context

There are many tests to address whether a dataset is consistent with having been drawn from some continuous distribution.

- Histograms are built from such data. So why not work directly with “the data”? Good question!
- Maybe original data not readily available.
- May be able to use features of binned data, eg, normal approximation ⇒ $\chi^2$.

These tests may often be adapted to address whether two datasets have been drawn from the same continuous distribution, often called “two-sample” tests.

These tests may then be further adapted to the present problem, of determining whether two histograms have the same shape. Also discussed as comparing whether two (or more) rows of a “table” are consistent.
Notation, Conventions

- Assume both histograms have $k$ bins, with identical bin boundaries.
- The sampling distributions are:

$$P(U = u) = \prod_{i=1}^{k} \frac{\mu_i^{u_i} e^{-\mu_i}}{u_i!},$$

$$P(V = v) = \prod_{i=1}^{k} \frac{\nu_i^{v_i} e^{-\nu_i}}{v_i!},$$

- Define:

$$N_u \equiv \sum_{i=1}^{k} U_i, \quad \mu_T \equiv \langle N_u \rangle = \sum_{i=1}^{k} \mu_i;$$

$$N_v \equiv \sum_{i=1}^{k} V_i, \quad \nu_T \equiv \langle N_v \rangle = \sum_{i=1}^{k} \nu_i;$$

$$t_i \equiv u_i + v_i, \quad i = 1, \ldots, k$$

Will drop distinction between random variable and realization.
Power, Confidence Level, \( P \)-value

Interested in the power of a test, for given confidence level. The power is the probability that the null hypothesis is rejected when it is false. Power depends on the true sampling distribution.

\[
\text{Power} \equiv 1 - \text{probability of a Type II error.}
\]

The confidence level is the probability that the null hypothesis is accepted, if the null hypothesis is correct.

\[
\text{Confidence level} \equiv 1 - \text{probability of a Type I error.}
\]

In physics, we don’t specify the confidence level of a test in advance, at least not formally. Instead, we quote the \( P \)-value for our result. This is the probability, under the null hypothesis, of obtaining a result as “bad” or worse than our observed value. This would be the probability of a Type I error if our observation were used to define the critical region of the test.
Large Statistics Case

If all of the bin contents of both histograms are large, we use the approximation that the bin contents are normally distributed.

Under $H_0$, 

$$\langle u_i \rangle = \langle v_i \rangle \equiv \mu_i, \quad i = 1, \ldots, k.$$ 

More properly, it is $\langle U_i \rangle = \mu_i$, etc., but we are permitting $u_i$ to stand for the random variable as well as its realization. Let the difference for the contents of bin $i$ between the two histograms be: 

$$\Delta_i \equiv u_i - v_i,$$

and let the standard deviation for $\Delta_i$ be denoted $\sigma_i$. Then the sampling distribution of the difference between the two histograms is:

$$P(\Delta) = \frac{1}{(2\pi)^{k/2}} \left( \prod_{i=1}^{k} \frac{1}{\sigma_i} \right) \exp \left( -\frac{1}{2} \sum_{i=1}^{k} \frac{\Delta_i^2}{\sigma_i^2} \right).$$
Large Statistics Case – Test Statistic

This suggests the test statistic:

\[ T = \sum_{i=1}^{k} \frac{\Delta_i^2}{\sigma_i^2}. \]

If the \( \sigma_i \) were known, this would simply be distributed according to the chi-square distribution with \( k \) degrees of freedom.

The maximum-likelihood estimator for the mean of a Poisson is just the sampled number. The mean of the Poisson is also its variance, and we will use the sampled number also as the estimate of the variance in the normal approximation.

We’ll refer to this approach as a “\( \chi^2 \)” test.
Large Statistics Algorithm – absolute

We suggest the following algorithm for this test:

1. For $\sigma_i^2$ form the estimate

$$\hat{\sigma}_i^2 = (u_i + v_i).$$

2. Statistic $T$ is thus evaluated according to:

$$T = \sum_{i=1}^{k} \frac{(u_i - v_i)^2}{u_i + v_i}.$$

If $u_i = v_i = 0$ for bin $i$, the contribution to the sum from that bin is zero.

3. Estimate the $P$-value according to a chi-square with $k$ degrees of freedom. Note that this is not an exact result.
Large Statistics Algorithm – shapes

If only comparing shapes, then scale both histogram bin contents:

1. Let
   \[ N = 0.5(N_u + N_v). \]

   Scale \( u \) and \( v \) according to:
   \[ u_i \rightarrow u_i' = u_i(N/N_u) \]
   \[ v_i \rightarrow v_i' = v_i(N/N_v). \]

2. Estimate \( \sigma_i^2 \) with:
   \[ \hat{\sigma}_i^2 = \left( \frac{N}{N_u} \right) u_i + \left( \frac{N}{N_v} \right) v_i. \]

3. Statistic \( T \) is thus evaluated according to:
   \[ T = \sum_{i=1}^{k} \left( \frac{u_i}{N_u} - \frac{v_i}{N_v} \right)^2 \frac{u_i}{N_u^2} + \frac{v_i}{N_v^2}. \]

4. Estimate the \( P \)-value according to a chi-square with \( k - 1 \) degrees of freedom. Note that this is not an exact result.
Application to Example

With some small bin counts, might not expect this method to be especially good for our example, but can try it anyway:

<table>
<thead>
<tr>
<th>Type of test</th>
<th>$T$</th>
<th>NDOF</th>
<th>$P(\chi^2 &gt; T)$</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$ absolute comparison</td>
<td>53.4</td>
<td>36</td>
<td>0.031</td>
<td>0.023</td>
</tr>
<tr>
<td>$\chi^2$ shape comparison</td>
<td>38.5</td>
<td>35</td>
<td>0.31</td>
<td>0.28</td>
</tr>
<tr>
<td>Likelihood Ratio shape comparison</td>
<td>38.7</td>
<td>35</td>
<td>0.31</td>
<td>0.35</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov shape comparison</td>
<td>0.050</td>
<td>35</td>
<td>NA</td>
<td>0.22</td>
</tr>
<tr>
<td>Bhattacharyya shape comparison</td>
<td>0.986</td>
<td>35</td>
<td>NA</td>
<td>0.32</td>
</tr>
<tr>
<td>Cramér-Von-Mises shape comparison</td>
<td>0.193</td>
<td>35</td>
<td>NA</td>
<td>0.29</td>
</tr>
<tr>
<td>Anderson-Darling shape comparison</td>
<td>0.944</td>
<td>35</td>
<td>NA</td>
<td>0.42</td>
</tr>
<tr>
<td>Likelihood value shape comparison</td>
<td>90</td>
<td>35</td>
<td>NA</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Column “$P$-value” attempts a more reliable estimate of the probability, under the null hypothesis, that a value for $T$ will be as large as observed. Compare with column $P(\chi^2 > T)$ column, the probability assuming $T$ follows a $\chi^2$ distribution with NDOF degrees of freedom.
The absolute comparison yields much poorer agreement than the shape comparison.

This is readily understood:

– The total number of counts in one dataset is 783; in the other it is 633.
– Treating these as samplings from a normal distribution with variances 783 and 633, we find a difference of 3.99 standard deviations or a two-tailed $P$-value of $6.6 \times 10^{-5}$.
– This low probability is diluted by the bin-by-bin test to 0.023.
– **Lesson:** The more you know about what you want to test, the better (more powerful) the test you can make.
An Issue: We don’t know $H_0$!

Evaluation of the probability under the null hypothesis is in fact problematic, since the null hypothesis actually isn’t completely specified.

- The problem is the dependence of Poisson probabilities on the absolute numbers of counts. Probabilities for differences in Poisson counts are not invariant under the total number of counts.

- Unfortunately, we don’t know the true mean numbers of counts in each bin under $H_0$. Thus, we must estimate these means.

- The procedure adopted here has been to use the maximum likelihood estimators (see later) for the mean numbers, in the null hypothesis.

We’ll have further discussion of this procedure below – It does not always yield valid results.
Use of chi-square Probability Distribution

In our example, the probabilities estimated according to our simulation and the $\chi^2$ distribution probabilities are fairly close to each other. Suggests the possibility of using the $\chi^2$ probabilities – if we can do this, the problem that we haven’t completely specified the null hypothesis is avoided.

Conjecture: Let $T$ be the “$\chi^2$” test statistic, for either the absolute or the shape comparison, as desired. Let $T_c$ be a possible value of $T$ (perhaps the critical value to be used in a hypothesis test). Then, for large values of $T_c$, under $H_0$ (or $H'_0$):

$$P(T < T_c) \geq P \left( T < T_c | \chi^2(T, \text{ndof}) \right),$$

where $P \left( T < T_c | \chi^2(T, \text{ndof}) \right)$ is the probability that $T < T_c$ according to a $\chi^2$ distribution with ndof degrees of freedom (either $k$ or $k - 1$, depending on which test is being performed).
Is it Useful?

The conjecture tells us that if we use the probabilities from a $\chi^2$ distribution in our test, the error that we make is in the “conservative” direction.

- That is, we’ll reject the null hypothesis less often than we would with the correct probability.

- This conjecture is independent of the statistics of the sample, bins with zero counts are fine. In the limit of large statistics, the inequality approaches equality.

Unfortunately, it isn’t as nice as it sounds. The problem is that, in low statistics situations, the power of the test according to this approach can be dismal. We might not reject the null hypothesis in situations where it is obviously implausible.
Comparison of actual (cumulative) probability distribution for $T$ (solid blue curves) with chi-square distribution (dashed red curves). All plots are for 100 bin histograms, and null hypothesis:
(a) Each bin has mean 100.
(b) Each bin has mean 1.
(c) Bin $j$ has mean $30/j$. 
General (Including Small Statistics) Case

If the bin contents are not large, then the normal approximation may not be good enough and the "\( \chi^2 \) statistic" may not follow a \( \chi^2 \) distribution.

A simple approach is to combine bins until the normal approximation is valid. In many cases this doesn’t lose too much statistical power.

A common rule-of-thumb is to combine until each bin has at least 7 counts.

Try this on our example, as a function of the minimum number of events per bin. The algorithm is to combine corresponding bins in both histograms until both have at least "\texttt{minBin}" counts in each bin.
Combining bins for $\chi^2$ test

Left: The example pair of histograms.

Middle: The solid curve shows the value of the test statistic $T$, and the dashed curve shows the number of histogram bins for the data in the example, as a function of the minimum number of counts per bin.

Right: The $P$-value for consistency of the two datasets in the example. The solid curve shows the probability for a chi-square distribution, and the dashed curved shows the probability for the actual distribution, with an estimated null hypothesis.
Working with the Poissons - Normalization Test

Alternative: Work with the Poisson distribution. Separate the problem of the shape from that of the overall normalization.

