



## Chapter 6

# Hypothesis Tests

Often, we want to address questions such as whether the possible observation of a new effect is really “significant”, or merely a chance fluctuation. Also, we frequently ask whether a given model provides a good description of the data, or whether two results are consistent. This is the domain of “hypothesis tests”. Hypothesis tests and interval estimation are much the same thing, but with different perspectives. Again, we could adopt either frequentist or Bayesian approaches. We’ll concentrate mostly on frequentist methods here.

The questions we ask may involve values of parameters in which case we have parametric tests, or they may involve more general features of the distribution, corresponding to non-parametric tests.

While a relatively few methods are used in point and interval estimation, a very large number of tests have been developed for testing hypotheses, in part due to the large variety of questions that can be asked, but also due to the difficulty in obtaining generally optimal properties. We illustrate some of the most commonly used tests, while discussing issues more generally.

### 6.1 The Simple Hypothesis Test

Sampling is performed from a probability distribution,  $f_X(x; \theta)$ , with unknown parameter  $\theta$ .<sup>1</sup> For example, often  $X$  is the estimator for  $\theta$ . In our experiment, we observe  $X = x$ . We wish to test the hypothesis:

$$H_0 : \theta = \theta_0,$$

against the alternative:

$$H_1 : \theta = \theta_1.$$

The question being posed is: With what confidence does our data rule out  $H_0$ ? Of course, this is in the category of a parametric test.

---

<sup>1</sup>I’ll use language as if  $X$  and  $\theta$  are one-dimensional quantities. However, they may also be multiple dimensional.

Both  $H_0$  and  $H_1$  specify completely hypothetical sampling distributions, they are known as “simple” hypotheses. This problem is completely soluble with a most “powerful” test in classical statistics. The solution involves an ordering principle for the likelihood ratio statistic, already familiar from chapter 5. We’ll formulate things at the start in the way statistics texts do it, by supposing we have a specified confidence level,  $\alpha$ , for which we reject the  $H_0$  hypothesis, but then we’ll modify slightly to answer the stated question.

We set up the test by saying we are going to “reject” (i.e., consider unlikely)  $H_0$  in favor of  $H_1$  if the observation  $x$  lies in some region  $R$  of the sample space. This is known as the “critical region” for the test. When we reject one hypothesis in favor of another hypothesis, there are two types of error we could make:

Type I error: Reject  $H_0$  when  $H_0$  is true.  
 Type II error: Accept  $H_0$  when  $H_1$  is true.

The probability of making a Type I error is:

$$\alpha = \text{Prob}(x \in R|H_0) \quad (6.1)$$

$$= \int_R f(x; \theta_0) dx. \quad (6.2)$$

The probability  $\alpha$  is typically called the “confidence level”, or will become the “ $P$ -value” in our later discussion.

The probability of making a Type II error is:

$$\beta = \text{Prob}(x \in \tilde{R}|H_1) \quad (6.3)$$

$$= 1 - \int_R f(x; \theta_1) dx, \quad (6.4)$$

where  $\tilde{R}$  denotes the complement of  $R$  in the sample space. The quantity  $1 - \beta$  is called the “power” of the test; it is the probability that  $H_0$  is correctly rejected.

There are many possible critical regions  $R$  which give the same value for  $\alpha$ . We wish to pick the “best critical region”, by finding that region for which the power is greatest. Fortunately, this is straightforward for a simple test. We wish to maximize:

$$1 - \beta = \int_R f(x; \theta_1) dx \quad (6.5)$$

$$= \int_R \frac{f(x; \theta_1)}{f(x; \theta_0)} f(x; \theta_0) dx, \quad (6.6)$$

subject to the constraint:

$$\alpha = \int_R f(x; \theta_0) dx.$$

Notice that

$$\frac{1 - \beta}{\alpha} = \left\langle \frac{f(x; \theta_1)}{f(x; \theta_0)} \right\rangle_{(R; H_0)},$$

where the subscript on the  $\langle \rangle$  denotes an average restricted to the critical region, under hypothesis  $H_0$ . Thus, we wish to build the critical region by including those values of  $x$  for which the ratio  $\frac{f(x; \theta_1)}{f(x; \theta_0)}$  is largest. This is the “ordering principle”. The region  $R$  contains all values  $x$  for which:

$$\frac{f(x; \theta_1)}{f(x; \theta_0)} \geq \Lambda_\alpha,$$

where  $\Lambda_\alpha$  is determined by the  $\alpha$  constraint.

We may re-express this in the context of likelihood functions. Let

$$\lambda(x) = \frac{L(\theta_1; x)}{L(\theta_0; x)}$$

be the “likelihood ratio” for the two hypotheses, for a sampled value  $x$ . Note that  $\lambda$  is itself a random variable. If  $\lambda \geq \Lambda_\alpha$ , the sample  $x$  is in the critical region. The likelihood ratio test is **uniformly most powerful** (UMP) for the test between simple hypotheses, meaning that for a given confidence level  $\alpha$ , it has the greatest power against any alternative  $\theta_1$ .

Now, we can turn this around, given a sampling  $x$  (and hence  $\lambda$ ), and ask what the confidence level, or P-value for  $H_0$  is, according to the value  $x$ . That is, what is the probability to get the observed value, or more extreme (where “extreme” is defined in the sense of being in the direction toward favoring  $H_1$ ), if  $H_0$  is true? This is the probability with which we “rule out”  $H_0$ .

Let’s try an example: Suppose

$$f(x; \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}.$$

We wish to test:

$$H_0 : \theta = \theta_0 = -1,$$

against the alternative:

$$H_1 : \theta = \theta_1 = +1.$$

We sample a value  $x^*$  and form the likelihood ratio (we’ll also take the logarithm for convenience):

$$\ln \lambda^* = \frac{1}{2} [(x^* - \theta_1)^2 - (x^* - \theta_0)^2] \quad (6.7)$$

$$= 2x^*. \quad (6.8)$$

This defines the critical region:  $\ln \Lambda_\alpha = 2x^*$ . The critical region is thus given by  $\ln \lambda \geq 2x^*$ . That is, we are trying to determine the probability for  $\lambda$  to exceed the observed value. Since  $\ln \lambda = 2x$ , we want the probability that  $x > x^*$ :

$$\alpha = \int_R f(x; \theta_0) dx = \int_{x^*}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta_0)^2} dx.$$

The greater the value of  $x^*$ , the smaller is  $\alpha$ , and thus the more likely we are to rule out  $H_0$ . Note that our result is intuitive in this situation since everything is nicely monotonic.

Consider a specific example, with,

$$f(x; \theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2},$$

with  $\sigma = 0.83$ . The sample value is  $x^* = 2.72$ . The hypotheses being compared are  $\theta = \theta_0 = -0.68$  and  $\theta = \theta_1 = 0.68$ . We already know from our example above that  $\alpha$  will be given by the probability that  $x > x^*$ :

$$\alpha = \int_{x^*}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\theta_0}{\sigma}\right)^2} dx. \quad (6.9)$$

$$= 2 \times 10^{-5}. \quad (6.10)$$

Likewise, the power of the test is:

$$1 - \beta = \int_{x^*}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\theta_1}{\sigma}\right)^2} dx. \quad (6.11)$$

$$= 0.01. \quad (6.12)$$

The rather low power is just telling us that  $\theta_0$  and  $\theta_1$  are pretty close together on the scale of  $\sigma$  and the value of  $\alpha$ . That is, for  $\alpha = 2 \times 10^{-5}$ , even if  $H_1$  is correct, we will accept  $H_0$  with 99% probability. In a sense, we were “lucky” with our actual sampling, to get such a small value for  $\alpha$  ( $x^*$  is unlikely even if  $H_1$  is true).

In general, we may not be able to analytically evaluate the sampling distribution for  $\lambda$ , under the hypothesis  $H_0$ . In this case, we resort to Monte Carlo simulation to evaluate  $\alpha$ . Care must be taken to simulate enough experiments to learn how the tails behave, since that is where the action lies.

Note that the procedure described here is not the same as the procedure a Bayesian applies. The Bayesian would form posterior probabilities:

$$L_0 \equiv \frac{L(\theta_0, x^*)P(\theta_0)}{L(\theta_0, x^*)P(\theta_0) + L(\theta_1, x^*)P(\theta_1)}, \quad (6.13)$$

$$L_1 \equiv \frac{L(\theta_1, x^*)P(\theta_1)}{L(\theta_0, x^*)P(\theta_0) + L(\theta_1, x^*)P(\theta_1)}, \quad (6.14)$$

where the prior probabilities are given by  $P(\theta_0)$  and  $P(\theta_1)$  (with  $P(\theta_0) + P(\theta_1) = 1$ ). Then one may compare the values of the posterior likelihoods (“degrees of belief”) at the two hypotheses to get a relative degree of belief for which of the two hypotheses is preferred. This may be contrasted with the strictly frequentist method. As with interval estimation, it is desirable to give the frequentist answer, since that describes the data independent of prior beliefs. If a Bayesian interpretation is desired, that may be provided in addition.

## 6.2 The Run Test

The above parametric discussion may be contrasted with an example of a quite general non-parametric test, applicable to any sampling distribution. A simple test for independence of a sequence of random variables is a **runs test**. There are variants using this name; a common one is the following: Consider a sequence,  $X = X_1, \dots, X_n$ , of random variables. If these are truly independent, there should be no correlation between successive values. Let  $M$  be the median of the sample. Set up a test vector  $T$  by assigning  $T_i = +$  if  $X_i > M$  and  $T_i = -$  if  $X_i < M$ . If any values equal the median, drop them from the sample (and adjust  $n$ ). A sequence composed purely of +'s or of -'s is called a run. The entire sequence of +'s and -'s can be described as a sequence of runs of varying lengths.

Let  $R$  be the number of runs, a random variable. If  $R$  is near  $n$ , the sequence  $X$  is not random, since the  $X_i$  tend to systematically alternate: If  $X_i$  is greater than the median, then  $X_{i+1}$  is usually less than the median. Alternatively, if  $R$  is near 2, then again the sequence is not random, since in this case if  $X_i$  is greater than the median then  $X_{i+1}$  is usually also greater than the median.

What values of  $R$  are most likely if the sequence really is random, that is under the null hypothesis? To answer this in detail, we need the distribution of  $R$  [1]:

$$P(R = 2s) = 2 \binom{\frac{n}{2} - 1}{s - 1}^2 / \binom{n}{\frac{n}{2}} \quad (6.15)$$

$$P(R = 2s - 1) = 2 \binom{\frac{n}{2} - 1}{s - 2} \binom{\frac{n}{2} - 1}{s - 1} / \binom{n}{\frac{n}{2}}. \quad (6.16)$$

The expectation value of  $R$  is

$$\langle R \rangle = 1 + \frac{n}{2}, \quad (6.17)$$

which is reasonably intuitive since the probability is 1/2 that the sign will change from one element in the sequence to the next. The variance of  $R$  is

$$\text{Var}(R) = \frac{1}{4} \frac{n(n-2)}{n-1}. \quad (6.18)$$

It may be remarked that the runs test is robust in the sense that it is completely insensitive to tails in the sampling distribution for  $X$ . However, by ignoring the actual sizes of fluctuations about the median, possibly relevant information is lost, and the test may not be as powerful as a test one might devise to use this information. On the other hand, the familiar  $\chi^2$  test is completely oblivious to runs, hence is less powerful than the runs test when the possibility of runs is the issue.

## 6.3 Significance of an Effect

The discovery of a new phenomenon must be critically evaluated before acceptance. One criterion for acceptance is “statistical significance”: the observation is unlikely to be due to statistical fluctuation of known processes. It is desirable to quantify the measure of statistical significance in terms of probabilities, and there are standard approaches for this. However, there are several difficulties, both in principle and in execution. We review these difficulties, and suggest approaches to mitigate them.

The significance of a possible effect may be defined as the probability of an observed deviation from null, under the null hypothesis. That is, as the  $P$ -value discussed above. Basing a calculation of significance on a result for a 68% confidence interval does not always give a reliable estimate of significance. The tails may be non-normal. A separate analysis is generally required, which models the tails appropriately.

### 6.3.1 Significance as Hypothesis Test

It is conventional to use the language of hypothesis testing when discussing tests of significance. The “null hypothesis”,  $H_0$ , is the hypothesis that there is no new effect. The “alternative hypothesis”,  $H_1$ , is the hypothesis that there is a new effect. If the observation is sufficiently unlikely to occur in the null hypothesis, then we reject  $H_0$  in favor of the alternative.

Statisticians define a “rejection region” corresponding to a given significance level,  $\alpha$ . This is a region of sampling space which has probability  $\alpha$  under the null hypothesis. As before,  $\alpha$  is the probability of making a Type I error, that is, the probability of rejecting the null hypothesis if the null hypothesis is true. In physics practice, we usually quote the “rejection region” based on the observation, by taking it to be the region for which an observation is no more likely than the actual observation. In this case,  $\alpha$  is called the “ $P$ -value”.

Consider a histogram drawn from a sampling distribution with a flat background with independent bin contents distributed according to  $N(100, 10)$ , as a normal approximation to Poisson sampling, plus a Gaussian signal of (exactly) 100 counts centered at  $x = 0$  and standard deviation one. A particular sampling from this distribution is shown in Fig. 6.1a.

