

Statistical Hypothesis Testing for Assessing Monte Carlo Estimators: Applications to Image Synthesis

Kartic Subr
kartic@ics.uci.edu
University of California, Irvine

James Arvo
arvo@uci.edu
University of California, Irvine

Abstract

Image synthesis algorithms are commonly compared on the basis of running times and/or perceived quality of the generated images. In the case of Monte Carlo techniques, assessment often entails a qualitative impression of convergence toward a reference standard and severity of visible noise; these amount to subjective assessments of the mean and variance of the estimators, respectively. In this paper we argue that such assessments should be augmented by well-known statistical hypothesis testing methods. In particular, we show how to perform a number of such tests to assess random variables that commonly arise in image synthesis such as those estimating irradiance, radiance, pixel color, etc. We explore five broad categories of tests: 1) determining whether the mean is equal to a reference standard, such as an analytical value, 2) determining that the variance is bounded by a given constant, 3) comparing the means of two different random variables, 4) comparing the variances of two different random variables, and 5) verifying that two random variables stem from the same parent distribution. The level of significance of these tests can be controlled by a parameter. We demonstrate that these tests can be used for objective evaluation of Monte Carlo estimators to support claims of zero or small bias and to provide quantitative assessments of variance reduction techniques. We also show how these tests can be used to detect errors in sampling or in computing the density of an importance function in MC integrations.

1 Introduction

Novel rendering algorithms are often proposed in order to compute a given image faster or to allow effective trade-offs between speed and accuracy. In either case, the question naturally arises as to how one can demonstrate that a proposed algorithm meets the stated criteria. Presently it is widespread practice within the rendering community to employ a combination of objective and subjective criteria;

running time is an objective criterion that is easy to measure and compare, while image quality, which presents a much greater challenge, generally rests upon subjective criteria such as visual inspection of two images or variance-plots.

In the context of Monte Carlo image synthesis one is often faced with the task of supporting an assertion that a given algorithm is superior in that it can produce images with the same first-order statistics (generally the expected value at each pixel), while exhibiting different second-order statistics (generally a reduction in variance). For example, algorithms for importance sampling or stratified sampling, when properly implemented, will exhibit precisely these characteristics; that is, reducing variance while leaving the mean intact. On the other hand, biased estimators are sometimes specifically constructed, primarily to reduce variance in the estimate or to simplify the algorithm. Such results are commonly demonstrated with comparison images showing a reduction in the “graininess” of the image and/or a reduction in running time by virtue of the proposed algorithm. Plots of the first- and second- order statistics of the estimators are used to help in the assessment.

There are numerous disadvantages to relying on subjective assessments such as visual comparison of images or plots: 1) they are only weakly quantitative, since comparisons are usually binary 2) the absolute variance is not a useful indicator of the quality of the estimator unless some assertions can be made about the mean 3) subtle errors can go undetected, and 4) the comparison cannot be automated.

While completely automatic ranking of estimators is an enormous challenge, in this paper we present initial steps in that direction. We propose the use of well-known statistical methods to make objective comparisons among Monte Carlo estimators, and in some cases quantitatively. Specifically, we employ *hypothesis tests* to provide objective answers to several very basic queries about random variables (r.v.’s). If X and Y are r.v.’s, we answer queries such as “Is the mean value of X equal to μ_0 ?” or “Is the mean value of X equal to the mean value of Y ?” or “Is the variance of X less than that of Y ?”. The structure of such queries is to first

pose a *null hypothesis*, such as $\langle X \rangle = \langle Y \rangle$ and competing *alternative hypotheses* such as $\langle X \rangle \neq \langle Y \rangle$, $\langle X \rangle < \langle Y \rangle$ and $\langle X \rangle > \langle Y \rangle$. Then, solely based on samples drawn from the parent distributions of X and Y the null hypothesis is either *accepted* or *rejected* with a given level of confidence. The null hypothesis is only accepted if the data do not provide enough evidence to reject it. If the null hypothesis is rejected, further tests are made to decide which alternative hypothesis may be accepted. See, for example, Freund and Walpole [9] for a concise introduction to hypothesis testing.

Previous work in computer graphics has drawn upon similar tools, such as the Chi-Square and Student-t distributions, although previous applications have focused on the problem of estimating true variance using sample variance for the purpose of stochastic anti-aliasing [6, 16, 17]. Our approach differs in that we set up a variety of significance tests for assessing both the mean and variance of the r.v.'s themselves for the purpose of verifying that they are indeed estimating what they are intended to estimate; that is, we do not directly assess the accuracy of an approximation, but the correctness and efficiency of the estimator.

2 Review: Hypothesis Tests

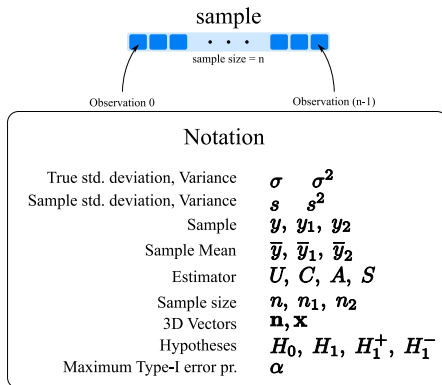


Figure 1. Notation

There are numerous types of statistical tests, associated with different forms of application problems, such as significance tests that determine whether a hypothesis ought to be rejected, parametric tests to verify hypotheses concerning parameter values, goodness of fit tests to determine whether an observed distribution is compatible with a theoretical one, etc. Statistically *significant* results are those that are unlikely to have occurred by chance. *Significance Tests* are procedures for establishing the probability of an outcome, on a null hypothesis of no effect or relationship. In contrast to the Bayesian approach to inductive inference which is based on the inverse probability $Pr(H|x)$ of a hy-

pothesis H given the data x , Fisher urged the adoption of direct probability $Pr(x|H)$ in an attempt to argue “from observations to hypotheses” [7]. If the data deviated from what was expected by more more than a specified criterion, the level of significance, the data was used to reject the null hypothesis. However, Fisher’s significance tests are difficult to frame in general since often there exist no natural or well-defined complements to null hypotheses eg. H_0 : The sample was drawn from the unit normal distribution.