To test normalization, compare totals over all bins between the histograms. Distribution is

\[ P(N_u, N_v) = \frac{\mu_T^{N_u} \nu_T^{N_v}}{N_u! N_v!} e^{-(\mu_T + \nu_T)}. \]

The null hypothesis is \( H_0 : \mu_T = \nu_T \), to be tested against alternative \( H_1 : \mu_T \neq \nu_T \). We are interested in the difference between the two means; the sum is a nuisance parameter. Hence, consider:

\[
P(N_v|N_u + N_v = N) = \frac{P(N|N_v)P(N_v)}{P(N)} = \frac{\frac{\mu_T^{N-N_v} e^{-\mu_T} \nu_T^{N_v} e^{-\nu_T}}{(N-N_v)! N_v!} \left(\frac{\mu_T + \nu_T}{N} \right)^N e^{-(\mu_T + \nu_T)}}{N!}.
\]

Example of use of conditional likelihood.

\[
= \left( \frac{N}{N_v} \right)^{N_v} \left( \frac{\nu_T}{\mu_T + \nu_T} \right)^{N_v} \left( \frac{\mu_T}{\mu_T + \nu_T} \right)^{N-N_v}.\]
Normalization test (general case)

This probability permits us to construct a uniformly most powerful test of our hypothesis.* It is a binomial distribution, for given $N$. The uniformly most powerful property holds independently of $N$, although the probabilities cannot be computed without $N$.

The null hypothesis corresponds to $\mu_T = \nu_T$, that is:

$$P(N_v|N_u + N_v = N) = \left( \frac{N}{N_v} \right) \left( \frac{1}{2} \right)^N.$$

For our example, with $N = 1416$ and $N_v = 633$, the $P$-value is $7.4 \times 10^{-5}$, for a two-tailed probability. Compare with our earlier estimate of $6.6 \times 10^{-5}$ in the normal approximation. For our binomial test we have “conservatively” included the endpoints (633 and 783). Mimicing more closely the normal estimate by including one-half the probability at the endpoints, we obtain $6.7 \times 10^{-5}$, very close to the normal number.

---

Many possible tests. We’ll consider 7:

- “Chi-square test” \( (\chi^2) \)
- Bhattacharyya distance measure (BDM)
- Kolmogorov-Smirnov test (KS)
- Cramér-von-Mises test (CVM)
- Anderson-Darling test (AD)
- Likelihood ratio test (\( \ln \lambda \))
- Likelihood value test (\( \ln \mathcal{L} \))

There are many other possible tests that could be considered, for example, schemes that “partition” the \( \chi^2 \) to select sensitivity to different characteristics D. J. Best, *Nonparametric Comparison of Two Histograms*, Biometrics **50** (1994) 538.].
Chi-square test for shape

Even though we don’t expect it to follow a $\chi^2$ distribution in the low statistics regime, we may evaluate the test statistic:

$$\chi^2 = \sum_{i=1}^{k} \left( \frac{u_i}{N_u} - \frac{v_i}{N_v} \right)^2 \frac{u_i}{N_u^2} + \frac{v_i}{N_v^2}.$$ 

If $u_i = v_i = 0$, the contribution to the sum from that bin is zero.
Geometric (BDM) test for shape

Geometric motivation: Let the bin contents of a histogram define a vector in a $k$-dimensional space. If two vectors are drawn from the same distribution (null hypothesis), they will tend to point in the same direction. If we represent each histogram as a unit vector with components:

\[ \{u_1/N_u, \ldots, u_k/N_u\}, \text{ and } \{v_1/N_v, \ldots, v_k/N_v\}, \]

we may form the “dot product” test statistic:

\[ T_{\text{BDM}} = \sqrt{\frac{u}{N_u} \cdot \frac{v}{N_v}} = \left( \sum_{i=1}^{k} \frac{u_i v_i}{N_u N_v} \right)^{1/2}. \]

This is known as the “Bhattacharyya distance measure” (BDM).

This statistic is related to the $\chi^2$ statistic – the $\frac{u}{N_u} \cdot \frac{v}{N_v}$ dot product is close to the cross term in the $\chi^2$ expression.
Sample application of BDM test

Apply this formalism to our example. The sum over bins gives 0.986. According to our estimated distribution of this statistic under the null hypothesis, this gives a $P$-value of 0.32, similar to the $\chi^2$ test result (0.28).

Bin-by-bin contributions to the BDM test statistic for the example.

Estimated distribution of the BDM statistic for the null hypothesis in the example.
Kolmogorov-Smirnov test

Another approach to a shape test may be based on the Kolmogorov-Smirnov (KS) idea: Estimate the maximum difference between observed and predicted cumulative distribution functions and compare with expectations.

Modify the KS statistic to apply to comparison of histograms: Assume neither histogram is empty. Form the “cumulative distribution histograms” according to:

\[ u_{ci} = \sum_{j=1}^{i} \frac{u_j}{N_u} \quad v_{ci} = \sum_{j=1}^{i} \frac{v_j}{N_v}. \]

Then compute the test statistic (for a two-tail test):

\[ T_{KS} = \max_{i} |u_{ci} - v_{ci}|. \]
Sample application of KS test

Apply this formalism to our example. The maximum over bins is 0.050. Estimating the distribution of this statistic under $H_0$ gives a $P$-value of 0.22, somewhat smaller than for the $\chi^2$ test result, but indicating consistency of the histograms. The smaller KS $P$-value is presumably a reflection of the emphasis this test gives to the region near the peak of the distribution (i.e., the region where the largest fluctuations occur in Poisson statistics), where we indeed have a largish fluctuation.

Bin-by-bin distances for the KS test statistic for the example.

Estimated PDF of the KS distance under $H_0$ in the example.
Cramér-von-Mises test

The idea of the Cramér-von-Mises (CVM) test is to add up the squared differences between the cumulative distributions being compared. Used to compare an observed distribution with a presumed parent continuous probability distribution. Algorithm is adaptable to the two-sample comparison, and to the case of comparing two histograms.