We do a simple cut-and-count fit to the data in Fig. 6.1a, using the sidebands ( $|x| > 3$ ) to estimate the background,  $B$ . The background is subtracted from the observed counts in the signal region ( $|x| < 3$ ) yielding a signal estimate of  $S = 194 \pm 39$  events. The uncertainty has been computed here as  $\sqrt{S + B} = 39$ . As the sideband statistics is large, the contribution to the error estimate from the sidebands is neglected. The significance ( $P$ -value) of this signal is given by the probability of obtaining a signal estimate at least as large (in absolute value for a two-tailed test) as that observed, where this probability is computed according to the null hypothesis that the data is sampled entirely from the

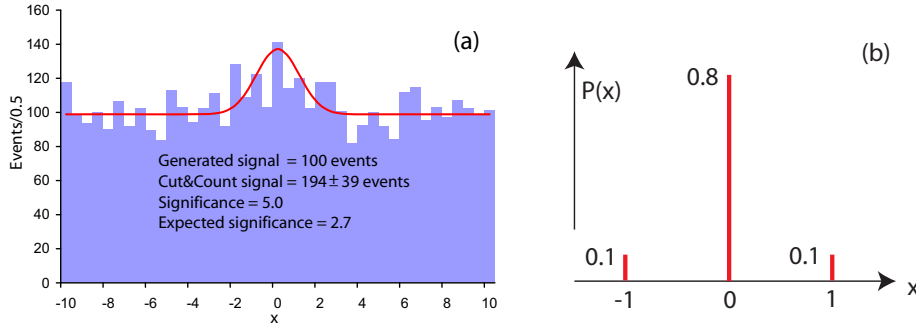


Figure 6.1: (a) An example to demonstrate the computation of significance. (b) A sampling distribution with mean zero and standard deviation  $\sigma = 0.2$ , with a probability of 20% to encounter a  $5\sigma$  fluctuation from the mean.

background distribution:

$$H_0 : N_{\text{signal}} = 0; \quad (6.19)$$

$$H_1 : N_{\text{signal}} \neq 0. \quad (6.20)$$

In this example, this probability is  $P = 5.7 \times 10^{-7}$ , the probability of having a  $> 5$  standard deviation fluctuation of a normal distribution.

Note that we have had an upward fluctuation of the estimated signal in Fig. 6.1a, due to a background fluctuation (since the signal is actually fixed). The expected signal is 100 events, corresponding to an expected significance of  $2.7\sigma$ .

It should be stressed, since there is a common confusion, that the significance is not obtained by dividing the signal estimate ( $S = 194$ ) by the uncertainty in the signal ( $\delta S = 39$ ),  $194/39 = 5.0$ . That would be addressing how likely a signal of the estimated size would be to fluctuate to zero, not the probability of the background to fluctuate to give us the signal. It turns out to be a good approximation in this example only because the background-to-signal is large.

### 6.3.2 Significance as “ $n\sigma$ ”

Common parlance is to say an effect has, for example, “ $5\sigma$ ” significance. Quoting significance as “ $n\sigma$ ” implies that the observation is  $n$  standard deviations away from the value expected under the null hypothesis, where a standard deviation is computed according to  $\sigma = \sqrt{\langle(x - \langle x \rangle)^2\rangle}$ . But we often don’t really mean this. Fig. 6.1b shows a distribution for which the probability of finding a  $5\sigma$  effect is 20%. In spite of our intuition,  $5\sigma$  fluctuations are not necessarily improbable.

Instead, when we quote a significance as  $n\sigma$ , we sometimes mean that the null probability ( $P$ -value) is given by the probability of a  $n\sigma$  fluctuation of a normal distribution, i.e.,  $P = P(|x| > 5)$  for  $x$  from  $N(0, 1)$ , or  $P = 5.7 \times 10^{-7}$ , assuming two-tailed probability.

But sometimes, we really do mean  $5\sigma$ , usually presuming that the sampling distribution is approximately normal. This may not be an accurate presumption when far out in the tails. It has also become popular in recent times to compute the change in log-likelihood,  $\lambda \equiv \sqrt{-2\Delta \ln L}$ , and call this “ $n\sigma$ ”. The two likelihoods compared are the value from the best fit assuming the alternative hypothesis  $H_1$  and the maximum under the null hypothesis. The distribution of this statistic may be computed under the null hypothesis. But often it is quoted without investigating this probability, in the hope that it corresponds approximately to the normal case. That is, in the normal case we have  $L_0(\theta = 0; x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}(x/\sigma)^2\right]$  and  $L_{\max}(\hat{\theta} = x; x) = \frac{1}{\sqrt{2\pi}\sigma}$ , giving  $\sqrt{-2\Delta \ln L} = \sqrt{\Delta\chi^2} = x/\sigma = n$ .

Given the confusion, it is desirable to be more concise by quoting probabilities, or “ $P$ -values” as is common in the statistics world. In any event, it is important to say what you mean.

The example of Fig. 6.1a was known to be normal sampling. Often this is true approximately, by the central limit theorem. However, in computing significances we may be far into the tails of the distribution, and the assumption of normality may be unjustified even if the distribution is close to normal near the mean.

If there is any doubt, the actual distribution of the significance statistic under the null hypothesis should be computed. This is usually done by simulation, with a “toy Monte Carlo”. To compute the tails may require a large amount of computing time, and sometimes the best one can do is insure that the significance is more than some amount, as limited by MC statistics. The possibility that even the simulation does not give an accurate representation of the tail distribution should be considered as well.

### 6.3.3 Systematic unknowns and nuisance parameters

Often a measurement of interest depends on the measurement of auxillary quantities, such as backgrounds and efficiencies. These are called “nuisance parameters”; we don’t really care about them, but they are needed for the result of interest. The uncertainties due to the uncertain values of the nuisance parameters are treated as “systematic errors” and quoted separately. We might see a branching fraction measurement quoted as:  $B(\text{new effect}) = 10 \pm 1 \pm 5$ , where the  $\pm 1$  is the “statistical error”, and the  $\pm 5$  is the “systematic error”. This may represent a highly significant observation or not, depending on the source of the systematic uncertainty.

Perhaps the systematic uncertainty is from an uncertainty in the background subtraction. In this case, it is an additive uncertainty – the significance of the deviation from zero branching fraction is only “ $2\sigma$ ”. Alternatively, the systematic error could be due to uncertainty in the signal efficiency. In this case, it is a multiplicative uncertainty, and has little bearing on the significance of the deviation from zero, since  $5 \pm 0.5$  is as significant as  $10 \pm 1$ .

A difficulty is that even if the sampling distribution for the nuisance param-

eter estimator is known (except for the value of the nuisance parameter), it is not easy in general to derive exact  $P$ -values in a lower-dimensional parameter space. A notable exception is the multivariate normal distribution, where it is possible. However, not all of our measurements are normal to a good enough approximation, and the problem becomes more difficult. An approach to coping with this difficulty is to compute the distribution of the significance statistic assuming other true values of the nuisance parameter(s), besides the best estimate value.

Still more problematic is when the distribution of the nuisance estimators is not even known. An example here is “theoretical” uncertainties, for which no distribution exists (in the frequency sense), only some judgement. In this case, it may be important to repeat the significance computation for the range of plausible theoretical values. The worst-case value could then be used in quoting the significance, or alternatively the dependence of the significance on the uncertain parameter(s) may be described.

### 6.3.4 The Pitfalls

There are a variety of pitfalls that we encounter in trying to evaluate whether we have observed something new or not. All share a common theme: They involve a lack of understanding of the sampling distribution. We have already touched on the problem of modeling the tails of the distribution according to the null hypothesis and the problem of systematic uncertainties. We discuss some additional issues here.

#### The Stopping Problem

There is a strong tendency to work on an analysis until we are convinced that we got it “right”, then we stop. We return to the example of Section 5.5 to illustrate how this can lead to erroneous hypothesis tests.

Recall that this example involved an experiment to measure a parameter  $\theta$  corresponding to the mean of a Gaussian distribution of standard deviation one. The experimenter has a prejudice that  $\theta > 1$ , and takes a second sample if his first sample does not support this prejudice. The sample mean is used as an estimate of  $\theta$ .

In terms of the random variables  $m$  and  $n$  (the number of samples), the sampling distribution of the experiment is:

$$f(m, n; \theta) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(m-\theta)^2}, & n = 1, m > 1 \\ 0, & n = 1, m < 1 \\ \frac{1}{\pi} e^{-(m-\theta)^2} \int_{-\infty}^1 e^{-(x-m)^2} dx, & n = 2. \end{cases} \quad (6.21)$$

This distribution is shown in Fig. 6.2a.

The likelihood function, as a function of  $\theta$ , has the shape of a normal distribution, given any experimental result. The peak is at  $\theta = m$ , hence the sample mean is the maximum likelihood estimator for  $\theta$ . However, in spite of the form

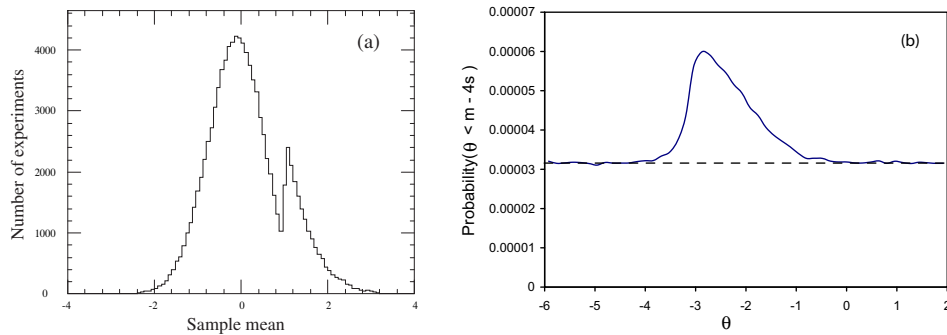


Figure 6.2: (a) Histogram of the sampling distribution for the sample mean, according to Eqn. 6.21, for  $\theta = 0$ . (b) The probability to observe a “ $4\sigma$ ” fluctuation in the stopping problem example as a function of the distribution parameter  $\theta$ . The dashed line shows the probability for a  $4\sigma$  fluctuation of a normal distribution.

of the likelihood, the sample mean is not sampled from a normal distribution. This is a not widely-appreciated distinction: The form of the likelihood function does not imply the form of the sampling distribution. In frequency statistics, it is the sampling distribution that is crucial to computing probabilities. The common suggestion that showing the likelihood function provides a complete picture is not generally valid.

The implication for computing significance is that treating the sample mean as sampled from a normal distribution will give an incorrect result. This is illustrated in Fig. 6.2b, in which the probability of an apparent  $4\sigma$  fluctuation is plotted as a function of the value of the parameter  $\theta$ . Depending on  $\theta$ , the actual probability may be as much as twice as likely as the experimenter thinks.

In our scenario we think we are taking  $n$  samples from a normal distribution, and make probability statements (about significance) according to a normal distribution for the sample mean. We get an erroneous result because of the mistake in the distribution.

There is a related phenomenon: First “observations” of a new process tend to be biased high. Especially in exploratory investigations, null results tend not to be reported quantitatively. In addition, many people mix the question of “significance” with the choice of confidence interval (one- or two-sided) to quote to describe the result. The first claim of a new effect will preferentially occur with a positive fluctuation. This suggests the importance, in summarizing knowledge, of averaging in earlier null results to get best estimates. It further calls for reporting of results permitting averaging, rather than only quoting upper limits. In any event, avoidance of bias calls for designing the complete experiment, including how the results will be quoted, prior to carrying it out.

### The Exploratory Analysis

In a typical exploratory search for unexpected phenomena, a histogram is made and examined for unexplained structure. We call this sort of search “bump hunting”. When we see an interesting structure (e.g., Fig. 6.3a), the first question is usually “Is it significant?”.

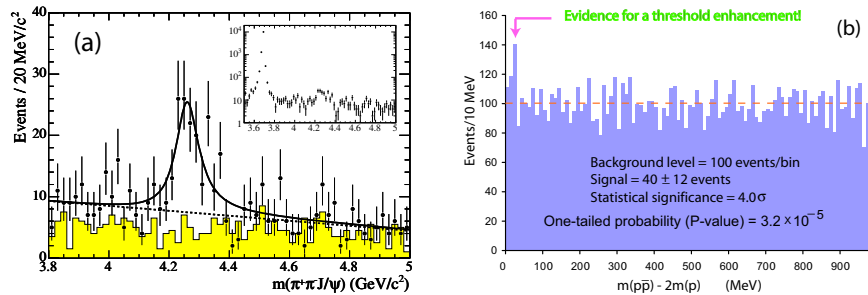


Figure 6.3: Bump hunting: (a) The  $\pi^+\pi^-J/\psi$  mass spectrum in the initial state radiation process  $e^+e^- \rightarrow \gamma\pi^+\pi^-J/\psi$  from the BaBar experiment.  $P(H_0) \sim 10^{-11} - 10^{-16}$  [3] (b) A simulated spectrum of the  $p\bar{p}$  mass from threshold to 1 GeV. The dotted line indicates the estimated background level.

Quite often, the analysis in a bump hunt is developed concurrently with looking at the data. In this case, it is impossible to compute the significance of an effect. The problem, once again, is that we don’t know the sampling distribution under the null hypothesis, and thus we cannot compute probabilities.

Recall the example of Fig. 6.1. We assume that we know the (flat) sampling distribution under the null hypothesis, and that we are looking for an effect described by a Gaussian of mean zero and standard deviation one. As long as these assumptions are correct, we make correct inferences of significance.