The terminology *Hypothesis Testing* was made popular by Neyman and Pearson [10, 11] who formulated two competing hypotheses called the null hypothesis (H_0) and the *alternative hypothesis* (H_1). Given a sample ¹ from an arbitrary population, the goal of hypothesis testing is to test H_0 against H_1 according to the given data. Hypothesis tests are carried out with the aid of a *test statistic* which is a prescription according to which a number is computed from a given sample; that is, a real-valued function of the sample. Sometimes the test statistic could be a function of two samples, and in such cases the test is called a *two sample test*. Given a sample, its associated value of the test statistic is used to decide between accepting the null and the alternative hypotheses. Thus there exist probabilities associated with false rejection (Type I) and false acceptance (Type II) errors which are typically denoted by α and β respectively. Although the Neyman-Pearson theory was criticised [8] for only being suited to situations in which repeated random sampling has meaning, it fits well in the context of assessing MC estimators used in image synthesis. While Fisher’s view of inductive inference focused on the rejection of the null hypothesis, the Neyman-Pearson theory sought to establish rules for making decisions between two hypotheses. This fundamental difference is exploited in all the tests that are discussed in this paper.

The general algorithm for testing hypotheses proceeds in a number of steps. The first step involves formalization of the null hypothesis. After stating the hypothesis in a way that allows the probabilities of samples to be calculated assuming that the hypothesis is true, the next step is to set up a statistical test that will aid in likely reject the null hypothesis in favour of the alternative hypothesis. An acceptable α along with the test statistic defines a region of the parent distribution where H_0 is rejected in favor of H_1 ; this region is called the *critical region*. α defines the maximum probability of the test statistic falling in the critical region despite the null hypothesis being true and corresponds to the fraction of the time that the null hypothesis is erroneously rejected. If the critical region is chosen to lie either completely at the left tail of the parent distribution or completely at the right tail, the test is called a *one-tailed test* or asymmetrical or one-

¹Here we shall use the term *sample* as it is used in statistics; that is, to refer to a set of observations of a population, not a single observation, as it is commonly used in the graphics literature.

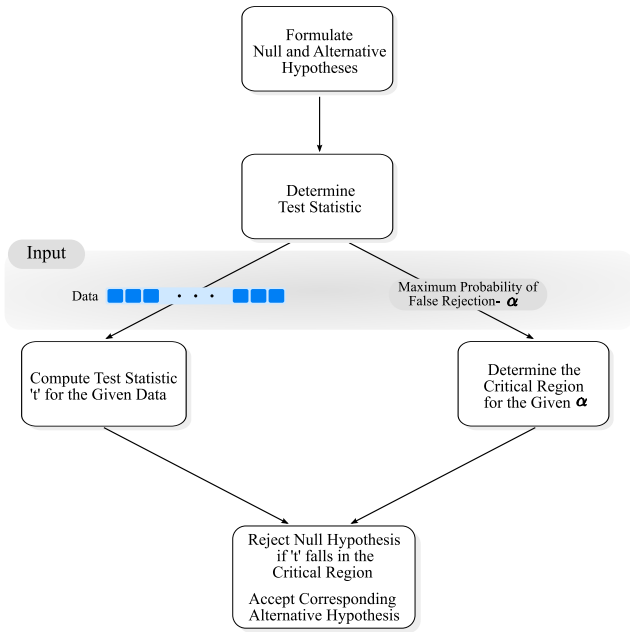


Figure 2. General Algorithm for Hypothesis Testing

sided test. If the critical region is chosen to equally cover the left and right tails, the test is called a *two-tailed test* or symmetrical or two-sided test. α is an input parameter and is typically chosen to be low.

With the hypothesis and test statistic set up and having identified the critical region, the data is examined for evidence to reject the null hypothesis. The test statistic is calculated for the given sample data and tested to check if it lies in the critical region. If this is the case, then the conclusion is that either the null hypothesis is incorrect or an erroneous result of probability less than α has occurred and in either case we accept the alternate hypothesis. Parametric hypothesis tests that hypothesize about parameters of the parent distribution, such as mean and variance, are intimately tied to the distribution of the population under study and most of the existing techniques only apply to distributions of a restricted type. In fact, the vast majority of the existing theory has been developed for populations with normal distributions.

One-tailed Tests : Tests in which the critical region lies at either the left or right of the distribution $p(x)$ followed by the test statistic. Given the max probability of false rejection α , the two critical values are obtained as $P^{-1}(\alpha)$ and $P^{-1}(1 - \alpha)$ which are the the inverse cumulative distribution evaluated at α and $1 - \alpha$ respectively. The null hypothesis is rejected if the test statistic that is computed from the data lies below or above the critical values respectively. The appropriate alternate hypothesis may be accepted.

Two-tailed Tests : Tests in which the critical region

is equally distributed at both ends of the distribution $p(x)$ followed by the test statistic. Given the max probability of false rejection α , two critical values are obtained as $P^{-1}(\alpha/2)$ and $P^{-1}(1 - \alpha/2)$. The null hypothesis is rejected if the test statistic that is computed from the data does not lie between these two critical values.

