Plan for the Statistics Lectures

- Lecture I (Wednesday, September 18, 11:45-12:30)
  1. Important probability concepts
  2. Point estimation
- Lecture II (Thursday, September 19, 10:45-12:30)
  1. Frequency and Bayes interpretations
  2. Interval estimation
  3. Systematic uncertainties
- Lecture III (Friday, September 20, 10:45-12:30)
  1. Hypothesis tests
  2. Resampling methods
- Lecture IV (Saturday, September 21, 10:45-12:30)
  1. Density estimation

Hypothesis tests

1. Basics (test vs p-value)
2. Nuisance parameters
3. Tests converging to χ^2
4. Other tests
5. Issues in significance

Hypothesis test basics

Consider a test, T, for hypotheses

$$H_0: \text{Statement } A$$
$$H_1: \text{Statement } B$$

- Assumes that exactly one of A or B must be true.
- H_0 is called the null hypothesis; H_1 is the alternative
- Test $T \in \{0, 1\}$ defines a statistic (RV) such that we accept H_0 if $T = 0$ and accept H_1 (reject H_0) if $T = 1$
- The critical region, $\mathcal{R}$ of a test is the set of observations for which $T = 1$
- The significance level is the probability to reject H_0, if H_0 is true. This is the probability of a Type I error:

$$\alpha \equiv P(X \in \mathcal{R}|H_0)$$

- The power of a test is the probability to reject H_0 if H_1 is true Power is one minus the probability of a Type II error.

$$\beta \equiv P(X \in \bar{\mathcal{R}}|H_1)$$

The power depends on the alternative distribution

Example: Likelihood Ratio Test

- We have already made use of the likelihood ratio (LR) test in our discussion on CIs.
- Let us show that this test is Uniformly Most Porwerful (UMP) for a simple test
  - A simple test tests simple hypotheses, that is H_0 and H_1 are completely specified
  - For example, we wish to test

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

    θ could be a parameter vector, but θ_0 and θ_1 give it completely
  - A hypothesis that is not simple is composite (e.g., $\theta > 0$)
- We wish to construct a test of H_0 against H_1 that is most powerful against any alternative θ_1 (uniformly most powerful)

Example: Likelihood Ratio Test

- Of the many possible critical regions, we wish to construct the one for which the power is greatest. We wish to maximize:

$$1-\beta = \int_{\mathcal{R}} f(x;\theta_1)dx$$
$$= \int_{\mathcal{R}} \frac{f(x;\theta_1)}{f(x;\theta_0)} f(x;\theta_0)dx$$

subject to constraint:

$$\alpha = \int_{\mathcal{R}} f(x;\theta_0)dx$$

- Notice that:

$$\frac{1-\beta}{\alpha} = E\left[\frac{f(x;\theta_1)}{f(x;\theta_0)}\right]_{(\mathcal{R};H_0)}$$

where the expectation value is restricted to the critical region, under H_0

Example: Likelihood Ratio Test

- Thus, build our critical region by selecting those values of X for which $f(x;\theta_1)/f(x;\theta_0)$ is largest, $\geq \Lambda_\alpha$
- Express in likelihood functions. Form likelihood ratio:

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

- If $\lambda \geq \Lambda_\alpha$, then x is in the critical region
- Thus, by construction LR test is UMP
- Note that the likelihood ratio $\lambda = \lambda(X)$ is an RV
- Can turn around and ask, given sampling x, what the probability is that $\lambda < \lambda(x)$. This is called a *p*-value

p-values

- Have seen notion of a hypothesis test as testing a null hypothesis against an alternative hypothesis
- Physicists make use of this paradigm, as well as another:
  - Given an observartion, the p-value is defined as the probability that the null hypothesis will produce a result as "extreme", or more, as the observed result
- The p-value only refers to the null hypothesis – there is no explicit alternative
- Users of p-values call the p-value the significance level
- This is as if α had been set equal to p and a critical region defined by p

p-values and hypothesis tests – Historical note

- Practice of statistics is tied up with philosophical issues
- These extend beyond the Bayes versus frequency debate
- Subject of hypothesis tests is also fertile ground, dating at least to a debate between Fisher and Neyman
- Fisher: p-values, notion of testing a given theory (H_0). We say goodness-of-fit (GOF)
- Neyman-Pearson: H_0 vs H_1, notion of testing between two theories
- Physicists do both
- p-value is a RV; critical region $\mathcal{R}$ is not (only decision T is)
- Even GOF has an implied alternative (ie, "anything else")
- Neyman-Pearson, Fisher both non-Bayesian; goal was to eliminate dependence on priors

Goodness of fit

- The goodness of fit (GOF) problem refers to whether a dataset is consistent with sampling from a model for the distribution. If data set $X = (X_1, \ldots, X_N)$, is sampled iid from some cdf $F_X(x)$ and the model is denoted M, this is a hypothesis test of the form:

$$H_0 : F = M$$
$$H_1 : F \neq M$$

- This test is called a one-sample test. It is "one-sample" because we are comparing a dataset with a given (theoretical) distribution
- The two-sample test is also commonly encountered. In this case, we compare two datasets, X and Y, sampled iid from cdf's F_X and G_Y, to see whether they are consistent with being drawn from the same distribution:

$$H_0 : F_X = G_Y$$
$$H_1 : F_X \neq G_Y$$

Goodness of fit

- Goodness of fit tests may also be categorized as tests on binned data (e.g., histograms) or unbinned data
- Statisticians call histograms tables. A simple histogram is a table with one row
- A scatterplot may be binned into a table with multiple rows
- Categorizing a problem as binned or unbinned is somewhat artificial – a test designed to be used on unbinned data may usually be adapted to binned data (though the reverse may not be possible)
- A large number of goodness of fit tests exist. There is no "one size fits all" test. The choice of a good test depends on the details
- "Goodness" of a test is measured in terms of its power for a specified significance level
- An alternative must be assumed in order to compute the power of a test

Goodness of fit – caution

- Commonly (but not always), only the asymptotic distribution (under H_0) of the test statistic is known
- This does not mean that the test cannot be used when the asymptotic condition is not met
- With sufficient computing capability, the distribution of the test statistic may be determined via simulations.
- Must be done with some care, as H_0 is often not completely known.
  - For example, might wish to test null hypothesis that two histograms are sampled from the same distribution.
  - The distribution itself may not be known. Instead, it must be estimated somehow from the available data.
  - If suitable care is not taken, the estimate may not be robust against fluctuations, and badly erroneous results obtained.
  - When this may be the case, suitable studies (e.g., with different estimates of H_0) should be undertaken to determine this sensitivity.

Binned goodness of fit tests

- Commonly used goodness of fit test is the chi-square test
- Motivation in the LS fitting process: We have a set of measurements, sampled from a multivariate normal, $x_1, \ldots, x_D$, and a model to predict the means, $\mu_1, \ldots, \mu_D$. The model may depend on zero or more unknown parameters, $\theta_1, \ldots, \theta_R$. Sampling distribution is:

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left\{ -\frac{1}{2} \left[x - \mu(\theta)\right]^{\mathsf{T}} \Sigma^{-1} \left[x - \mu(\theta)\right] \right\} \\ &= \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-\frac{1}{2} \chi^2 \right) \end{aligned}$$

- The LS fitting procedure is to find those values of θ such that the χ^2 is minimized. If no additional conditions are applied, the value of the minimum χ^2, $\chi^2_{\min}$, is drawn from a chi-square distribution with $D - R$ DOF. Comparing the observed $\chi^2_{\min}$ with the chi-square distribution provides the chi-square goodness of fit test.

Binned goodness of fit tests – Pearson χ^2

- Consider fitting a model to a histogram
- N is the number of histogram bins and x_n is the content of bin n
- If the bin contents are counts from a Poisson process, then the bins are independent and the covariance matrix is $\mathrm{cov}(X) = \mathrm{diag}(\mu_1, \ldots, \mu_N)$
- A GOF test statistic known as the Pearson chi-square is defined by:

$$\chi^2_P = \sum_{n=1}^{N} \frac{(x_n - \mu_n)^2}{\mu_n},$$

- This is asymptotically $\chi^2(N)$ distributed (or $\chi^2(N-1)$ if the normalization is fixed to the observed total counts). Usually we have R parameters to estimate and replace μ with $\hat{\mu}_i = \mu_i(\hat{\theta})$ where $\hat{\theta}$ is determined by minimizing χ^2_P. This reduces the degrees of freedom, by one for each estimated parameter, subject to regularity conditions discussed below

*Binned goodness of fit tests – Pearson χ^2

- The Pearson chi-square test should not be used if any of the histogram bins have small statistics, since the $\chi^2(N-R)$ distribution will not apply
- Various rules-of-thumb for how many counts are needed in each bin for a good-enough approximation, usually around 5-10
- May combine bins to meet the minimum requirement, though at the loss of sensitivity to possible structure in the data
- Another approach is to use bins that have equal probability content under H_0, ensuring a uniform weighting of the intervals

*Binned goodness of fit tests – Neyman modified χ^2

- Especially at low statistics there is a tendency for χ^2_P minimization to overestimate μ_n.
- May be avoided by instead estimating the parameters via maximum likelihood (using Poisson statistics)
- An alternative approach defines a statistic (Neyman modified chi-square):

$$\chi^2_N = \sum_{n=1}^{N} \frac{(x_n - \mu_n)^2}{x_n}$$

- For large statistics, this also works, but for low statistics it suffers a similar disease: fluctuations toward small values of x_n will be more highly weighted, tending to bias towards small values of μ_n.

Statistics converging to chi-square
Binning not required

- Three important test statistics that follow a χ^2 distribution for large samples, under certain assumptions:
  - Likelihood ratio test
  - The Wald test
  - The score test

Statistics converging to chi-square – Likelihood ratio

- Already discussed this test for simple test and CIs
- For composite hypotheses we maximize the likelihoods under H_0 and H_1 before taking the ratio
- Taking twice the the logarithm, we have the statistic

$$2\log\lambda = 2\log\max_{H_1} L(H_1; X) - 2\log\max_{H_0} L(H_0; X).$$

- We suppose H_0 defines a region of dimension $V < R$ in a parameter space for θ, and H_1 to the entire remaining R-dimensional space
- Then, under H_0, $2\log\lambda$ is asymptotically $\chi^2(R-V)$ distributed, under some regularity conditions (later)

*Statistics converging to chi-square – Wald statistic

- Following above scenario, we may write
$$H_0 : \theta = q(\vartheta),$$
where ϑ is a parameter vector of dimension $V < R$, and q provides the mapping onto the R-dimensional θ. May rewrite H_0 as the $R - V$ equations giving the kernel of the mapping, that is as $Q(\theta) = 0$. Eg, if $H_0 : \theta = \theta_0$, then $Q(\theta) = \theta - \theta_0$. Let the MLE under H_0 be $\hat{\vartheta}$, and the MLE under H_1 be $\hat{\theta}$
- The Wald statistic is:
$$W \equiv \left[Q(\hat{\theta})\right]^{\mathsf{T}}_{\theta=\hat{\theta}} \left\{ \left[\frac{\partial Q}{\partial \theta}\right]^{\mathsf{T}}_{\theta=\hat{\theta}} I(\hat{\theta})^{-1} \left[\frac{\partial Q}{\partial \theta}\right]_{\theta=\hat{\theta}} \right\}^{-1} Q(\hat{\theta}),$$
where $I(\hat{\theta})$ is the Fisher information matrix, estimated at $\theta = \hat{\theta}$:
$$I_{ij}(\theta = \hat{\theta}) = E\left[\frac{\partial^2 \log L(\theta; X)}{\partial \theta_i \partial \theta_j}\right]_{\theta=\hat{\theta}}.$$

*Statistics converging to chi-square – Wald statistic

- For a normal distribution $I(\theta)$ is the inverse of the covariance matrix, and is independent of θ if the parameters are functions of location only
- Eg, suppose the null hypothesis is $H_0 : \theta = \theta_0$. Then $V = 0$ and $Q(\theta) = \theta - \theta_0$, and we have the familiar-looking statistic:

$$W = (\hat{\theta} - \theta_0)^{\mathsf{T}} I(\hat{\theta})(\hat{\theta} - \theta_0)$$

- The likelihood ratio compares two likelihood values; the Wald statistic compares two values of θ
- Notice that the Wald statistic does not require evaluation of $\hat{\vartheta}$

*Statistics converging to chi-square – Score statistic

- Alternatively, a score statistic based on the score may be computed, measuring the gradient of the likelihood under H_0
- In this case it is not necessary to obtain the MLEs for θ. Instead, we compute:

$$U = S(\theta_0)^{\mathsf{T}} I^{-1}(\theta_0) S(\theta_0),$$

  where S is the score function evaluated at θ_0 (H_0)
- Comparing with the Wald statistic, we see that the evaluation is made at H_0 rather than at the peak of the likelihood, and the deviation measure is replaced by a slope

*Statistics converging to chi-square – Comments

- Which of these approaches is best?
- Asymptotically, they are equivalent!
- More generally, there is no universal answer
- One ingredient in deciding may be the different computational requirements
- (Much) more can be said, eg. Rayner and Best, *Smooth tests of Goodness of Fit*, Oxford Univ. Press (1989); Cressie and Read, J. R. Stat. Soc. B **46** (1984) 440; as well as NarskyPorter(2014), Wiley

Goodness of fit – Counting DOF

- Common confusion: How many DOF do I use?
- Requires care. There are conditions for the validity of the $\chi^2(N-R)$ or $\chi^2(R-V)$ distribution of the test statistic
- Often arises when using χ^2 to evaluate the statistical significance of a possible signal
- For example: We do two fits to the same dataset (say a histogram with N bins):
  - Fit A has R_A parameters, with χ^2_A
  - Fit B has a subset R_B of the parameters in fit A, with χ^2_B, where the $R_A - R_B$ other parameters (call them θ) are fixed at zero
- What is the distribution of $\Delta\chi^2 = \chi^2_B - \chi^2_A$?
  - More carefully, what is the distribution under H_0, corresponding to fit B?

Counting DOF

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

$$\Delta\chi^2 \equiv \chi^2_B - \chi^2_A$$

is the same as a likelihood ratio ($2\log\lambda$) statistic for 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 $\chi^2(R_A - R_B)$ distribution under the conditions:

1. Parameter estimates in computing λ are consistent under H_0
2. Parameter values under H_0 are not boundary points of $H_0 \cup H_1$ (the maintained hypothesis). For example, if there is a single parameter θ, with $H_0 : \theta = 0$ and $H_1 : \theta > 0$, then the maintained hypothesis is $\theta \geq 0$ and the parameter value in H_0 is a boundary point
3. There are no nuisance parameters under the alternative hypothesis other than those present in H_0

Unfortunately, commonly encountered situations violate these requirements

Counting DOF - Example I

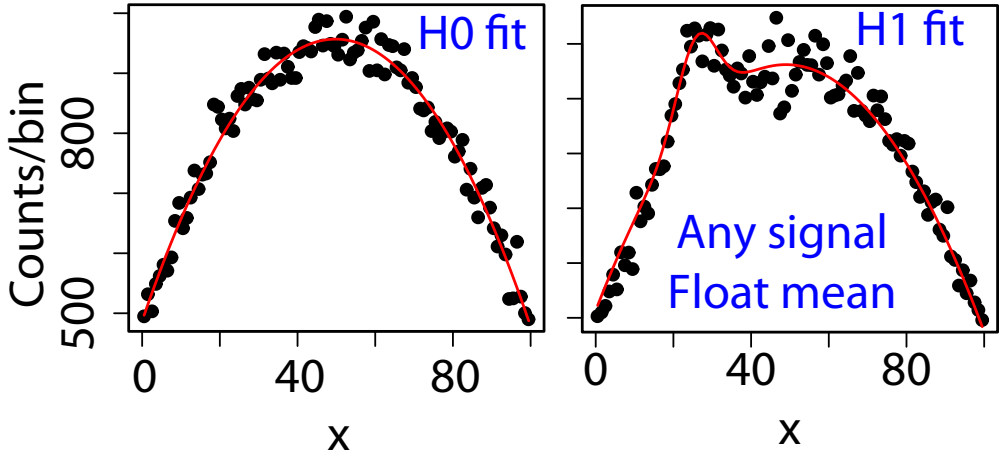
- Consider fitting a spectrum to decide whether a bump is significant
- The parameter of greatest interest is the signal strength
- Compare fits (i.e., χ^2 values) with and without a signal component to estimate significance of the signal
- Under H_0, the signal is zero. Under H_1, the signal is non-zero
- If the signal fit has, e.g., a parameter for location, this constitutes an additional nuisance parameter under H_1, that is a nuisance parameter that is not defined under H_0
- If the fit for signal constrains the signal yield to be non-negative, this violates the interior point requirement

Let us illustrate this by example

Counting DOF - Example I

Sample ML fits:

- Left: Distribution generated and fit under background-only H_0
- Right: Distribution generated under H_1

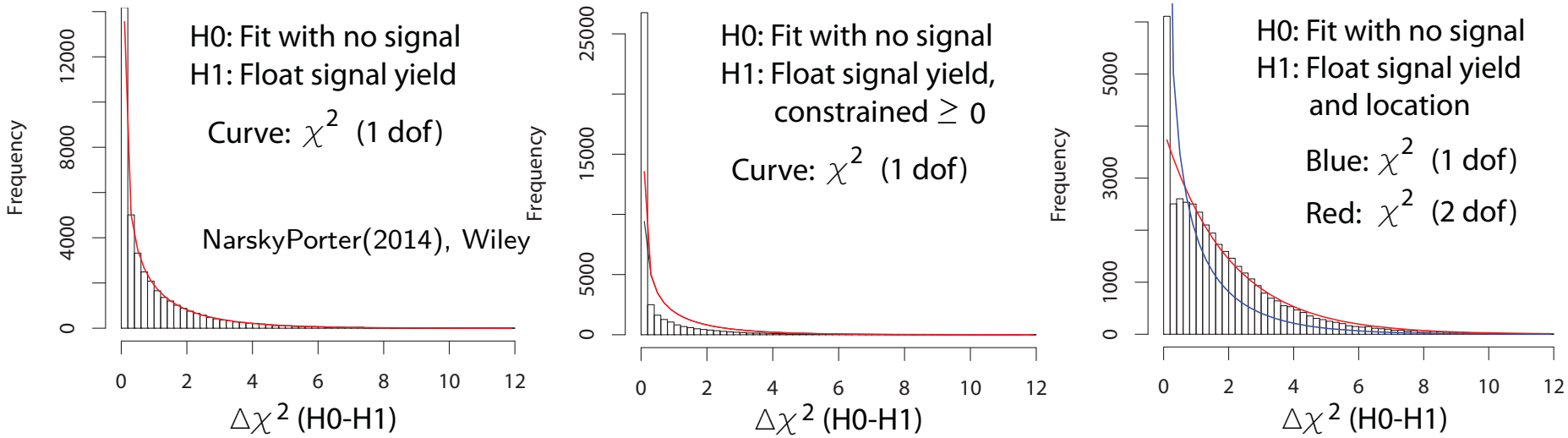


NarskyPorter(2014), Wiley

(R function `optim` used for fits)

- In order to compute the distribution of $2 \ln \lambda \sim \Delta\chi^2$, we consider the spectrum generated under H_0, the background only hypothesis
- The difference in χ^2 bewteen the H_0 and H_1 fits is calculated for this spectrum
- The distribution of the $\Delta\chi^2$ statistic is estimated by simulating each "experiment" many times, under H_0

Counting DOF - Example I results



- When the location parameter is fixed in the fit, and the signal yield is allowed to be positive or negative, the distribution follows a $\chi^2(1)$
- When we constrain the yield to be non-negative, the distribution becomes more peaked towards zero than $\chi^2(1)$
- When both signal yield and location are unconstrained, distribution is somewhere between the curves for 1 and 2 DOF. This is because the location parameter is an additional nuisance parameter under H_1

*Counting DOF – Comment

- A similar issue arises when the unknown parameters of a distribution are determined according to a ML fit to the unbinned observations
- If the data is then binned and a χ^2 statistic computed, this statistic is not in general $\chi^2(N-R-1)$ (assuming normalization taken from the data) distributed
- This is because the unbinned fit produces more efficient estimators than a binned fit in general
- Under regularity conditions, the asymptotic distribution lies between $\chi^2(N-1)$ and $\chi^2(N-R-1)$. If the $\chi^2(N-R-1)$ distribution is assumed, the result will be to reject the null hypothesis too often
- Depending on the distribution, the error made may be minor or substantial, so this should be considered when taking this approach

Counting DOF - Example II

- We have a histogram of an observed mass distribution
- We are interested in the possibility that there is a resonance at $x = m - m_R = 0$ on a flat background
- Including possibility of interference, model the pdf for the mass distribution by:

$$f(x; a, \theta) = B|1 + ae^{i\theta}/(x + i)|^2,$$

  where B is a normalization constant, and a and θ are resonance parameters. Assume that the mass and width of the resonance are known, and the mass resolution is negligible
- Want to test:

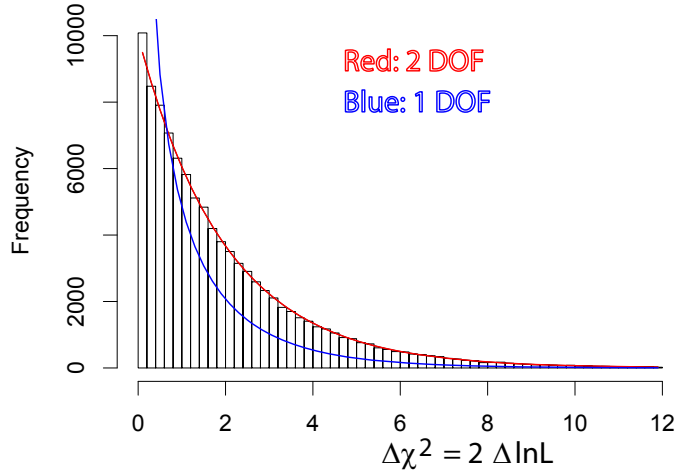
$$\begin{aligned} H_0 : &\quad a = 0 \\ H_1 : &\quad a > 0 \end{aligned}$$

- To do this, perform two fits to the histogram, one with $a = 0$, and one with a and θ allowed to float. We compute the change in log likelihood, $\Delta = \Delta \ln L$, between the two fits. Assume large bin contents

Counting DOF - Example II

- Considering the χ^2 distribution, in order to compute a p-value for your test, how many degrees of freedom should we use for statistic 2Δ?
- There are two parameters under H_1 (a and θ) specified or not present under H_0. Suggests 2 DOF
- But θ looks like a nuisance parameter not present under H_0
- And $a = 0$ doesn't look like an interior value if amplitude ≥ 0
- So try simulation to check:

- Generate a large number of MC datasets under H_0, and repeat analysis on each of them.
- Compute 2Δ for each experiment and make a histogram of 2Δ.
- For concreteness, assume $x \in (-10, 10)$, and an experiment has 50,000 events on the average.