The test statistic for comparing the two samples $x_1, x_2, \ldots, x_N$ and $y_1, y_2, \ldots, y_M$ is*:

$$T = \frac{NM}{(N + M)^2} \left\{ \sum_{i=1}^{N} [E_x(x_i) - E_y(x_i)]^2 + \sum_{j=1}^{M} [E_x(y_j) - E_y(y_j)]^2 \right\},$$

where $E_x$ is the empirical cumulative distribution for sampling $x$. That is, $E_x(x) = n/N$ if $n$ of the sampled $x_i$ are less than or equal to $x$.

---

Cramér-von-Mises test – Adaptation to comparing histograms

Adapt this for the present application of comparing histograms with bin contents $u_1, u_2, \ldots, u_k$ and $v_1, v_2, \ldots, v_k$ with identical bin boundaries: Let $z$ be a point in bin $i$, and define the empirical cumulative distribution function for histogram $u$ as:

$$E_u(z) = \sum_{j=1}^{i} u_i / N_u.$$ 

Then the test statistic is:

$$T_{CVM} = \frac{N_u N_v}{(N_u + N_v)^2} \sum_{j=1}^{k} (u_j + v_j) \left[ E_u(z_j) - E_v(z_j) \right]^2.$$ 

27 Frank Porter, Caltech Seminar, February 26, 2008
Sample application of CVM test

Apply this formalism to our example, finding $T_{CVM} = 0.193$. The resulting estimated distribution under the null hypothesis is shown below. According to our estimated distribution of this statistic under the null hypothesis, this gives a $P$-value of 0.29, similar with the $\chi^2$ test result.

Example.

Estimated PDF of the CVM statistic under $H_0$ for the example.
Anderson-Darling (AD) test for shape

The Anderson-Darling (AD) test is another non-parametric comparison of cumulative distributions. It is similar to the Cramér-von-Mises statistic, but is designed to be sensitive to the tails of the CDF. The original statistic was designed to compare a dataset drawn from a continuous distribution, with CDF $F_0(x)$ under the null hypothesis:

$$A_m^2 = m \int_{-\infty}^{\infty} \frac{[F_m(x) - F_0(x)]^2}{F_0(x)[1 - F_0(x)]} dF_0(x),$$

where $F_m(x)$ is the empirical CDF of dataset $x_1, \ldots, x_m$. 
Anderson-Darling (AD) test, adaptation to comparing histograms

Scholz and Stephens* provide a form of this statistic for a $k$-sample test on grouped data (e.g., as might be used to compare $k$ histograms). The expression of interest for two histograms is:

$$T_{AD} = \frac{1}{N_u + N_v} \sum_{j=k_{\text{min}}}^{k_{\text{max}}-1} \frac{t_j}{\Sigma_j (N_u + N_v - \Sigma_j)} \{ [(N_u + N_v)\Sigma_{uj} - N_u\Sigma_j]^2 / N_u$$

$$+ [(N_u + N_v)\Sigma_{vj} - N_v\Sigma_j]^2 / N_v \},$$

where $k_{\text{min}}$ is the first bin where either histogram has non-zero counts, $k_{\text{max}}$ is the number of bins counting up the the last bin where either histogram has non-zero counts, and

$$\Sigma_{uj} \equiv \sum_{i=1}^{j} u_i, \quad \Sigma_{vj} \equiv \sum_{i=1}^{j} v_i, \quad \text{and} \quad \Sigma_j \equiv \sum_{i=1}^{j} t_i = \Sigma_{uj} + \Sigma_{vj}.$$
Sample application of AD test for shape

We apply this formalism to our example. The sum over bins gives 0.994. According to our estimated distribution of this statistic under the null hypothesis, this gives a \textit{P-value of 0.42}, somewhat larger than the $\chi^2$ test result. The fact that it is larger is presumably due to the greater focus of this test on the tails, with less emphasis on the peak region (where the large fluctuation is seen in the example).

![Anderson-Darling statistic](image)

Example.

Estimated distribution of the AD test statistic for the null hypothesis for the example.
Likelihood ratio test for shape

Base a shape test on the same conditional likelihood idea as for the normalization test. Now there is a binomial associated with each bin. The two histograms are sampled from the joint distribution:

\[ P(u, v) = \prod_{i=1}^{k} \frac{\mu_i^{u_i}}{u_i!} e^{-\mu_i} \frac{\nu_i^{v_i}}{v_i!} e^{-\nu_i}. \]

With \( t_i = u_i + v_i \), and fixing the \( t_i \) at the observed values, we have the multi-binomial form:

\[ P(v|u + v = t) = \prod_{i=1}^{k} \binom{t_i}{v_i} \left( \frac{\nu_i}{\nu_i + \mu_i} \right)^{v_i} \left( \frac{\mu_i}{\nu_i + \mu_i} \right)^{t_i - v_i}. \]
Likelihood ratio test for shape (continued)

The null hypothesis is $\nu_i = a\mu_i, \ i = 1, \ldots, k$. We want to test this, but there are two complications:

1. The value of “$a$” is not specified;
2. We still have a multivariate distribution.

For $a$, we substitute the maximum likelihood estimator:

$$\hat{a} = \frac{N_v}{N_u}.$$

Use a likelihood ratio statistic to reduce the problem to a single variable. This will be the likelihood under the null hypothesis (with $a$ given by its maximum likelihood estimator), divided by the maximum of the likelihood under the alternative hypothesis.
Likelihood ratio test for shape (continued)

We form the ratio:

\[
\lambda = \frac{\max_{H_0} \mathcal{L}(a|v; u + v = t)}{\max_{H_1} \mathcal{L}(\{a_i \equiv v_i/\mu_i\}|v; u + v = t)} = \prod_{i=1}^{k} \left( \frac{\hat{a}_i}{1+\hat{a}_i} \right)^{v_i} \left( \frac{1}{1+\hat{a}_i} \right)^{t_i-v_i}
\]

The maximum likelihood estimator, under \(H_1\), for \(a_i\) is just \(\hat{a}_i = v_i/u_i\).