Now consider the example in Fig. 6.3b. It shows a “mass distribution” plotted in bins that are larger than the resolution. The bump hunting here consists of looking for one-bin peaks, as signs of narrow structure. The observed histogram looks flat except for a bump near threshold. It is desired to know the significance of this peak.

The usual approach to computing the significance is to estimate the background level under the peak, and then compute, under the null hypothesis of no signal, the probability of a fluctuation to the observed level. Here, the background level is estimated from the sidebands to be 100 events, with negligible statistical uncertainty. The excess in the peak (“signal”) is  $40 \pm 12$  events. With a background standard deviation of 10 events (the counts in each bin were generated from a Gaussian distribution approximating a Poisson), the probability of a  $\geq 40$  event upward fluctuation is  $P = 3.2 \times 10^{-5}$ .

So, our result looks very significant. But did we really model the sampling distribution properly? No. First, we admit that we would have accepted a

fluctuation as interesting in any bin. Thus, we should divide our probability estimate by 100. Our probability is no longer so spectacular, although it is still small. However, our second admission is that this was not the only histogram we looked at. Perhaps we varied the cuts, or looked at other masses, until we found something interesting.<sup>2</sup> We also might have considered a feature wider than one bin to be interesting.

The significances we quote for bumps in exploratory analyses usually aren't  $P$ -values. We still give these numbers, generally as a number of "sigma"s. What we really mean is: "If I had done a controlled analysis, and had been interested in the observed values for the mean and width, then the null hypothesis would require a fluctuation of this number of standard deviations of a Gaussian distribution to produce a bump as large as I see." With this understanding, it is perhaps not utterly useless, as we can interpret it in the context of experience. However, we are occasionally reminded that experience is that we may be fooled. An interesting collection of some historical anecdotes may be found in [4].

One way we mitigate this pitfall is "conservatism", applied to the interpretation rather than to the computation. That is, we make up for the unreliability of our probability calculation by requiring very great computed significance before claiming a new effect. Thus, " $3\sigma$ " isn't regarded as especially unlikely, even though the implied probability under the null hypothesis is only 0.13% for a one-tailed test.

Another common approach is to try to take into account the effective number of bins that have been looked at by applying a "trials factor" to the  $P$ -value estimate. This is what we were doing when we suggested multiplying our  $P$ -value by 100, above, to take into account the fact that we would have accepted a fluctuation as interesting anywhere in the histogram. However, this is only a partial mitigation, as we admitted that in fact we looked at an indeterminate number of histograms.

## 6.4 Blinding the Analysis

An important technique to avoid the pitfalls in the previous section is to "blind" the analysis. The goal of blinding is to ensure a known sampling distribution under the null hypothesis. Several approaches to blinding exist, which may be chosen as appropriate to the problem (for a review, see Ref. [5]). We list several used in physics:

1. Hide the answer in a box, don't look inside until the analysis procedure is defined.
2. You can look, but keep the answer hidden via an unknown transformation, again until the procedure is fixed.

---

<sup>2</sup>In this example, I re-generated the simulated experiment until I got a " $4\sigma$ " effect somewhere, and then stopped. The sampling distribution was  $N(100, 10)$  for each bin.

3. Obscure the real data. For example, let the data be visible, but add simulated signal to it, to be removed only when the analysis methodology is final.
4. Design the analysis on a dataset that will be discarded.
5. “Divide and conquer”: The idea is to separate the analysis into pieces that will be combined to get the quantity of interest only when the pieces are final. A nice example of this methodology is the muon anomalous magnetic moment measurement ( $g_\mu - 2$ ) [6]. In this experiment, teams independently measured the magnetic field and the muon precession; neither alone gives a clue to the value of  $g_\mu - 2$ .

We’ll elaborate on a couple of these to give the flavor.

### 6.4.1 Blinding the box

In this approach, the analysis is designed with the help of simulations, control samples, and sidebands. The data that will be fit for the result is kept invisible, until the analysis is deemed fixed. An illustration of this is the measurement of the process  $B^\pm \rightarrow K^\pm e^+ e^-$  in BaBar.[7] Figure 6.4a shows a scatterplot of simulated signal data in the relevant two kinematic variables. The “large sideband” region is the only visible region of the data prior to fixing the analysis. Note that all of the data that will be used in the fit for the results is kept blind, including the (smaller) sidebands around the signal region. The final unblinded data is shown in Fig. 6.4b.

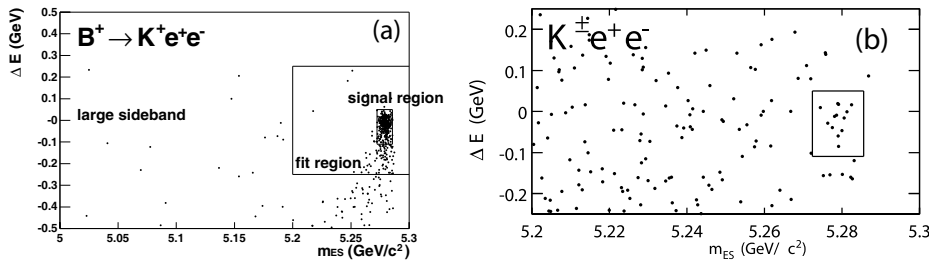


Figure 6.4: Illustration of the blinding methodology in an early BaBar  $B \rightarrow Kee$  analysis [7]. (a) Signal Monte Carlo in the two relevant kinematic variables, showing the large sideband region, the fit region, and the region where most of the signal is concentrated. (b) Data after unblinding, in the fit region. The small box is the signal region as defined in the left plot.

### 6.4.2 Hiding the answer

Sometimes we can let the data be visible in graphical form, but obscure the numerical result of interest. For example, a hidden, perhaps random, offset may be applied to the real answer to prevent it from being seen during the analysis design. An example is the BaBar  $CP$ -violation measurement.[5] Fig. 6.5 shows the  $\Delta t$  distributions, in which  $CP$  violation appears as a difference between the  $B^0$  and  $\bar{B}^0$  “tags” and as an asymmetry about 0. These asymmetries are obscured from view while the analysis is being tuned, by applying an unknown transformation to the two  $\Delta t$  distributions.

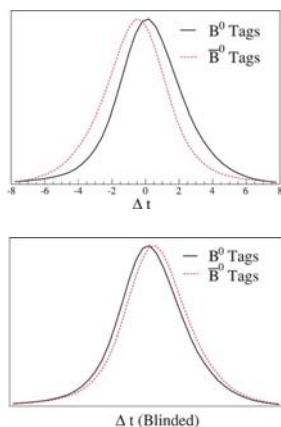


Figure 6.5: Example of blinding an analysis by hiding the answer via a transformation on the data. [5] Top: unblinded result; Bottom: blinded data.

### 6.4.3 Blinding a Bump Hunt

In section 6.3.4 we discussed the difficulty of evaluating significance in a “bump hunting” exploratory analysis. For example, consider the search for structure in some invariant mass spectrum. Can the methods of blind analysis be applied to this situation so that meaningful measures of significance may be computed?

There is no difference in principle in this case from our other blinding examples. The difficulty arises from the broad range of phenomena one may be interested in discovering. For example in a particle physics experiment, we might not know the location, width, kinematic regime, or even the particle content of interest. The analysis must encompass all interesting possibilities.

One relatively straightforward approach to this problem is to divide the data up into two samples, one used to design the analysis and one to use for the results. The dataset used for the analysis design is discarded once the analysis is designed. This costs sensitivity, but can be an effective method. Note that the details of how one uses the design sample don’t much matter; one can even tune cuts in ways that enhance apparent peaks. If there is actually no effect, then

the most probable result will be that nothing significant is observed when the blinded sample is observed. Figure 6.6 provides an illustration of what happens when an earlier example is followed with a second dataset. Of course, systematic effects will show up in both samples, but that is another matter.

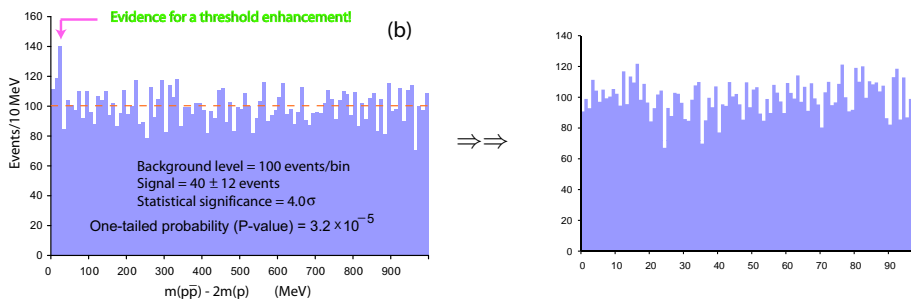


Figure 6.6: Left: The result of an exploratory analysis on a dataset; Right: The same analysis, as fixed on the first dataset, applied to a new dataset.

## 6.5 Goodness-of-Fit

A very common question that is asked is, with variations: “Does my model provide a good description of my data?” We refer to tests addressing this question as “goodness-of-fit” tests. A large number of such tests exist; a consequence of the fact that there is no perfect general goodness-of-fit test. Given a dataset generated under null hypothesis, it is generally possible to find a test which rejects the null hypothesis (and choosing the test after you see the data is dangerous). Alternatively, given a dataset generated under alternative hypothesis, it is possible to find a test for which the null passes (i.e., should put thought into what you want to test for).

The more specific a question you can ask, the better (meaning, loosely, more powerful) a test you can choose. Very popular tests are  $\chi^2$ , likelihood ratio, and Kolmogorov-Smirnov [See section 6.7]. Monte Carlo simulation may be used to evaluate the distribution of the test statistic. But some thought should go into whether the test is really sensitive to the question. Indeed, statisticians are not overly-fond of the tests just mentioned, and one can often do better. Section 6.7 provides a case study comparing several “general-purpose” goodness-of-fit tests.

### 6.5.1 Counting Degrees of Freedom

There is often confusion about how many degrees-of-freedom to use in tests such as the  $\chi^2$  test. In fact, in some common situations, the answer is complicated.

The following situation arises with some frequency (with variations). We do two fits to the same dataset (say a histogram with  $N$  bins): Fit  $A$  has  $n_A$

parameters, with  $\chi_A^2$ . Fit  $B$  has a subset  $n_B$  of the parameters in fit  $A$ , with  $\chi_B^2$ , where the  $n_A - n_B$  other parameters (call them  $\theta$ ) are fixed at zero. What is the distribution of  $\Delta\chi^2 = \chi_B^2 - \chi_A^2$ ?

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

$$\Delta\chi^2 \equiv \chi_B^2 - \chi_A^2$$

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

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

In this case, the  $\Delta\chi^2$  is distributed according to a  $\chi^2$  (DOF =  $n_A - n_B$ ) distribution under the following conditions:

1. Parameter estimates in  $\lambda$  are consistent (converge to the correct values) under  $H_0$ .
2. Parameter values in the null hypothesis are interior points of the maintained hypothesis (union of  $H_0$  and  $H_1$ ).
3. There are no additional nuisance parameters under the alternative hypothesis.

Unfortunately, commonly-encountered situations may violate these requirements.

For example, consider fitting a spectrum to decide whether a bump is significant or not. In this case, the parameter of greatest interest is the signal strength. We compare fits with and without a signal component to estimate significance of the signal. Under the null hypothesis, the signal is zero. Under the alternative hypothesis, the signal is non-zero, or possibly greater than zero. If the signal fit has, e.g., a parameter for location (maybe a mass), this constitutes an additional nuisance parameter under the alternative hypothesis. If the fit for signal constrains the signal yield to be non-negative, this violates the interior point requirement. In such cases, the number of degrees of freedom is not a single number. Let us illustrate this by example.

In Figs. 6.7 and 6.8 are shown the results of various fits to two possible mass spectra, one consisting only of background, and one with a signal component.<sup>3</sup> We compare each alternative hypothesis fit with the null hypothesis fit, calculate the change in  $\chi^2$ . The  $\chi^2$  is calculated as the value of  $-2 \ln L$  at the maximum likelihood, but it is readily checked that this is consistent with an evaluation of the sum of the squared deviations between the fit and the histogram bins.

In order to obtain the distributions of the  $\Delta\chi^2$  statistics, each “experiment” is simulated many times, under the null hypothesis. The results are shown in Fig. 6.9. When the location parameter is fixed in the fit, and the signal yield is allowed to be positive or negative, the distribution follows a  $\chi^2$  for 1 DOF, reflecting the difference of 1 parameter between the fits, with all of the above conditions satisfied. However, when the fit constrains the yield to be

<sup>3</sup>The R function `optim` is used for the fits here. It is a general-purpose minimizer that works in a multi-dimensional parameter space.