All is well: Test is for origin, an interior point of complex plane

Univariate unbinned goodness of fit tests

- If we have an iid sample of size N, $x_1, x_2, \ldots x_N$, we have a univariate unbinned dataset, where we assume a continuous sampling distribution. This dataset may be used to test hypotheses concerning the sampling distribution
- We list a few (additional) univariate test statistics that could be considered
  - Kolmogorov-Smirnov
  - Cramér–von Mises
  - Anderson-Darling
  - Watson (supplemental material)
  - Neyman smooth (supplemental material)

Kolmogorov-Smirnov test

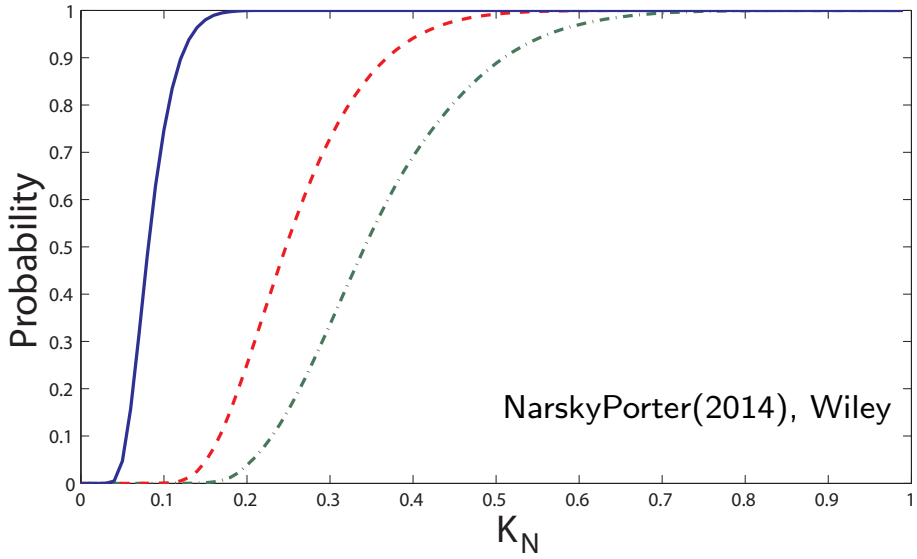
- Besides the χ^2 and LR tests, the Kolmogorov-Smirnov (KS) is familiar among physicists
- This tests for difference between two cumulative distributions
- Given any pair of cdf's, F and G on a sample space, it may be possible to define a distance or metric, $\rho(F, G)$, that returns a non-negative number satisfying all the normal properties of a distance on a metric space
- In particular we may define the distance

$$\rho(F, G) \equiv \sup_x |F(x) - G(x)|$$

When $G = F_N$ is the empirical cdf of our dataset, and F is the H_0 cdf, $\rho = K_N(F)$ provides the Kolmogorov-Smirnov GOF statistic

Kolmogorov-Smirnov test

- The distribution of $K_N(F)$ is independent of F (exercise), for continuous F
- Thus, the distribution of the Kolmogorov-Smirnov statistic has the convenient property that it is known and depends only on the sample size N



Cumulative distribution for the Kolmogorov-Smirnov statistic. From left to right, $N = 100, 10, 5$

*Cramér–von Mises test

- An obvious variation on the Kolmogorov-Smirnov approach is to replace the supremum distance function with another common measure of distance, the average squared deviation:

$$C_N^2(F) = \int_{-\infty}^{\infty} [F_N(y) - F(y)]^2 \, dF(y).$$

This is known as the Cramér–von Mises test. The distribution of $C_N^2(F)$ likewise does not depend on F

*Anderson-Darling test

- There are many test statistics one could invent based on the difference between cumulative distributions
- The KS test just described has the property that it tends to emphasize the region of most rapid change in the cdf (that is, the region of the peak of the pdf), as that is where the maximum difference under H_0 is likely to occur
- The Anderson-Darling test (AD) gives more weight to the tails of the distribution. For example, sampling is often approximately normal in the central region, but the tails may be significantly non-Gaussian. The Anderson-Darling test is powerful in detecting such cases
- The AD statistic is defined as:

$$A_N^2(x) = N \int_{-\infty}^{\infty} \frac{[F_N(y) - F(y)]^2}{F(y)\,[1 - F(y)]} dF(y),$$

where F_N is the empirical cdf of our dataset $x = x_1, \ldots, x_N$ and F is the cdf under H_0

Multivariate tests

- We have listed a number of univariate test statistics
- But maybe the x_n are multidimensional, say with dimension D
- Could also construct a D-dimensional generalization of the histogram, and apply the χ^2 test
- As long as the sample is large enough, we'll have an approximate χ^2 distribution
- If the appropriate in each dimension is 100 intervals, then we have a total of 100^D bins. It doesn't take many dimensions to get sparse bin populations, even with large datasets
- May mitigate with an adaptive binning procedure
- However, the power of the test suffers if there is important information in the distribution within a bin

Multivariate tests

- A variety of unbinned methods for dealing with the multivariate GOF problem have been proposed, eg,
  - Energy tests
  - Transformation to uniform distribution
  - Local density tests
  - Nearest neighbor methods
  - Kernel-based tests
  - Mixed sample tests
  - Using a classifier
- Williams, arXiv:1006.3019v2 (2010) studies several methods using the example of a high energy physics Dalitz plot analysis
- We defer discussion of these to the supplemental material

Significance

- When asking for the "significance" of an observation (of, perhaps a new effect), you ask for a test of the hypotheses:

$$H_0 : \text{There is no new effect}$$
$$H_1 : \text{There is a new effect}$$

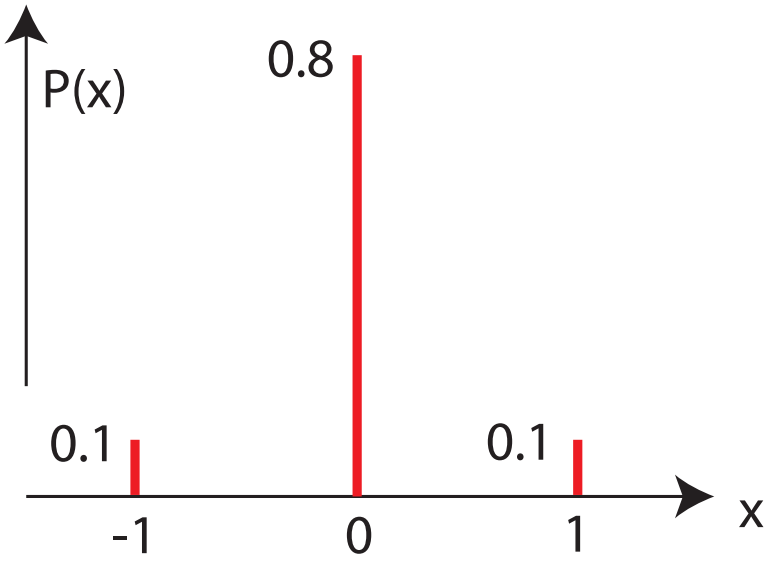
- Not really different from GOF and CIs
- Significance is quoted as the p-value for H_0
- A 68% confidence interval does not always tell you much about significance
  - The tails may be non-normal
  - A separate analysis is generally required, which models the tails appropriately.

Aside: Significance as "$n\sigma$"

HEP parlance is to say an effect has, e.g., "5σ" significance. At face value, this means the observation is "5 standard deviations" away from the mean (under H_0):

$$\sigma \equiv E\left[(x-\bar{x})^2\right].$$

But we often don't really mean this. Note that a 5σ effect of this sort may not be improbable:



$$P(|x-\bar{x}| = 5\sigma) = 20\% \ !$$

Aside: Significance as "$n\sigma$" (continued)

- Instead, we often mean that the probability (p-value) for the effect is given by the probability of a fluctuation in a normal distribution 5σ from the mean, i.e.,

$$P = P(|x| > 5), \text{ for } x \sim \mathrm{N}(0,1)$$
$$= 5.7 \times 10^{-7} \text{ (two-tailed probability)}$$

- Sometimes we really do mean 5σ, usually presuming that the sampling distribution is approximately normal. [May not be an accurate presumption when out in the tails!]
- Also popular to call $\sqrt{-2\Delta \ln \mathcal{L}}$ the "n" in "$n\sigma$".
  From: $\mathcal{L}_0(\theta = 0; x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(x/\sigma\right)^2\right]$,
  $\mathcal{L}_{\max}(\hat{\theta} = x; x) = \frac{1}{\sqrt{2\pi}\sigma}$, giving
  $\sqrt{-2\Delta \ln \mathcal{L}} = \sqrt{\Delta\chi^2} = x/\sigma = n$
- Desirable to be more concise by quoting probabilities, or p-values as is common in the statistics world. At least say what you mean!

Estimating significance: Pitfalls

What are the dangers? In a nutshell:

Unknown or unknowable sampling distributions

Ways to not know the distribution:

- The Improbable Tails
- Systematic Unknowns
- The Stopping Problem Example
- The exploratory Bump Hunt

The Stopping Problem

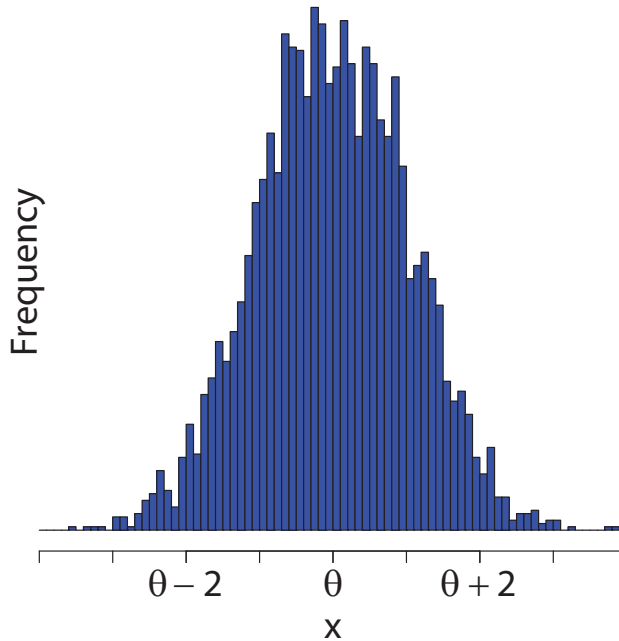
There is a strong tendency to work on an analysis until we are convinced that we got it "right", then we stop
Simple example: "Keep sampling" until we are satisfied
Motivate our example:

- Ample historical evidence that experimental measurements are sometimes biased by some preconception of what the answer "should be". For example, a preconception could be based on the result of another experiment, or on some theoretical prejudice
- A model for such a biased experiment is that the experimenter works "hard" until s/he gets the expected result, and then quits. Let's Consider a simple example of a distribution which could result from such a scenario [NIM A **368** (1996) 793]

Stopping Problem: Normal likelihood function example

Consider an experiment in which a measurement of a parameter θ corresponds to sampling from a Gaussian distribution of standard deviation one:

$$\mathrm{N}(x;\theta,1)dx = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} dx$$



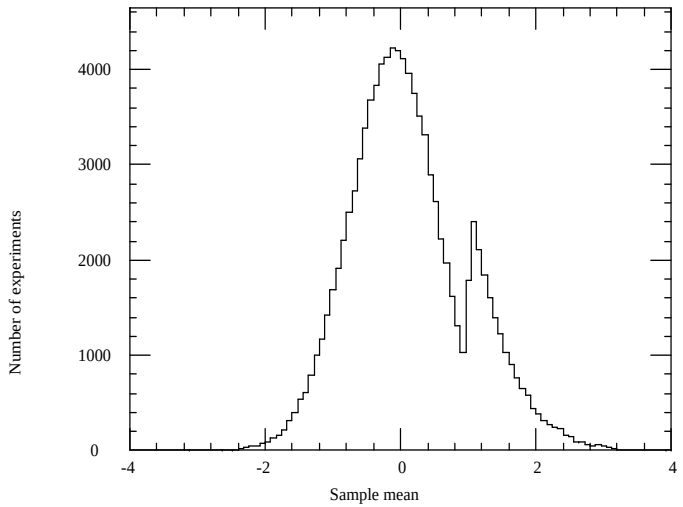
- Suppose the experimenter has a prejudice that θ is greater than one
- Subconsciously, s/he makes measurements until the sample mean, $m = \frac{1}{N}\sum_{n=1}^{N} x_n$, is greater than one, or until s/he becomes convinced (or tired) after a maximum of N measurements
- The experimenter then uses m to estimate θ

Stopping Problem: Normal likelihood function example

For illustration, assume that $N = 2$. In terms of the random variables m and n, the pdf is:

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

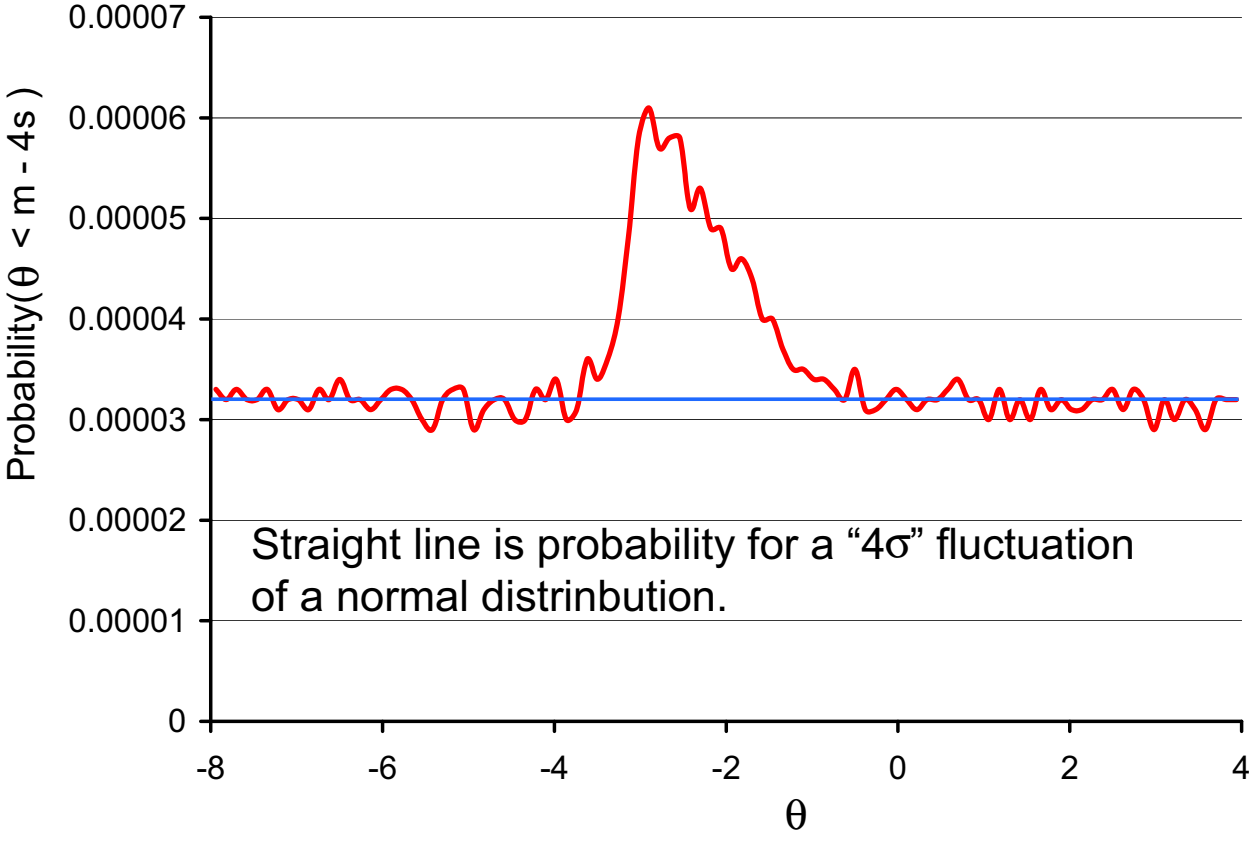
Histogram of sampling distribution for m, with pdf given by above equation, for $\theta = 0$



The likelihood function, as a function of θ, has the shape of a normal distribution, given any experimental result. The peak is at $\theta = m$, so m is the MLE for θ

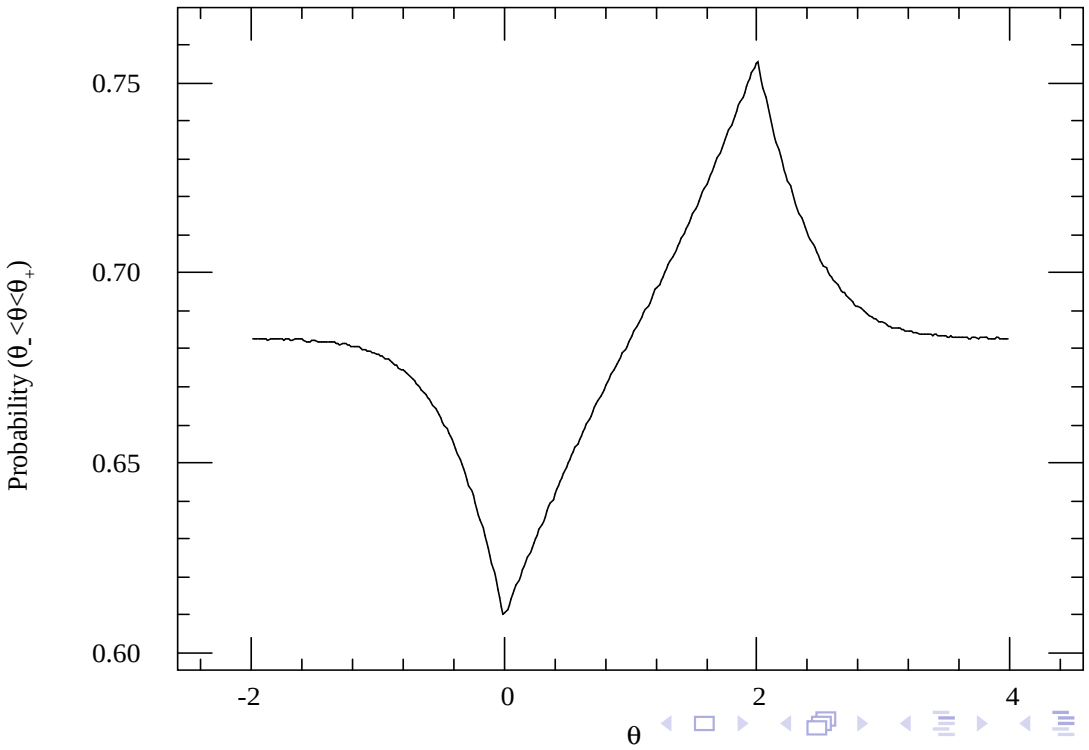
Stopping Problem: Normal likelihood function example

In spite of the normal form of the likelihood function, the sample mean is not sampled from a normal distribution. The "4σ" tail is more probable (for some θ) than the experimenter thinks.