Two important hurdles in trying to apply statistical tests to populations defined as the outputs of MC estimators are :

- dealing with estimators whose estimates are not distributed normally
- formulating the null hypothesis and setting up the statistical tests

By the central limit theorem, the distribution of the estimated means of samples of MC estimator E rapidly approaches a normal distribution as the size of each sample is increased. To overcome the first of the two hurdles, rather than assess the primary estimator, we simply use distributions obtained from secondary estimators E_s (see Figure 5) in our assessment.

To overcome the latter hurdle, we first need to define the goal of the test. In the context of MC estimators two parameters are of interest– mean and variance. Our goal is to hypothesize about each of these parameters in two distinct settings: comparing an estimator with analytically obtained results and comparing two estimators (one- and two- sample tests). We address each of the four different combinations of problems describing the null hypotheses and describe the corresponding well-known statistical tests. In addition, we describe a non-parametric two-sample goodness of fit (GoF) test which tests that two samples stem from the same parent distribution. The rest of this section is simply a review of

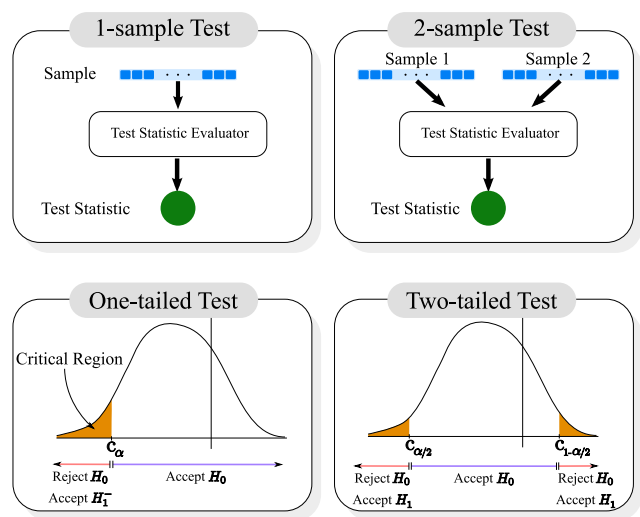


Figure 3. Different tests

the above tests, while the applications of these tests in the context of MC estimators in image synthesis are presented in Section 3.

2.1 One Sample Mean Test

The goal of this test is to assert with some confidence that the mean of the distribution from which a sample y of size n is drawn, is a specific value μ_0 . The test assumes that the distribution from which the sample is drawn is normal but does not make any assumption about its true variance. The null and alternative hypotheses for this test are

$$\begin{aligned} H_0 &: \bar{y} = \mu_0, \\ H_1 &: \bar{y} \neq \mu_0, \\ H_1^+ &: \bar{y} > \mu_0, \\ H_1^- &: \bar{y} < \mu_0. \end{aligned}$$

The test statistic is

$$t_\nu = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \quad (1)$$

which follows the Student's t-distribution with $\nu = n - 1$ degrees of freedom. The null hypothesis is tested against the first alternative hypothesis with a two-tailed test and against the other two alternative hypotheses with the appropriate one-tailed tests. If the data do not provide enough evidence, at the given α probability of false rejection, to reject the null hypothesis in favour of any of the alternate hypotheses then we accept that the mean of the sample is not significantly different from μ_0 .

2.2 One Sample Variance Test

This test allows the variance of the distribution from which a sample y of size n is drawn, to be compared with some confidence against a specific value σ_0^2 . The test assumes that the distribution from which the sample is drawn is normal but does not make any assumption about its true mean. The null and alternative hypotheses for this test are

$$\begin{aligned} H_0 &: \sigma^2 = \sigma_0^2, \\ H_1^+ &: \sigma^2 > \sigma_0^2, \\ H_1^- &: \sigma^2 < \sigma_0^2. \end{aligned}$$

The distribution of observed variances s^2 for samples drawn from some numerical population follows the chi-square distribution, which we use as the test statistic in this case. The test statistic is

$$\chi_\nu^2 = \frac{\nu s^2}{\sigma_0^2} \quad (2)$$

where again the degrees of freedom $\nu = n - 1$. An interesting property of this distribution is that the s^2 values average σ^2 , the actual (usually unknown) variance of the distribution. Two one-tailed tests are performed to test if the data provides enough evidence to reject the null hypothesis in favour of either of the alternative hypotheses.

2.3 Comparing Means of Two Samples

This test compares the means of two distributions, each of which is represented by one sample, to check for equality without making any assumptions about the variances of the distributions. If the two samples are y_1 and y_2 of sizes n_1 and n_2 respectively, the null and alternative hypotheses are

$$\begin{aligned} H_0 &: \bar{y}_1 = \bar{y}_2, \\ H_1 &: \bar{y}_1 \neq \bar{y}_2. \end{aligned}$$

The test statistic is

$$T_\nu = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \quad (3)$$

which follows the Student's t-distribution with

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

degrees of freedom. A two-tailed test is used to determine whether the samples provide enough evidence to reject the null hypothesis in favour of the alternative hypothesis.

2.4 Comparing Variances of Two Samples

To compare the variances of two distributions, each of which is represented by one sample, we use the standard F-test. If the two samples are y_1 and y_2 of sizes n_1 and n_2 respectively, the null and alternative hypotheses are

$$\begin{aligned} H_0 &: s_1^2 = s_2^2, \\ H_1^+ &: s_1^2 > s_2^2, \\ H_1^- &: s_1^2 < s_2^2. \end{aligned}$$

The test statistic is

$$F_{\nu_1, \nu_2} = \frac{s_1^2}{s_2^2} \quad (4)$$

which follows the F-distribution with ($\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$) degrees of freedom. The null hypothesis is tested against the alternative hypotheses using two one-tailed tests.