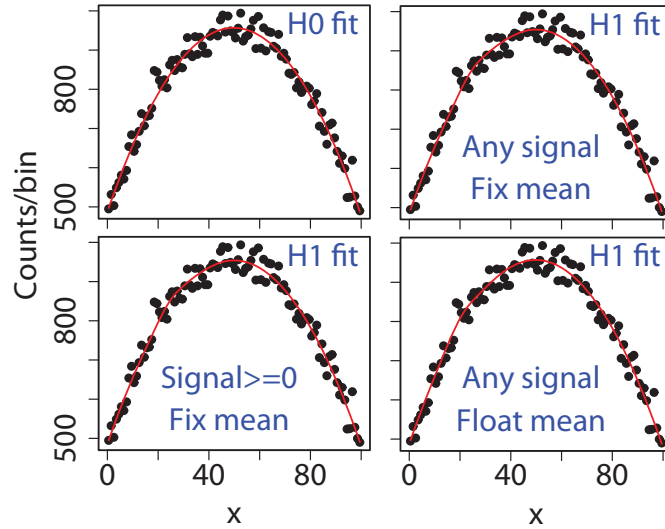


Figure 6.7: Examples of maximum likelihood fits to a distribution generated under the null hypothesis. Upper left: Fit to the null hypothesis; Upper right: Fit to the alternative hypothesis with unconstrained signal yield and fixed location; Lower left: Fit to the alternative hypothesis with signal yield constrained to be non-negative and fixed location; Lower right: Fit to the alternative hypothesis with unconstrained signal yield and location.

non-negative, the distribution becomes more sharply peaked towards zero than a  $\chi^2$  for 1 DOF. When both the signal yield and location are unconstrained, the distribution is somewhere between the curves for 1 and 2 DOF, and does not follow the 2 DOF that is often assumed. This is because the location parameter is an additional nuisance parameter under the alternative hypothesis.

## 6.6 Consistency of Correlated Results

We sometimes encounter the question of whether a new analysis is consistent with an old analysis. Often, the new analysis is a combination of additional data plus changed (improved...) analysis of original data. The stickiest issue is handling the correlation in testing for consistency in the overlapping data. We'll assume here that the data are in the form of “events” that may be complicated objects consisting of many random variables each. It should be understood that statistical differences can arise even comparing results based on the same events, due to differences in the analysis of those events.

Given a sampling  $\hat{\theta}_1, \hat{\theta}_2$  from a bivariate normal distribution  $N(\theta, \sigma_1, \sigma_2, \rho)$ , with  $\langle \hat{\theta}_1 \rangle = \langle \hat{\theta}_2 \rangle = \theta$ , the difference  $\Delta\theta \equiv \hat{\theta}_2 - \hat{\theta}_1$  is  $N(0, \sigma)$ -distributed with  $\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$ .

If the correlation is unknown, all we can say is that the variance of the

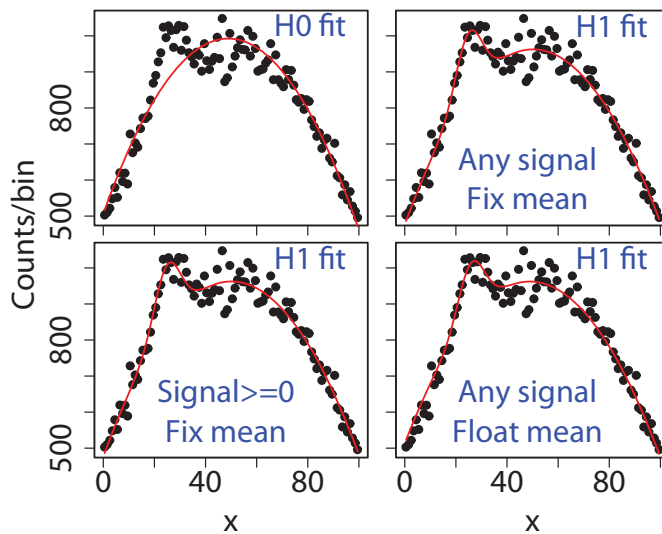


Figure 6.8: Examples of maximum likelihood fits to a distribution generated under the alternative hypothesis. Upper left: Fit to the null hypothesis; Upper right: Fit to the alternative hypothesis with unconstrained signal yield and fixed location; Lower left: Fit to the alternative hypothesis with signal yield constrained to be non-negative and fixed location; Lower right: Fit to the alternative hypothesis with unconstrained signal yield and location.

difference is in the range  $(\sigma_1 - \sigma_2)^2 \dots (\sigma_1 + \sigma_2)^2$ . If we at least believe  $\rho \geq 0$  then the maximum variance of the difference is  $\sigma_1^2 + \sigma_2^2$ .

### 6.6.1 Consistency – Simple example of two analyses on same events

Suppose we measure a neutrino mass,  $m$ , in a sample of  $n = 10$  independent events. The measurements are  $x_i, i = 1, \dots, 10$ . Assume the sampling distribution for  $x_i$  is  $N(m, \sigma_i)$ .

We may form unbiased estimator,  $\hat{m}_1$ , for  $m$ :

$$\hat{m}_1 = \frac{1}{n} \sum_{i=1}^n x_i \pm \sqrt{\frac{1}{n^2} \sum_{i=1}^n \sigma_i^2}.$$

The result (from a MC) is  $\hat{m}_1 = 0.058 \pm 0.039$ .

Then we notice that we have some further information which might be useful: we know the experimental resolutions,  $\sigma_i$  for each measurement. We form another unbiased estimator,  $\hat{m}_2$ , for  $m$ :

$$\hat{m}_2 = \frac{\sum_{i=1}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}} \pm \frac{1}{\sqrt{\sum_{i=1}^n \frac{1}{\sigma_i^2}}}.$$

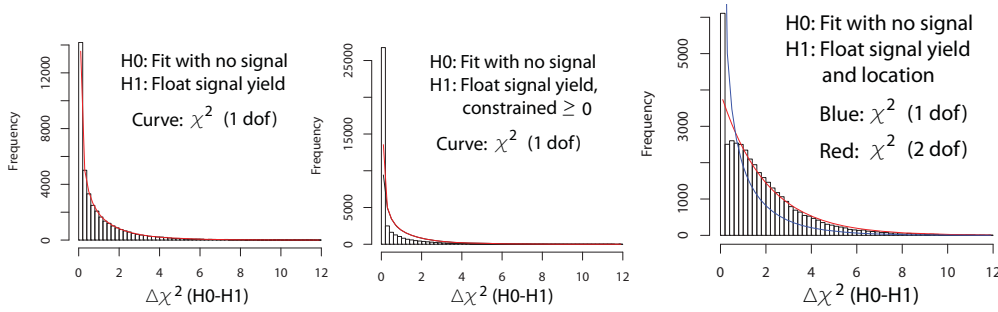


Figure 6.9: Distributions of  $\Delta\chi^2$  for different alternative hypotheses fits. Left: Alternative fit for a fixed location, and signal yield unconstrained. The red curve is the  $\chi^2$  distribution for 1 DOF; Middle: Alternative fit for a fixed location, but with signal yield constrained to be positive. The red curve is the  $\chi^2$  distribution for 1 DOF; Right: Alternative fit for a variable location, and signal yield unconstrained. The blue curve is the  $\chi^2$  distribution for 1 DOF, and the red curve is for 2 DOF.

The result (from the same MC) is  $\hat{m}_1 = 0.000 \pm 0.016$ .

The results are certainly correlated, so the question of consistency arises (we know the error on the difference is between 0.023 and 0.055). Both  $\hat{m}_1$  and  $\hat{m}_2$  are normally distributed in this example, so the  $P$ -value for consistency may be computed. Fortunately, we have enough information to compute the correlation coefficient, or more directly compute the  $\chi^2$  to be used in a  $\chi^2$  test for consistency. That is, we know the moment matrix. As the reader is encouraged to demonstrate, it is:

$$M = \begin{pmatrix} \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 & 1 / \sum_{i=1}^n \frac{1}{\sigma_i^2} \\ 1 / \sum_{i=1}^n \frac{1}{\sigma_i^2} & 1 / \sum_{i=1}^n \frac{1}{\sigma_i^2} \end{pmatrix}.$$

Note that the form of this matrix is indicative of the fact that all of the information in the first analysis is used in the second analysis. The difference between the two estimators is distributed according to the normal distribution with standard deviation:

$$\sigma_{\Delta\hat{m}} = \sqrt{\frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 - \frac{1}{\sum_{i=1}^n \frac{1}{\sigma_i^2}}}.$$

In this example, the difference between the results is  $0.058 \pm 0.036$ , where the 0.036 error includes the correlation ( $\rho = 0.41$ ).

## 6.7 Case Study: Testing Consistency of Two Histograms

As a thorough illustration of the choices, considerations, and techniques of performing hypothesis tests, we consider a case study: testing two histograms for consistency. Along the way, we introduce and contrast several standard tests. However, this is by no means an exhaustive set.

Several approaches to testing the hypothesis that two histograms are drawn from the same distribution are investigated. We note that single-sample continuous distribution tests may be adapted to this two-sample grouped data situation. The difficulty of not having a fully-specified null hypothesis is an important consideration in the general case, and care is required in estimating probabilities with “toy” Monte Carlo simulations. The performance of several common tests is compared; no single test performs best in all situations.

Each histogram represents a sampling from a multivariate Poisson distribution. The question is whether the means are bin-by-bin equal between the two distributions. Or, if we are only interested in “shape”, are the means related by the same scale factor for all bins? We investigate this question in the context of frequency statistics.

For example, consider Fig. 6.10. Are the two histograms consistent or can we conclude that they are drawn from different distributions?

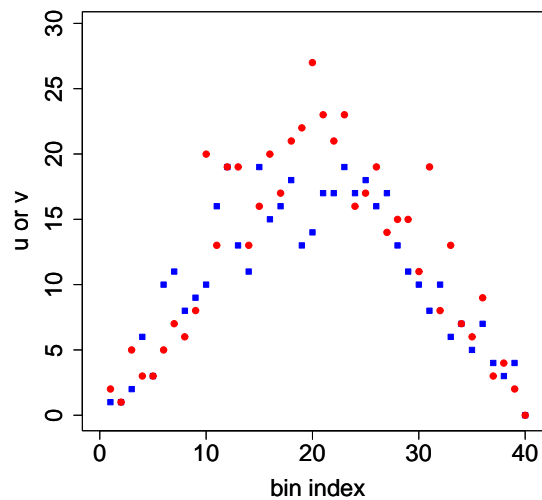


Figure 6.10: Two histograms (blue squares and red circles) to be compared for consistency.

There are at least two variants of interest to this question:

1. We wish to test the hypothesis:

- $H_0$ : The means of the two histograms are bin-by-bin equal, against
- $H_1$ : The means of the two histograms are not bin-by-bin equal.

2. We wish to test the hypothesis:

- $H'_0$ : The densities of the two histograms are bin-by-bin equal, against
- $H'_1$ : The densities of the two histograms are not bin-by-bin equal.

In the second case, the relative normalization of the two histograms is not an issue: we only compare the shapes.

It may be noted that there are a large variety of tests that attempt to answer the question of whether a given dataset is consistent with having been drawn from some specified continuous distribution. These tests may typically be adapted to address the question of whether two datasets have been drawn from the same continuous distribution, often referred to as “two-sample” tests. These tests may further be adapted to the present problem, that of determining whether two histograms have the same shape. This situation is also referred to as comparing whether two (or more) rows of a “table” are consistent. The datasets of this form are also referred to as “grouped data”.

Although we keep the discussion focussed on the comparison of two histograms, it is worth remarking that many of the observations apply also to other situations, such as the comparison of a histogram with a model prediction.

### 6.7.1 Notation

We assume that we have formed our two histograms with the same number of bins,  $k$ , with identical bin boundaries. The bin contents of the “first” histogram are given by realization  $u$  of random variable  $U$ , and of the second by realization  $v$  of random variable  $V$ . Thus, the sampling distributions are:

$$P(U = u) = \prod_{i=1}^k \frac{\mu_i^{u_i}}{u_i!} e^{-\mu_i}, \quad (6.22)$$

$$P(V = v) = \prod_{i=1}^k \frac{\nu_i^{v_i}}{v_i!} e^{-\nu_i}, \quad (6.23)$$

where the vectors  $\mu$  and  $\nu$  are the mean bin contents of the respective histograms.

We define:

$$N_u \equiv \sum_{i=1}^k U_i, \quad \text{total contents of first histogram,} \quad (6.24)$$

$$N_v \equiv \sum_{i=1}^k V_i, \quad \text{total contents of second histogram,} \quad (6.25)$$

$$\mu_T \equiv \langle N_u \rangle = \sum_{i=1}^k \mu_i, \quad (6.26)$$

$$\nu_T \equiv \langle N_v \rangle = \sum_{i=1}^k \nu_i, \quad (6.27)$$

$$t_i \equiv u_i + v_i, \quad i = 1, \dots, k. \quad (6.28)$$

We are interested in the power of a test, at any given confidence level. The power is the probability that the null hypothesis is rejected when it is false. Of course, the power depends on the true sampling distribution. In other words, the power is one minus the probability of a Type II error. The confidence level is the probability that the null hypothesis is accepted, if the null hypothesis is correct. Thus, the confidence level is one minus the probability of a Type I error. In physics, we usually 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.

Note that we are dealing with discrete distributions here, and exact statements of frequency are problematic, though not impossible. Instead of attempting to construct exact statements, our treatment of the discreteness will be such as to err on the "conservative" side. By "conservative", we mean that we will tend to accept the null hypothesis with greater than the stated probability. It is important to understand that this is not always the "conservative" direction, for example it could mislead us into accepting a model when it should be rejected.

We will drop the distinction between the random variable (upper case symbols  $U$  and  $V$ ) and a realization (lower case  $u$  and  $v$ ) in the following, but will point out where this informality may yield confusion.

The computations in this note are carried out in the framework of the R statistics package[2].