*Stopping Problem: Normal likelihood function example

- The likelihood function, as a function of θ, is a Gaussian, given any experimental result.
- In spite of the normal form of the likelihood function, the sample mean is not sampled from a normal distribution.
- The interval defined by where the likelihood function falls by $e^{-1/2}$ does not correspond to a 68% CI



Next: Resampling methods

Resampling methods

- Permutation sampling
- Bootstrap
- Jackknife
- Cross-validation

Resampling – Introduction

- With modern computing resources, new methods of statistical analysis are practical
- Consider here approach to estimation afforded by resampling
- Avoid deriving formulas for the properties (eg, variance) of a statistic (eg, a point estimator). Avoid imposing potentially invalid assumptions (a model) about the distribution
- Instead, use the data in hand to produce a statistical ensemble
- Often, MC modeling is used to answer questions about the distribution of a statistic. Resampling offers an alternative approach that avoids assumptions in the MC model
- Resampling methods have become popular, and may be used to address different problems

Resampling – Introduction

- Resampling methods overlap considerably in application. Broadly, the permutation methods are generally used for hypothesis testing, the bootstrap and jackknife are used to estimate bias, variance and confidence intervals, while cross-validation is used to estimate the accuracy of a predictive model
- Theoretical basis for these methods is in asymptotic properties such as consistency and convergence. A thorough discussion is well outside our scope, and indeed this remains an active area of research. While the basic ideas are rather simple and elegant, care should be exercised in the execution. References for further study are noted

Permutation sampling

- Consider two datasets, A and B, of sample sizes N_A and N_B
- Wish to test whether the two populations are consistent with arising from the same underlying distribution by comparing some statistic computed on each dataset
- Call this statistic S_A or S_B; eg, it could be the sample mean, or the median, or the variance, etc.
- Under H_0 that the sampling distribution for A and B is the same, must have $E(S_A) = E(S_B)$. Thus the difference $\Delta S = S_B - S_A$ is a measure of the difference between the distributions
- In order to apply a test based on the observed ΔS, we need to know its distribution under H_0
- The method of permutation resampling permits us to estimate this distribution without any assumptions (e.g., normality) of the sampling distribution

Permutation sampling

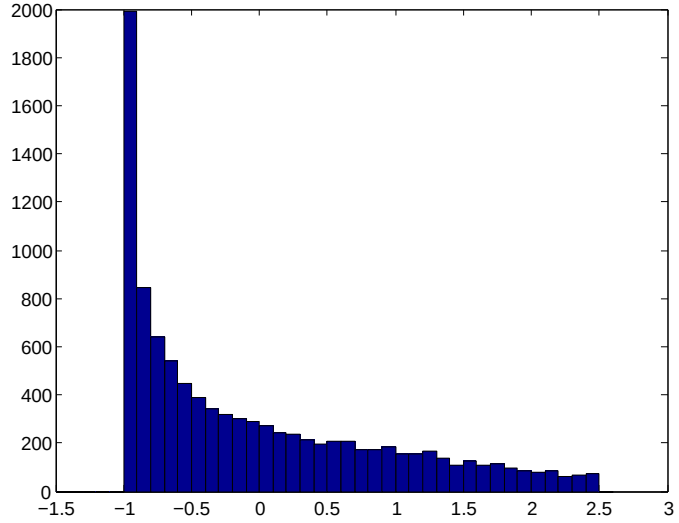
- Under H_0, all samples in A and B are equivalent, and therefore, any permutation of the labeling A and B has equal probability
- Thus, the distribution of ΔS is determined by considering all possible permutations of the A and B labels, grouping the $N_A + N_B$ samplings into sets of size N_A and N_B and computing ΔS for each permutation
- The set of values of ΔS from all permutations provides the estimated distribution of ΔS
- The actual ΔS may be compared with this distribution to obtain a p-value for H_0
- With additional effort, the test may be inverted to derive confidence intervals

Permutation sampling

- Permutations within a given A and B repartition are unnecessary. Even so, the computing required to evaluate all remaining $\binom{N_A + N_B}{N_A}$ permutations may be prohibitive
- For larger samples, we may randomly sample permutations without being exhaustive (providing a MC permutation test). This is called a conditional MC because it is a Monte Carlo simulation conditioned on the empirical data set

Permutation sampling

- For an example, we apply a MC permutation test to the distribution shown
- A sample of size $N_A = 1000$ is compared with a sample of size $N_B = 20$. We ask whether the mean of sample A is greater than the mean of sample B



NarskyPorter(2014), Wiley

- If the sampling were from a normal distribution, the statistic

$$t = (\bar{x}_A - \bar{x}_B) / \sqrt{\frac{(N_A - 1)s_A^2 + (N_b - 1)s_B^2}{N_A + N_B - 2} \frac{N_A + N_B}{N_A N_B}}$$

is distributed as the Student t-distribution with $N_A + N_B - 2$ degrees of freedom and may be used to test the desired hypothesis. Here, $\bar{x}$ refers to the sample mean, and s refers to the sample variance (computed with $N - 1$)

Permutation sampling

- However, for non-normal distributions the t-distribution may not apply
- In the case of the example, a test at nominal 1% significance rejects the (true) H_0 with only 0.57% probability
- The permutation test rejects H_0 at closer to the desired 1% probability (1.07% in a simulation using 10,000 permutations)
- We see that large errors may result from erroneous model assumptions, and the permutation test avoids making such assumptions
- The permutation test is a type of non-parametric test. It is thus more robust than parametric tests which depend on the validity of the sampling model
- Such an advantage may come at the price of power. If a reliable model is available, more powerful tests can generally be constructed

Bootstrap

A popular resampling technique is the bootstrap [Efron, Ann.Stat. **7** (1979) 1], motivated as a means to estimate the variance of an estimator for a population parameter. Can be especially useful when the sampling distribution is unknown, but is not limited to this situation. The basic bootstrap algorithm is as follows:

- Suppose we wish to estimate parameter(s) θ with an iid sample of size N, $X_1, \ldots, X_N$. Each of the X_n may be a vector of RVs. Eg, X_n could be an event in a particle physics dataset.
- Denote the estimator (eg, MLE) for θ by $\hat{\theta}(X)$
- Now form a set of B bootstrap replicas by randomly sampling sets of size N from X, with replacement. For example, in R:
  `for (b in 1:B) xr[b] = sample(x,replace=TRUE)`
- For replication b, form replicated estimator $\hat{\theta}(b)$, where the argument is now the replication index. This procedure may be called the MC bootstrap, because of the MC approach to choosing replications

Bootstrap

- These replications can be used to estimate the variance of the estimator. Simply take the sample mean and variance of the bootstrap estimators:

$$\bar{\theta} = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}(i),$$

$$s_\theta^2 = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}(i) - \bar{\theta})^2$$

- Then s_θ is the estimated standard deviation of the estimator $\hat{\theta}$
- If we have multiple parameters, the covariances may also be estimated with the bootstrap:

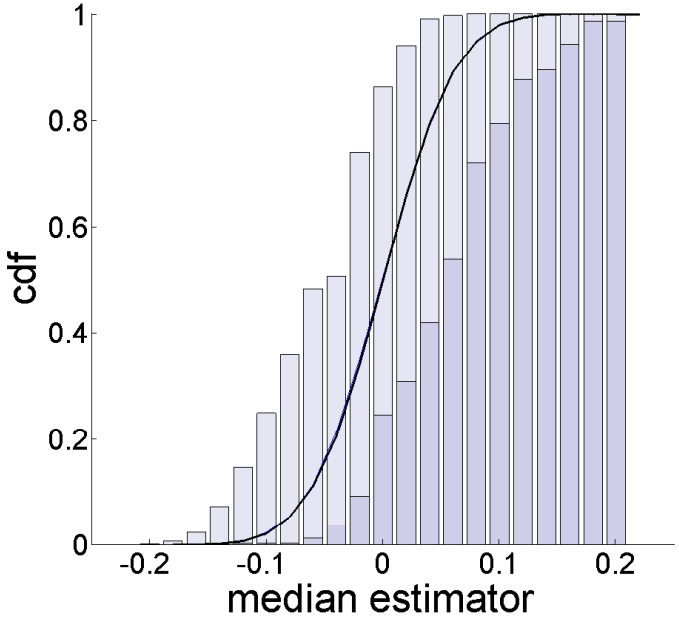
$$\mathrm{cov}(\hat{\theta}_m, \hat{\theta}_n) = \frac{1}{B-1}\sum_{i=1}^{B}\left[\hat{\theta}_m(i) - \bar{\theta}_m(i)\right]\left[\hat{\theta}_n(i) - \bar{\theta}_n(i)\right]$$

Bootstrap

- The bootstrap provides a way to approximately sample from the acutal parent distribution for X. Instead of repeatedly sampling from the actual distribution, which is likely to be impractical, we sample from the empirical distribution
- Example: Consider the estimation of the median parameter of a BW distribution
  - We use the sample median as our estimator
  - We will use the boostrap to estimate the variance of the sample median

Bootstrap

- Bootstrap estimate for the cdf of the median estimator, for two different samplings of size 1001 from a BW
- Curve shows the actual cdf
- Shapes of the distributions are similar, illustrating applicability of the bootstrap for estimating variance
- Translations of the bootstrap cdf's are expected from fluctuations in the median estimation, but do not affect the estimation of variance
- The variance estimates will also fluctuate around the true variance

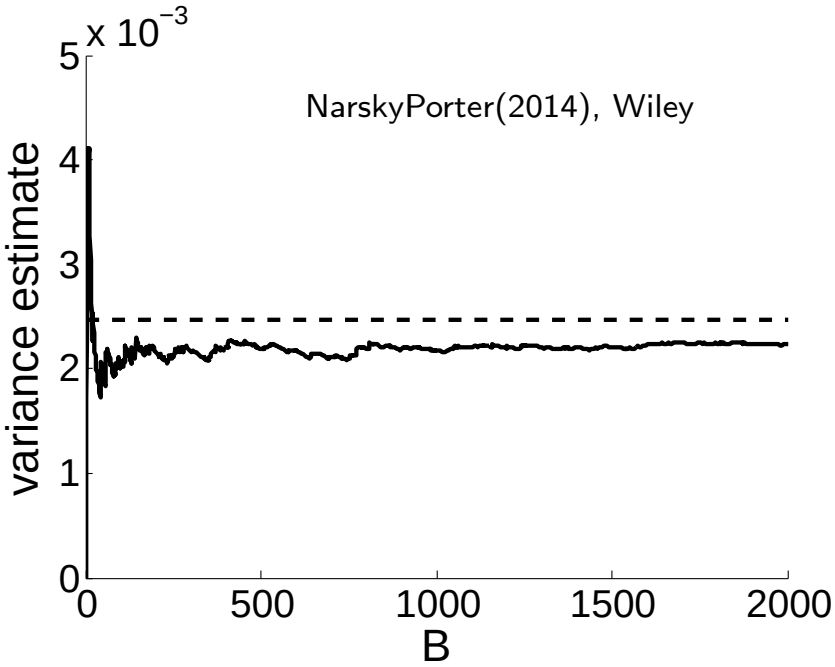


NarskyPorter(2014), Wiley

Bootstrap

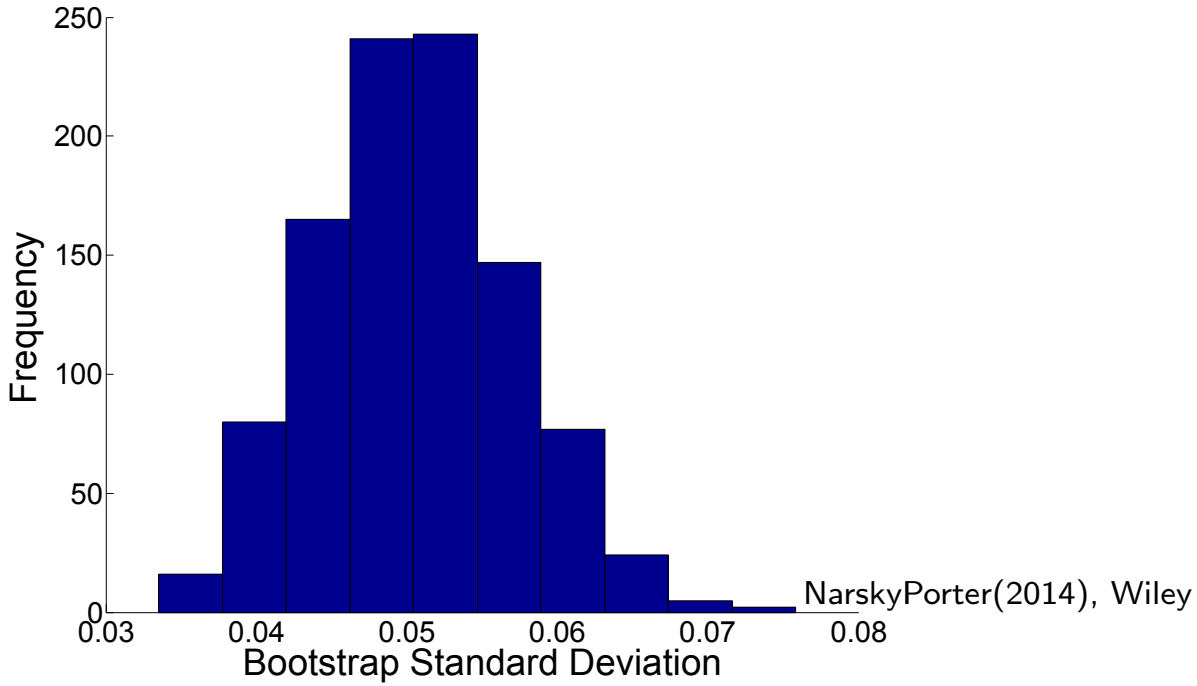
- The convergence of the bootstrap estimator for variance may be examined by plotting the estimator against the number of bootstrap samples

Convergence of bootstrap estimator for the variance of the median of a Cauchy distribution, as a function of number of bootstrap samples. The dashed line shows the actual variance



Bootstrap

- Bootstrap samples are based on the empirical distribution, hence will reflect any fluctuations that may be present
- Figure below shows the distribution of the bootstrap estimator for the standard deviation of the median, for our BW example
- The actual value of the standard deviation is 0.050



Bootstrap

- In this example, the sampling distribution is known, and the variance of the median can be precisely calculated
- This should of course be done when possible
- However, when the sampling distribution is not known the bootstrap provides a straightforward method for estimating the variance

Bootstrap – Estimation of bias

- The bootstrap may also be used to estimate the bias of a parameter estimator
- Suppose $\theta = \theta(F)$ is a parameter (e.g., the variance) of distribution F
- Estimator $\hat{\theta} = \hat{\theta}(x)$ is a statistic computed from a dataset of iid $x_1, \ldots, x_N$
- Denote the bias of $\hat{\theta}$ by:

$$b_F(\theta) = E_F\left[\hat{\theta}(x)\right] - \theta(F),$$

  where the subscript F denotes expectation value with respect to distribution F
- If we don't know what F is, even up to unknown θ, then we cannot evaluate the bias
- Let's see what the bootstrap can do...

Bootstrap – Estimation of bias

- Our data provides an approximation to F, the empirical distribution $\hat{F}$; use this to get a bootstrap estimate of bias:

$$b_{\hat{F}} = E_{\hat{F}}\left[\hat{\theta}(x^*)\right] - \theta(\hat{F})$$

  x^* is a bootstrap sample drawn from empirical distribution $\hat{F}$
- Since the empirical distribution $\hat{F}$ is used, the bias estimate depends only on that, and not on θ itself
- Eg, if we estimate the variance (θ) of F with $\hat{\theta} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2$, where $\bar{x}$ is the sample mean, we find (exercise) a bootstrap bias estimate of $-\frac{1}{N}\hat{\theta}$, which approximates the known bias of $-\theta/N$ for this estimator

Bootstrap – Estimation of bias

- In general, cannot evaluate the above expectation analytically, and must resort to MC bootstrap sampling
- Thus, we obtain a sequence of boostrap estimators $\hat{\theta}^*(1), \ldots, \hat{\theta}^*(B)$, where B is the number of bootstrap replications
- Approximate the desired expectation with the average of these,

$$\hat{\theta}^*(\cdot) \equiv \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}^*(i)$$

- Obtaining bias estimate

$$\hat{b}_{\hat{F};B} = \hat{\theta}^*(\cdot) - \theta(\hat{F})$$

Bootstrap confidence intervals

- The bootstrap may be used to estimate confidence intervals
- Useful when the sampling distribution is not known, and our familiar methods don't apply
- Wish to find a confidence interval for a population parameter, θ, where we have a statistic $\hat{\theta} = \hat{\theta}(X)$ to estimate θ
- Bootstrap sampling from X corresponds to obtaining an iid sample X^* of size N from the empirical distribution $\hat{F}$, our surrogate for the actual distribution. Given a set of B such bootstrap replicas, we compute the corresponding estimators $\hat{\theta}^*(1), \ldots, \hat{\theta}^*(B)$
- Let $P_B(u)$ be the emprical cdf for this set of $\hat{\theta}^*$s. This is a monotonic step function
- To obtain an estimated upper confidence bound at the $1 - \alpha$ confidence level, we solve for $\hat{u}_\alpha$ in

$$\hat{u}_\alpha = P_B^{-1}(\alpha)$$

Bootstrap confidence intervals

- Because of the discontinuities, this may be only approximately soluble, so let us be more explicit
  - Begin by ordering all of the $\hat{\theta}^*$'s
  - Then count up until reaching a fraction of at least $1 - \alpha$ of them
  - The smallest such $\hat{\theta}^*$ value is $\hat{u}_\alpha$
- That is, the bootstrap percentile method corresponds to finding the appropriate percentile of P_B
- This method can readily be adapted to estimating two-sided confidence intervals

