**Solution: chapter 3, problem 8:**

In this exercise, we wish to compare performance of the Anderson-Darling (AD) and Kolmogorov-Smirnov (KS) tests (sections 3.3.1 and 3.3.2 in Narsky and Porter (2014a), referred to hereafter as NP). The null hypothesis is sampling from a standard normal distribution:

$$H_0 : f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}. \tag{3.1}$$

This is completely specified, hence a simple hypothesis. However, the alternative hypothesis is "not standard normal" so the test is composite. We are asked to evaluate the power of the specified test algorithms in rejecting data sampled from a Cauchy distribution with location parameter 0 and full width at half maximum (FWHM) equal to that of a standard normal. The FWHM of the standard normal is $\Gamma = 2\sqrt{-2\log\frac{1}{2}} \approx 2.3548$. Thus we'll be measuring power to reject sampling from the following Cauchy:

$$H_1 : f(x) = \frac{\Gamma}{2\pi} \frac{1}{x^2 + (\Gamma/2)^2}. \tag{3.2}$$

Let us assume we have a sample of size $N$ ($= 100$ in the problem statement, but we'll look at $N = 10$ as well), $\mathbf{x} = x_1, \ldots, x_N$. Denote the empirical cdf (ecdf) by $F_N$. The Kolmogorov-Smirnov test statistic is (Eq. 3.31 in NP):

$$\rho(F, F_N) = \sup_{x \in (-\infty, \infty)} |F(x) - F_N(x)|. \tag{3.3}$$

It is noted in NP that the distribution of the KS statistic is independent of the sampling distribution for given sample size (NP exercise 3.6), with distribution in NP Eqs. (3.32) and (3.33), and shown in NP Fig. 3.4.

The Anderson-Darling test statistic is defined in Eq. 3.35 of NP:

$$A^2_N(\mathbf{x}) = N \int_{y=-\infty}^{y=\infty} \frac{[F_N(y) - F(y)]^2}{F(y)\,[1 - F(y)]} dF(y). \tag{3.4}$$

We may do the integral and obtain an expression that is readily computed. First, let $\mathbf{z}$ be the ordered set of sampled values $\mathbf{x}$, i.e., $z_1 < z_2 < \cdots < z_N$. Notice that

$$F_N(x) = \frac{n}{N}, \quad \text{for } z_n \le x < z_{n+1} \tag{3.5}$$

where $z_0 \equiv -\infty$ and $z_{N+1} \equiv \infty$. Define

$$G_n \equiv F(z_n), \tag{3.6}$$

with $G_0 \equiv 0$. We may then express the integral as:

$$A^2_N = N \sum_{n=0}^{N} \int_{G_n}^{G_{n+1}} \frac{(\frac{n}{N} - F)^2}{F(1-F)} dF, \tag{3.7}$$

with $G_{N+1} \equiv 1$. Integrating yields

$$A_N^2 = N \sum_{n=0}^{N} \left[ \left(\frac{n}{N}\right)^2 \log F - \left(\frac{N-n}{N}\right)^2 \log(1-F) + 1 - F \right]_{G_n}^{G_{n+1}}. \quad (3.8)$$

After some manipulation, a convenient form for computation is

$$A_N^2 = -\frac{1}{N} \sum_{n=1}^{N} (2n-1) \left\{ \log F(z_n) + \log\left[1 - F(z_{N+1-n})\right] \right\} - N. \quad (3.9)$$

By definition this statistic is non-negative.

MATLAB has both Kolmogorov-Smirnov and Anderson-Darling tests coded. The KS test is provided by function `kstest`. The default is to test for the standard normal distribution, which is what we want. The AD test is available as function `adtest`. The default is to test for normality, but for unknown mean and variance, which are estimated from the data. This isn't what we want, so we specify the distribution. The function makes use of 3.9, so we needn't code it ourselves when using MATLAB. Our MATLAB code for this problem is provided in Narsky and Porter (2014b).