2.5 2-Sample Goodness of Fit

Given two samples y_1 and y_2 of sizes n_1 and n_2 , we would like to test if they were drawn from the same parent distribution. We use the 2-sample Kolmogorov-Smirnov (K-S) test for this purpose. This is the only non-parametric test that we use in this paper and doesn't make any assumptions about the distributions so long as they are continuous. The null and alternative hypotheses are $H_0 : \{y_1 \text{ and } y_2 \text{ come from the same distribution}\}$ and $H_1 : \{y_1 \text{ and } y_2 \text{ come from different distributions}\}$. The test statistic for the 2-sample K-S test is

$$D_2 = \max\{n W(F_{n1}(x), G_{n2}(x)) |F_{n1}(x) - G_{n2}(x)|\}$$

where $n = (n_1 n_2 / (n_1 + n_2))^{1/2}$ and $W(u, v)$ is a two-sample weighting function. $F_{n1}(x)$ and $G_{n2}(x)$ are the cumulative distributions computed from the samples y_1 and y_2 ,

$$F_{n1}(x) = \sum_{x_i < x, x_i \in y_1} x_i,$$

$$G_{n2}(x) = \sum_{x_i < x, x_i \in y_2} x_i.$$

The inclusion of the weighting function allows for a family of K-S tests, of which we choose the one described by Canner [4] where

$$W(u, v) = [z(1 - z)]^{-1/2}$$

$$z = (n_1 u + n_2 v) / (n_1 + n_2).$$

We use the critical values provided by Canner in his paper and compare the test statistic computed from the data with the appropriate critical value to decide whether the null hypothesis is to be rejected.

An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test (the chi-square GoF test depends on an adequate sample size for the approximations to be valid). The K-S test has received criticism for possessing some important limitations:

1. It only applies to continuous distributions.
2. It tends to be more sensitive near the center of the distribution than at the tails.
3. Perhaps the most serious limitation is that the distribution must be fully specified. That is, if location, scale, and shape parameters are estimated from the data, the critical region of the K-S test is no longer valid. It typically must be determined by simulation.

We will use this test in a context where none of these limitations prove to be very important, making this an effective tool for testing GoF in our application.

3 Testing Image Synthesis Estimators

Testing for bias: The bias of an estimator is defined as the difference between the estimator's expectation and the actual value of the estimand, which is the quantity being estimated. Given a new estimator E , often we would like to test that E is unbiased at a certain level of significance. If we can either compute the estimand μ_0 of E from an analytic expression or from well converged simulation, then we can draw a sample of estimates y using E and apply the one-sample test to compare \bar{y} with μ_0 .

Testing variance of an estimator: For a newly proposed estimator E , we may verify that its variance is less than an allowable variance limit σ_0^2 by drawing a sample of estimates y using E and applying the one-sample test to compare its variance with σ_0^2 .

Comparing Means of two estimators: There are at least two scenarios when we would like to compare the mean of an estimator E with that of an estimator that has already proven unbiased. First, if there exists no analytic expression for the estimand of E or if obtaining well-converged estimates using existing techniques is impractical, we could not use a one-sample test to test if E is biased. Second, this test could be used to detect erroneous implementation like non-uniform sampling, missing cosine factors, etc. The test is performed by drawing samples of estimates from each estimator and performing the two-sample test for comparing means. Rejection of the null hypothesis indicates that the means are not equal.

Comparing Variances of two estimators: If one has access to an unbiased estimator, such as a brute-force Monte Carlo estimator that is trivial to verify, one can automate the testing of new lower-variance estimators to verify that they are unbiased or nearly unbiased. While checks of this nature can often be performed "by eye," either through visual comparison of images, or by comparing numbers, the latter techniques are subjective and not conducive to either automation or quantitative testing. To compare the variances of two estimators, we draw samples from each and perform the two-sample test for comparing variances. Failure to reject the null hypothesis allows us to conclude that the variance of the new technique is not worse than that of the existing technique. If the new technique is easier to implement, or executes faster than existing techniques, asserting that its variance is not demonstrably worse can be useful. On the other hand, if the null hypothesis is rejected and $H_1^+ : s_1^2 < s_2^2$ is accepted for some α , we are justified in asserting that the new technique has lower variance.

The rest of this section presents multiple scenarios where the properties of popular MC estimators used by the rendering community are assessed. The applications presented include testing an estimator for bias, comparison of the

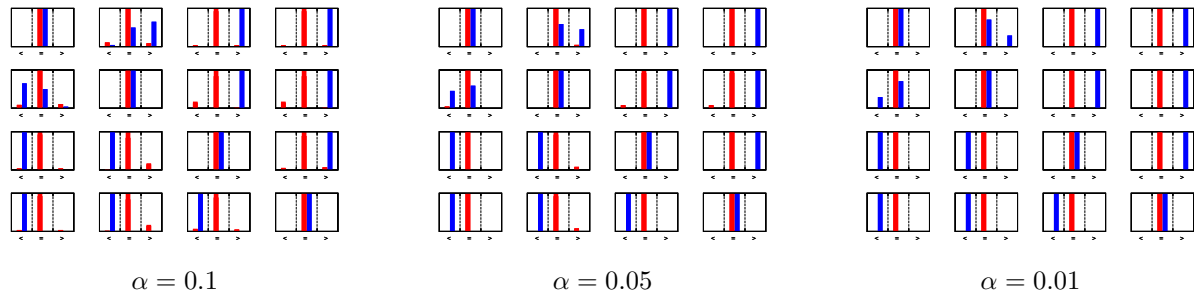


Figure 4. All combinations of four estimators were tested (see Section 3.1) to compare their means and variances. Rows and columns in each matrix of plots correspond to estimators U , C , A and S respectively. Frequencies of the results “less than”, “equal to” and “greater than” for 2-sample mean (red) and variance (blue) tests are shown in each cell of the matrix from a sequence of 100 runs of each. The results clearly confirm that the means of all the estimators are equal and that $\sigma_U > \sigma_C > \sigma_A > \sigma_S$. The diagonals correspond to testing an estimator against itself and, as expected, we see that the mean and variance tests report equality. When the comparisons were repeated for lower values of α , we see fewer false rejections. Note that there is no clear winner in the test for variance between U and C but on average $\sigma_U > \sigma_C$.

means of two estimators, comparison of the variances in estimates due to different sampling schemes, verification of reflectance function sampling schemes and detecting common errors in implementations. We set up the testing scenario, in each of the subsections below, in a way that allows us to demonstrate the benefits of objective assessment using hypothesis testing.

