Are Two Distributions Different From Each Other?

There are many situations in which you would like to determine whether two distributions are distinct. Do the properties of two types of galaxies match? Are there multiple categories of gamma-ray bursts? Is the initial mass function of stars universal, or does it have systematic differences between different types of environment?

In this lecture we will focus on a common, and useful, approach to determining whether two distributions are different: the Kolmogorov-Smirnov, or K-S, test. This is a simple test that I absolutely recommend that you put in your arsenal, but it also has certain weaknesses that we will discuss.

Before we discuss the test itself: I do, of course, hope that whenever you do such a test you apply common sense. For example, if the samples you are comparing have been obtained in two different ways (e.g., using two different instruments or even on two different nights with the same instrument) you need to be very cautious! Selection effects, or systematic differences in the observations, could easily lead you to conclude that two samples are different when in fact they are drawn from the same distribution.

The K-S test

This is a test of the distinction between two *one-dimensional* distributions. This is important to stress: sometimes people will try to extend the test to more than one dimension, but there is no unique way to perform this test in more than one dimension. The test can be performed between two tabulated distributions, or between one tabulated distribution and a theoretical distribution (which must be fully specified; for example, it is illegitimate to *fit* a Gaussian to data, and then compare the data with the fitted Gaussian). We will focus on the first category, where we compare two tabulated distributions (e.g., lists of data). It is assumed in this test that the two samples are mutually independent, and the test works strictly for *continuous* variables (i.e., ones in which what you're measuring can take on a continuous set of values, such as distances, rather than a discrete set of values, such as rolls of a die).

With those assumptions, the procedure is to take the measurements and from them form two cumulative distributions, $C_1(x)$ and $C_2(x)$. When you plot them, cumulative probability distributions run from 0 to 1 on the vertical axis, and over the range of values of x on the horizontal axis. Cumulative distribution plots are a somewhat underappreciated method of representing distributions; we are more used to differential probability plots (e.g., showing $P(x)$ as a function of x), but those require binning to be meaningful to the eye. In any case, the statistic used in the K-S test is just the maximum of the absolute magnitude of the deviation between those cumulative distributions:

$$d \equiv \max |C_1(x) - C_2(x)| \ . \tag{1}$$

The maximum is over x; of course d can't be greater than 1 or less than 0. Tables exist to calculate the probability of the distributions being the same, given d and the number of points in distribution 1 and distribution 2. Codes exist in many languages to compute the K-S probability given two tabulated distributions.

One of the convenient aspects of the K-S test is that it makes no assumption about the distributions. This is therefore a very broadly applicable test. But it is a rule that tests that are very general lack optimal power for specific situations. To see how that works, we'll do two analyses to determine whether the samples in data9_1.txt (on the website; this has 100 points drawn from a zero-centered Gaussian with a standard deviation $\sigma_1 = 0.8$) are drawn from a distribution different from the samples in data9_2.txt (also on the website; this has 150 points drawn from a zero-centered Gaussian with a standard deviation $\sigma_2 = 1.2$).

As usual, we start by plotting the data, in Figure 1. The distributions do seem pretty clearly different to our eyes. Indeed, as expected, the width of the second distribution is greater than the width of the first distribution.

When we perform the K-S test (I used the C code "kstwo.c" from Numerical Recipes; use your favorite language), we find that the maximum distance between the cumulative probability distributions is $\max |C_1(x) - C_2(x)| = 0.103$, and with 100 points in one distribution and 150 in the other, the probability that the distributions are the same is calculated in this test to be 0.521. Thus, based only on the K-S test, we would conclude that the samples have a good chance of having been drawn from the same distribution.

One thing you can note from this test, and from Figure 1, is that the K-S test isn't very sensitive to the wings of the distributions. To our eye, the clearest difference between the two distributions is that Distribution 2 extends more to negative x, and to positive x, than Distribution 1. But because the *vertical* separation is not large there, the K-S test doesn't pick it up. For that reason, some people advocate other tests, such as the Anderson-Darling test, but that test (1) only compares an empirical distribution (formed from samples) with a specified theoretical distribution, rather than being able to compare two empirical distributions with each other, and (2) has different probability percentiles for different specified theoretical distributions, unlike the K-S test (which has the same percentiles independent of the distributions). As is common, what you apply will depend on your particular task and tastes.