Thus, we rewrite our test statistic as:

\[
\lambda = \prod_{i=1}^{k} \left( \frac{1 + v_i/u_i}{1 + N_v/N_u} \right)^{t_i} \left( \frac{N_v u_i}{N_u v_i} \right)^{v_i}
\]

In practice, we’ll work with

\[
-2 \ln \lambda = -2 \sum_{i=1}^{k} \left[ t_i \ln \left( \frac{1 + v_i/u_i}{1 + N_v/N_u} \right) + v_i \ln \left( \frac{N_v u_i}{N_u v_i} \right) \right].
\]

If \(u_i = v_i = 0\), the bin contributes zero.

If \(v_i = 0\), contribution is \(-2 \ln \lambda_i = -2t_i \ln \left( \frac{N_u}{N_u + N_v} \right)\).

If \(u_i = 0\), the contribution is \(-2 \ln \lambda_i = -2t_i \ln \left( \frac{N_v}{N_u + N_v} \right)\).
Sample application of ln λ test

Apply this test to example, obtaining \(-2 \ln \lambda = \sum_{i=1}^{k} -2 \ln \lambda_i = 38.7\).

Asymptotically, \(-2 \ln \lambda\) should be distributed as a \(\chi^2\) with \(N_{DOF} = k - 1\), or \(N_{DOF} = 35\). If valid, this gives a \(P\)-value of 0.31, to be compared with a probability of 0.35 according to the estimated actual distribution.

We obtain nearly the same answer as the application of the chi-square calculation with no bins combined, a result of nearly bin-by-bin equality of the two statistics.

Value of \(-2 \ln \lambda_i\) (circles) or \(\chi_i^2\) (squares) as a function of histogram bin in the comparison of the two distributions of the example.
Comparison of $\ln \lambda$ and $\chi^2$

To investigate when this might hold more generally, compare the values of $-2\ln \lambda_i$ and $\chi^2_i$ as a function of $u_i$ and $v_i$, below. The two statistics agree when $u_i = v_i$ with increasing difference away from that point. This agreement holds even for low statistics. However, shouldn’t conclude that the chi-square approximation may be used for low statistics – fluctuations away from equal numbers lead to quite different results when we get into the tails at low statistics. Our example doesn’t really sample these tails – the only bin with a large difference is already high statistics.

Value of $-2\ln \lambda_i$ or $\chi^2_i$ as a function of $u_i$ and $v_i$ bin contents. This plot assumes $N_u = N_v$. 

![Graph showing comparison of $-2\ln \lambda_i$ and $\chi^2_i$ for different values of $u_i$ and $v_i$.]
Likelihood value test

- Let $\ln \mathcal{L}$ be the value of the likelihood at its maximum value under the null hypothesis.

- $\ln \mathcal{L}$ is an often-used but controversial goodness-of-fit statistic.

- Can be demonstrated that this statistic carries little or no information in some situations.

- However, in the limit of large statistics it is essentially the chi-square statistic, so there are known situations were it is a plausible statistic to use.

- Using the results in the previous slides, the test statistic is:

$$\ln \mathcal{L} = \sum_{i=1}^{k} \left[ \ln \left( \frac{t_i}{v_i} \right) + t_i \ln \left( \frac{N_u}{N_u + N_v} \right) + v_i \ln \left( \frac{N_v}{N_u} \right) \right].$$

If either $N_u = 0$ or $N_v = 0$, then $\ln \mathcal{L} = 0$. 
Sample application of the ln $\mathcal{L}$ test

- Apply this test to the example. The sum over bins is 90. Using our estimated distribution of this statistic under the null hypothesis, this gives a $P$-value of 0.29, similar to the $\chi^2$ test result.

- To be expected, since our example is reasonably well-approximated by the large statistics limit.

Estimated distribution of the ln $\mathcal{L}$ test statistic under $H_0$ in the example.
Distributions Under the Null Hypothesis

Our example may be too easy... 

When the asymptotic distribution may not be good enough, we would like to know the probability distribution of our test statistic under the null hypothesis.

Difficulty: our null hypothesis is not completely specified!

The problem is that the distribution depends on the values of $\nu_i = a\mu_i$. Our null hypothesis only says $\nu_i = a\mu_i$, but says nothing about what $\mu_i$ might be.

[It also doesn’t specify $a$, but we have already discussed that complication, which appears manageable.]
Estimating the null hypothesis

- Turn to the data to make an estimate for $\mu_i$, to be used in estimating the distribution of our test statistics.

- Straightforward approach: use the M.L. parameter estimators (under $H_0$):

  $$\hat{\mu}_i = \frac{1}{1 + \hat{a}}(u_i + v_i),$$
  where $\hat{a} = N_v/N_u$

  $$\hat{\nu}_i = \frac{\hat{a}}{1 + \hat{a}}(u_i + v_i).$$

- Repeatedly simulate data using these values for the parameters of the sampling distribution.

- For each simulation, obtain a value of the test statistic.

- The distribution so obtained is then an estimate of the distribution of the test statistic under $H_0$, and $P$-values may be computed from this.
But it isn’t guaranteed to work...

– We have just described the approach that was used to compute the estimated probabilities for the example. The bin contents are reasonably large, and this approach works well enough for this case.

– Unfortunately, this approach can do very poorly in the low-statistics realm.

– Consider a simple test case: Suppose our data is sampled from a flat distribution with a mean of 1 count in each of 100 bins.
Algorithm to check estimated null hypothesis

We test how well our estimated null hypothesis works for any given test statistic, $T$, as follows:

1. Generate a pair of histograms according to some assumed $H_0$.
   (a) Compute $T$ for this pair of histograms.
   (b) Given the pair of histograms, compute the estimated null hypothesis according to the specified prescription above.
   (c) Generate many pairs of histograms according to the estimated null hypothesis in order to obtain an estimated distribution for $T$.
   (d) Using the estimated distribution for $T$, determine the estimated $P$-value for the value of $T$ found in step (a).