3.1 Irradiance Tests

Consider the irradiance at a point \mathbf{x} with normal \mathbf{n} due to a triangular uniform, lambertian emitter in the absence of occluders. The existence of an analytical solution, commonly known as *Lambert’s formula* [1], combined with the availability of several MC solutions for comparison make this problem a good candidate for case study. The irradiance at point \mathbf{x} is given by

$$E(\mathbf{x}) = \int_{\mathcal{H}^2} L(\mathbf{x}, \omega) (\mathbf{n} \cdot \omega) d\omega, \quad (5)$$

where $L(\mathbf{x}, \omega)$ is the incident radiance at \mathbf{x} along ω and \mathcal{H}^2 is the hemisphere of directions defined by \mathbf{n} . $E(\mathbf{x})$ is estimated using the following methods:

1. Estimator U : uniformly sampling the hemisphere of directions and averaging the cosine weighted incident radiance along those directions.
2. Estimator C : sampling the projected hemisphere and averaging the incident radiance along those directions.
3. Estimator A : sampling the area of the triangle uniformly and averaging the estimates of irradiance due to each individual area element.

4. Estimator S : uniformly sampling the solid angle subtended by the triangle and averaging the estimates of irradiance along each direction.

We compare means and variances of the above estimators against each other and also compare against the analytical mean obtained using Lambert’s formula. The tests are valid in this setting because the secondary estimators for the above yield roughly normal distributions (see Figure 5). Thus, each of the tests is repeated a number of times and the average result is reported. All the above estimators are known to be unbiased and the mean tests confirm this on average. We observe that sometimes, depending on the data, the mean test fails. By reducing the value of α , we can verify that the failures approximately correspond to false rejections allowed by the factor α . The result of the variance tests confirm that on average, $\sigma_U > \sigma_C > \sigma_A > \sigma_S$ (see Figure 4).

3.2 Testing BRDF Sampling Schemes

One of the many desirable properties of a BRDF is its suitability to be used in a MC rendering setup. This usually involves being able to sample from the reflectance function or an approximation of this function. In the latter case, so long as the exact density associated with each direction in the sample is known there is no risk of introducing a bias while estimating reflected radiance using the sample, regardless of how weak the approximation. However, the closer the approximation, the lower the variance in the estimated reflected radiance.

The goal of this case study is to use two popular BRDF models proposed by Ashikhmin and Shirley [3] and Ward [26, 25] and test whether the distributions sampled by the two techniques significantly differ from their corre-