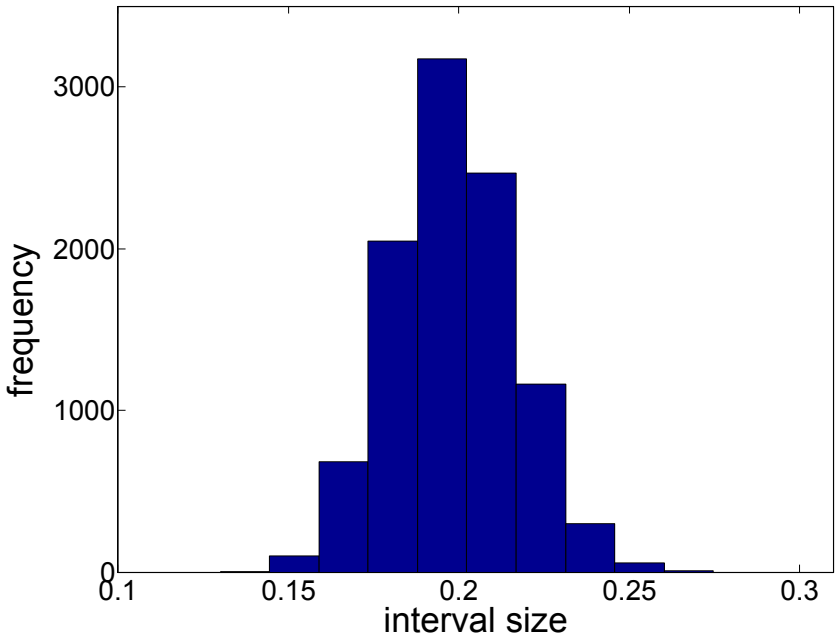
Bootstrap confidence intervals – example

- For example, we estimate the population mean θ and estimate a 68% CI
- Our example uses a dataset of size $N = 100$ from a $\mathrm{N}(0, 1)$ distribution. The estimator is the sample mean, $\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} x_i$. If we knew that we were sampling from a Gaussian, we would typically quote the interval $(\hat{\theta} - 0.1, \hat{\theta} + 0.1)$
- However, supposing that we don't know anything about the sampling distribution, we use the bootstrap. In MATLAB:
  ```
  ci = bootci(nbootstrap,{@mean,x},'alpha',alpha,'type','
  ```
  Here, `@mean` is a function that computes the sample mean, `alpha` ≈ 0.32, and `per` specifies that we use the percentile method described above
- We perform the calculation with `nbootstrap` $= 300$ bootstrap samples (sufficient for illustration, a bit small in practice)

Bootstrap confidence intervals – example

- The method is found to cover with a probability of ∼67.8%, in good agreement with the desired 68.3%

- Histogram of the size of the CI, for 10000 experiments
- Compare with the fixed size of 0.2 in the usual approach when the distribution is known to be Gaussian



NarskyPorter(2014), Wiley

- Find that the bootstrap may be used to estimate confidence intervals with accurate coverage
- The cost of not knowing the sampling distribution shows up as variation of the interval size
- Refinements exist, eg, BCa CI (see, eg, NarskyPorter(2014), Wiley)

Bootstrap – Use in particle physics

- The largest use of the bootstrap in particle physics has so far been in classification
- However, it is beginning to be used in parametric error analysis as well
  - Estimation of uncertainty in changes in parameters when including $\rho(1700)$ contribution in a $B \to \rho\pi$ time-dependent Dalitz plot analysis (arXiv:1304.3503v1)
  - Evaluation of uncertainty in pdf from limited MC sample size (arXiv:1303.0571v1)
- I expect that this usage will grow

Jackknife

An algorithm known as the jackknife may also be used to estimate both variance and bias of a parameter estimator

- Consider situation in which we use our sample $x_1, \ldots, x_N$ to estimate the mean of the parent population. Naturally, we'll use the sample mean for our estimator:

$$\hat{\theta}(x) = \frac{1}{N} \sum_{n=1}^{N} x_n$$

- What is the variance, $\sigma^2_{\hat{\theta}}$, of our estimator? We may estimate this using $1/N$ times the sample variance:

$$s^2 = \frac{1}{N(N-1)} \sum_{n=1}^{N} (x_n - \hat{\theta})^2$$

- Now let's try something. Let $x_{-i} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N\}$ be the dataset obtained by removing x_i from our original sample. The set x_{-i} is called a jackknife sample

Jackknife

- The sample mean for set x_{-i} is $\hat{\theta}_{-i} \equiv \frac{1}{N-1}\sum_{n\neq i} x_n$
- Then we may write $s^2 = \frac{N-1}{N}\sum_{i=1}^{N}(\hat{\theta} - \hat{\theta}_{-i})^2$
- Define $\hat{\theta}'$ as the sample mean over all the $\hat{\theta}_{-i}$'s:

$$\hat{\theta}' \equiv \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_{-i}$$

- Use this to rewrite the estimated variance in the form:

$$s^2 = \frac{N-1}{N}\sum_{i=1}^{N}(\hat{\theta}' - \hat{\theta}_{-i})^2 + (N-1)(\hat{\theta}' - \hat{\theta})^2$$

- First term is the variance with respect to the (jackknife) sample mean, that is an estimate of the variance of $\hat{\theta}$ about its mean with the $(N-1)/N$ scale factor. Second term compares $\hat{\theta}'$ with $\hat{\theta}$, and is related to the bias

Jackknife

- We can make this concrete as follows: Assume that our estimator is consistent in the sense that $E\hat{\theta} \xrightarrow{N\to\infty} \theta$. If this limit is approached at leading order as $1/N$, then $b_{N-1} \propto \frac{N}{N-1} b_N$, for large enough N, where b_N denotes the bias for a sample of size N:

$$b_N(\theta) \equiv E\hat{\theta} - \theta,$$

Hence, we can use our jackknife samples to estimate the bias of $\hat{\theta}$, noting that:

$$E(\hat{\theta}' - \hat{\theta}) = b_{N-1} - b_N = \frac{1}{N-1} b_N$$

Jackknife

- We summarize the jackknife method for estimating variance and bias: The jackknife estimate for the variance of estimator $\hat{\theta}$ with respect to its expectation value is

$$s'^2 = \frac{N-1}{N} \sum_{i=1}^{N} (\hat{\theta}' - \hat{\theta}_{-i})^2$$

The jackknife estimate for the bias of estimator $\hat{\theta}$ with respect to θ is

$$\hat{b}_N = (N-1)(\hat{\theta}' - \hat{\theta})$$

A bias correction may be applied to estimator $\hat{\theta}$ to obtain the improved (bias-corrected jackknife estimate) estimate for θ:

$$\hat{\theta}^* = N\hat{\theta} - (N-1)\hat{\theta}'$$

Jackknife

- The algorithm readily generalizes beyond our construction. Instead of the sample mean, we may substitute any statistic of the form:

$$\hat{\theta}(x) = a + \sum_{n=1}^{N} b(x_n)$$

- This is called a linear statistic – it is here simply a linear function of a transformation from our original set of iid RVs to another set of iid RVs
- Our above discussion goes through without difficulty for this case
- For non-linear statistics, we may still apply the method, but with due caution

Jackknife

- Example with known properties: estimation of a population variance, σ^2. Consider estimate

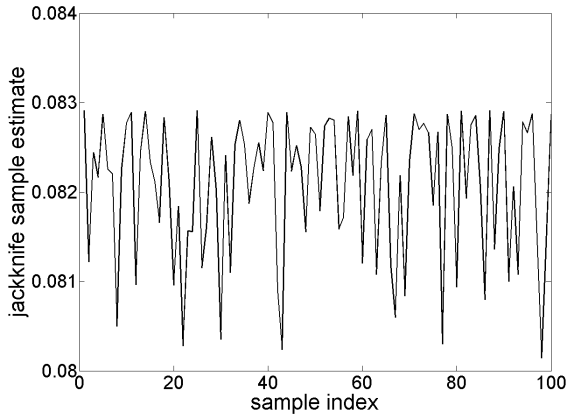
$$\hat{\sigma}^2(X) = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2,$$

  where $\bar{x}$ is the sample mean
- We then construct the jackknife samples according to our prescription, and calculate the jackknife estimates of variance and bias for $\hat{\sigma}^2$
- We carry out this exercise on a sample of size $N = 100$ drawn from a uniform distribution on $(0, 1)$
- Our sample has an estimated variance of 0.0821, according to the equation above

Jackknife

- The jackknife estimates the bias of this estimator as -0.000829, compared with the known bias of $-1/12N = -0.000833$
- The jackknife estimate for the variance of the estimator gives 0.0077^2
- Hence bias-corrected estimate of the variance of the sampling distribution is $0.0821 + 0.000829 = 0.0829 \pm 0.0077$, compared with the known variance $1/12 = 0.0833$

NarskyPorter(2014), Wiley



Estimation of variance using the jackknife

In this case the bias correction is small compared with the statistical uncertainty. In fact, this statistic is not linear, but our specific example is "close" to linear, so it is not unexpected that the method works

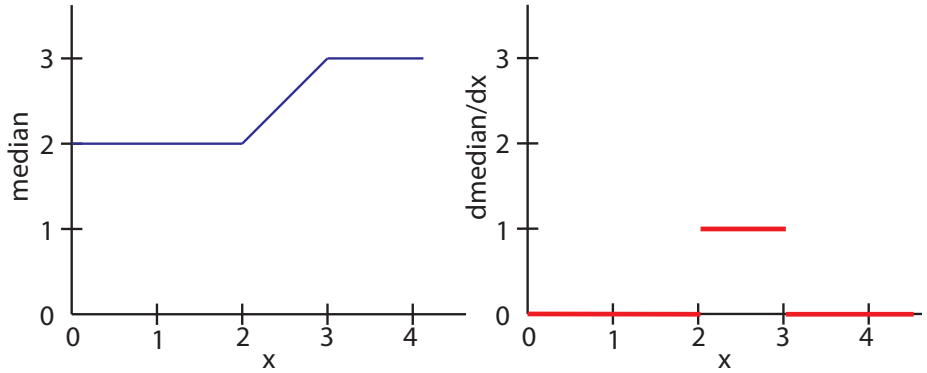
Jackknife vs Bootstrap?

- Jackknife and bootstrap can both be used to estimate variance and bias. It is thus natural to ask which is better. As "better" is vague, the answer is "it depends". Some considerations:
- The jackknife does better in the estimation of bias at least for linear statistics, such as the mean. The sample mean is an unbiased estimator for this parameter, and that is what the jackknife tells us. Estimating this bias with a set of B randomly drawn bootstrap samples will in general produce a non-zero bias estimate due to the random fluctuations
- Jackknife requires examining N samples of size $N-1$, while the bootstrap requires examining some large number B of bootstrap datasets each of size N. Unless N is large, the computation required in the jackknife is more manageable
- The jackknife does not make use of all of the available information in the case of nonlinear statistics – it represents a linear approximation to the (exhaustive) bootstrap and may be comparatively inefficient

Jackknife vs Bootstrap?

- The jackknife runs into trouble with non-smooth statistics. Smoothness captures a notion of continuity on a dataset – small changes in the data are reflected as small changes in the statistic

Eg, the median is a non-smooth statistic: Suppose we have a dataset of size 3 with $x, a, b = x, 2, 3$



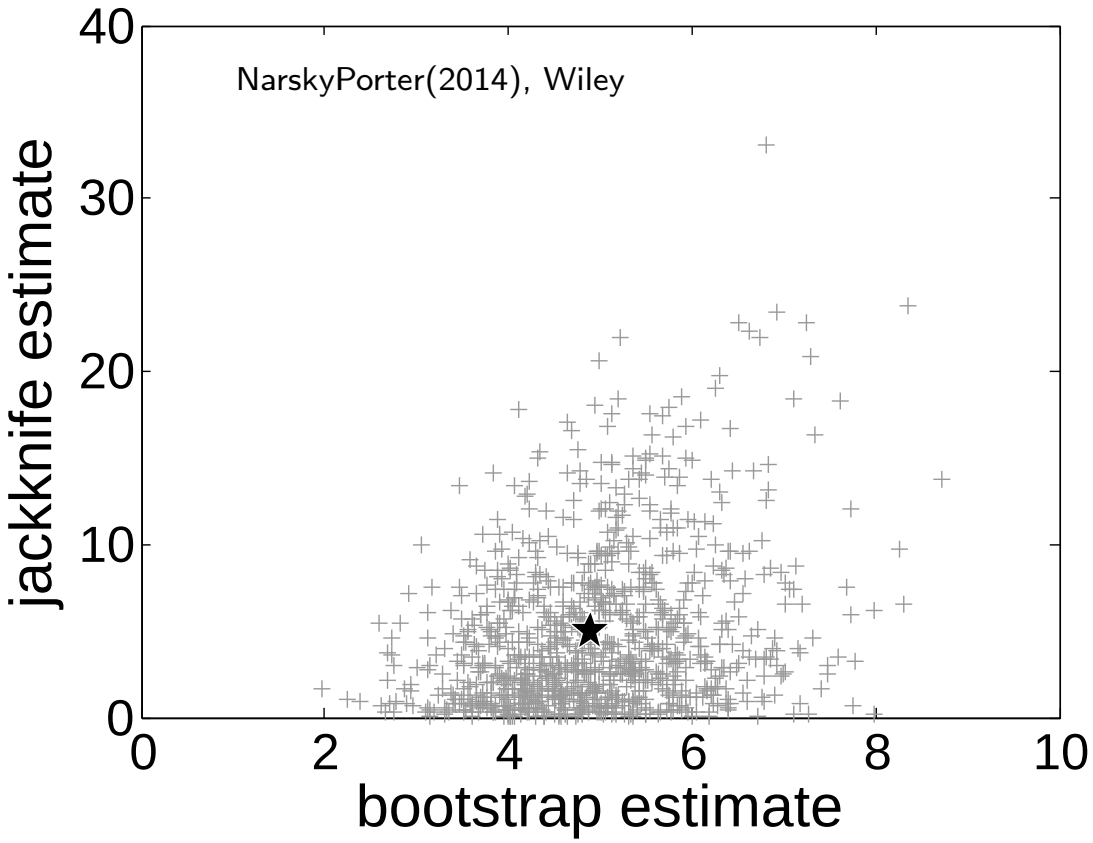
- Gives trouble in jackknife estimation. In a dataset of any size, the leave-one-out jackknife samples will have at most three different values for the medians. If we attempt to estimate the variance of the median estimator using the jackknife, we obtain unreliable (i.e., inconsistent) results. The bootstrap does much better

Jackknife vs Bootstrap? – Example

Estimation of standard deviation for the median of a $N = 100$ sample size from a $U(0, 100)$ sample. Star is true value.

Note scale difference!

High estimates occur when the two samplings on either side of the median are far apart



- This problem can be mitigated, at the cost of additional computing power (and perhaps bias and variance in more typical situations), with the delete-d jackknife

Cross-validation

Suppose in a regression analysis we want to know whether adding another resonance, or another term in an angular distribution fits the data significantly better

- If we can assume the sampling errors are Gaussian, we simply compare the residual sum-of-squares (RSS) values in a Fisher-Snedecor F-test
- Otherwise Cross-validation can be used
- Consider a bivariate dataset $\mathcal{D} \equiv \{(X_n, Y_n), n = 1, \ldots, N\}$
- Suppose we are interested in finding the best straight line fit. In this case, our regression function is $r(x) = ax + b$. We estimate parameters a and b by finding $\hat{a}$ and $\hat{b}$ that minimize (assuming equal weights for simplicity):

$$\sum_{n=1}^{N} (Y_n - \hat{a}X_n - \hat{b})^2$$

Cross-validation

- The value of a new sampling is predicted given X_{N+1}:

$$\hat{Y}_{N+1} = \hat{a}X_{N+1} + \hat{b}$$

- We wish to estimate the expected prediction error (EPE) for Y_{N+1}:

$$\mathrm{EPE}(r) = E\left\{[Y - r(X)]^2\right\} = \int [y - r(x)]^2 f(x,y)\, dxdy$$

  The expectation is over both X and Y, it is the expected error (squared) over the joint distribution
- We don't know $f(x,y)$ so we use the data to estimate it

Cross-validation

- A simple approach is to divide our $\{(X_i, Y_i), i = 1, \ldots, N\}$ dataset into two pieces, perhaps two halves
- One piece (the training set) could be used to determine the regression function, and the other piece (the testing set) could be used to estimate the EPE
- This seems a bit wasteful, since we are only using half of the available data to obtain our regression function, and we could do a better job with all of the data
- The next thing that occurs to us is to reverse the roles of the two pieces and somehow average the results, and this is a pretty good idea
- But let's take this to an extreme, known as leave-one-out cross-validation

Leave-one-out cross-validation

The algorithm for leave-one-out cross-validation is as follows:

1. Form N subsets of the dataset $\mathcal{D}$, each one leaving out a different datum, say (X_k, Y_k). We'll use subscript $-k$ to denote quantities obtained omitting datum (X_k, Y_k). Likewise, we let $\mathcal{D}_{-k}$ be the dataset leaving out (X_k, Y_k)
2. Do the regression on dataset $\mathcal{D}_{-k}$, obtaining regression function r_{-k}
3. Using this regression predict the value for the missing point:

$$\hat{Y}_k = r_{-k}(X_k)$$

4. Repeat this process for $k = 1, \ldots, N$. Estimate the EPE according to:

$$\frac{1}{N}\sum_{k=1}^{N}(\hat{Y}_k - Y_k)^2$$

Cross-validation example

- Let's try an example application: We have a dataset and wish to consider whether to use the straight line relation

$$Y = aX + b,$$

  or the quadratic relation

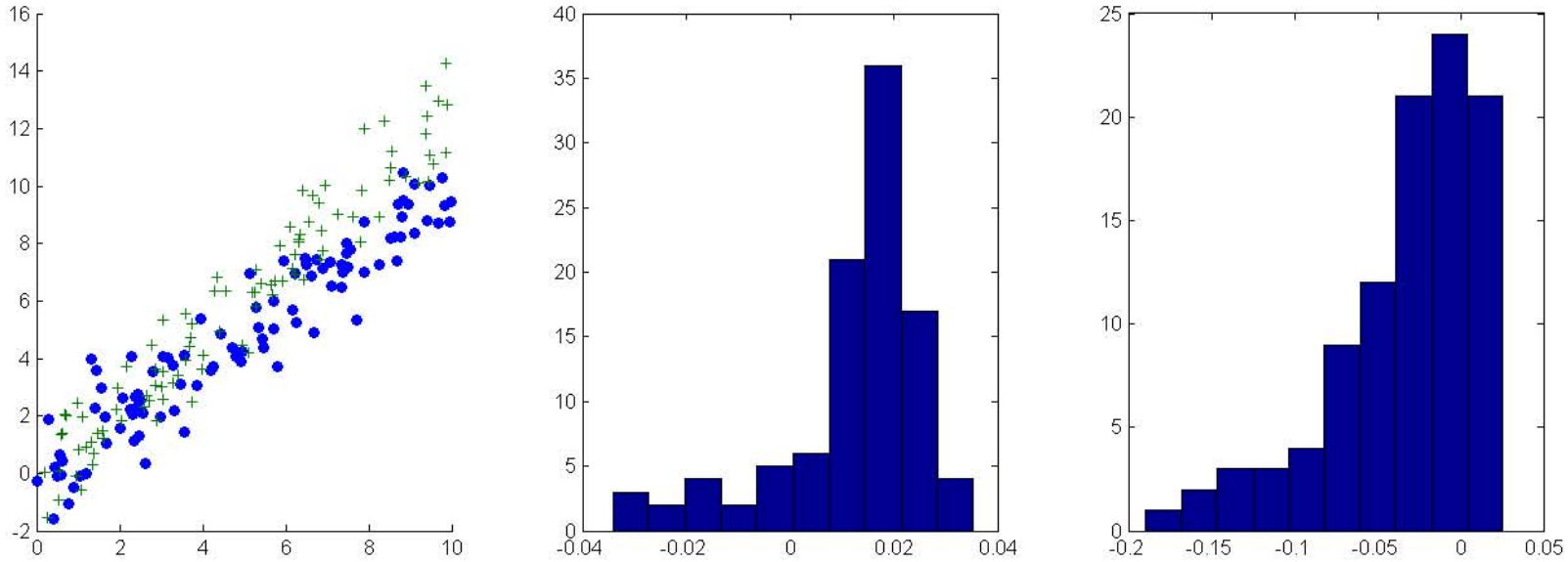
$$Y = a'X^2 + b'X + c$$

- We know that the fitted residuals for the quadratic model will always be smaller than for the linear model
- The predictive error is not necessarily smaller with the additional adjustable parameters. We thus use our EPE as a means to decide between models
- We'll try this on a simulated dataset of size $N = 100$
  - Where the linear model is correct
  - Where a quadratic term is present
- MATLAB function `crossval` is used to perform the EPE estimates, with calls of the form:

```
crossval('mse',x,y,'Predfun',@linereg,'leaveout',1);
```

Cross-validation example

- Left: Data samples generated according to a linear model $Y = X$ (filled circles) or a quadratic model $Y = X + 0.03X^2$ (plus symbols)
- Middle: Distribution of quadratic model minus linear model EPE for data generated according to a linear model
- Right: Distribution of quadratic model minus linear model EPE for data generated according to a quadratic model



(from NarskyPorter(2014), Wiley)

Cross-validation example

- When the linear model is correct, choosing the linear model when the difference is larger than zero gets it right in 84 out of 100 cases
- When the quadratic model is correct, choosing the quadratic model when the difference is less than zero gets it right in 79 out of 100 cases

Remark

- Leave-one-out cross-validation is suitable for small datasets. For large N, required computer time may be prohibitive
- For large dataasets, we may use $\mathcal{K}$-fold cross-validation. The dataset is divided into $\mathcal{K}$ disjoint subsets of size $m \approx N/\mathcal{K}$. Leave-one-out cross-validation corresponds to $\mathcal{K} = N$
- See, eg, NarskyPorter(2014), Wiley, and references therein for further development

Next: Density Estimation

Supplemental material

Watson test

- Another variation on the Cramér–von Mises approach is the Watson test

$$U_N^2 = N \int_{-\infty}^{\infty} \left\{ F_N(x) - F(x) - \int_{-\infty}^{\infty} [F_N(y) - F(y)]\, dF(y) \right\}^2 dF(x)$$

- Here, the difference between empirical and theoretical distributions is "corrected" by subtracting the mean difference.
- Thus, this test ignores a simple shift and concentrates on higher order differences.