### 6.7.2 Large Statistics Case

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

Under  $H_0$ ,

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

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, as noted above. Let the difference in the contents of bin  $i$  between the two histograms be:

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

6.7. CASE STUDY: TESTING CONSISTENCY OF TWO HISTOGRAMS 127

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

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

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 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.

If it is desired to only compare shapes, then the suggested algorithm is to scale both histogram bin contents:

1. Let

$$N = 0.5(N_u + N_v).$$

Scale  $u$  and  $v$  according to:

$$u_i \rightarrow u'_i = u_i(N/N_u) \tag{6.29}$$

$$v_i \rightarrow v'_i = v_i(N/N_v). \tag{6.30}$$

2. Estimate  $\sigma_i^2$  with:

$$\hat{\sigma}_i^2 = \left( \frac{N}{N_u} \right)^2 u_i + \left( \frac{N}{N_v} \right)^2 v_i.$$

Table 6.1: Results of tests for consistency of the two datasets in Fig. 6.10. The tests below the  $\chi^2$  lines are described in Section 3.

Type of test	$T$	NDOF	$P(\chi^2 > T)$	$P$ -value
$\chi^2$ Absolute comparison	29.8	40	0.88	0.86
$\chi^2$ Shape comparison	24.9	39	0.96	0.95
Likelihood Ratio Shape comparison	25.3	39	0.96	0.96
Kolmogorov-Smirnov Shape comparison	0.043	39	NA	0.61
Bhattacharyya Shape comparison	0.986	39	NA	0.97
Cramér-Von-Mises Shape comparison	0.132	39	NA	0.45
Anderson-Darling Shape comparison	0.849	39	NA	0.45
Likelihood value shape comparison	79	39	NA	0.91

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

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

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

Due to the presence of bins with small bin counts, we might not expect this method to be especially good for the data in Fig. 1, but we can try it anyway. Table 6.1 gives the results of applying this test, both including the normalization and only comparing shapes.

In the column labeled “ $P$ -value” an attempt is made to compute (by simulation) a more reliable estimate of the probability, under the null hypothesis, that a value for  $T$  will be as large as that observed. This may be compared with the  $P(\chi^2 > T)$  column, which is the probability assuming  $T$  follows a  $\chi^2$  distribution with NDOF degrees of freedom.

Note that the absolute comparison yields slightly poorer agreement between the histograms than the shape comparison. The total number of counts in one dataset is 492; in the other it is 424. Treating these as samplings from a normal distribution with variances 492 and 424, we find a difference of 2.2 standard deviations or a two-tailed  $P$ -value of 0.025. This low probability is diluted by the bin-by-bin test. Using a bin-by-bin test to check whether the totals are consistent is not a powerful approach. In fact, the two histograms were generated with a 10% difference in expected counts.

The evaluation by simulation 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. Thus, we must estimate these means. The procedure adopted here is to use the maximum likelihood estimators (see below)

for the mean numbers, in the null hypothesis. We'll have further discussion of this procedure below – it does not always yield valid results.

### 6.7.3 General Case

If the bin contents are not necessarily large, then the normal approximation may not be good enough. There are various approaches we could take in this case. We'll discuss and compare several possibilities.

#### Combining Bins

A simple approach is to combine bins until the normal approximation is good enough. In many cases this doesn't lose too much statistical power. It may be necessary to check with simulations that probability statements are valid. Figure 6.11 shows the results of this approach on the data in Figure 6.10, as a function of the minimum number of events per bin. The comparison being made is for the shapes. The algorithm is to combine corresponding bins in both histograms until both have at least “`minBin`” counts in each bin.

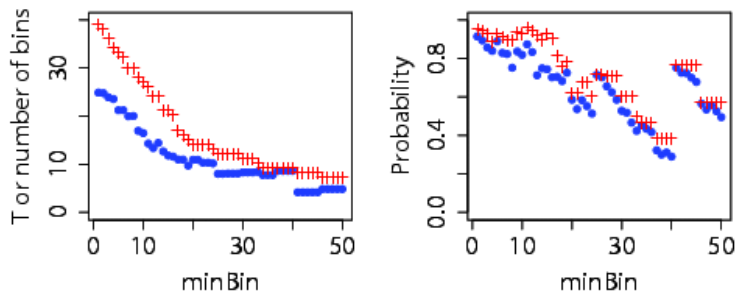


Figure 6.11: Left: The blue dots show the value of the test statistic  $T$ , and the red pluses shows the number of histogram bins for the data in Fig. 1, as a function of the minimum number of counts per histogram bin. Right: The  $P$ -value for consistency of the two datasets in Fig. 6.10. The red pluses show the probability for a chi-square distribution, and the blue circles show the probability for the actual distribution, with an estimated null hypothesis.

#### Testing for Equal Normalization

An alternative is to work with the Poisson distributions. Let us separate the problem of the shape from the problem of the overall normalization. In the case of testing equality of overall normalization, there is a well-motivated choice for the test statistic, even for low statistics.

To test the normalization, we simply compare totals over all bins between the two histograms. Our 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 thus interested in the difference between the two means; the sum is effectively a nuisance parameter. That is, we are interested in

$$\begin{aligned} P(N_v | N_u + N_v = N) &= \frac{P(N | N_v) P(N_v)}{P(N)} & (6.31) \\ &= \frac{\mu_T^{N-N_v} e^{-\mu_T} \nu_T^{N_v} e^{-\nu_T}}{(N-N_v)! N_v!} \bigg/ \frac{(\mu_T + \nu_T)^N e^{-(\mu_T + \nu_T)}}{N!} \\ &= \binom{N}{N_v} \left( \frac{\nu_T}{\mu_T + \nu_T} \right)^{N_v} \left( \frac{\mu_T}{\mu_T + \nu_T} \right)^{N-N_v}. & (6.32) \end{aligned}$$

This probability now permits us to construct a uniformly most powerful test of our hypothesis [8]. Note that it is simply 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:

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

For our example, with  $N = 916$  and  $N_v = 424$ , the  $P$ -value is 0.027, assuming a two-tailed probability is desired. This may be compared with our earlier estimate of 0.025 in the normal approximation. Note that for our binomial calculation we have “conservatively” included the endpoints (424 and 492). If we try to mimic more closely the normal estimate by subtracting one-half the probability at the endpoints, we obtain 0.025, essentially the normal number we found earlier. The `dbinom` function [9] in the R package has been used for this computation.

### 6.7.4 Shape Comparison Statistics

There are many different possible statistics for comparing the shapes of the histograms. We investigate several choices. Table 6.1 summarizes the result of each of these tests applied to the example in Fig. 6.10. We list the statistics here, and discuss performance in the following sections.

#### Chi-square test for shape

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

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

## 6.7. CASE STUDY: TESTING CONSISTENCY OF TWO HISTOGRAMS 131

If  $u_i = v_i = 0$ , the contribution to the sum from that bin is zero. We have already discussed application of this statistic to the example of Fig. 6.10.

### Geometric test for shape

Another test statistic we could try may be motivated from a geometric perspective. We consider the bin contents of a histogram to define a vector in a  $k$ -dimensional space. If two such vectors are drawn from the same distribution (the null hypothesis), then they will tend to point in the same direction (we are not interested in the lengths of the vectors here). Thus, if we represent each histogram as a unit vector with components:

$$\{u_1/N_u, \dots, u_k/N_u\}, \text{ and } \{v_1/N_v, \dots, v_k/N_v\},$$

we may form the 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”. We’ll refer to it as the “BDM” statistic for short. We assume that neither histogram is empty for this statistic. All vectors lie in the positive direction in all coordinates, so there is no issue with taking the square root.

It may be noticed that 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.

We apply this formalism to the example in Fig. 6.10. The resulting terms in the sum over bins are shown in Fig. 6.12. The sum over bins gives 0.986 (See Table 6.1 for a summary). According to our estimated distribution of this statistic under the null hypothesis, this gives a  $P$ -value of 0.97, similar to the  $\chi^2$  test result.

### Kolmogorov-Smirnov test

Another approach to a shape test may be based on the Kolmogorov-Smirnov (KS) idea. Recall that the idea of the KS test is to estimate the maximum difference between observed and predicted cumulative distribution functions (CDFs) and compare with expectations. We may adapt this idea to the present case. It should be remarked that if we have the actual data points from which the histograms are derived, then we may use the Kolmogorov-Smirnov (“KS”) procedure directly on those points. This would incorporate additional information and yield a potentially more powerful test. However, if the bin widths are small compared with possible structure it may be expected to not make much difference.

We modify the KS statistic to apply to comparison of histograms as follows. We assume that neither histogram is empty. Form the “cumulative distribution histograms” according to:

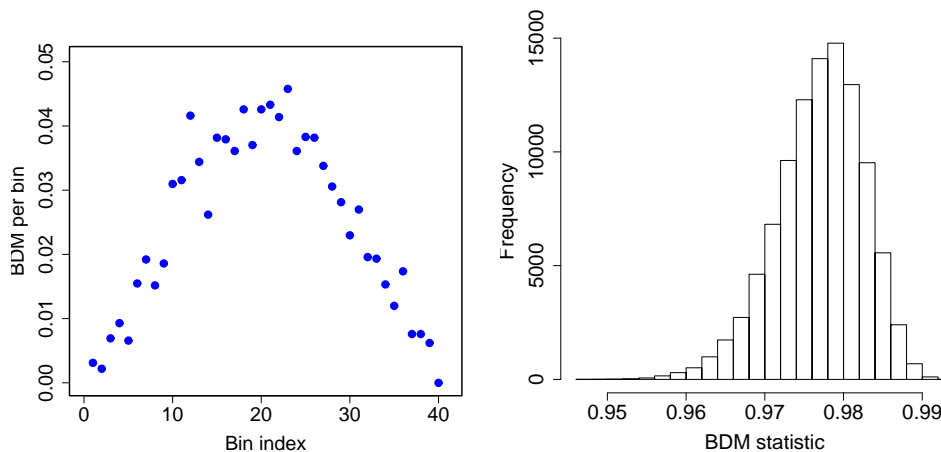


Figure 6.12: Left: Bin-by-bin contributions to the geometric (“BDM”) test statistic for the example of Fig. 6.10. Right: Estimated distribution of the BDM statistic for the null hypothesis in the example of Fig. 6.10.

$$u_{ci} = \sum_{j=1}^i u_j / N_u \quad (6.33)$$

$$v_{ci} = \sum_{j=1}^i v_j / N_v. \quad (6.34)$$

Then compute the test statistic:

$$T_{KS} = \max_i |u_{ci} - v_{ci}|.$$

Test statistics may also be formed for one-tail tests, but we consider only the two-tail test here.

We apply this formalism to the example in Fig. 6.10. The bin-by-bin distances are shown in Fig. 6.13. The maximum over bins gives 0.043 (See Table 6.1 for a summary). According to our estimated distribution of this statistic under the null hypothesis, this gives a  $P$ -value of 0.61, somewhat smaller than for the  $\chi^2$  test result, but still indicating consistency of the two histograms. Note that the KS test tends to emphasize the region near the peak of the distribution, that is the region where the largest fluctuations are expected in Poisson statistics.

### 6.7.5 Cramér-von-Mises test

Somewhat similar to the Kolmogorov-Smirnov test is the Cramér-von-Mises (CVM) test. The idea in this test is to add up the squared differences between

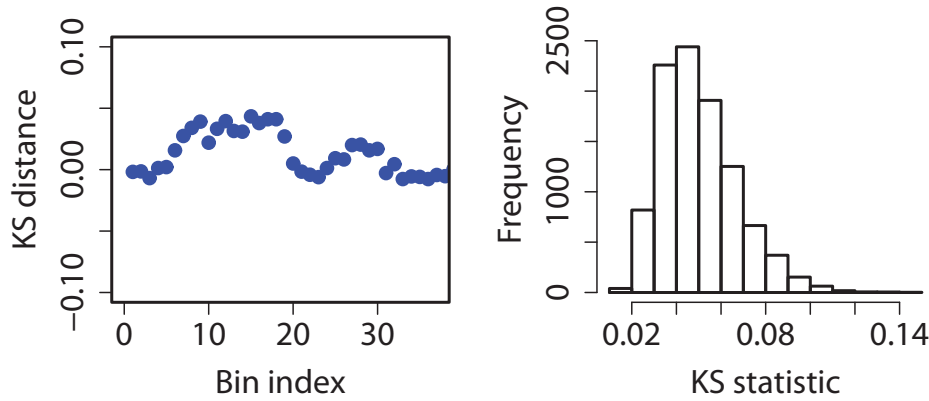


Figure 6.13: Left: Bin-by-bin distances for the Kolmogorov-Smirnov test statistic for the example of Fig. 6.10. Right: Estimated distribution of the Kolmogorov-Smirnov distance for the null hypothesis in the example of Fig. 6.10.

the cumulative distributions being compared. Again, this test is usually thought of as a test to compare an observed distribution with a presumed parent continuous probability distribution. However, the algorithm can be adapted to the two-sample comparison, and to the case of comparing two histograms.

The test statistic for comparing the two samples  $x_1, x_2, \dots, x_N$  and  $y_1, y_2, \dots, y_M$  is [10]:

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

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  $x$ .

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

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

Then the test statistic is:

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

We apply this formalism to the example in Fig. 6.10, finding  $T_{\text{CVM}} = 0.132$ . The resulting estimated distribution under the null hypothesis is shown in

Fig. 6.14. According to our estimated distribution of this statistic under the null hypothesis, this gives a  $P$ -value of 0.45 (See Table 6.1 for a summary), somewhat smaller than the  $\chi^2$  test result.