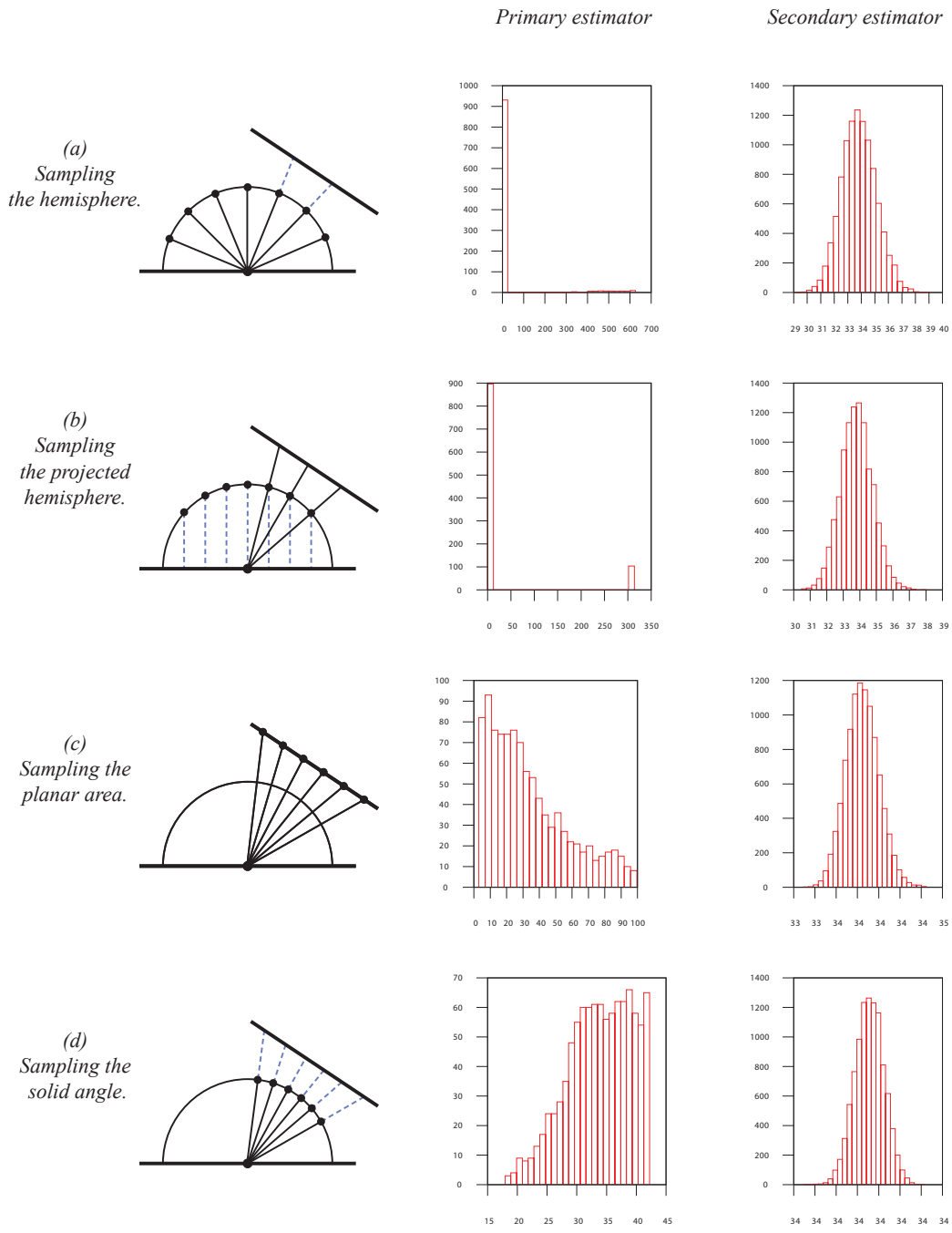


Figure 5. Comparing four different Monte Carlo estimators for computing irradiance. The histograms show frequency vs irradiance for a large number of estimates. The distribution characteristic of the secondary estimators is observed to be close to normal with the same mean but different variances depending on the sampling scheme.

sponding reflectance functions. We select an input direction arbitrarily and obtain a sample containing many output directions according to the BRDF. We bin these samples and visualize the 2D histogram as a greyscale image where image intensity is proportional to bin frequency. For comparison, we visualize the histograms obtained by sampling each BRDF using rejection sampling. The test is set up so that the size of the sample obtained using rejection is equal to the size of the sample obtained by sampling the BRDF.

Visual inspection of the histograms is sufficient to assert that the sampling of the Ward’s BRDF does not match the actual reflectance distribution. In the case of the Ashikhmin-Shirley BRDF however, it is not obvious. To assess the Ashikhmin-Shirley BRDF sampling algorithm we use the 2-sample GoF test. Since the test is applicable only to univariate distributions and we have a 2D distribution for a fixed outgoing direction, we linearize this 2D space by using a space filling curve such as Morton-order [21].

Since the GoF test is not a parametric test, we do not make any assumptions about the distribution other than that it is continuous [3]. Also since we can afford to repeat the experiment multiple times, two of the three major limitations of the K-S test are no longer major concerns in our ap-

plication. The third limitation of the K-S test suggests that it will be less likely to detect sampling anomalies near the pole or near the horizon. We show that this is not a major concern in practice. If need be this decreased sensitivity to the tails may be made insignificant by adopting a parameterization scheme for the BRDF such as the half-angle parameterization [19] in conjunction with a linearization scheme, thus keeping the interesting changes of the BRDF in the middle of the distribution. The fact that the K-S test does not make assumptions about the distribution from which the samples are drawn is key.

The results of the 2-sample K-S test for a sample directly drawn from Ward’s BRDF against one drawn using rejection failed for all levels of significance and any numbers of samples drawn. On the other hand, a similar test for the Ashikhmin-Shirley BRDF passed with $\alpha = 0.005$ for a sample size of less than 100. For larger samples the Ashikhmin-Shirley BRDF failed the test indicating that the distribution being drawn from does not match the reflectance distribution exactly. This is consistent with the sampling technique [3] which derives the scheme for a distribution that is very close to the reflectance function but not identical.

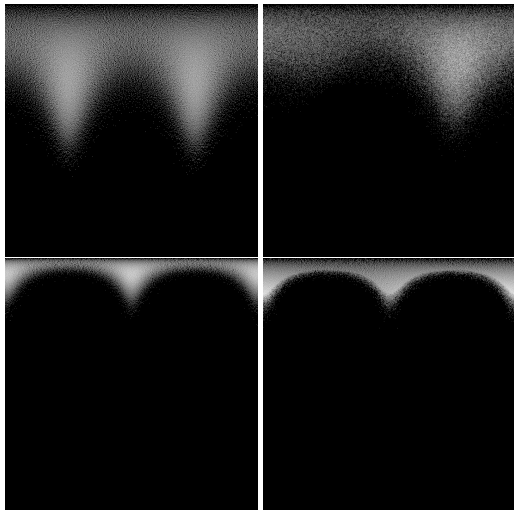


Figure 6. Histograms of sample directions for two anisotropic BRDF’s (Ward and Ashikhmin-Shirley) are shown, for a given outgoing direction. Multiple peaks are observed due to the anisotropy. Rows and columns in the image correspond to polar and azimuthal angles respectively. Sampling from the reflectance distribution (left) vs sampling using rejection (right) is shown. While it is evident that the distributions do not match for Ward’s BRDF (top row), it is not obvious from visual inspection if the two samples for the Ashikhmin-Shirley BRDF (bottom row) represent the same distribution.

3.3 Reflected Radiance

Testing the BRDF sampling using GoF tests can provide useful insight into the potential variance in the estimates of reflected radiance off a glossy surface. If it has been confirmed that the sampled distribution does not exactly match the reflectance distribution, it is of interest to know whether the correct weights are being used with each direction while estimating reflected radiance using the sample. Since we have already verified that the sampling of Ward’s BRDF does not follow the actual reflectance distribution, we perform a test to verify that the true function sampled from can be used as an importance function without introducing bias.

The reflected radiance from a surface with Ward’s BRDF was estimated m times along an outgoing direction by (1)sampling the BRDF and (2) sampling the BRDF using rejection. That is, for a given outgoing direction, we obtained m estimates of the reflected radiance using each sampling scheme from which we constructed two samples of size m . By performing a 2-sample test for the means of these two samples, we tested that they have the same mean. The process was repeated k times along each outgoing direction, for 1000 outgoing directions uniformly distributed over the hemisphere with $m = 50$. 98.6% of the tests with $\alpha = 0.01$ reported that the means were equal. Therefore, this random variable can be used for unbiased importance sampling.