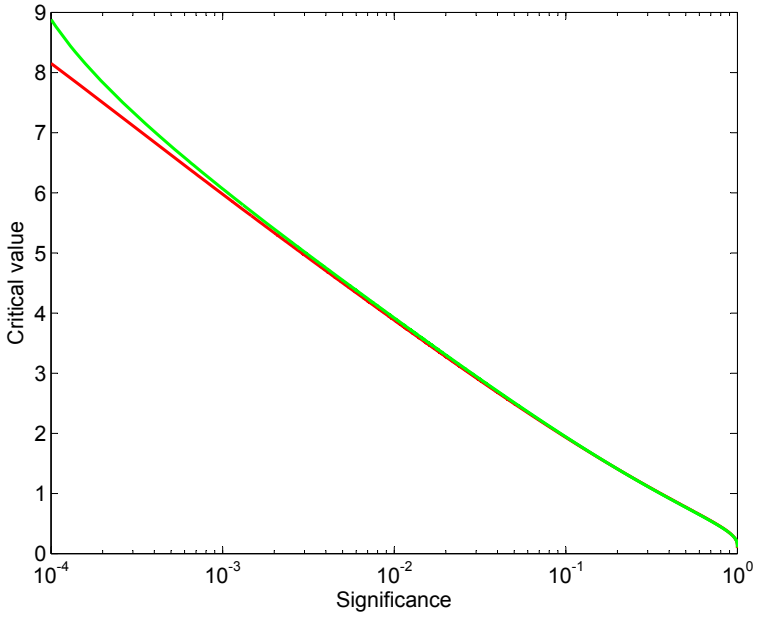


**Figure 3.1** The critical value for the AD test as a function of significance. The red curve is for $N = 100$ and the green for $N = 10$.

MATLAB's `adtest` package uses the empirical formula of Marsaglia and Marsaglia (2004) in order to compute the $p$ value corresponding to a given value

| $N$ | KS power | AD power |
|---|---|---|
| 10 | 0.063 | 0.80 |
| 100 | 0.85 | 1 |

**Table 3.1** Comparison of power for the KS and AD tests of standard normality on a dataset drawn from a Cauchy distribution. All entries are for a significance level of 0.01.

of $A_N^2$. The dependence of critical value on significance is shown in Fig. 3.1. For small significances, the variation with sample size may be seen. Some further discussion of the reliability of the $p$ values computed in `adtest` is given below.

We first check whether the tests (and our code) are behaving as we expect. Fig. 3.2 shows the distribution of $p$ values for data sampled from $H_0$. There are 100,000 entries. These distributions should be uniform and indeed they look uniform. The bin size is 0.01, so we are getting expected behavior at least down to the desired 0.01 significance level. That is, the first bin has approximately 1000 counts, or 1% of the distribution. The $\chi^2$ test for uniformity yields $p$ values of 36% for the KS test and 30% for the AD test.

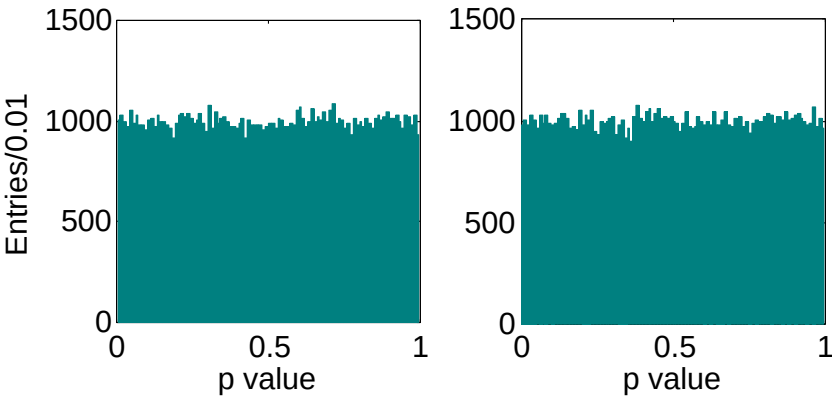


**Figure 3.2** The distribution of test $p$ values for data sampled from $H_0$ for the KS test (left) and the AD test (right). Each entry is for a sample of size $N = 100$.

Next, we compute the power as a function of significance for both tests, for data sampled from the specified Cauchy. The results are shown in Fig. 3.3, both for $N = 10$ and $N = 100$. As expected, for both KS and AD the power is greater at a given significance for larger sample sizes. The AD test does substantially better than the KS test for both sample sizes. This is also seen in Table 3.1, where the power is tabulated for a significance of 0.01.