2. Repeat step 1 many times and make a histogram of the estimated $P$-values. This histogram should be uniform if the estimated $P$-values are good estimates.
Checking the estimated null distribution

The next two slides show tests of the estimated null distribution for each the 7 test statistics.

Shown are distributions of the simulated $P$-values. The data are generated according to $H_0$, consisting of 100 bin histograms for:

- A mean of 100 counts/bin (left column), or
- A mean of 1 count/bin (other 3 columns).

The estimates of $H_0$ are:

- Weighted bin-by-bin average (left two columns),
- Each bin mean given by the average bin contents in each histogram (third column),
- Estimated with a Gaussian kernel estimator (right column) based on the contents of both histograms.

The $\chi^2$ is computed without combining bins.
<counts/bin> = 100

- H0 is weighted bin-by-bin average
- H0 from overall average
- H0 from Gaussian kernel estimator

The diagrams illustrate the frequency distribution for different estimation methods: χ², BDM, KS, and CVM. The x-axis represents the estimated probability, and the y-axis represents the frequency.
Summary of tests of null distribution estimates

Probability that the null hypothesis will be rejected with a cut at 1% on the estimated distribution. $H_0$ is estimated with the bin-by-bin algorithm in the first two columns, by the uniform histogram algorithm in the third column, and by Gaussian kernel estimation in the fourth column.

<table>
<thead>
<tr>
<th>Test statistic</th>
<th>Probability (%)</th>
<th>Probability (%)</th>
<th>Probability (%)</th>
<th>Probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin mean = $H_0$ estimate</td>
<td>100</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Bin-by-bin</td>
<td>0.97 ± 0.24</td>
<td>18.5 ± 1.0</td>
<td>1.2 ± 0.3</td>
<td>1.33 ± 0.28</td>
</tr>
<tr>
<td>BDM</td>
<td>0.91 ± 0.23</td>
<td>16.4 ± 0.9</td>
<td>0.30 ± 0.14</td>
<td>0.79 ± 0.22</td>
</tr>
<tr>
<td>KS</td>
<td>1.12 ± 0.26</td>
<td>0.97 ± 0.24</td>
<td>1.0 ± 0.2</td>
<td>1.21 ± 0.27</td>
</tr>
<tr>
<td>CVM</td>
<td>1.09 ± 0.26</td>
<td>0.85 ± 0.23</td>
<td>0.8 ± 0.2</td>
<td>1.27 ± 0.28</td>
</tr>
<tr>
<td>AD</td>
<td>1.15 ± 0.26</td>
<td>0.85 ± 0.23</td>
<td>1.0 ± 0.2</td>
<td>1.39 ± 0.29</td>
</tr>
<tr>
<td>$\ln \lambda$</td>
<td>0.97 ± 0.24</td>
<td>24.2 ± 1.1</td>
<td>1.5 ± 0.3</td>
<td>2.0 ± 0.34</td>
</tr>
<tr>
<td>$\ln L$</td>
<td>0.97 ± 0.24</td>
<td>28.5 ± 1.1</td>
<td>0.0 ± 0.0</td>
<td>0.061 ± 0.061</td>
</tr>
</tbody>
</table>
Conclusions from tests of null distribution estimates

- In the “large statistics” case, where sampling is from histograms with a mean of 100 counts in each bin, all test statistics display the desired flat distribution.

- The $\chi^2$, $\ln \lambda$, and $\ln L$ statistics perform essentially identically at high statistics, as expected, since in the normal approximation they are equivalent.

- In the “small statistics” case, if the null hypothesis were to be rejected at the estimated 0.01 probability, the bin-by-bin algorithm for estimating $a_i$ would actually reject $H_0$: 19% of the time for the $\chi^2$ statistic, 16% of the time for the BDM statistic, 24% of the time for the $\ln \lambda$ statistic, and 29% of the time for the $L$ statistics, all unacceptably larger than the desired 1%. The KS, CVM, and AD statistics are all consistent with the desired 1%.
Problem appears for low statistics

Intuition behind the failure of the bin-by-bin approach at low statistics: Consider the likely scenario that some bins will have zero counts in both histograms. Then our algorithm for the estimated null hypothesis yields a zero mean for these bins. The simulation to determine the probability distribution for the test statistic will always have zero counts in these bins, ie, there will always be agreement between the two histograms. Thus, the simulation will find that low values of the test statistic are more probable than it should.

The AD, CVM, and KS tests are more robust under our estimates of $H_0$ than the others, as they tend to emphasize the largest differences and are not so sensitive to bins that always agree. For these statistics, our bin-by-bin procedure for estimating $H_0$ does well even for low statistics, although we caution that we are not examining the far tails of the distribution.
Obtaining better null distribution estimates

A simple approach to salvaging the situation in the low statistics regime involves relying on the often valid assumption that the underlying $H_0$ distribution is “smooth”. Then only need to estimate a few parameters to describe the smooth distribution, and effectively more statistics are available.

Assuming a smooth background represented by a uniform distribution yields correct results. This is cheating a bit, since we aren’t supposed to know that this is really what we are sampling from.

The $\ln \mathcal{L}$ and to a lesser extent the BDM statistic, do not give the desired 1% result, but now err on the “conservative” side. May be possible to mitigate this with a different algorithm. Expect the power of these statistics to suffer under the approach taken here.
More “honest” – Try a kernel estimator

Since we aren’t supposed to know that our null distribution is uniform, we also try a kernel estimator for $H_0$, using the sum of the observed histograms as input. Try a Gaussian kernel, with a standard deviation of 2. It works pretty well, though there is room for improvement. The bandwidth was chosen here to be rather small for technical reasons; a larger bandwidth might improve results.