Neyman smooth test

- Unsatisfied with limitations of the χ^2 test (in particular, its inability to detect runs, for example, successive histogram bins improbably "running" higher than the model, as deviations from H_0), Neyman devised the Neyman smooth test
- The basic idea is to transform the data to quantities uniformly-distributed under H_0, and then use Legendre polynomials to frame H_1 as a "smooth" pdf with degree of smoothness determined by the order of the polynomials used
- The observed moments with respect to the Legendre polynomials may then be compared with the values of zero expected under H_0
- One feature of this approach is that it provides a framework to investigate in more detail the reason for a bad fit to the model

Neyman smooth test

- If the density under the null hypothesis is $g(x|H_0)$, we make the transformation $x \to u$ according to:

$$u = \int_{-\infty}^{x} g(x'|H_0)dx'$$

- Under H_0, u is uniformly distributed on $(0, 1)$, with pdf $f(u) = 1$
- The GOF test of H_0 is thus one of testing uniformity of the distribution for u

Neyman smooth test

- This is framed as a test against H_1:

$$H_1 : f(u|\theta) = \exp\left[-C(\theta) + \sum_{k=1}^{K} \theta_k P_k(2u-1)\right], \quad u \in (0,1),$$

  where $\theta = (\theta_1, \ldots, \theta_k)$, $C(\theta)$ provides normalization, and P_k is the k-th Legendre polynomial
- The parameters θ are expansion coefficients in a truncated Legendre series.
- The highest degree polynomial included, K, is called the order of the test.
- Notice that the test may be rephrased as:

$$\begin{aligned} H_0 : &\quad \theta = 0, \\ H_1 : &\quad \theta \neq 0 \end{aligned}$$

Neyman smooth test

- To complete the construction of a test statistic, a set of "optimal" values for θ may be chosen, typically by maximizing the likelihood under H_1, yielding parameter estimates $\hat{\theta}(x)$
- Then either a likelihood ratio statistic or a Wald test statistic may be constructed. For example, the likelihood ratio statistic is:

$$\lambda = \frac{L(\theta = 0; x)}{L(\theta = \hat{\theta}; x)}$$

- H_0 has been treated as a simple hypothesis. However, the method may be applied to composite H_0 as well. According to the above approach, the test may be expressed in terms of a statistic with χ^2 asymptotic distribution

*Multivariate tests – Energy tests

- An approach to the multivariate goodness of fit problem with a physical appeal is an energy test
- The test is based on the quantity

$$\phi = \frac{1}{2} \int \int g(x) g(x') R(\|x - x'\|) dx dx',$$

  where $R(y)$ is a "potential energy" function
- If $R = 1/|x - x'|$ we see that ϕ looks like the Coulomb energy of a charge distribution given by g
- Here, let g be the difference between a pdf being tested and the pdf under H_0
- Smaller values of ϕ correspond to better agreement with H_0

*Multivariate tests – Energy tests

- Idea may be implemented with a MC approach, in order to deal with the difficulty of multi-dimensional integrals
- Thus, we compare our observations $x_1, \ldots, x_N$ with simulated data under H_0, $y_1, \ldots y_M$
- See Aslan and Zech, NIM A **537** (2005) 626; Aslan and Zech, arXiv:hep-ex/0203010v5 for details
- The distribution of ϕ under H_0 depends on R as well as H_0. It is not readily calculated, but may be estimated via simulations

*Multivariate tests – Energy tests

- Various choices for function R have been proposed, with different choices suitable for different distrubutions.
- For example, a logarithmic form:

$$R_{\log}(r) = \begin{cases} -\log r & r > a, \\ -\log a & r < a \end{cases}$$

  may appropriate for slowly varying distributions.
- A Gaussian form may be more optimal for rapidly varying distributions. For example, Williams (2010) chooses

$$R(\|x_i - y_j\|) = \exp\left[-\frac{\|x_i - y_j\|^2}{2\sigma(x_i)\sigma(x_j)}\right],$$

  with $\sigma(x)$ vaying as $1/f_0(x)$ where f_0 is the null hypothesis density, so that areas of high density are relatively highly weighted. With this weight function, Williams obtains powerful results for a Dalitz plot analysis with rapid variations

*MV tests – Transformation to a uniform distribution

- An step that is useful in some approaches to multivariate GOF testing is to first make a transformation of the distribution under the null hypothesis.
- Idea is to transform the distribution to a uniform distribution on the unit D-cube.
- Then one can work on tests of uniformity in multidimensions.
- See Rosenblatt, Ann.Math.Stat., **23** (1952) 470; NarskyPorter(2014), Wiley for details
- The transformation is not unique and GOF on the different choices may not be equivalent. Good practice is to first put it into approximately factorized form, perhaps with a rotation
- If not factorizable, at least look for factorizable subsets that can be investigated further

*MV tests – Transformation to a uniform distribution

- With our transformation complete, we may design tests for uniformity on the D-cube.
- Whatever statistic is defined, its distribution under the null hypothesis of uniformity is completely determined by D and N.
- Many tests can be imagined using the distances between sampled data in this D-cube.
- See, e.g., NarskyPorter(2014), Wiley for specific implementations, including SLEUTH (Abbott et al. Phys.Rev. D **62** (2000)092004) and nearest neighbors (Narsky, arXiv:physics/0306171v1 (2003)).

*MV tests – Local density tests

- The MV GOF problem can be thought of as testing whether the spatial distribution of observed points in our observation space is consistent with the hypothesized model.
- Thus, we may estimate the local density of observations around any given point and compare with the prediction of the model
- Idea is similar to the notion of nearest neighbors, except that we compare densities of observations, not distributions of distances.
- Can apply to either transformed or non-transformed variables.
- Let $I(\text{statement})$ be an indicator function, that is:

$$I(\text{statement}) = \begin{cases} 1 & \text{if statement is true,} \\ 0 & \text{if statement is false.} \end{cases}$$

*MV tests – Local density tests

- Let $|x_i - x_j|$ be the distance between observtions x_i and x_j. Any metric could be tried, Euclidean distance is probably a good choice.
- Then

$$N_i \equiv \sum_{j \neq i}^{N} I(\|x_i - x_j\| < r)$$

  counts the number of observations in our dataset within a distance r of observation i.
- The quantity N_i/V_r, where V_r is the volume of the (hyper-) sphere of radius r, is thus a measure of the local density of observations in the vicinity of observation i.
- If the underlying sampling distribution is uniform, then we have expectation value $E(N_i) = (N-1)V_r/V$, where V is the total volume of our sampling space (assumed finite)

*MV tests – Local density tests

- If the data are sampled from a uniform distribution, the observations will tend to be "maximally spread out" in a statistical sense, compared with a distribution with peaks. [Another extreme, where the data may be sampled from a regular grid of values, will be more spread out than for a uniform distribution, and may also be of interest]
- That is, the values of N_i/V_r will tend to cluster around the average density of points (excluding one), $(N-1)/V$.
- If the sampling distribution is not uniform, there must be clustering around other values than $(N-1)/V$, that is, regions of higher density.
- This results in N_i values that are higher than for the uniform case, on average. Hence, a candidate statistic is the sum of the N_i's over the dataset, for a given r.
- For details of computing a suitable test statistic K, including the boundary issue, see Williams (2010); Ripley, J.R.Stat.Soc. B **39** (1977) 172; NPwiley

*MV tests – Local density tests

- Statistic K may be computed for different values of r, giving sensitivity at different scales, and a plot of K against r provides a useful visual tool.
- This method provides a simple way to test for uniformity. It can be generalized for other distributions by "dividing out" the the hypothethical local density of observations.
- Williams (2010) applies this to the example of the Dalitz plot. The method is especially powerful when there are large local deviations from the model. It is not so useful for small datasets, where the local density estimates have large variance.
- For a review of the intimately related subject of nearest neighbor methods, see http://www.public.iastate.edu/~pdixon/stat406/NearestNeighbor.pdf

*MV tests – Other methods

- Kernel based tests:
  - We will discuss kernel density estimation later
  - Provides a means to GOF test by comparing the estimator with the H_0 model
- Mixed sample tests
  - May wish to compare two datasets to see whether they are consistent with being drawn from the same population.
  - For example, comparing an experimentally observed dataset with a Monte Carlo simulation.
  - One approach to the multivariate problem combines the nearest neighbor idea with pooling the data, that is combining the two datasets
  - A statistic is formed that is sensitive to whether neighbors are from the same or different dataset
- Using a classifier
  - Many classifiers are well-suited to MV problems
  - Run a classifier on the two samples, getting scores
  - Compare scores with a univariate test (due care to determining expected distribution under H_0)

Chi-square test for shape

Even though we don't expect it to follow a χ^2 distribution, we may evaluate the test statistic:

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

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

Geometric (BDM) test for shape

Geometric motivation: Let the bin contents of a histogram define a vector in a k-dimensional space. If two vectors are drawn from the same distribution (null hypothesis), they will tend to point in the same direction (not interested in the lengths of the vectors here). If we represent each histogram as a unit vector with components:

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

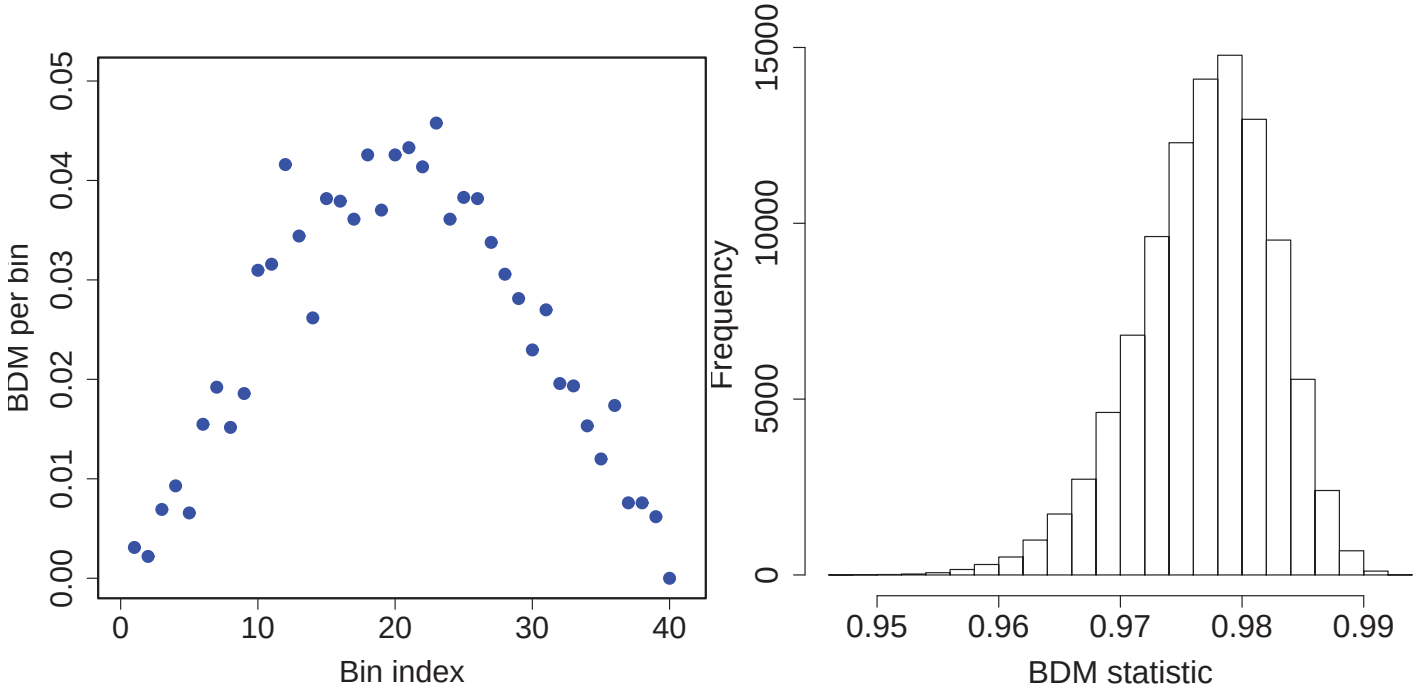
we may form the "dot product" test statistic:

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

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

Sample application of BDM test

Apply this formalism to our example. The sum over bins gives 0.986. According to our estimated distribution of this statistic under the null hypothesis, this gives a P-value of 0.97, similar to the χ^2 test result (0.95).



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

Right: Estimated distribution of the BDM statistic for the null hypothesis in the example.

Kolmogorov-Smirnov test

Another approach to a shape test may be based on the Kolmogorov-Smirnov (KS) idea: Estimate the maximum difference between observed and predicted cumulative distribution functions and compare with expectations.
Modify the KS statistic to apply to comparison of histograms as follows. Assume neither histogram is empty. Form the "cumulative distribution histograms" according to:

$$u_{ci} = \sum_{j=1}^{i} u_j / N_u \qquad v_{ci} = \sum_{j=1}^{i} v_j / N_v .$$

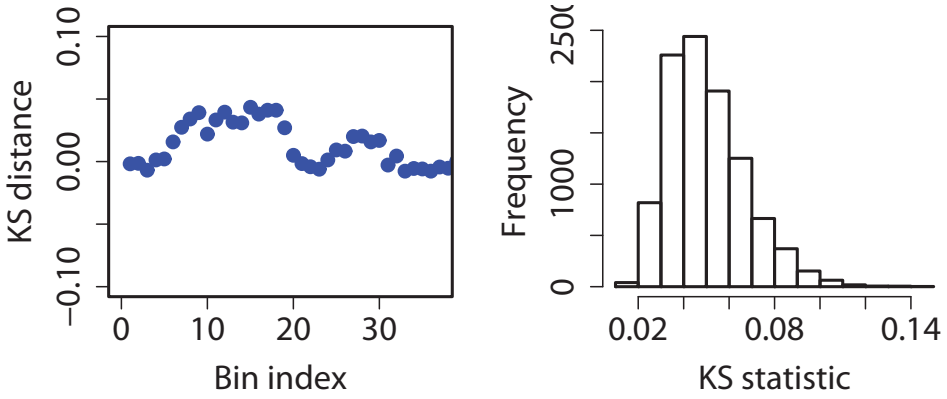
Then compute the test statistic:

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

(We consider only the two-tail test here.)

Sample application of KS test

Apply this formalism to our example. The maximum over bins is 0.043. Estimating the distribution of this statistic under H_0 gives a p-value of 0.61, somewhat smaller than for the χ^2 test result, but indicating consistency of the histograms. KS will tend to emphasize differences near the peak of the distribution, since that is where the Poisson fluctuations are greatest.



Left: Bin-by-bin distances for the KS test statistic for the example. Right: Estimated PDF of the KS distance under H_0 in the example.

Cramér-von-Mises test

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

The test statistic for comparing the two samples $x_1, x_2, \ldots, x_N$ and $y_1, y_2, \ldots, y_M$ is [T. W. Anderson, *On the Distribution of the Two-Sample Cramér-Von Mises Criterion*, Ann. Math. Stat. **33** (1962) 1148]:

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

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

Cramér-von-Mises test – Adaptation to comparing histograms

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

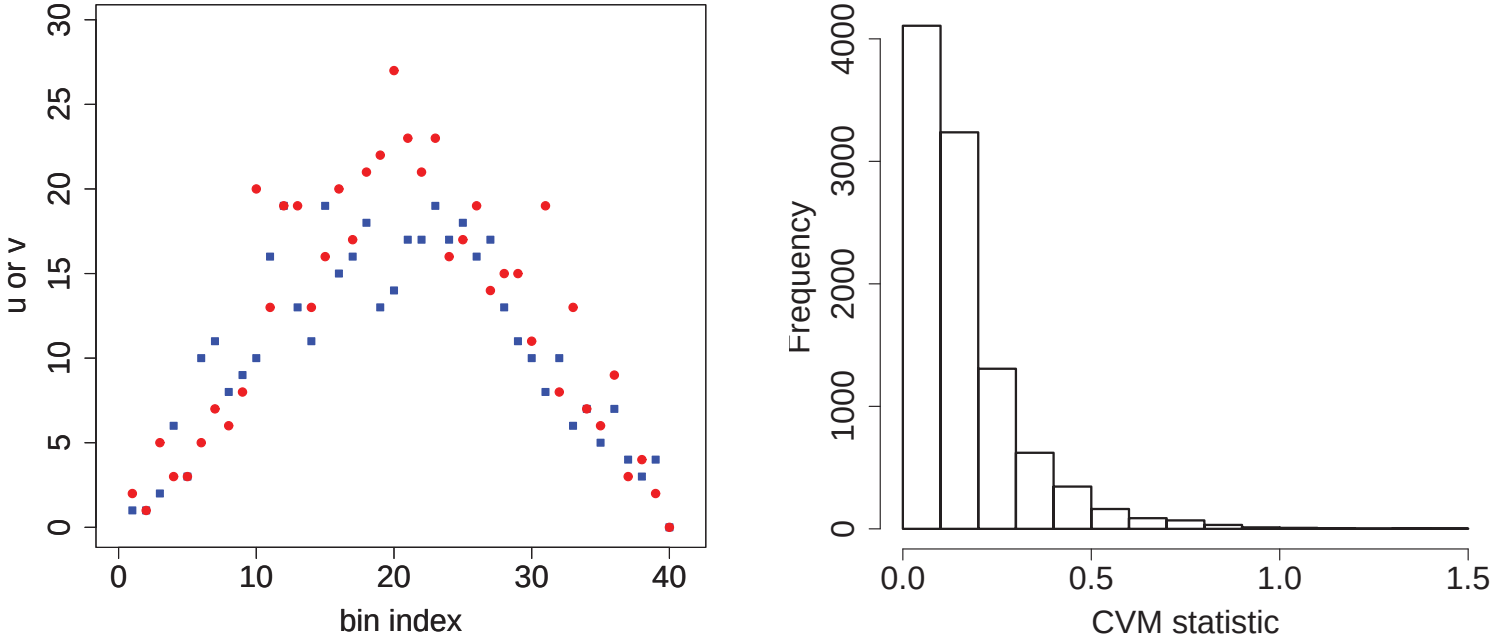
$$E_u(z) = \sum_{j=1}^{i} u_i / N_u.$$

Then the test statistic is:

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

Sample application of CVM test

Apply this formalism to our example, finding $T_{\mathrm{CVM}} = 0.132$. The resulting estimated distribution under the null hypothesis is shown below. According to our estimated distribution of this statistic under the null hypothesis, this gives a P-value of 0.45, somewhat smaller than the χ^2 test result.