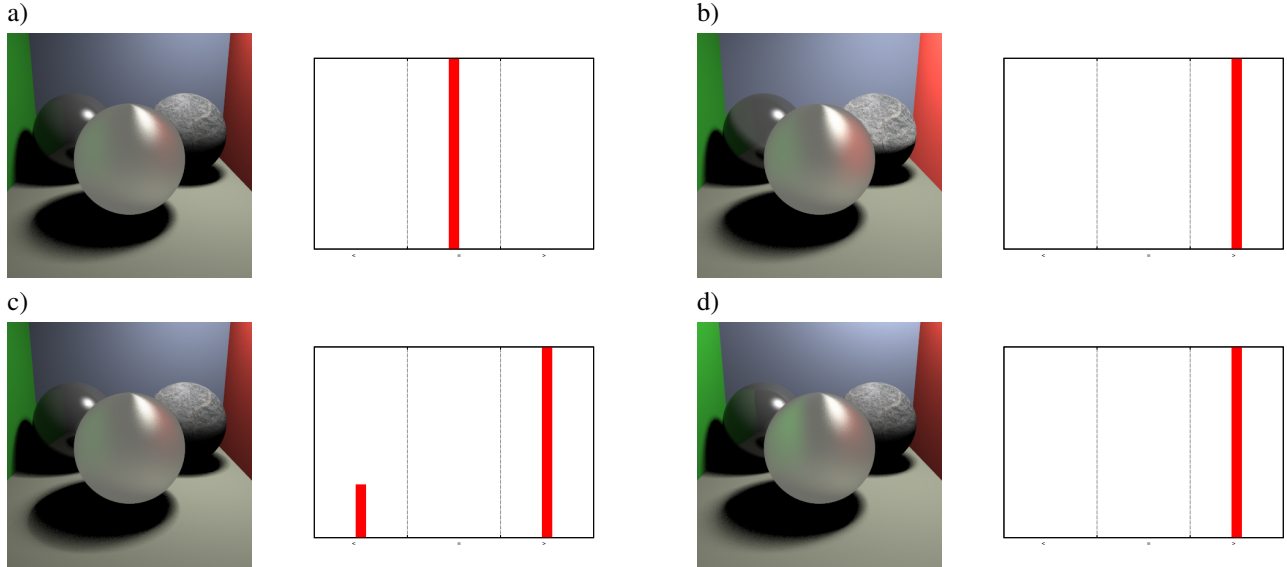


Figure 7. Results of the 2-sample tests comparing the mean of an estimator against a trusted estimator before and after three errors were introduced in the former. The tests were performed with $\alpha = 0.01$ and detected the difference in means after introduction of the erroneous when tested against the trusted estimator. Images generated using the erroneous estimators are shown for a scene with shiny, textured and glossy (Ward’s BRDF) spheres. a) Before introducing errors b) Missing cosine term; c) Non-uniform sampling of the illuminant; d) Incorrect change of variables in Equation (3.4). The errors are not always be obvious from just visual inspection.

3.4 Detecting Errors

One of the applications of the hypothesis testing approaches we have described is catching unintended sources of bias, and determining whether an experimental variance reduction technique is in fact effective.

As graphics researchers often discover, it is difficult to construct low-variance estimators that remain unbiased, either because of the intrinsic difficulty of correctly normalizing the probability density functions, or simply because there are so many opportunities for error. For example, it is easy to forget a factor of a cosine or π , or incorrectly perform a change of variables (e.g. cosine over distance squared) which will lead to erroneous results that nevertheless look plausible and may therefore go unnoticed. Indeed, many sources of bias would be nearly impossible to detect without an objective comparison with either an analytic solution, or a trusted Monte Carlo estimator. For example, if stratified sampling over a 2-manifold is used with a mapping that is not uniform (i.e. a mapping that does not map equal areas in the parameter domain to equal areas on the manifold), there will be a systematic bias unless the strata are weighted according to their respective areas. Similarly, if samples are used both to estimate the mean and to guide adaptive sampling, the result is systematically biased downward [14]. In both cases, the bias may be arbitrarily large, yet offers no obvious visual clue of its existence. Such er-

rors are relatively easy to catch with hypothesis testing.

We intentionally introduce three common unintended sources of bias in the estimator A (see Section 3.1) and demonstrate that they could be detected by using the tests described in Section 2. In constructing A , Equation (3.1) is rewritten, using a change of variables, as

$$E(\mathbf{x}) = \int_{Area(\Delta)} L(\mathbf{x}, \mathbf{z}) \frac{\mathbf{n} \cdot \mathbf{z}}{\|\mathbf{z}\|} \frac{\mathbf{n}_\Delta \cdot \mathbf{z}}{\|\mathbf{z}\|^3} d\mathbf{y}, \quad (6)$$

where the integral is now over the area of the triangle as opposed to the sphere of directions, with \mathbf{y} as the variable of integration. \mathbf{n}_Δ is the triangle’s normal and $\mathbf{z} = \mathbf{x} - \mathbf{y}$ is a vector along ω . The term $(\mathbf{n}_\Delta \cdot \mathbf{z} / \|\mathbf{z}\|^3)$ is a factor that appears in the integral due to the change of variables. Specifically, we made the following three alterations

1. *Omitting the cosine term* $(\mathbf{n} \cdot \mathbf{z} / \|\mathbf{z}\|)$ in Equation (3.4)
2. *Non-uniform sampling* of the area of the triangle by using uniform random variables in $[0, 1]$ as barycentric coordinates.
3. *Incorrect change of variables* by omitting the $(\mathbf{n}_\Delta \cdot \mathbf{z} / \|\mathbf{z}\|^3)$ in Equation (3.4).