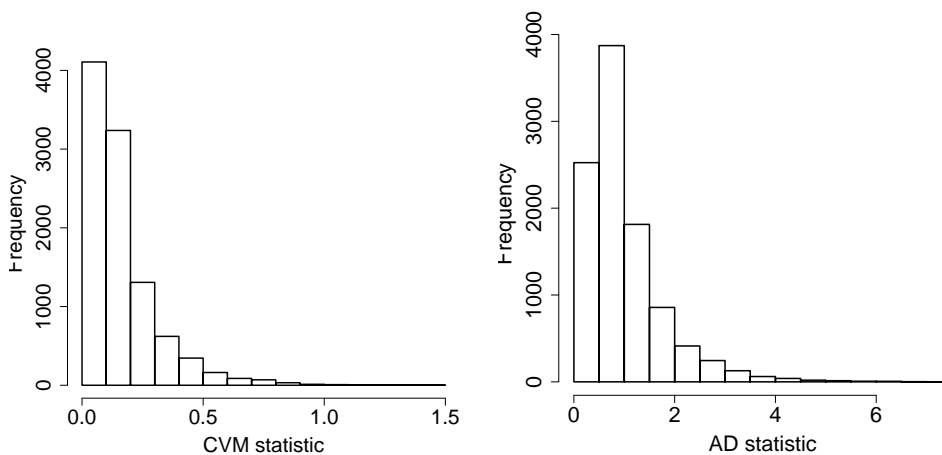


Figure 6.14: Estimated distributions of the test statistic for the null hypothesis in the example of Fig. 6.10 Left: The Cramér-von-Mises statistic. Right: The Anderson-Darling statistic.

### 6.7.6 Anderson-Darling test for shape

The Anderson-Darling test is another variant on the theme of 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, once again, 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  $x_1, \dots, x_m$ . Scholz and Stephens [11] provide a form of this statistic for a  $k$ -sample test on grouped data (e.g., as might be used to compare  $k$  histograms). Based on the result expressed in their Eq. 6, the expression of interest to us for two histograms is:

$$T_{AD} = \frac{1}{N_u + N_v} \sum_{j=k_{\min}}^{k_{\max}-1} \frac{t_j}{\Sigma_j(N_u + N_v - \Sigma_j)} \left\{ \begin{aligned} & [(N_u + N_v)\Sigma_{uj} - N_u\Sigma_j]^2 / N_u \\ & + [(N_u + N_v)\Sigma_{vj} - N_v\Sigma_j]^2 / N_v \end{aligned} \right\}, \quad (6.35)$$

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

$$\Sigma_{uj} \equiv \sum_{i=1}^j u_i, \quad (6.36)$$

$$\Sigma_{vj} \equiv \sum_{i=1}^j v_i, \quad \text{and} \quad (6.37)$$

$$\Sigma_j \equiv \sum_{i=1}^j t_i = \Sigma_{uj} + \Sigma_{vj}. \quad (6.38)$$

We apply this formalism to the example in Fig. 6.10. The resulting estimated distribution under the null hypothesis is shown in Fig. 6.14. The sum over bins gives 0.849 (See Table 6.1 for a summary). According to our estimated distribution of this statistic under the null hypothesis, this gives a  $P$ -value of 0.45, somewhat smaller than the  $\chi^2$  test result, but similar with the CVM result.

### 6.7.7 Likelihood ratio test for shape

We may base a test whether the histograms are sampled from the same shape distribution on the same binomial idea as we used for the normalization test. In this case, however, there is a binomial associated with each bin of the histogram. We start with the null hypothesis, that the two histograms are sampled from the joint distribution:

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

where  $\nu_i = a\mu_i$  for  $i = 1, 2, \dots, k$ . That is, the “shapes” of the two histograms are the same, although the total contents may differ.

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

$$P(v|u+v=t) = \prod_{i=1}^k \binom{t_i}{v_i} \left( \frac{\nu_i}{\nu_i + \mu_i} \right)^{v_i} \left( \frac{\mu_i}{\nu_i + \mu_i} \right)^{t_i - v_i}.$$

The null hypothesis is that  $\nu_i = a\mu_i$  for all values of  $i$ . We would like to test this, but there are now two complications:

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

For  $a$ , we will substitute an estimate from the data, namely the maximum likelihood estimator:

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

Note that this estimate is a random variable; its use will reduce the effective number of degrees of freedom by one.

We propose to 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. Thus, we form the ratio:

$$\lambda = \frac{\max_{H_0} \mathcal{L}(a|v; u + v = t)}{\max_{H_1} \mathcal{L}(\{a_i \equiv \nu_i/\mu_i\}|v; u + v = t)} \quad (6.39)$$

$$= \prod_{i=1}^k \frac{\left(\frac{\hat{a}}{1+\hat{a}}\right)^{v_i} \left(\frac{1}{1+\hat{a}}\right)^{t_i-v_i}}{\left(\frac{\hat{a}_i}{1+\hat{a}_i}\right)^{v_i} \left(\frac{1}{1+\hat{a}_i}\right)^{t_i-v_i}}. \quad (6.40)$$

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 according to:

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

In practice, we'll work with

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

Before attempting to apply this, we investigate how to handle zero bin contents. It is possible that  $u_i = v_i = 0$  for some bin. In this case,  $P(v_i|u_i + v_i = t_i) = 1$ , under both  $H_0$  and  $H_1$ , and this bin contributes zero to the sum. It is also possible that  $t_i \neq 0$ , but  $v_i = 0$  or  $u_i = 0$ . If  $v_i = 0$ , then

$$P(0|t_i) = \left( \frac{\mu_i}{\nu_i + \mu_i} \right)^{t_i}.$$

Under  $H_0$ , this is

$$\left( \frac{1}{1+a} \right)^{t_i},$$

and under  $H_1$  it is

$$\left( \frac{1}{1+a_i} \right)^{t_i}.$$

The maximum likelihood estimator for  $a_i$  is  $\hat{a}_i = 0$ . Thus, the likelihood ratio for bin  $i$  is

$$\lambda_i = \left( \frac{1}{1+\hat{a}} \right)^{t_i},$$

6.7. CASE STUDY: TESTING CONSISTENCY OF TWO HISTOGRAMS 137

and this contributes to the sum an amount:

$$-2 \ln \lambda_i = -2t_i \ln \left( \frac{N_u}{N_u + N_v} \right).$$

If instead  $u_i = 0$ , then

$$P(t_i|t_i) = \left( \frac{\nu_i}{\nu_i + \mu_i} \right)^{t_i}.$$

and the contribution to the sum is

$$-2 \ln \lambda_i = -2t_i \ln \left( \frac{N_v}{N_u + N_v} \right).$$

We apply this formalism to the example in Fig. 6.10. The resulting terms in the sum over bins are shown in Fig. 6.15. The sum over bins gives 25.3 (See Table I for a summary). This statistic should asymptotically be distributed according to a  $\chi^2$  distribution with the number of degrees of freedom equal to one less than the number of bins, or  $N_{\text{DOF}} = 39$  in this case. If valid, this gives a  $P$ -value of 0.96 in this example. This may be compared with a probability of 0.96 according to the estimated actual distribution. In this example we obtain nearly the same answer as the naive application of the chi-square calculation with no bins combined.

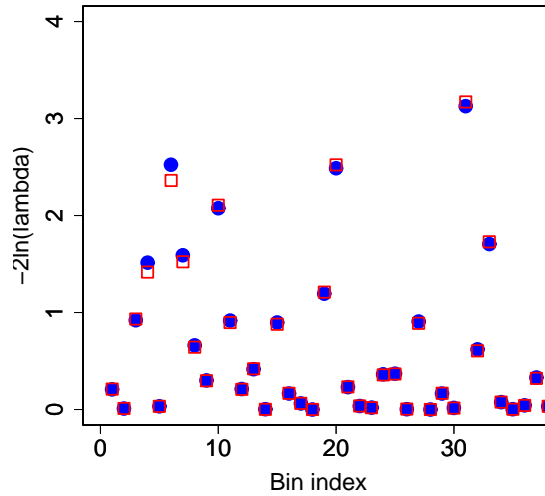


Figure 6.15: Value of  $-2 \ln \lambda_i$  or  $\chi_i^2$  as a function of histogram bin in the comparison of the two distributions of Fig. 1. Blue circles are  $-2 \ln \lambda_i$ ; red squares are  $\chi_i^2$ .

We may see that this close agreement is a result of nearly bin-by-bin equality of the two statistics, see Fig. 6.15. To investigate when this might hold more

generally, we compare the values of  $-2 \ln \lambda_i$  and  $\chi_i^2$  as a function of  $u_i$  and  $v_i$ , Fig. 6.16. We observe that the two statistics agree when  $u_i = v_i$  with increasing difference away from that point. This observation is readily verified analytically. This agreement holds even for low statistics. However, we 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 precise value of the probability should not be taken too seriously, except to conclude that the two distributions are consistent according to these tests. For example, when we combine bins to improve expected  $\chi^2$  behavior, we see fairly large fluctuations in the probability estimate just due to the re-binning (Fig. 6.11).

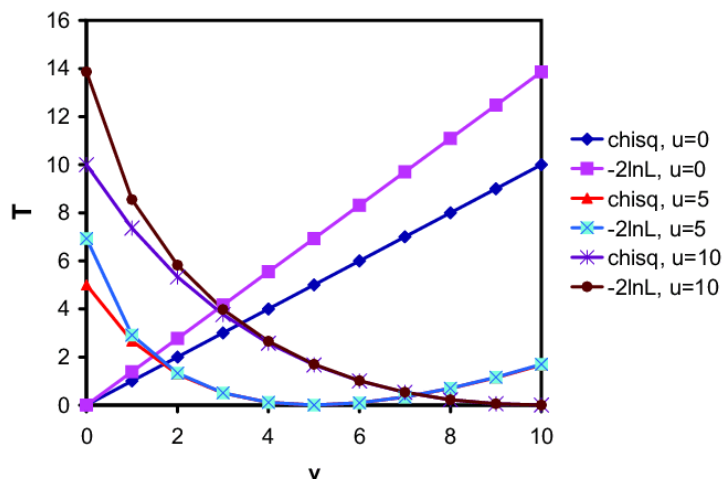


Figure 6.16: Value of  $-2 \ln \lambda_i$  or  $\chi_i^2$  as a function of  $u_i$  and  $v_i$  bin contents. This plot assumes  $N_u = N_v$ . The  $i$  subscript is dropped, with the understanding that this comparison is for a single bin.

### 6.7.8 Likelihood value test for shape

An often-used but controversial goodness-of-fit statistic is the value of the likelihood at its maximum value under the null hypothesis. It 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 where it is a plausible statistic to use. We thus look at it here.

Using the results in the previous section, the test statistic is:

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

If either  $N_u = 0$  or  $N_v = 0$ , then  $T = 0$ .

We apply this formalism to the example in Fig. 6.10. The resulting estimated distribution under the null hypothesis is shown in Fig. 6.17. The sum over bins gives 90 (See Table 6.1 for a summary). According to 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. The fact that it is similar may be expected from the fact that our example is reasonably well-approximated by the large statistics limit.

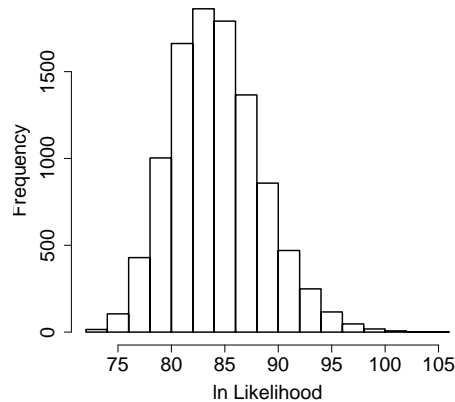


Figure 6.17: Fig. 6.17. Estimated distribution of the  $\ln \mathcal{L}$  test statistic for the null hypothesis in the example of Fig. 6.10.

There are many other possible tests that could be considered, for example, schemes that “partition” the  $\chi^2$  to select sensitivity to different characteristics [12]

### 6.7.9 Distributions Under the Null Hypothesis

For the situation where the asymptotic distribution may not be good enough, we would like to know the probability distribution of our test statistic under the null hypothesis. However, we encounter a 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. Note that it also doesn’t specify  $a$ , but we have already discussed that complication, which appears manageable (although in extreme situations one might need to check for dependence on  $a$ ).

We turn once again to the data to make an estimate for  $\mu_i$ , to be used in estimating the distribution of our test statistics. The straightforward approach is to use the maximum likelihood parameter estimators (under  $H_0$ ):

$$\hat{\mu}_i = \frac{1}{1 + \hat{a}}(u_i + v_i), \quad (6.41)$$

$$\hat{\nu}_i = \frac{\hat{a}}{1 + \hat{a}}(u_i + v_i), \quad (6.42)$$

where  $\hat{a} = N_v/N_u$ . The data is then repeatedly simulated using these values for the parameters of the sampling distribution. For each simulation, a value of the test statistic is obtained. The distribution so obtained is then an estimate of the distribution of the test statistic under the null hypothesis, and  $P$ -values may be computed from this. Variations in the estimates for  $\hat{\mu}_i$  and  $\hat{a}$  may be used to check robustness of the probability estimates obtained in this way.