But how else might we assess whether the distributions are likely to be the same or different? If we are specifically interested in the case where (say) we know that both samples are zero-centered Gaussians and we'd like to know whether the standard deviations are equal, then we have an easy Bayesian test. In that test, all we need to do is (1) fit both data sets with a zero-centered Gaussian of the same standard deviation, (2) fit each data set with zero-centered Gaussians of potentially different standard deviations, and (3) do a model

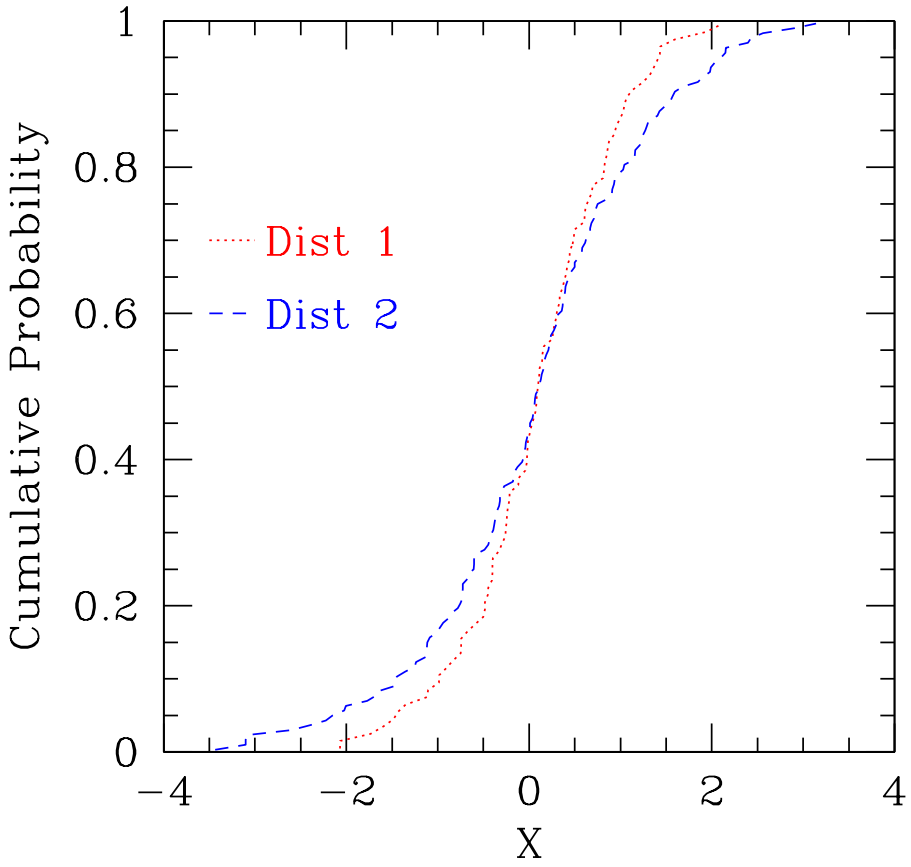

Fig. 1.— Cumulative distribution functions of the two distributions that we are comparing. Distribution 1 is drawn from a Gaussian with a mean of zero and a standard deviation of 0.8; Distribution 2 is drawn from a Gaussian with a mean of zero and a standard deviation of 1.2. To the eye, these seem different; what will the K-S test tell us?

comparison in which we take the ratio of the evidences in the two models.

If we want a quick check we can, for example, *maximize* the likelihoods of the single-stdev and double-stdev models, and then multiply the difference between the maximum log likelihoods by 2 to use Wilks' theorem with a chi squared table. Doing that, we find that $\Delta \ln \mathcal{L} = 8.17$, so the equivalent $\Delta\chi^2 = 16.34$; the more complex model does incorporate the less complex model, so we effectively have $\Delta\chi^2 = 16.34$ for 1 degree of freedom, and thus we find that the extra complexity is required at the 4σ level.

This result is expected: a *specific* test (such as: assuming that the distributions are Gaussians with means of zero, are the standard deviations different?) will have more power than a general test such as the K-S test. But in many cases you won't have a specific model in mind, and then something like a K-S test is very handy to have. It's fast and easy, so as long as you have continuous variables and your data are one-dimensional, I'm happy in the spirit of practical astrostatistics to suggest that you use this test. Just remember that it tends to go one way: if the K-S test says that your distributions are *different*, they probably are, but if the K-S test says that your distributions are likely to be the *same*, then it could just be that this relatively weak test has insufficient power to tell the difference.

Now it's your turn. On the website I have put lists of estimated masses of neutron stars in double neutron star binaries, and estimated masses of neutron stars *not* in double neutron star binaries. Major caveats apply: for example, the companion types are different, and the mass measurement techniques vary over these samples. Nonetheless, do a K-S comparison. Is there evidence that the mass distributions are different?