Left: Example histograms.
Right: Estimated PDF of the CVM statistic under H_0 for the example.

Anderson-Darling (AD) test for shape

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

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

where $F_m(x)$ is the empirical CDF of dataset $x_1, \ldots x_m$.

Anderson-Darling (AD) test, adaptation to comparing histograms

Scholz and Stephens [*k-Sample Anderson-Darling Tests*, J. Amer. Stat. Assoc. **82** (1987) 918] provide a form of this statistic for a k-sample test on grouped data (e.g., as might be used to compare k histograms). The expression of interest for two histograms is:

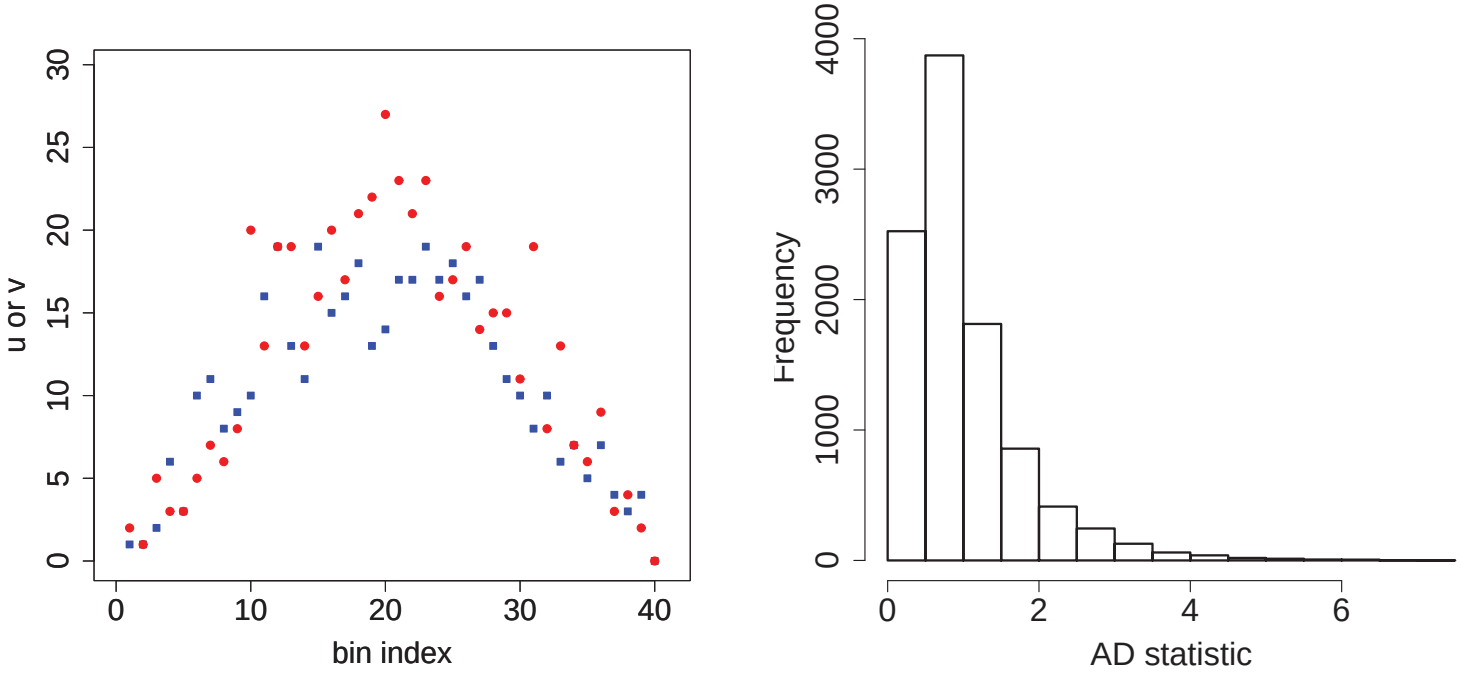
$$T_{\mathrm{AD}} = \frac{1}{N_u + N_v} \sum_{j=k_{\min}}^{k_{\max}-1} \frac{t_j}{\Sigma_j \left(N_u + N_v - \Sigma_j\right)} gg\{ \left[(N_u + N_v)\Sigma_{uj} - N_u \Sigma_j\right]^2 / N_u + \left[(N_u + N_v)\Sigma_{vj} - N_v \Sigma_j\right]^2 / N_v gg\},$$

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

$$\Sigma_{uj} \equiv \sum_{i=1}^{j} u_i, \quad \Sigma_{vj} \equiv \sum_{i=1}^{j} v_i, \quad \text{and} \quad \Sigma_j \equiv \sum_{i=1}^{j} t_i = \Sigma_{uj} + \Sigma_{vj}.$$

Sample application of AD test for shape

We apply this formalism to our example. The sum over bins gives 0.849. According to our estimated distribution of this statistic under the null hypothesis, this gives a P-value of 0.45, somewhat smaller than the χ^2 test result



Left: Example histograms.
Right: Estimated distribution of the AD test statistic for the null hypothesis for the example.

Likelihood ratio test for shape

Base a shape test on the same conditional likelihood idea as for the normalization test. Now there is a binomial associated with each bin. 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 $i = 1, 2, \ldots, k$. 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:

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

Likelihood ratio test for shape (continued)

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

- The value of " a" is not specified;
- We still have a multivariate distribution.

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

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

We use a likelihood ratio statistic to reduce the problem to a single variable. This will be the likelihood under the null hypothesis (with a given by its maximum likelihood estimator), divided by the maximum of the likelihood under the alternative hypothesis.

Likelihood ratio test for shape (continued)

We form the ratio:

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

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

Thus, we rewrite our test statistic as:

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

In practice, we'll work with

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

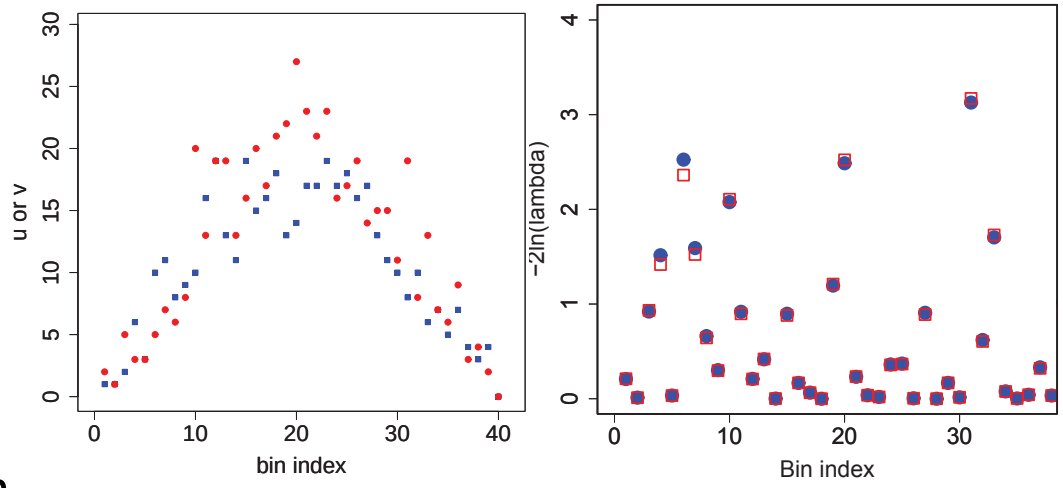
If $u_i = v_i = 0$, the bin contributes zero.

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

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

Sample application of ln λ test

Apply this test to example, obtaining
$-2\ln\lambda = \sum_{i=1}^{k} -2\ln\lambda_i = 25.3$.
Asymptotically, $-2\ln\lambda$ should be distributed as a χ^2 with $N_{\mathrm{DOF}} = k - 1$, or $N_{\mathrm{DOF}} = 39$. If valid, this gives a p-value of 0.96, to be compared with a probability of 0.96 according to the estimated actual distribution.
Obtain nearly the same answer as the application of the chi-square calculation with no bins combined, a result of nearly bin-by-bin equality of the two statistics.



Left: Example histogram
Right: Value of $-2\ln\lambda_i$ (circles) or χ_i^2 (squares) as a function of histogram bin in the comparison of the two distributions

Likelihood value test

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 many cases in the limit of large statistics it is essentially the chi-square statistic, so there are known situations were it is a plausible statistic to use. We look at it here.

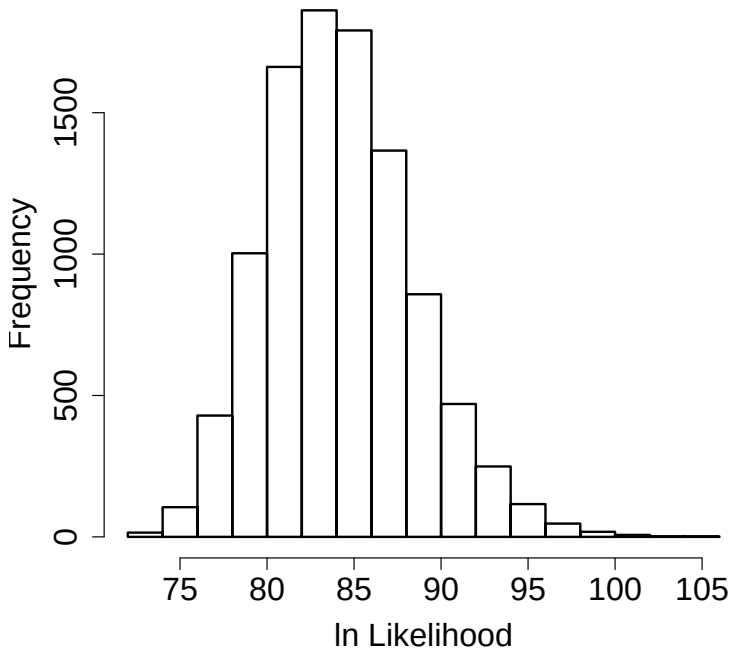
Using the results in the previous section, the test statistic is:

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

If either $N_u = 0$ or $N_v = 0$, then $\ln \mathcal{L} = 0$.

Sample application of the ln $\mathcal{L}$ test

Apply this test to the example. The sum over bins gives 79. According to our estimated distribution of this statistic under the null hypothesis, this gives a p-value of 0.91, similar to the χ^2 test result. The fact that it is similar may be expected because our example is reasonably approximated by the large statistics limit.



Estimated distribution of the ln $\mathcal{L}$ test statistic for the null hypothesis in our example.

Consistency of two correlated results

E.g., question of whether a new analysis is consistent with an old analysis

- Often, new analysis is a combination of additional data plus changed (improved...) analysis of original data
- The issue is handling the correlation in testing for consistency in the overlapping data
- Statistical differences can arise even comparing results based on the same events

Simple check: Given a sampling $\hat{\theta}_1, \hat{\theta}_2$ from a bivariate normal distribution $\mathrm{N}(\theta, \sigma_1, \sigma_2, \rho)$, with $\langle\hat{\theta}_1\rangle = \langle\hat{\theta}_2\rangle = \theta$, the difference $\Delta\theta \equiv \hat{\theta}_2 - \hat{\theta}_1$ is $\mathrm{N}(0, \sigma)$-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 difference is in the range $(\sigma_1 - \sigma_2)^2 \ldots (\sigma_1 + \sigma_2)^2$. If we at least believe $\rho \geq 0$ then the maximum variance of the difference is $\sigma_1^2 + \sigma_2^2$

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 $x_i, i = 1, \ldots, 10$.
Assume the sampling distribution for x_i is $\mathrm{N}(m, \sigma_i)$.
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 further information which might be useful: we know the experimental resolutions, σ_i for each measurement. We form another unbiased estimator, $\hat{m}_2$, for m:

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

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

Example continued

- The results are certainly correlated, so question of consistency arises (we know the error on the difference is between 0.023 and 0.055)
- In this example, the difference between the results is 0.058 ± 0.036, where the 0.036 error includes the correlation ($\rho = 0.41$).

Consistency – Evaluating the Correlation

- Art Snyder developed an approximate formula for evaluating the correlation in a comparison of ML analyses (eg, in one-dimensional case)
- Suppose we perform two ML analysis, with event likelihoods $\mathcal{L}_1$, $\mathcal{L}_2$, on the same set of N events [may use different information in each analysis]. Results are estimators $\hat{\theta}_1$, $\hat{\theta}_2$ for parameter θ
- The correlation coefficient ρ may be estimated according to:

$$\rho \approx \frac{\sum_{i=1}^{N} R_i \frac{d \ln \mathcal{L}_{1i}}{d\theta}\Big|_{\theta=\hat{\theta}_1} \frac{d \ln \mathcal{L}_{2i}}{d\theta}\Big|_{\theta=\hat{\theta}_2}}{\sqrt{\left(\sum_{i=1}^{N} \frac{d^2 \ln \mathcal{L}_{1i}}{d\theta^2}\Big|_{\theta=\theta_0}\right)\left(\sum_{i=1}^{N} \frac{d^2 \ln \mathcal{L}_{2i}}{d\theta^2}\Big|_{\theta=\theta_0}\right)}},$$

where (θ_0 is an expansion reference point)

$$R_i = \left[1 - (\hat{\theta}_1 - \theta_0)\frac{d^2 \ln \mathcal{L}_{1i}}{d\theta^2}\Big|_{\theta=\theta_0} gg / \frac{d \ln \mathcal{L}_{1i}}{d\theta}\Big|_{\theta=\hat{\theta}_0}\right]\left[1 - (\hat{\theta}_2 - \theta_0)\frac{d^2 \ln \mathcal{L}_{2i}}{d\theta^2}\Big|_{\theta=\theta_0} gg / \frac{d \ln \mathcal{L}_{2i}}{d\theta}\Big|_{\theta=\hat{\theta}_0}\right]$$

Consistency – Evaluating the Correlation

If $\theta_0 \approx \hat{\theta}_1 \approx \hat{\theta}_2$, then

$$\rho \approx \tilde{\sigma}_{\theta_1} \tilde{\sigma}_{\theta_2} \sum_{i=1}^{N} \frac{d \ln \mathcal{L}_{1i}}{d\theta}\Big|_{\theta=\hat{\theta}_0} \frac{d \ln \mathcal{L}_{2i}}{d\theta}\Big|_{\theta=\hat{\theta}_0},$$

where $\tilde{\sigma}^2_{\theta_k} \equiv 1/\sum_{i=1}^{N} \left(\frac{d\mathcal{L}_{ki}}{d\theta}\big|_{\theta=\theta_0} \right)^2$

Consistency – Example: $\sin 2\beta$

- 32×10^6 $B\bar{B}$ pairs – PRL, vol 87, 27 August 2001:

$$\sin 2\beta = 0.59 \pm 0.14(\mathrm{stat}) \pm 0.05(\mathrm{syst})$$

- 62×10^6 $B\bar{B}$ pairs – SLAC-PUB-9153, March 2002:

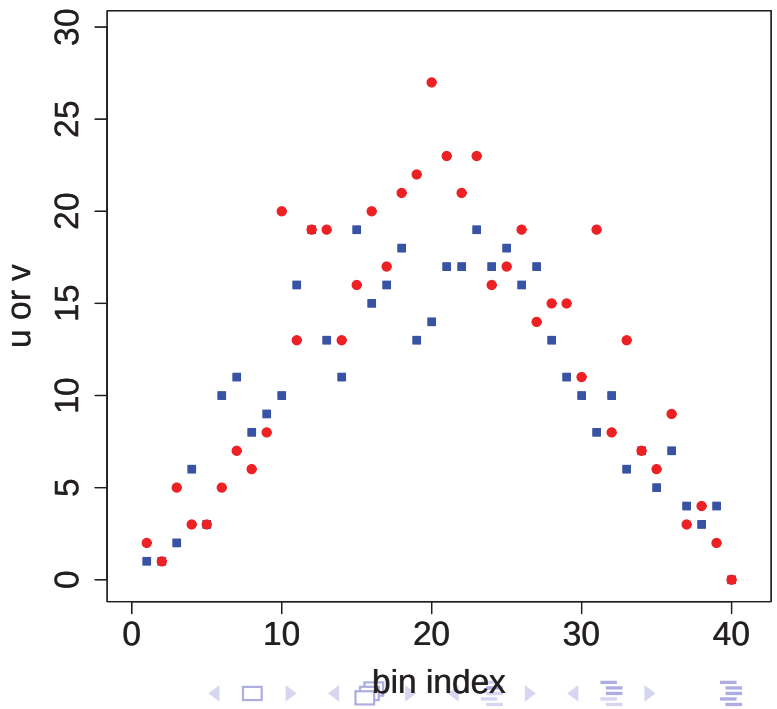
$$\sin 2\beta = 0.75 \pm 0.09(\mathrm{stat}) \pm 0.04(\mathrm{syst})$$

- Second result includes the earlier data, re-reconstructed. Analysis involves multivariate ML fits; reprocessing changes, eg, relative likelihood for an event to be signal or background. Not simply counting events. Are the two results statistically consistent?
- If these were independent data sets, a difference of 0.16 ± 0.17 would not be a worry. The issue is the correlation.
- A specialized analysis deriving from the previous formula is performed on the events in common between the two analyses. A correlation of $\rho = 0.87$ is deduced, yielding a difference of $\sim 2.2\sigma$.

Goodness-of-Fit – Considerations and Comparisons
Case Study: Testing Consistency of Two Histograms

- Sometimes we have two histograms and are faced with the question: "Are they consistent?"
- That is, are our two histograms consistent with having been sampled from the same parent distribution?
- Each histogram represents a sampling from a multivariate Poisson distribution.

Example of two histograms to be compared



arXiv:0804.0380

Testing Consistency of Two Histograms

There are two variants of interest to this question:

- 1. We wish to test the hypothesis (absolute equality):
  - H_0: 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 (shape equality):
  - 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

Testing Consistency of Two Histograms – Some Context

- We have listed several tests addressing whether a dataset is consistent with having been drawn from some continuous distribution
- These tests may often be adapted to address whether two datasets have been drawn from the same continuous distribution, called two-sample tests
- These tests may then be further adapted to the present problem, of determining whether two histograms have the same shape. Also discussed as comparing whether two (or more) rows of a table are consistent

Notation, Conventions

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:

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

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

where the vectors μ and ν are the mean bin contents of the respective histograms

Notation, Conventions (continued)

We define:

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

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

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

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

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

Will drop distinction between random variable and realization

*Large Statistics Case

If all of the bin contents of both histograms are large, we use the approximation that the bin contents are normally distributed. Under H_0,

$$E(u_i) = E(v_i) \equiv \mu_i, \;\; i = 1, \ldots, k$$

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

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

and let the standard deviation for Δ_i be denoted σ_i. Then the sampling distribution of the difference between the two histograms is:

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

*Large Statistics Case – Test Statistic

- This suggests the test statistic:

$$T = \sum_{i=1}^{k} \frac{\Delta_i^2}{\sigma_i^2}$$

- If the σ_i were known, this would simply be distributed according to the chi-square distribution with k degrees of freedom
- We'll use the Neyman modified χ^2 statistic, in which σ_i^2 is estimated by the sampled bin contents. We have already noted that it is asymptotially χ^2 distributed
- We'll refer to this approach as a "χ^2" test

*Large Statistics Algorithm – absolute

We suggest the following algorithm for this test:

1. For σ_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

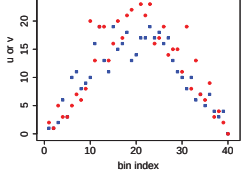
3. Estimate the p-value according to a χ^2 with k degrees of freedom. Note that this is not an exact result

*Large Statistics Algorithm – shapes

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

- Let
$$N = 0.5(N_u + N_v).$$
Scale u and v according to:
$$u_i \to u_i' = u_i(N/N_u)$$
$$v_i \to v_i' = v_i(N/N_v).$$
- Estimate σ_i^2 with:
$$\hat{\sigma}_i^2 = \left(\frac{N}{N_u}\right) u_i + \left(\frac{N}{N_v}\right) v_i.$$
- Statistic T is thus evaluated according to:
$$T = \sum_{i=1}^{k} \frac{\left(\frac{u_i}{N_u} - \frac{v_i}{N_v}\right)^2}{\frac{u_i}{N_u^2} + \frac{v_i}{N_v^2}}.$$
- Estimate the p-value according to a chi-square with $k-1$ degrees of freedom. Note that this is not an exact result.