We have just described the approach that was used to compute the estimated probabilities for the example of Fig. 6.10. The bin contents in this case are reasonably large, and this approach works well enough for this case.

Unfortunately, this approach does very poorly in the low-statistics realm. We 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. We test how well our estimated null hypothesis works for any given test statistic,  $T$ , as follows:

1. 1. Generate a pair of histograms according to the distribution just described.
  - (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 1a.
2. Repeat step 1 many times and make a histogram of the estimated  $P$ -values. Note that this histogram should be uniform if the estimated  $P$ -values are good estimates.

The distributions of the estimated probabilities for the seven test statistics under the null hypothesis are shown in the second column of Fig. 6.18. If the null hypothesis were to be rejected at the estimated 0.01 probability, this algorithm 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  $\mathcal{L}$  statistics, all unacceptably larger than the desired 1%. The KS, CVM, and AD statistics are all consistent with the desired 1%. For comparison, the first column of Fig. 6.18 shows the distribution for a “large statistics” case, where sampling is from histograms with a mean of 100 counts in each bin. We find that all test statistics display the desired flat distribution in this case. Table 6.2 summarizes these results.

6.7. CASE STUDY: TESTING CONSISTENCY OF TWO HISTOGRAMS 141

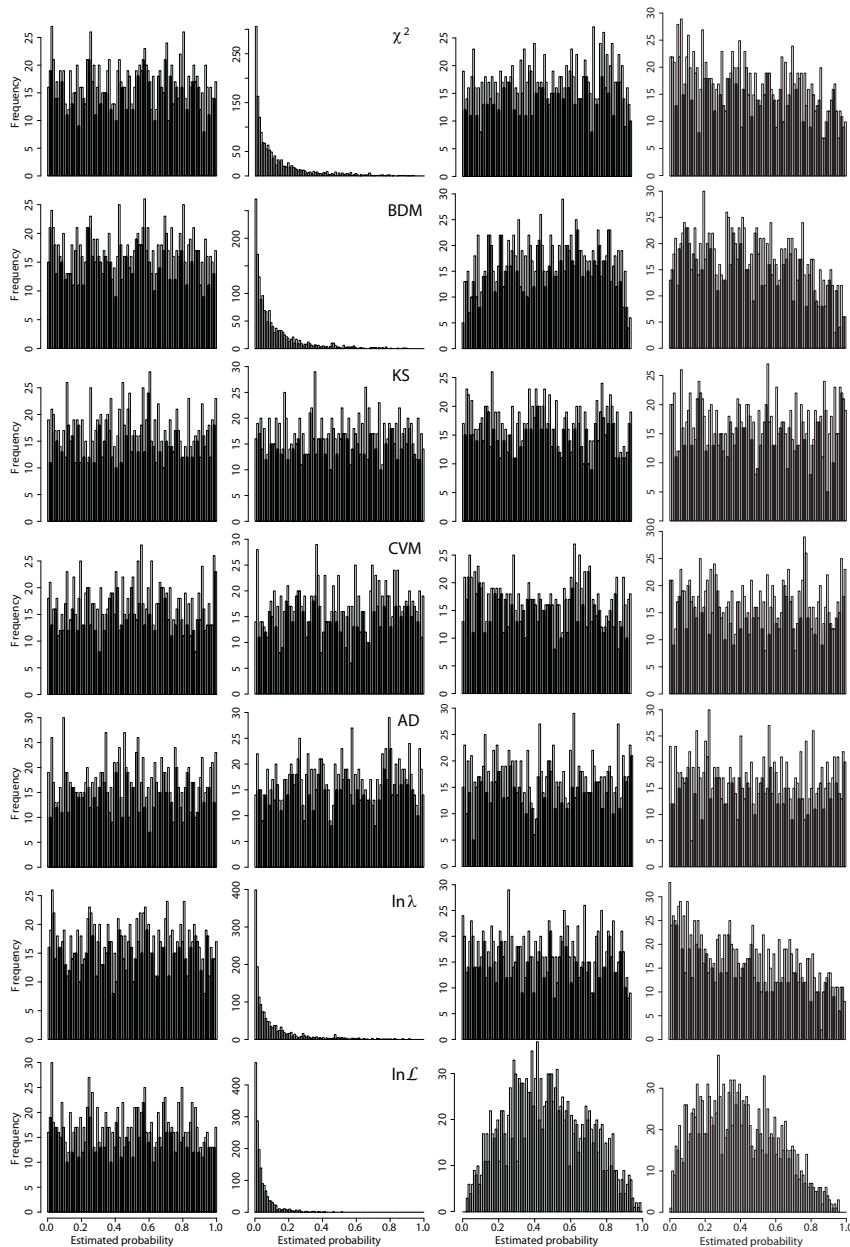


Figure 6.18: (Distribution of the estimated probability that the test statistic is worse than that observed, for seven different test statistics. The data are generated according to the null hypothesis, consisting of 100 bin histograms with a mean of 100 counts (left column) or one count (other columns). The first and second columns are for an estimated  $H_0$  computed as the weighted bin-by-bin average. The third column is for an estimated  $H_0$  where each bin is the average of the total contents of both histograms, divided by the number of bins. The rightmost column is for an estimated  $H_0$  estimated with a Gaussian kernel estimator using the contents of both histograms. The  $\chi^2$  is computed without combining bins.

Table 6.2: Probability that the null hypothesis will be rejected with a cut at 1% on the estimated distribution (see text).  $H_0$  is estimated with the bin-by-bin algorithm in the first two columns, by the uniform histogram algorithm in the third column, and with a Gaussian kernel estimation in the fourth column.

Test statistic	Probability (%)	Probability (%)	Probability (%)	Probability (%)
Bin mean = $H_0$ estimate	100 bin-by-bin	1 bin-by-bin	1 (uniform) uniform	1 (kernel) kernel
$\chi^2$	0.97 $\pm$ 0.24	18.5 $\pm$ 1.0	1.2 $\pm$ 0.3	1.33 $\pm$ 0.28
BDM	0.91 $\pm$ 0.23	16.4 $\pm$ 0.9	0.30 $\pm$ 0.14	0.79 $\pm$ 0.22
KS	1.12 $\pm$ 0.26	0.97 $\pm$ 0.24	1.0 $\pm$ 0.2	1.21 $\pm$ 0.27
CVM	1.09 $\pm$ 0.26	0.85 $\pm$ 0.23	0.8 $\pm$ 0.2	1.27 $\pm$ 0.28
AD	1.15 $\pm$ 0.26	0.85 $\pm$ 0.23	1.0 $\pm$ 0.2	1.39 $\pm$ 0.29
$\ln \lambda$	0.97 $\pm$ 0.24	24.2 $\pm$ 1.1	1.5 $\pm$ 0.3	2.0 $\pm$ 0.34
$\ln \mathcal{L}$	0.97 $\pm$ 0.24	28.5 $\pm$ 1.1	0.0 $\pm$ 0.0	0.061 $\pm$ 0.061

It may be noted that the issue really is one appearing at low statistics. We can give some intuition for the observed effect. Consider the likely scenario at low statistics that some bins will have zero counts in both histograms. In this case our algorithm for the estimated null hypothesis yields a zero mean for these bins. The simulation used to determine the probability distribution for the test statistic will always have zero counts in these bins, that is, there will always be agreement between the two histograms in these bins. Thus, the simulation will find that values of the test statistic are more probable than it should.

If we tried the same study with, say, a mean of 100 counts per bin, we would find that the probability estimates are valid, at least this far into the tails. The left column of Fig. 6.18 shows that more sensible behavior is achieved with larger statistics. The  $\chi^2$ ,  $\ln \lambda$ , and  $\ln \mathcal{L}$  statistics perform essentially identically at high statistics, as expected, since in the normal approximation they are equivalent.

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, we see that our procedure for estimating  $H_0$  does well even for low statistics, although we caution again that we are not examining the far tails of the distribution.

There are various possible approaches to salvaging the situation in the low statistics regime. Perhaps the simplest is to rely on the typically valid assumption that the underlying  $H_0$  distribution is “smooth”. Then instead of having an unknown parameter for each bin, we only need to estimate a few parameters to describe the smooth distribution, and effectively more statistics are available.

For example, we may repeat the algorithm for our example of a mean of one count per bin, but now assuming a smooth background represented by a uniform distribution. This is cheating a bit, since we perhaps aren’t supposed to know that this is really what we are sampling from, but we’ll pretend that we looked at the data and decided that this was plausible. As usual, we would

in practice want to try other possibilities to evaluate systematic effects.

Thus, we estimate:

$$\hat{\mu}_i = N_u/k, \quad i = 1, 2, \dots, k \quad (6.43)$$

$$\hat{\nu}_i = N_v/k, \quad i = 1, 2, \dots, k. \quad (6.44)$$

The resulting distributions for the estimated probabilities are shown in the third column of Fig. 6.18. These distributions are much more reasonable, at least at the level of a per cent (1650 sample experiments are generated in each case, and the estimated  $P$  value is estimated for each experiment with 1650 evaluations of the null hypothesis for that experiment).

It should be remarked that the  $\ln \mathcal{L}$  and, perhaps, to a much lesser extent the BDM statistic, do not give the desired 1% result, but now err on the “conservative” side. It may be possible to mitigate this with a different algorithm, but this has not been investigated. We may expect the power of these statistics to suffer under the approach taken here.

Since we aren’t supposed to know that our null distribution is uniform, we also try another approach to get a feeling for whether we can really do a legitimate analysis. Thus, we try a kernel estimator for the null distribution, using the sum of the observed histograms as input. In this case, we have chosen a Gaussian kernel, with a standard deviation of 2. The “density” package in R [??] is used for this. An example of such a kernel estimated distribution is shown in Fig. 6.19. The resulting estimated probability distributions of our test statistics are shown in the rightmost column of Fig. 6.18. In general, this works pretty well. The bandwidth was chosen here to be rather small; a larger bandwidth would presumably improve the results.

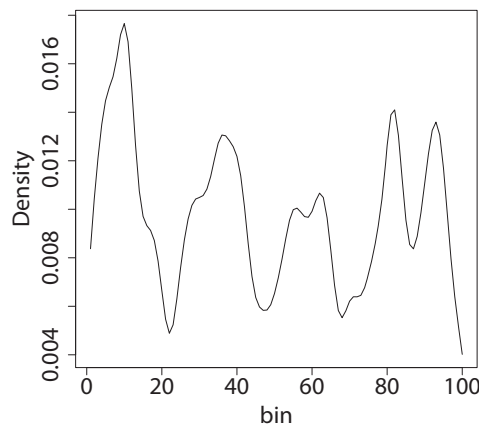


Figure 6.19: Sample Gaussian kernel density estimate of the null hypothesis (for sampling from a true null).

### 6.7.10 Comparison of Power of Tests

The power depends on what the alternative hypothesis is. Here, we mostly investigate adding a Gaussian component on top of a uniform background distribution. This choice is motivated by the scenario where one distribution appears to show some peaking structure, while the other does not. We also look briefly at a different extreme, that of a rapidly varying alternative.

The data for this study are generated as follows: The background (null distribution) 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. We vary the amplitude of the Gaussian and count how often the null hypothesis is rejected at the 1% confidence level. The amplitude is measured in percent, for example a 25% Gaussian has a total amplitude corresponding to an average of 25% of the total counts in the histogram, including the (small) tails extending beyond the histogram boundaries. The Gaussian counts are added to the counts from the null distribution. An example is shown in Fig. 6.20.

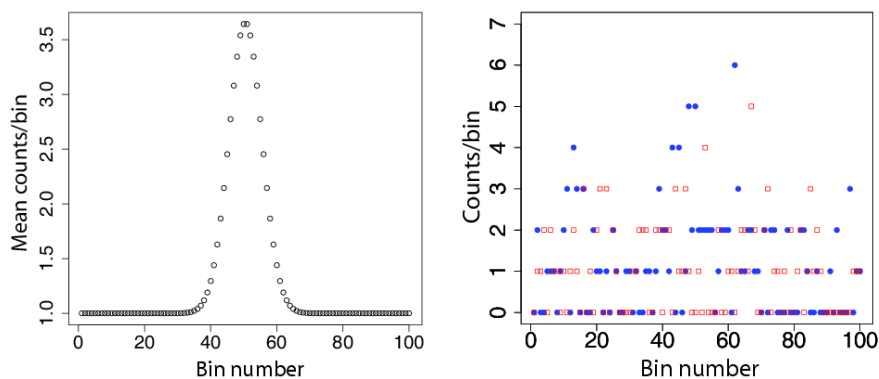


Figure 6.20: Left: The mean bin contents for a 25% Gaussian on a flat background of one count/bin (note the suppressed zero). Right: Example sampling from the 25% Gaussian (filled blue dots) and from the uniform background (open red squares).

The distribution of estimated probability, under  $H_0$ , that the test statistic is worse than that observed (i.e., the distribution of  $P$ -values) is shown in Fig. 6.21 for seven different test statistics. Three different magnitudes of the Gaussian amplitude are displayed. The power of the tests to reject the null hypothesis at the 99% confidence level is summarized in Table 6.3 and in Fig. 6.22 for several different alternative hypothesis amplitudes.