Sample Gaussian kernel density estimate of the null hypothesis.
Comparison of Power of Tests

- The power depends on what the alternative hypothesis is.
- Investigate adding a Gaussian component on top of a uniform background distribution. Motivated by the scenario where one distribution appears to show some peaking structure, while the other does not.
- Also look at a different extreme, a rapidly varying alternative.
Gaussian alternative

The Gaussian alternative data are generated as follows:

- The “background” has a mean of one event per histogram bin.

- The Gaussian has a mean of 50 and a standard deviation of 5, in units of bin number.

- As a function of the amplitude of the Gaussian, count how often the null hypothesis is rejected at the 1% confidence level. The amplitude is measured in percent, eg, a 25% Gaussian has a total amplitude corresponding to an average of 25% of the total counts in the histogram. The Gaussian counts are added to the counts from the background distribution.
Sample Gaussian alternative

Left: The mean bin contents for a 25% Gaussian on a flat background of one count/bin (note the suppressed zero).

Right: Example sampling from the 25% Gaussian (filled blue dots) and from the uniform background (open red squares).
Power estimates for Gaussian alternative

On the next two slides, we show, for seven test statistics, the distribution of the estimated probability, under $H_0$, that the test statistic is worse than that observed.

- Three different magnitudes of the Gaussian amplitude are displayed.
- The data are generated according to a uniform distribution, consisting of 100 bin histograms with a mean of 1 count, for one histogram, and for the other histogram with an added Gaussian peak of strength:
  - 12.5% (left column),
  - 25% (middle column), and
  - 50% (right column).

[The $\chi^2$ is computed without combining bins.]
Estimated probability

12.5%

25%

50%

AD

\ln \lambda

\ln L

Frank Porter, Caltech Seminar, February 26, 2008
Power comparison summary - Gaussian peak alternative

Estimates of power for seven different test statistics, as a function of $H_1$. The comparison histogram is generated with all $k = 100$ bins Poisson of mean 1. The selection is at the 99% confidence level, that is, the null hypothesis is accepted with (an estimated) 99% probability if it is true.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>H0</th>
<th>12.5</th>
<th>25</th>
<th>37.5</th>
<th>50</th>
<th>-25</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>$1.2 \pm 0.3$</td>
<td>$1.3 \pm 0.3$</td>
<td>$4.3 \pm 0.5$</td>
<td>$12.2 \pm 0.8$</td>
<td>$34.2 \pm 1.2$</td>
<td>$1.6 \pm 0.3$</td>
</tr>
<tr>
<td>BDM</td>
<td>$0.30 \pm 0.14$</td>
<td>$0.5 \pm 0.2$</td>
<td>$2.3 \pm 0.4$</td>
<td>$10.7 \pm 0.8$</td>
<td>$40.5 \pm 1.2$</td>
<td>$0.9 \pm 0.2$</td>
</tr>
<tr>
<td>KS</td>
<td>$1.0 \pm 0.2$</td>
<td>$3.6 \pm 0.5$</td>
<td>$13.5 \pm 0.8$</td>
<td>$48.3 \pm 1.2$</td>
<td>$91.9 \pm 0.7$</td>
<td>$7.2 \pm 0.6$</td>
</tr>
<tr>
<td>CVM</td>
<td>$0.8 \pm 0.2$</td>
<td>$1.7 \pm 0.3$</td>
<td>$4.8 \pm 0.5$</td>
<td>$35.2 \pm 1.2$</td>
<td>$90.9 \pm 0.7$</td>
<td>$2.7 \pm 0.4$</td>
</tr>
<tr>
<td>AD</td>
<td>$1.0 \pm 0.2$</td>
<td>$1.8 \pm 0.3$</td>
<td>$6.5 \pm 0.6$</td>
<td>$42.1 \pm 1.2$</td>
<td>$94.7 \pm 0.6$</td>
<td>$2.8 \pm 0.4$</td>
</tr>
<tr>
<td>$\ln \lambda$</td>
<td>$1.5 \pm 0.3$</td>
<td>$1.9 \pm 0.3$</td>
<td>$6.4 \pm 0.6$</td>
<td>$22.9 \pm 1.0$</td>
<td>$67.1 \pm 1.2$</td>
<td>$2.4 \pm 0.4$</td>
</tr>
<tr>
<td>$\ln \mathcal{L}$</td>
<td>$0.0 \pm 0.0$</td>
<td>$0.1 \pm 0.1$</td>
<td>$0.8 \pm 0.2$</td>
<td>$6.5 \pm 0.6$</td>
<td>$34.8 \pm 1.2$</td>
<td>$0.0 \pm 0.0$</td>
</tr>
</tbody>
</table>
Summary of power of seven different test statistics, for the alternative hypothesis with a Gaussian bump.
Left: linear vertical scale; Right: logarithmic vertical scale.
Power comparison - High statistics

Also look at the performance for histograms with large bin contents.