*Application to Example



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

Type of test	T	$P(\chi^2 > T)$	p-value
χ^2 absolute comparison	29.4	0.88	0.86
χ^2 shape comparison	24.9	0.96	0.95
Likelihood Ratio shape comparison	25.3	0.96	0.96
Kolmogorov-Smirnov shape comparison	0.043	NA	0.61
Bhattacharyya shape comparison	0.986	NA	0.97
Cramér-Von-Mises shape comparison	0.132	NA	0.45
Anderson-Darling shape comparison	0.849	NA	0.45
Likelihood value shape comparison	79	NA	0.91

Column "p-value" attempts a better estimate of the probability, under H_0, that a value for T will be as large as observed. Compare with column $P(\chi^2 > T)$ column, the probability assuming T follows a χ^2 distribution with NDOF degrees of freedom.

*Application to Example (continued)

- The absolute comparison yields slightly poorer agreement 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 severely diluted by the bin-by-bin test
- The two histograms were generated with a 10% difference in total expected counts
- Lesson:
  The more you know about what you want to test, the better (more powerful) the test you can make

*An Issue: We don't know H_0!

Evaluation of the probability under H_0 is in fact problematic, since H_0 isn't completely specified

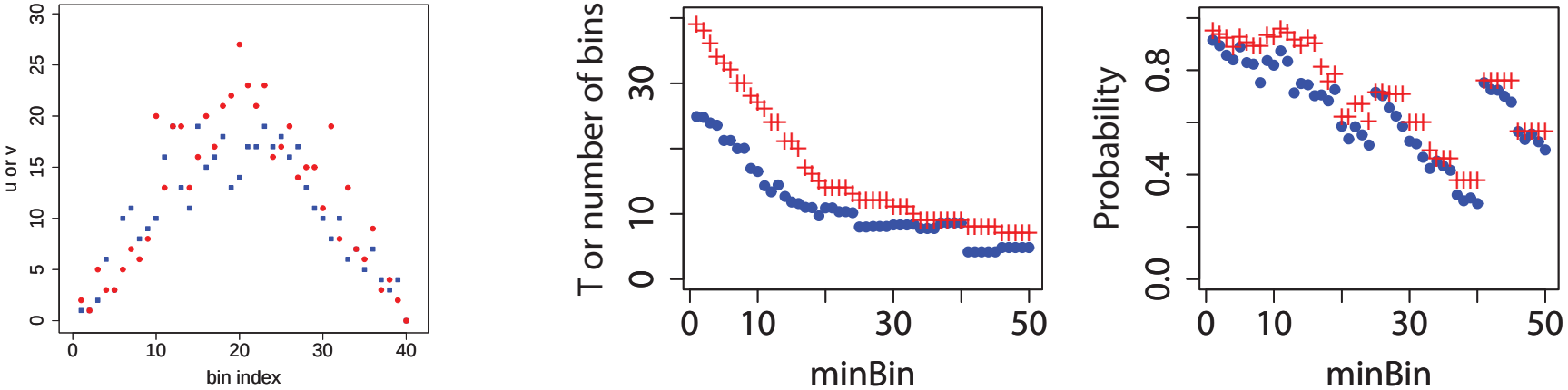
- The problem is the dependence of Poisson probabilities on the absolute numbers of counts. Probabilities for differences in Poisson counts are not invariant under the total expected number of counts
- Unfortunately, we don't know the true mean numbers of counts in each bin. Thus, we must estimate these means
- The procedure adopted here has been to use the MLE (see later) for the mean numbers, in the null hypothesis.

We'll have further discussion of this procedure below – It does not always yield valid results

*General (Including Small Statistics) Case

- If the bin contents are not large, then the normal approximation may not be good enough and the "χ^2 statistic" may not follow a χ^2 distribution
- Simple approach is to combine bins until the normal approximation is valid. In some cases this doesn't lose too much statistical power
- Try this on our example, as a function of the minimum number of events per bin. The algorithm is to combine corresponding bins in both histograms until both have at least "`minBin`" counts in each bin

*Combining bins for χ^2 test



Left: The example pair of histograms

Middle: The blue dots show the value of the test statistic T, and the red pluses show the number of histogram bins for the data in the example, as a function of the minimum number of counts per bin

Right: The p-value for consistency of the two datasets in the example The red pluses show the probability for a chi-square distribution, and the blue dots show the probability for the actual distribution, with an estimated null hypothesis.

*Working with the Poissons - Normalization Test

- Alternative: Work with the Poisson distribution. Separate the problem of the shape from that of the overall normalization.
- To test normalization, compare totals over all bins between the histograms. Distribution is

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

- The null hypothesis is $H_0 : \mu_T = \nu_T$, to be tested against alternative $H_1 : \mu_T \neq \nu_T$. We are interested in the difference between the two means; the sum is a nuisance parameter. Hence, consider: [Application of conditional likelihood!]

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

*Normalization test (general case)

- This probability permits us to construct a uniformly most powerful test of our hypothesis (E. L. Lehmann and Joseph P. Romano, *Testing Statistical Hypotheses*, 3rd ed., Springer, NY (2005)). It is a binomial distribution, for given N. The UMP 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, for a two-tailed probability. Compare with our earlier estimate of 0.025 in the normal approximation.
- Mimicing more closely the normal estimate by excluding one-half the probability at the endpoints, we obtain 0.025, essentially the normal number.

Testing the Shape – Catalog of Tests

Many possible tests. We consider 7:

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

We have introduced some of these in one-sample discussion; with suitable modification, we show the results for the present two-sample problem in the earlier table [Details in supplement]

Distributions Under the Null Hypothesis

- We have been overly glib so far – we still have some lessons to learn. . .
  - How do we know our p-values are right?
  - Which tests are better (powerful)?
- When the asymptotic distribution may not be good enough, we would like to know the probability distribution of our test statistic under H_0
- Difficulty: H_0 is not completely specified!
- The problem is that the distribution depends on the values of $\nu_i = a\mu_i$. H_0 only says $\nu_i = a\mu_i$, but says nothing about what μ_i might be
- Also doesn't specify a, but that complication appears manageable (although in extreme situations one might need to check for dependence on a)

Estimating the null hypothesis

- Turn again to the data to make an estimate for μ_i, to be used in estimating the distribution of our test statistic
- A straightforward approach is to use the ML parameter estimators (under H_0):

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

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

- Make repeated simulations using these values for the parameters of the sampling distribution. For each simulation, a value of the test statistic is obtained. This provides an estimate of the distribution of the test statistic under H_0, 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

Trouble in River City...

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

- Unfortunately, this approach does very poorly in the low-statistics realm
- Consider a simple test case: Suppose our data is sampled from a flat distribution with a mean of 1 count in each of 100 bins

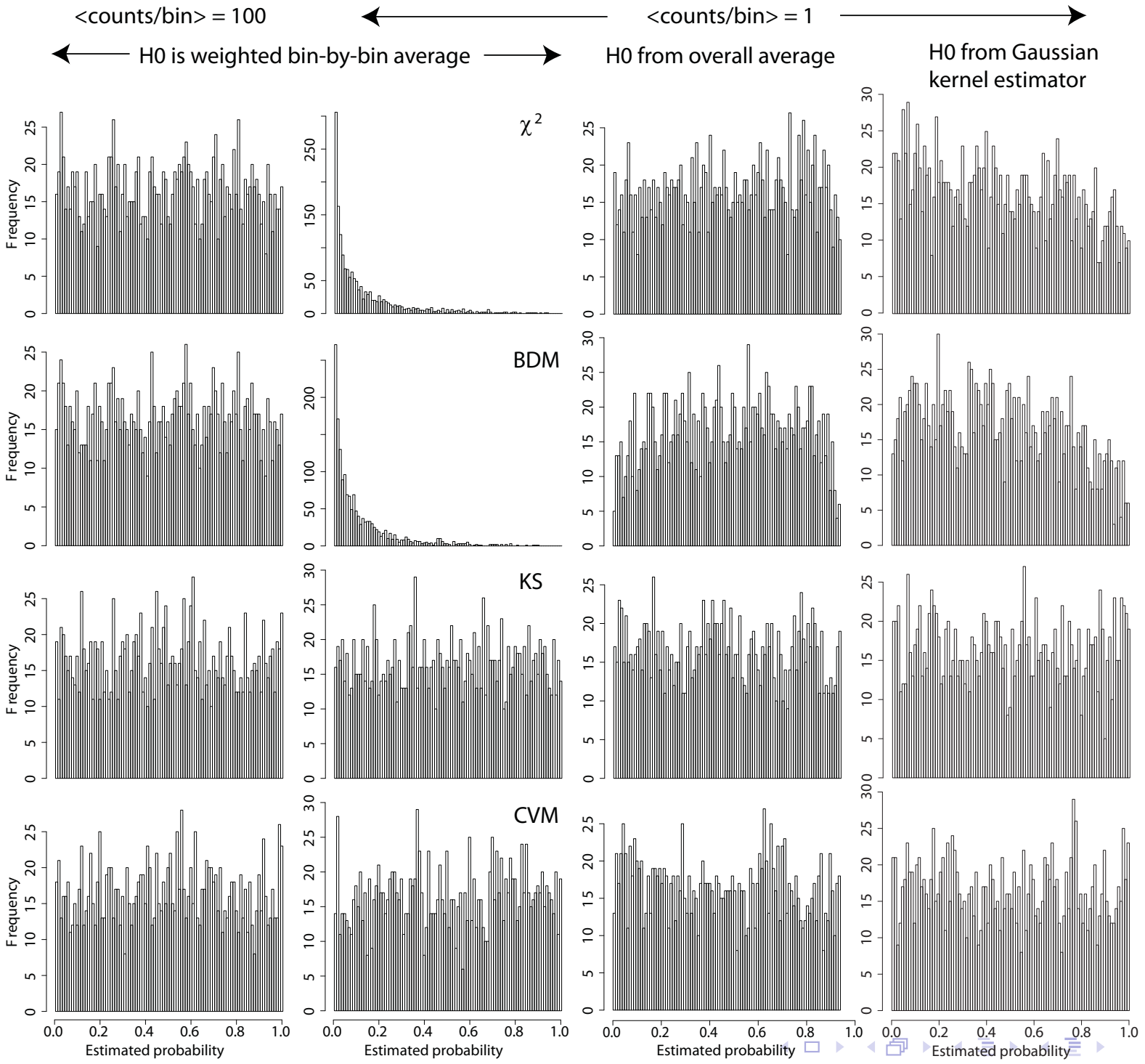
Algorithm to check estimated null hypothesis

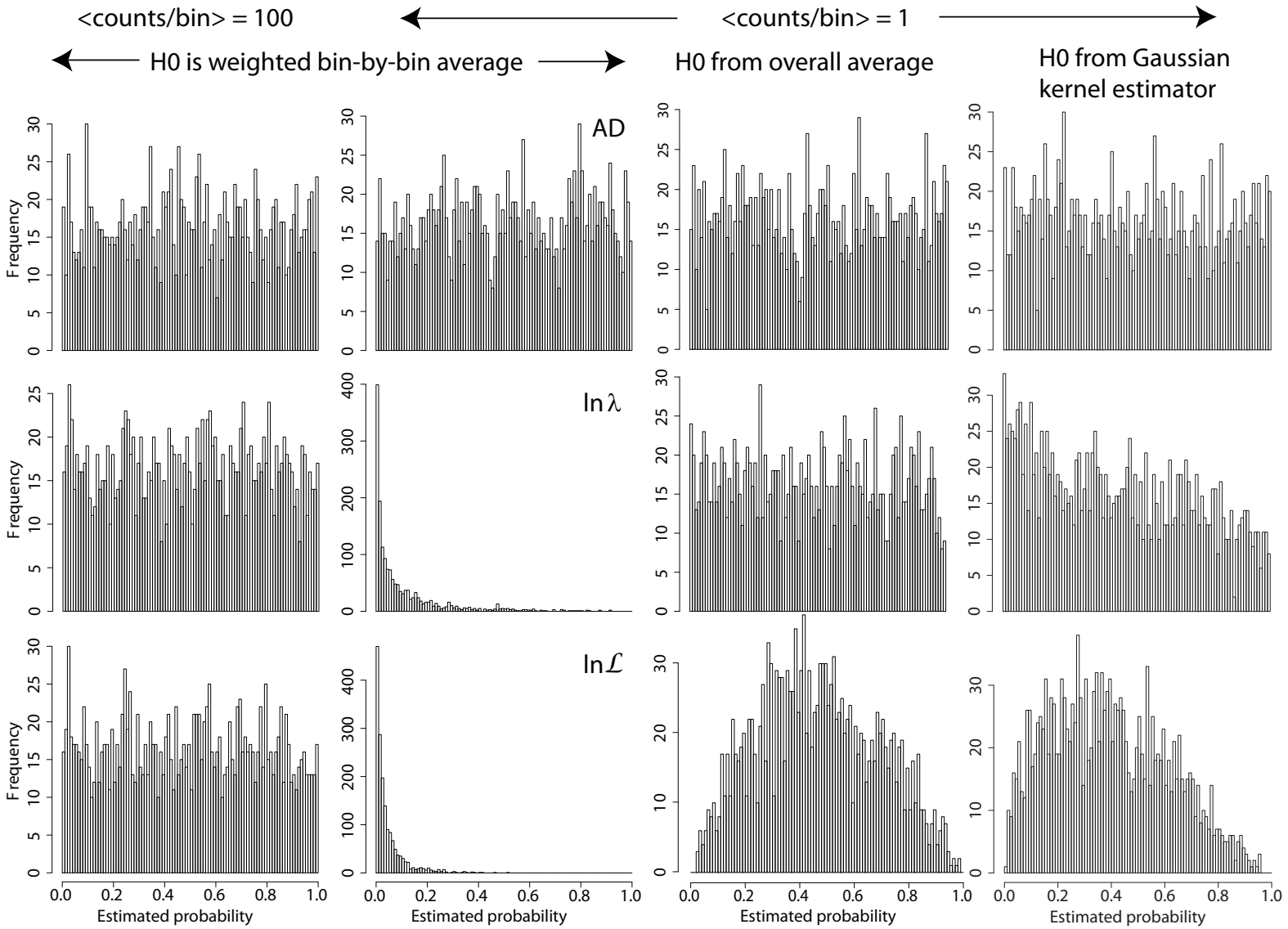
We test how well our estimated null hypothesis works for any given test statistic, T, as follows:

- Generate a pair of histograms according to the distribution above
- Compute T for this pair of histograms
- Given the pair of histograms, compute the estimated null hypothesis according to the specified prescription above
- Generate many pairs of histograms according to the estimated null hypothesis in order to obtain an estimated distribution for T.
- Using the estimated distribution for T, determine the estimated p-value for the value of T found in step 2.
- Repeat steps 2-6 many times and make a histogram of the estimated p-values. This histogram should be uniform if the estimated p-values are good estimates.

Checking the estimated null distribution

- The next two slides show tests of the estimated null distribution for each the 7 test statistics.
- Shown are distributions of the simulated p-values. The data are generated according to H_0, consisting of 100 bin histograms for:
  - A mean of 100 counts/bin (left column), or
  - A mean of 1 count/bin (other 3 columns).
- The estimates of H_0 are:
  - Weighted bin-by-bin average (left two columns),
  - Each bin mean given by the average bin contents in each histogram (third column),
  - Estimated with a Gaussian kernel estimator (right column) based on the contents of both histograms.
- The χ^2 is computed without combining bins


<counts/bin> = 100
<counts/bin> = 1
H0 is weighted bin-by-bin average
H0 from overall average
H0 from Gaussian kernel estimator
χ^2
BDM
KS
CVM
Frequency
Estimated probability


<counts/bin> = 100
<counts/bin> = 1
H0 is weighted bin-by-bin average
H0 from overall average
H0 from Gaussian kernel estimator
AD
ln λ
ln ℒ
Frequency
Estimated probability

Summary of tests of null distribution estimates

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

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

Conclusions from tests of null distribution estimates

- In the " small statistics" case, if the null hypothesis were to be rejected at the estimated 0.01 probability, the bin-by-bin algortihm for estimating a_i would actually reject H_0: 19% of the time for the χ^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%.
- In the " large statistics" case, where sampling is from histograms with a mean of 100 counts in each bin, all test statistics display the desired flat distribution.
- The χ^2, $\ln \lambda$, and $\ln \mathcal{L}$ statistics perform essentially identically at high statistics, as expected, since in the normal approximation they are equivalent.

Problem appears for low statistics

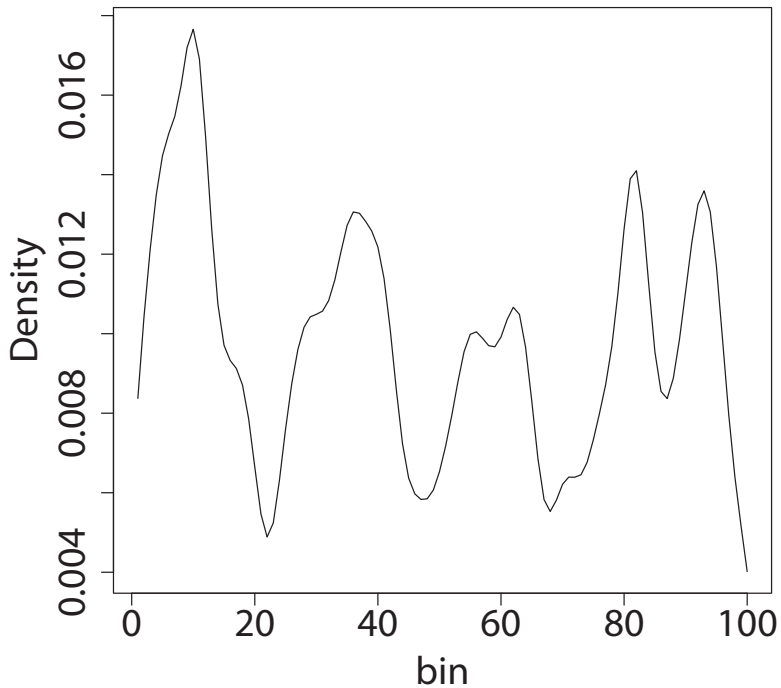
- Issue for the bin-by-bin approach appears at low statistics.
- Intuition: Consider the likely scenario at that some bins will have zero counts in both histograms. Then our algorithm for the estimated null hypothesis yields a zero mean for these bins. The simulation to determine the probability distribution for the test statistic will always have zero counts in these bins, ie, there will always be agreement between the two histograms. Thus, the simulation will find that low values of the test statistic are more probable than it should.
- The AD, CVM, and KS tests are more robust under our estimates of H_0 than the others, as they tend to emphasize the largest differences and are not so sensitive to bins that always agree. For these statistics, our bin-by-bin procedure for estimating H_0 does well even for low statistics, although we caution again that we are not examining the far tails of the distribution.

Obtaining better null distribution estimates

- A simple approach to salvaging the situation in the low statistics regime involves relying on the often valid assumption that the underlying H_0 distribution is "smooth". Then only need to estimate a few parameters to describe the smooth distribution, and effectively more statistics are available.
- Assuming a smooth background represented by a uniform distribution yields correct results. This is cheating a bit, since we perhaps aren't supposed to know that this is really what we are sampling from.
- 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. We may expect the power of these statistics to suffer under the approach taken here.

More "honest" – Try a kernel estimator

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



Sample Gaussian kernel density estimate of the null hypothesis.