All three errors were promptly detected by running the 2-sample test for means when tested against the unmodified trusted estimator S (see Figure 7).

4 Conclusion

We have demonstrated how the well-known idea of statistical hypothesis testing can be applied to Monte Carlo image synthesis. Specifically, we have shown its utility in testing whether a given estimator has the correct expected value or a variance bounded by a given value. We have also shown how to test whether two estimators have the same expected value, and whether one estimator has a smaller variance than another. At present, such conclusions are typically drawn in an informal way, either by subjective evaluation of images, or by comparing sample means and variances, subjectively allowing for statistical variation.

We have demonstrated how to set up the correct statistical tests in each of the scenarios mentioned above, and have illustrated their use in prototypical computations such as computing irradiance at a given point on a surface and computing reflected radiance at a given point along a given direction.

The techniques that we have described here are not limited in any way to the specific scenarios we have used as illustrations. They could be used to objectively compare a sophisticated path tracing technique (eg. Metropolis mutation strategy [24]) with a brute-force strategy (eg. brute-force Monte Carlo) that is guaranteed to produce the correct result, albeit very slowly. Other applications of the proposed techniques include objective comparison of different variance reduction techniques and statistical verification of sampling algorithms.

References

- [1] J. Arvo. The irradiance Jacobian for partially occluded polyhedral sources. In *Computer Graphics Proceedings, Annual Conference Series*, ACM SIGGRAPH, pages 343–350, July 1994.
- [2] J. Arvo. Stratified sampling of spherical triangles. In *Computer Graphics Proceedings, Annual Conference Series*, ACM SIGGRAPH, pages 437–438, Aug. 1995.
- [3] M. Ashikhmin and P. Shirley. An anisotropic Phong BRDF model. *Journal of Graphics Tools*, 5(2):25–32, 2000.
- [4] P. L. Canner. A simulation study of one- and two-sample kolmogorov-smirnov statistics with a particular weight function (in theory and methods). In *Journal of the American Statistical Association*, New York, NY, USA, 1975.
- [5] R. L. Cook. Stochastic sampling in computer graphics. *ACM Transactions on Graphics*, 5(1):51–72, 1986.
- [6] M. A. Z. Dippe and E. H. Wold. Antialiasing through stochastic sampling. *Computer Graphics*, 19(3):69–78, July 1985.
- [7] R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.
- [8] R. A. Fisher. *Statistical Methods and Scientific Inference*. Oliver and Boyd, second edition, 1959.
- [9] J. E. Freund and R. E. Walpole. *Mathematical Statistics*. Prentice-Hall, Englewood Cliffs, New Jersey, fourth edition, 1987.
- [10] N. J. and P. E.S. *On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference, Parts I, II*. Biometrika, 1928.
- [11] N. J. and P. E.S. *On the Problem of the Most Efficient Tests of Statistical Hypotheses*. Philosophical Transactions of the Royal Society of London, 1933.
- [12] J. T. Kajiya. The rendering equation. *Computer Graphics*, 20(4):143–150, Aug. 1986.
- [13] M. H. Kalos and P. A. Whitlock. *Monte Carlo Methods*, volume I, *Basics*. John Wiley & Sons, New York, 1986.
- [14] D. Kirk and J. Arvo. Unbiased sampling techniques for image synthesis. *Computer Graphics*, 25(4):153–156, July 1991.
- [15] D. Kirk and J. Arvo. Unbiased variance reduction for global illumination. In *Proceedings of the Second Eurographics Workshop on Rendering*, Barcelona, May 1991.
- [16] M. E. Lee, R. A. Redner, and S. P. Useton. Statistically optimized sampling for distributed ray tracing. *Computer Graphics*, 19(3):61–68, July 1985.
- [17] W. Purgathofer. A statistical method for adaptive stochastic sampling. In A. Requicha, editor, *Proceedings of Eurographics 86*, pages 145–152. Elsevier, North-Holland, 1986.
- [18] R. Y. Rubinstein. *Simulation and the Monte Carlo Method*. John Wiley & Sons, New York, 1981.
- [19] S. Rusinkiewicz. A new change of variables for efficient BRDF representation. In G. Drettakis and N. Max, editors, *Rendering Techniques '98 (Proceedings of Eurographics Rendering Workshop '98)*, pages 11–22, New York, NY, 1998. Springer Wien.
- [20] P. Shirley, C. Wang, and K. Zimmerman. Monte Carlo methods for direct lighting calculations. *ACM Transactions on Graphics*, 15(1):1–36, Jan. 1996.
- [21] H. Tropf and H. Herzog. Multidimensional range search in dynamically balanced trees. In *Angewandte Informatik*, pages 71–77, 1981.
- [22] G. Turk. Generating random points in triangles. In A. S. Glassner, editor, *Graphics Gems*, pages 24–28. Academic Press, New York, 1990.
- [23] E. Veach. Bidirectional path tracing. In *Ph.D. Dissertation*, pages 297–330, 1997.
- [24] E. Veach and L. J. Guibas. Optimally combining sampling techniques for Monte Carlo rendering. In *Computer Graphics Proceedings, Annual Conference Series*, ACM SIGGRAPH, pages 419–428, Aug. 1995.
- [25] B. Walter. Notes on the ward brdf. In *Technical Report, PCG-05-06CG*, New York, NY, USA, 2005.
- [26] G. J. Ward. Measuring and modeling anisotropic reflection. In *SIGGRAPH '92: Proceedings of the 19th annual conference on Computer graphics and interactive techniques*, pages 265–272, New York, NY, USA, 1992. ACM Press.