6.7. CASE STUDY: TESTING CONSISTENCY OF TWO HISTOGRAMS 145

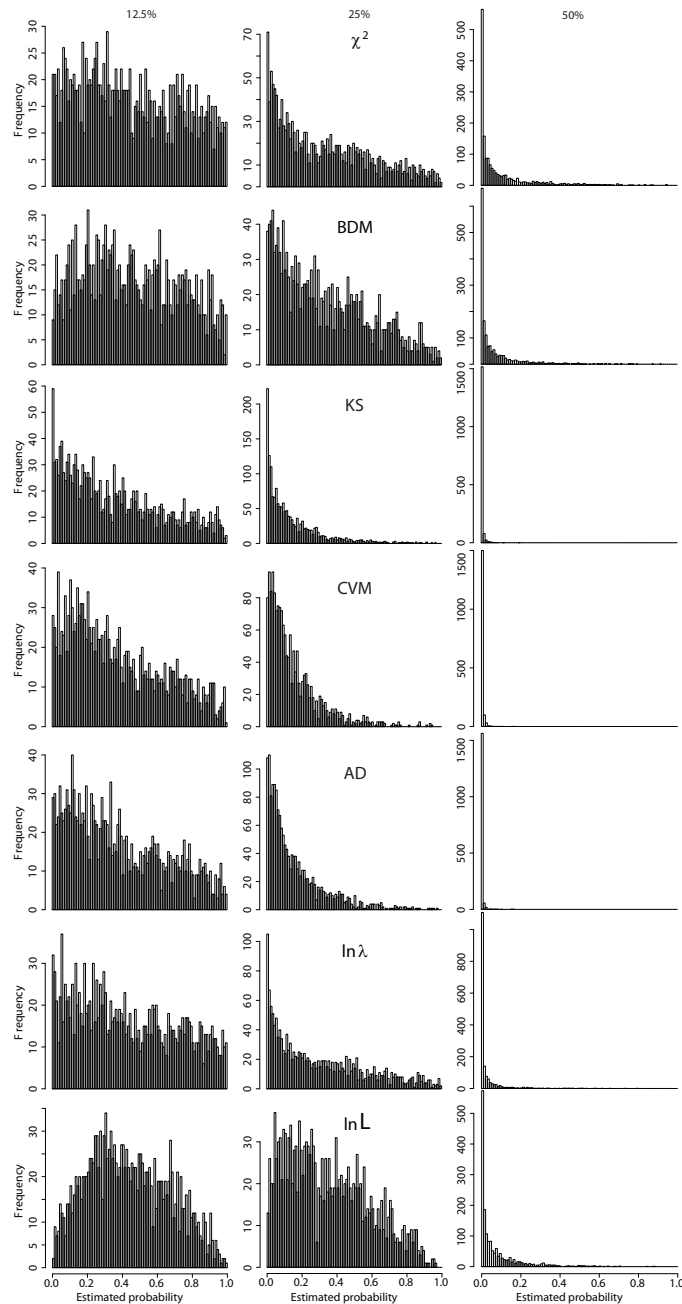


Figure 6.21: Distribution of estimated probability, under  $H_0$ , that the test statistic is worse than that observed, for seven different test statistics. The data are generated according to a uniform distribution, consisting of 100 bin histograms with a mean of 1 count, for one histogram, and for the other histogram with a uniform distribution plus a Gaussian of strength 12.5% (left column), 25% (middle column), and 50% (right column). The  $\chi^2$  is computed without combining bins.

Table 6.3: 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 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.

Statistic	H0 %	12.5 %	25 %	37.5 %	50 %	-25 %
$\chi^2$	$1.2 \pm 0.3$	$1.3 \pm 0.3$	$4.3 \pm 0.5$	$12.2 \pm 0.8$	$34.2 \pm 1.2$	$1.6 \pm 0.3$
BDM	$0.30 \pm 0.14$	$0.5 \pm 0.2$	$2.3 \pm 0.4$	$10.7 \pm 0.8$	$40.5 \pm 1.2$	$0.9 \pm 0.2$
KS	$1.0 \pm 0.2$	$3.6 \pm 0.5$	$13.5 \pm 0.8$	$48.3 \pm 1.2$	$91.9 \pm 0.7$	$7.2 \pm 0.6$
CVM	$0.8 \pm 0.2$	$1.7 \pm 0.3$	$4.8 \pm 0.5$	$35.2 \pm 1.2$	$90.9 \pm 0.7$	$2.7 \pm 0.4$
AD	$1.0 \pm 0.2$	$1.8 \pm 0.3$	$6.5 \pm 0.6$	$42.1 \pm 1.2$	$94.7 \pm 0.6$	$2.8 \pm 0.4$
$\ln \lambda$	$1.5 \pm 0.3$	$1.9 \pm 0.3$	$6.4 \pm 0.6$	$22.9 \pm 1.0$	$67.1 \pm 1.2$	$2.4 \pm 0.4$
$\ln \mathcal{L}$	$0.0 \pm 0.0$	$0.1 \pm 0.1$	$0.8 \pm 0.2$	$6.5 \pm 0.6$	$34.8 \pm 1.2$	$0.0 \pm 0.0$

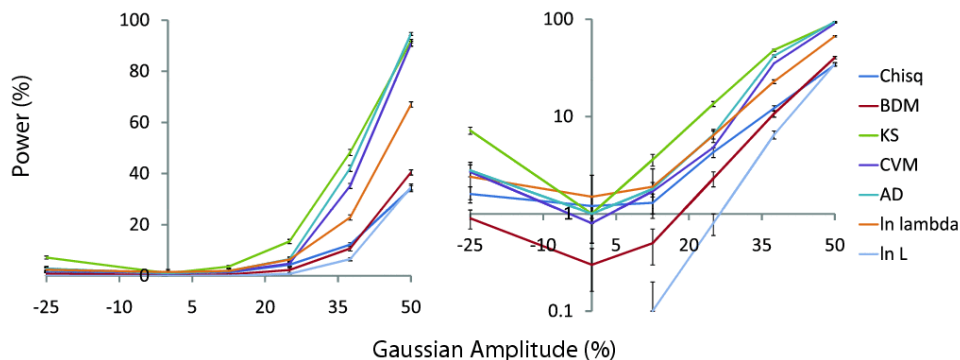


Figure 6.22: Summary of power of seven different test statistics, for the alternative hypothesis with a Gaussian bump. Left: linear vertical scale; Right: logarithmic vertical scale. [Best viewed in color. At an amplitude of 35%, the ordering, from top to bottom, of the curves is: KS, AD, CVM,  $\ln \lambda$ ,  $\chi^2$ , BDM,  $\ln \mathcal{L}$ .]

In Table 6.4 we take a look at the performance of our seven statistics for histograms with large bin contents. It is interesting that in this large-statistics case, for the  $\chi^2$  and similar tests, the power to reject a dip is greater than the power to reject a bump of the same area. This is 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. We 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.

To get an idea of what happens for a radically different alternative to the null distribution, we consider sensitivity to sampling from the “sawtooth” distribution as shown in figure 6.23. This is to be compared once again to samplings

6.7. CASE STUDY: TESTING CONSISTENCY OF TWO HISTOGRAMS 147

Table 6.4: 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.

Statistic	$H_0$ %	5 %	-5 %
$\chi^2$	$0.91 \pm 0.23$	$79.9 \pm 1.0$	$92.1 \pm 0.7$
BDM	$0.97 \pm 0.24$	$80.1 \pm 1.0$	$92.2 \pm 0.7$
KS	$1.03 \pm 0.25$	$77.3 \pm 1.0$	$77.6 \pm 1.0$
CVM	$0.91 \pm 0.23$	$69.0 \pm 1.1$	$62.4 \pm 1.2$
AD	$0.91 \pm 0.23$	$67.5 \pm 1.2$	$57.8 \pm 1.2$
$\ln \lambda$	$0.91 \pm 0.23$	$79.9 \pm 1.0$	$92.1 \pm 0.7$
$\ln \mathcal{L}$	$0.97 \pm 0.24$	$79.9 \pm 1.0$	$91.9 \pm 0.7$

from the uniform histogram. The results are tabulated in Table 6.5. The “percentage” sawtooth here refers to the fraction of the null hypothesis mean. That is, 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.

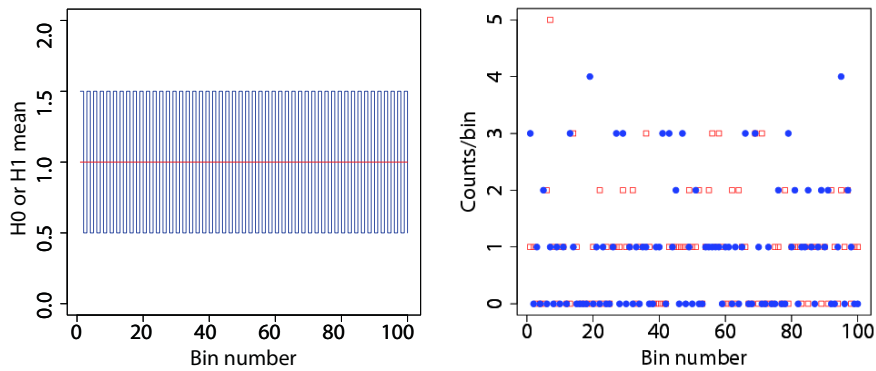


Figure 6.23: Left: The mean bin contents for a 50% sawtooth on a flat background of one count/bin (blue), compared with the flat background means (red). Right: Example sampling from the 50% sawtooth (filled blue dots) and from the uniform background (open red squares).

In this example, the  $\chi^2$  and likelihood ratio 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 average out local oscillations.

Table 6.5: Estimates of power for seven different test statistics, for a “sawtooth” alternative distribution.

Statistic	50 %	100 %
$\chi^2$	$3.7 \pm 0.5$	$47.8 \pm 1.2$
BDM	$1.9 \pm 0.3$	$33.6 \pm 1.2$
KS	$0.85 \pm 0.23$	$1.0 \pm 0.2$
CVM	$0.91 \pm 0.23$	$1.0 \pm 0.2$
AD	$0.91 \pm 0.23$	$1.2 \pm 0.3$
$\ln \lambda$	$4.5 \pm 0.5$	$49.6 \pm 1.2$
$\ln \mathcal{L}$	$0.30 \pm 0.14$	$10.0 \pm 0.7$

### 6.7.11 Conclusions

These studies have demonstrated some important lessons in “goodness-of-fit” testing:

1. There is no single “best” test for all applications. Statements such as “test X is better than test Y” are empty without giving more context. For example, the Anderson-Darling test is often very powerful in testing normality of data against alternatives with non-normal tails (such as the Cauchy distribution) [13]. However, we have seen that it is not always especially powerful in other situations. The more we know about what we wish to test for, the more reliably we can choose a powerful test. Each of the tests investigated here may be reasonable to use, depending on the circumstance. Even the controversial  $\mathcal{L}$  test works as well as the others sometimes. However, there is no known situation where it actually performs better than all of the others, and indeed the situations where it is observed to perform as well are here limited to those where it is equivalent to another test.
2. Computing probabilities via simulations is a very useful technique. However, it must be done with care. The issue of tests with an incompletely specified null hypothesis is particularly insidious. Simply generating a distribution according to some assumed null distribution can lead to badly wrong results. Where this could occur, it is important to verify the validity of the procedure. Note that we have only looked into the tails to the 1% level. The validity must be checked to whatever level of probability is needed for the results. Thus, we cannot blindly assume the results quoted here at the 1% level will still be true at, say, the 0.1% level.

We have concentrated on the specific question of comparing two histograms. However, the general 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,  $\ln \mathcal{L}$ , and  $\mathcal{L}$  tests may all be constructed for these other situations (as well as the  $\chi^2$  and BDM, if we bin the data).

# Bibliography

- [1] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, 3rd Ed., Hafner, New York (1973); volume 2, page 480, and references therein.
- [2] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna 2007, ISBN 3-900051-07-0, <http://www.r-project.org/>.
- [3] BaBar Collab. (B. Aubert et al.), *Phys. Rev. Lett.* **95**, 142001 (2005).
- [4] Scott Oser, [http://www.phas.ubc.ca/~oser/p509/Lec\\_18.pdf](http://www.phas.ubc.ca/~oser/p509/Lec_18.pdf).
- [5] J. Klein and A. Roodman, *Ann. Rev. Nucl. Part. Sci.* **55**, 141 (2005).
- [6] Muon g-2 Collab. (G. W. Bennett et al.), *Phys. Rev. D* **73**, 072003 (2006).
- [7] BaBar Collab. (B. Aubert et al.), *Phys. Rev. Lett.* **91**, 221802 (2003).
- [8] E. L. Lehmann and Joseph P. Romano, *Testing Statistical Hypotheses*, Third edition, Springer, New York (2005), Theorem 4.4.1.
- [9] <http://www.herine.net/stat/software/dbinom.html>.
- [10] T. W. Anderson, *On the Distribution of the Two-Sample Cramér-Von Mises Criterion*, *Annals Math. Stat.*, **33** (1962) 1148.
- [11] F. W. Scholz and M. A. Stephens, *k-Sample Anderson-Darling Tests*, *J. Amer. Stat. Assoc.* **82** (1987) 918.
- [12] D. J. Best, *Nonparametric Comparison of Two Histograms*, *Biometrics* **50** (1994) 538.
- [13] M. A. Stephens, *EDF Statistics for Goodness of Fit and Some Comparisons*, *Jour. Amer. Stat. Assoc.* **69** (1974) 730.