Comparison of Power of Tests

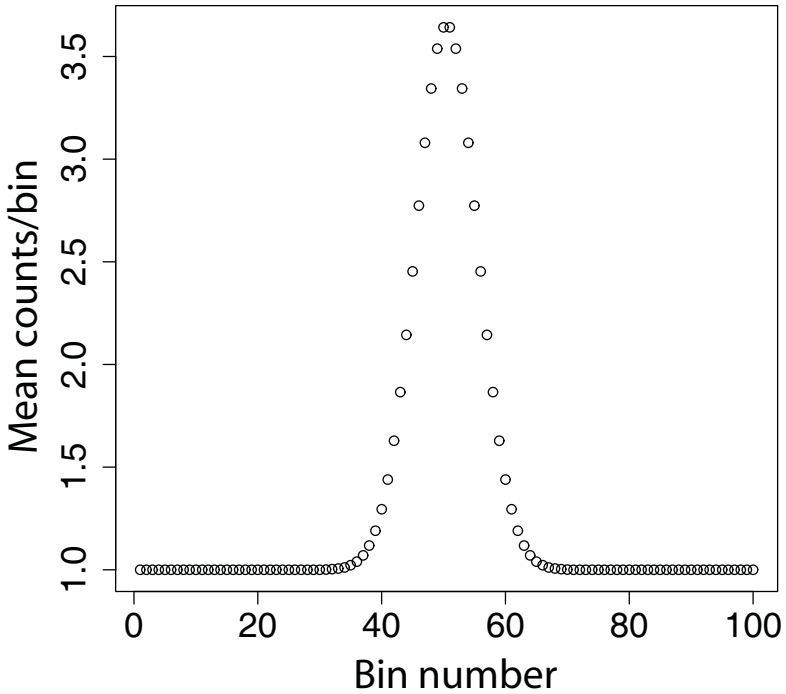
- Now that we have addressed how to estimate H_0, what test is best? Need to think about power
- The power depends on what the alternative hypothesis is.
- Investigate (eg) by adding a Gaussian component on top of a uniform background distribution in one histogram. Motivated by the scenario where one distribution appears to show peaking structure, while the other does not.
- We also look at a different extreme, a rapidly varying alternative.

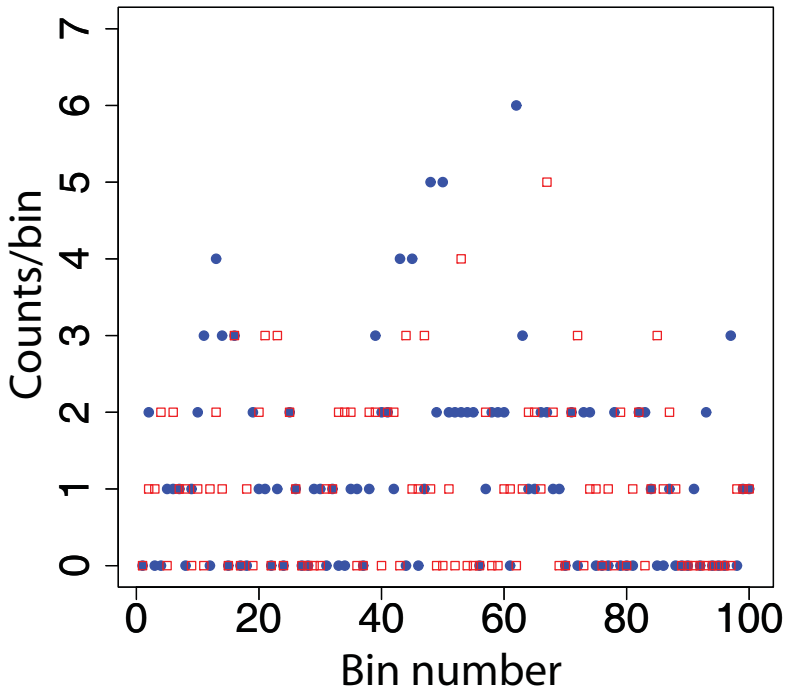
Gaussian alternaitve

The Gaussian alternative data 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, eg, a 25% Gaussian has a total amplitude corresponding to an average of 25% of the total counts in the histogram. The Gaussian counts are added to the counts from the null distribution.

Sample Gaussian alternative





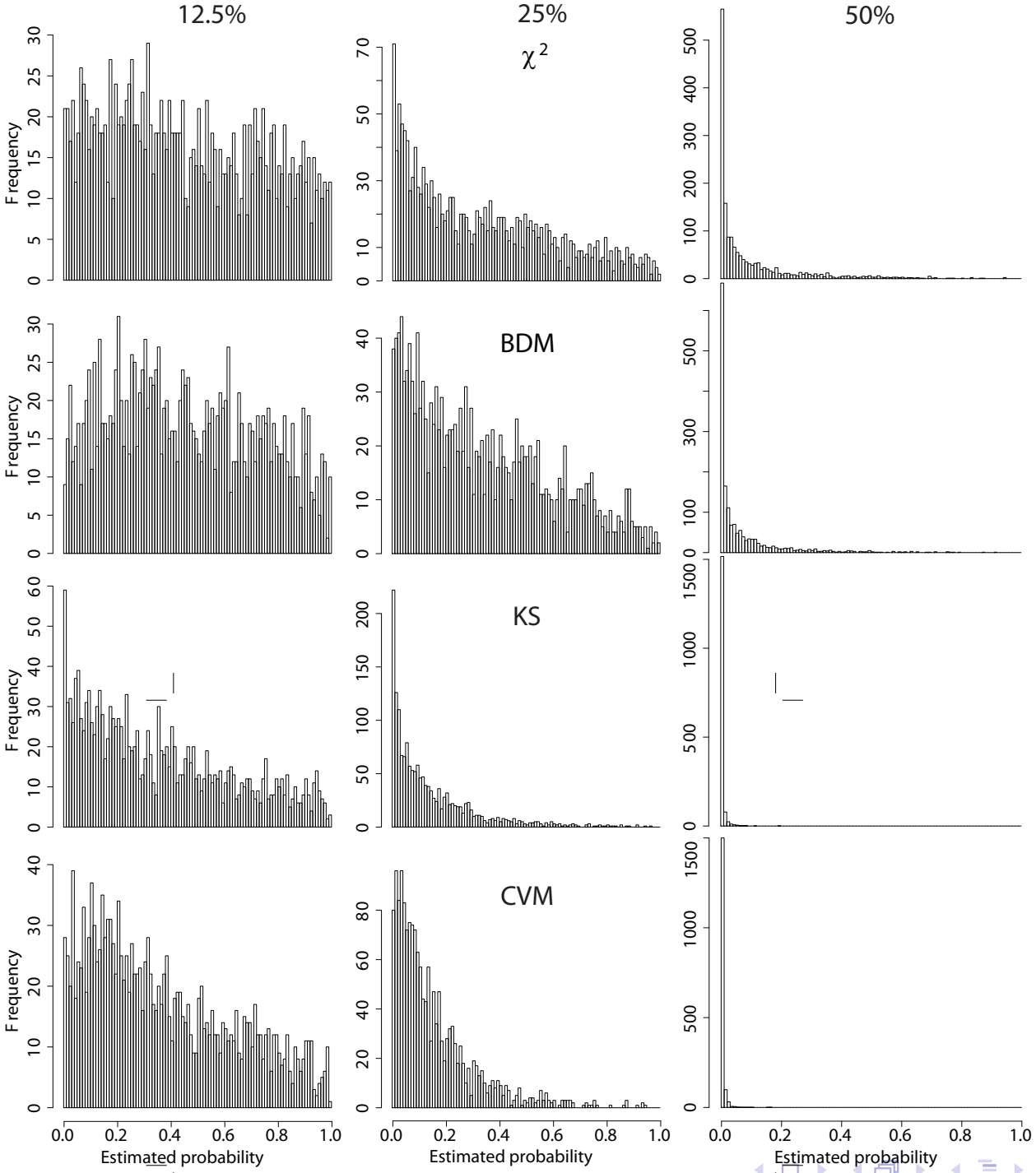
Left: The mean bin contents for a 25% Gaussian on a flat background of one count/bin (note the suppressed zero).
Right: Example sampling from the 25% Gaussian (filled blue dots) and from the uniform background (open red squares).

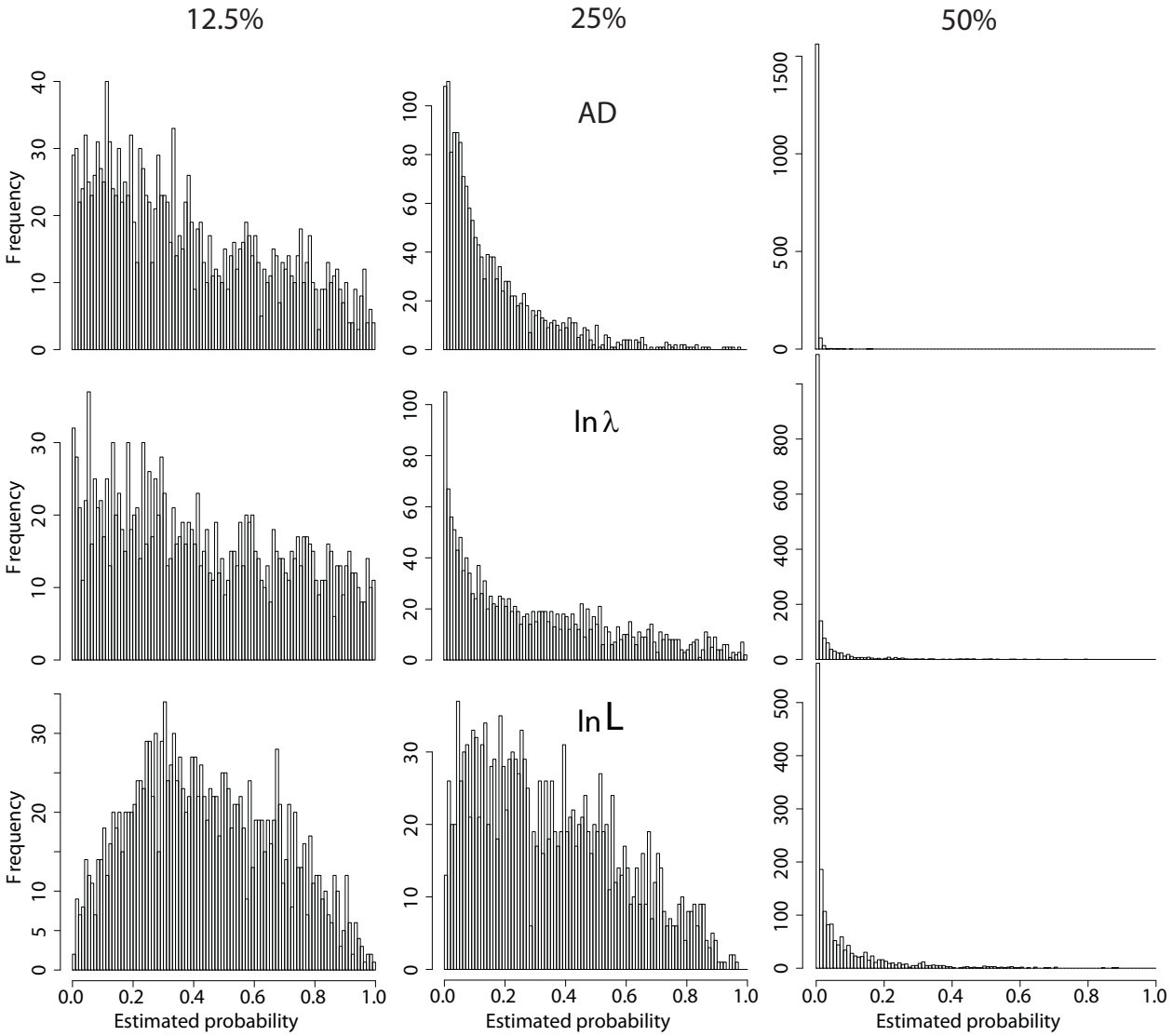
Power estimates for Gaussian alternative

On the next two slides, we show, for seven test statistics, the distribution of the estimated probability, under $H0$, that the test statistic is worse than that observed.

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

[The χ^2 is computed without combining bins.]


12.5%
25%
50%
χ^2
BDM
KS
CVM
Frequency
Estimated probability

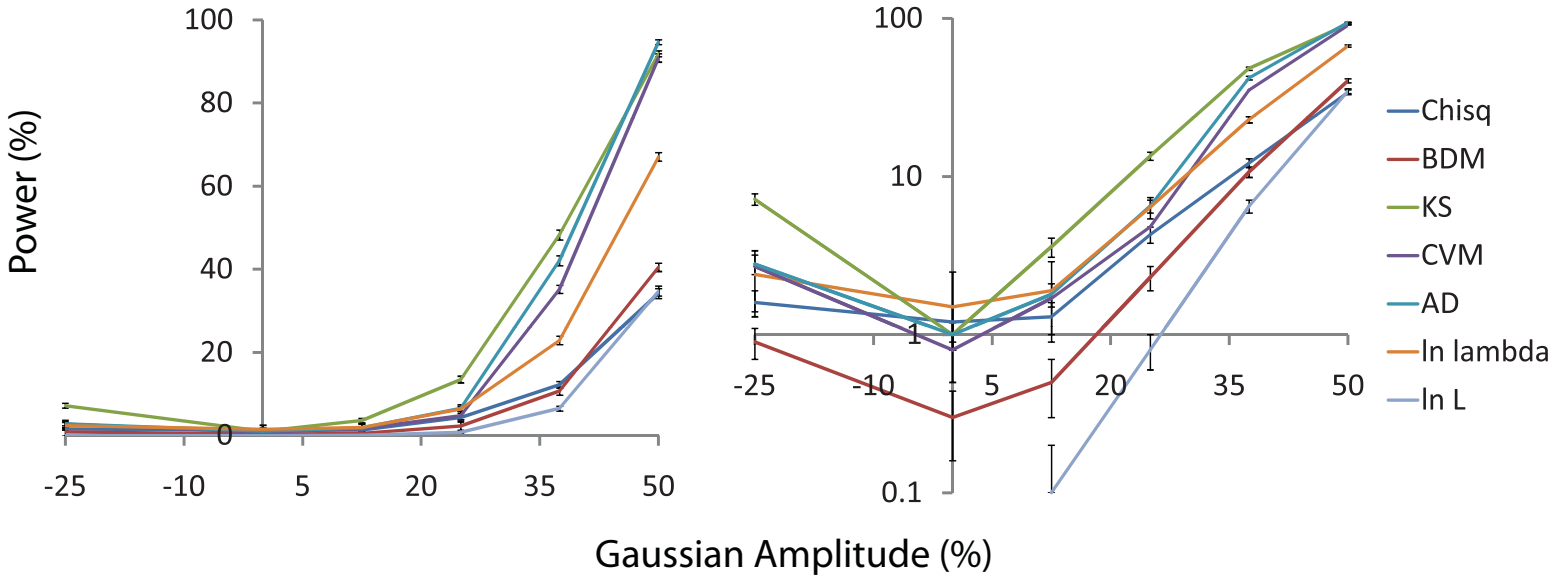

12.5%
25%
50%
AD
ln λ
ln L
Frequency
Estimated probability

Power comparison summary - Gaussian peak alternative

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. [red is most powerful]

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

Power comparison summary - Gaussian peak alternative



Summary of power of seven different test statistics, for the alternative hypothesis with a Gaussian bump.
Left: linear vertical scale; Right: logarithmic vertical scale.

Power comparison - High statistics

We also look at the performance of our seven tests for histograms with large bin contents.

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

Statistic	H0 %	5 %	-5 %
χ^2	0.91 ± 0.23	79.9 ± 1.0	92.1 ± 0.7
BDM	0.97 ± 0.24	80.1 ± 1.0	92.2 ± 0.7
KS	1.03 ± 0.25	77.3 ± 1.0	77.6 ± 1.0
CVM	0.91 ± 0.23	69.0 ± 1.1	62.4 ± 1.2
AD	0.91 ± 0.23	67.5 ± 1.2	57.8 ± 1.2
$\ln \lambda$	0.91 ± 0.23	79.9 ± 1.0	92.1 ± 0.7
$\ln \mathcal{L}$	0.97 ± 0.24	79.9 ± 1.0	91.9 ± 0.7

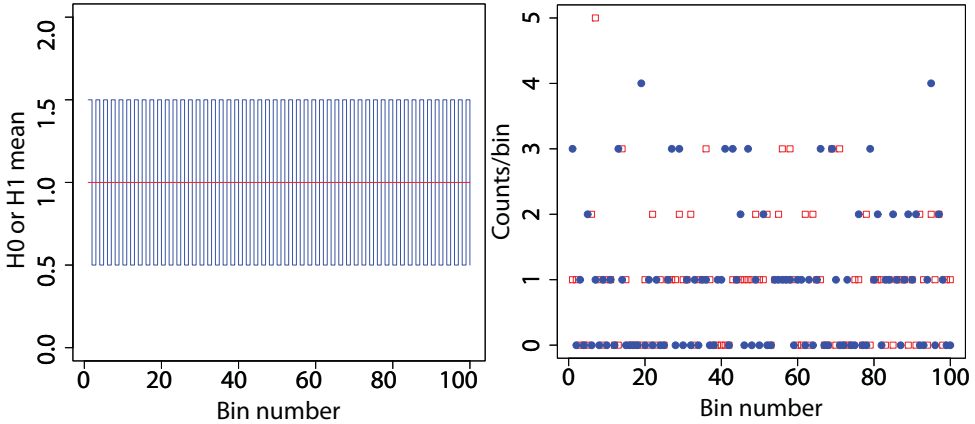
Comments on power comparison - High statistics

- In this large-statistics case, for the χ^2 and similar tests, the power to reject a dip is greater than the power to reject a bump of the same area.
- Presumably because the "error estimates" for the χ^2 are based on the square root of the observed counts, and hence give smaller errors for smaller bin contents.
- Also observe that the comparative strength of the KS, CVM, and AD tests versus the χ^2, BDM, $\ln\lambda$, and $\ln\mathcal{L}$ tests in the small statistics situation is largely reversed in the large statistics case.

Power study for a "sawtooth" alternative

For a very different alternative distribution, consider sampling from the "sawtooth" distribution. Compare again to samplings from the uniform histogram.

- The "percentage" sawtooth here is the fraction of the H_0 mean. A 100% sawtooth on a 1 count/bin background oscillates between a mean of 0 counts/bin and 2 counts/bin.
- The period of the sawtooth is two bins.



Left: The mean bin contents for a 50% sawtooth on a background of 1 count/bin (blue), and for the flat background (red))
Right: A sampling from the 50% sawtooth (blue) and from the uniform background (red))

Sawtooth alternative power results

Estimates of power for seven different test statistics, for a "sawtooth" alternative distribution.

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

Now the χ^2 and $\ln \lambda$ tests are the clear winners, with BDM next. The KS, CVM, and AD tests reject the null hypothesis with the same probability as for sampling from a true null distribution. This very poor performance for these tests is readily understood, as these tests are all based on the cumulative distributions, which average out local oscillations.

Conclusions from hypothesis test case study

These studies provide some lessons in hypothesis testing:

- No single best test for all applications. The statement "test X is better than test Y" is empty without more context. E.g., the Anderson-Darling test is often very powerful in testing normality of data against alternatives with non-normal tails (eg, a Cauchy distribution). It is not always especially powerful in other situations. The more we know about what we wish to test, the better we can choose a powerful test. Each of the tests here may be useful, depending on the circumstance.
- Even the controversial $\mathcal{L}$ test works as well as the others sometimes. However, the situations where it is observed to perform as well are here limited to those where it is equivalent to another test.

Conclusions (continued)

- Computing probabilities via simulations is a useful technique. However, must be done with care. Tests with an incompletely specified null hypothesis require care. Generating a distribution according to an assumed null distribution can lead to badly wrong results. It is important to verify the validity of the procedure. We have only looked into tails at the 1% level. Validity must be checked to whatever level of probability is needed. Should not assume that results at the 1% level will still be true at, say, the 0.1% level.
- Concentrated on the question of comparing two histograms. However, considerations apply more generally, to testing whether two datasets are consistent with being drawn from the same distribution, and to testing whether a dataset is consistent with a predicted distribution. The KS, CVM, AD, $\ln \mathcal{L}$, and $\mathcal{L}$ tests may all be constructed for these other situations (as well as the χ^2 and BDM, if we bin the data).