Estimates of power for seven different test statistics, as a function of $H_1$. The comparison histogram ($H_0$) is generated with all $k = 100$ bins Poisson of mean 100. The selection is at the 99% confidence level.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>H0</th>
<th>5</th>
<th>-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>0.91 ± 0.23</td>
<td>79.9 ± 1.0</td>
<td>92.1 ± 0.7</td>
</tr>
<tr>
<td>BDM</td>
<td>0.97 ± 0.24</td>
<td>80.1 ± 1.0</td>
<td>92.2 ± 0.7</td>
</tr>
<tr>
<td>KS</td>
<td>1.03 ± 0.25</td>
<td>77.3 ± 1.0</td>
<td>77.6 ± 1.0</td>
</tr>
<tr>
<td>CVM</td>
<td>0.91 ± 0.23</td>
<td>69.0 ± 1.1</td>
<td>62.4 ± 1.2</td>
</tr>
<tr>
<td>AD</td>
<td>0.91 ± 0.23</td>
<td>67.5 ± 1.2</td>
<td>57.8 ± 1.2</td>
</tr>
<tr>
<td>ln $\lambda$</td>
<td>0.91 ± 0.23</td>
<td>79.9 ± 1.0</td>
<td>92.1 ± 0.7</td>
</tr>
<tr>
<td>ln $\mathcal{L}$</td>
<td>0.97 ± 0.24</td>
<td>79.9 ± 1.0</td>
<td>91.9 ± 0.7</td>
</tr>
</tbody>
</table>
Comments on power comparison - High statistics

- In this large-statistics case, for the $\chi^2$ and similar tests, the power to
  for a dip is greater than the power for a bump of the same area.
  - Presumably because the “error estimates” for the $\chi^2$ are based
    on the square root of the observed counts, and hence give smaller
    errors for smaller bin contents.

- Also observe that the comparative strength of the KS, CVM, and AD
tests versus the $\chi^2$, BDM, $\ln \lambda$, and $\ln \mathcal{L}$ tests in the small statistics
situation is largely reversed in the large statistics case.
Power study for a “sawtooth” alternative

To see what happens for a radically different alternative distribution, consider sampling from the “sawtooth” distribution. Compare once again to sampling from the uniform histogram.

The “percentage” sawtooth here is the fraction of the null hypothesis mean. A 100% sawtooth on a 1 count/bin background oscillates between a mean of 0 counts/bin and 2 counts/bin.

The period of the sawtooth is always two bins.

Left: The mean bin contents for a 50% sawtooth on a background of 1 count/bin (blue), and for the flat background (red).
Right: A sampling from the 50% sawtooth (blue) and from the uniform background (red).
Now the $\chi^2$ and $\ln \lambda$ tests are the clear winners, with BDM next. The KS, CVM, and AD tests reject the null hypothesis with the same probability as for sampling from a true null distribution. This very poor performance for these tests is readily understood, as these tests are all based on the cumulative distributions, which smooth out local oscillations.
Conclusions

These studies provide some lessons in “goodness-of-fit” testing:

1. No single “best” test for all applications. The statement “test X is better than test Y” is empty without more context. E.g., the Anderson-Darling test is very powerful in testing normality of data against alternatives with non-normal tails (e.g., a Cauchy distribution)*. It is not always especially powerful in other situations. The more we know about what we wish to test, the better we can choose a powerful test. Each of the tests here may be useful, depending on the circumstance.

2. Even the controversial $L$ test works as well as the others sometimes. However, there is no known situation where it performs better than all of the others, and the situations where it is observed to perform as well are here limited to those where it is equivalent to another test.

Conclusions (continued)

3. Computing probabilities via simulations is a useful technique. However, must be done with care. Tests with an incompletely specified null hypothesis require care. Generating a distribution according to an assumed null distribution can lead to badly wrong results. It is important to verify the validity of the procedure. We have only looked into tails at the 1% level. Validity must be checked to whatever level of probability is needed. Should not assume that results at the 1% level will still be true at, say, the 0.1% level.

4. Concentrated on the question of comparing two histograms. However, considerations apply more generally, to testing whether two datasets are consistent with being drawn from the same distribution, and to testing whether a dataset is consistent with a predicted distribution. The KS, CVM, AD, ln $L$, and $L$ tests may all be constructed for these other situations (as well as the $\chi^2$ and BDM, if we bin the data).
ASIDE
Significance/GOF: Counting Degrees of Freedom

The following situation arises with some frequency (with variations):

I do two fits to the same dataset (say a histogram with $N$ bins):

- Fit $A$ has $n_A$ parameters, with $\chi^2_A$ [or perhaps $-2\ln \mathcal{L}_A$].
- Fit $B$ has a subset $n_B$ of the parameters in fit $A$, with $\chi^2_B$, where the $n_A - n_B$ other parameters (call them $\theta$) are fixed at zero.

What is the distribution of $\chi^2_B - \chi^2_A$?

Luc Demortier in:
http://phystat-lhc.web.cern.ch/phystat-lhc/program.html
Counting Degrees of Freedom (continued)

In the asymptotic limit (that is, as long as the normal sampling distribution is a valid approximation),

\[ \Delta \chi^2 \equiv \chi^2_B - \chi^2_A \]

is the same as a likelihood ratio \((-2 \ln \lambda)\) statistic for the test:

\[ H_0 : \theta = 0 \quad \text{against} \quad H_1 : \text{some } \theta \neq 0 \]

\(\Delta \chi^2\) is distributed according to a \(\chi^2(\text{NDOF} = n_A - n_B)\) distribution as long as:

- Parameter estimates in \(\lambda\) are consistent (converge to the correct values) under \(H_0\).
- Parameter values in the null hypothesis are interior points of the maintained hypothesis (union of \(H_0\) and \(H_1\)).
- There are no additional nuisance parameters under the alternative hypothesis.
Counting Degrees of Freedom (continued)

A commonly-encountered situation often violates these requirements: We compare fits with and without a signal component to estimate significance of the signal.

- If the signal fit has, e.g., a parameter for mass, this constitutes an additional nuisance parameter under the alternative hypothesis.
- If the fit for signal constrains the yield to be non-negative, this violates the interior point requirement.

![Graphs showing chi-square distributions with different hypotheses and constraints.](Image)
Counting Degrees of Freedom - Sample fits

Fits to data with no signal.  

Fits to data with a signal.