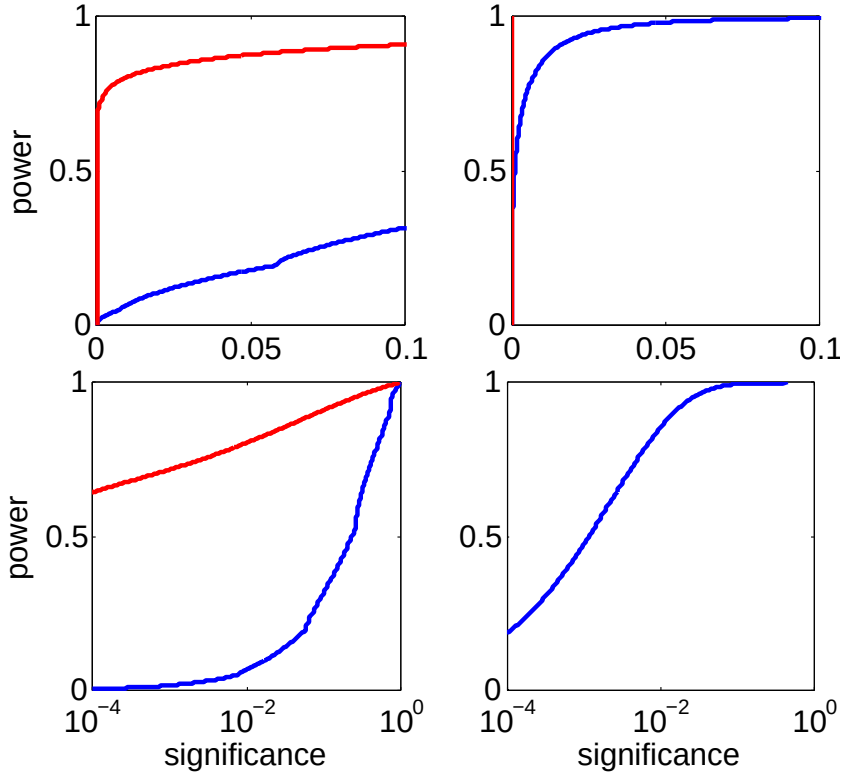


**Figure 3.3** Power vs significance for the KS test (blue) and the AD test (red). Left: $N = 10$. Right: $N = 100$. The top plots have linear scales in significance, the bottom logarithmic. The power for the AD test is essentially one for the entire range of the lower right plot.

**The fine print – Anderson-Darling**

We must insert a note of caution. The $A_N^2$ and $p$ values returned in performing the AD test are not necessarily valid, due to the limited numerical precision of the 64-bit floating point representation. The problem occurs at two levels. First, a problem appears when the cdf for the normal distribution approaches one. Because of the long tail of the Cauchy, there may be observations where the normal cdf evaluates to one due to the limited precision of the double precision (64-bit) floating point implementation. This happens when the cdf is bigger than about $1 - 10^{-16}$. In this case, Eq. 3.9 evaluates to `Inf` in MATLAB's `adtest`. Second, the Cauchy tails are so long that even for the low tail of the cdf, near zero, the evaluation of the normal cdf may be smaller than the smallest non-zero floating point number. We note that the smallest non-zero 64-bit floating point number in our case is about $2 \times 10^{-308}$. Then Eq. 3.9 evaluates to $-$`Inf` in `adtest`.

Faced with these problems, we have created a modified version of the MATLAB

`adtest` package. The first problem is avoided if we can we assume a symmetric distribution under $H_0$, as holds in the present case, and then evaluate $1 - F$ directly using the lower tail of the distribution. This avoids taking the difference between one and a number very near one. However, we then encounter the second problem when far enough out on the tail. The second problem is greatly mitigated by evaluating the logarithm of the cdf instead of the cdf itself. Our modified `adtest` does this, evaluating the AD statistic with a call to `adstatnormcdf`, in turn using `lognormcdf`. We provide these two functions at Narsky and Porter (2014c) and Narsky and Porter (2014d). With these modifications, we plot the distribution of the AD statistic for samples from the Cauchy distribution in Fig. 3.4. Note that without the modifications to `adtest` the values would nearly all be evaluated as infinity. The smallest sampled value is 9.3 and the largest is more than $10^{10}$, above the upper limit of the figure.

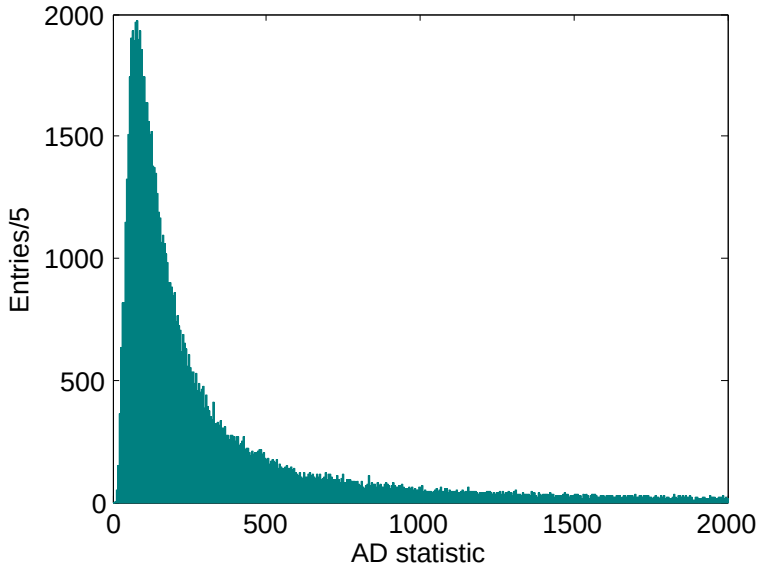


**Figure 3.4** The values of the AD statistic under H1 for $N = 100$. There are $10^5$ observations. The overflow tail extends past $10^{10}$.

Figures 3.1 and 3.3 shows significances down to $10^{-4}$. The Marsaglia and Marsaglia (2004) code implemented in `adtest` appears to be reliable down to this level, but not necessarily below. Marsaglia does not discuss $A_N^2$ values above 10, and indeed discussions of the AD test are typically concerned with significances only down to $\sim 0.01$. Thus, it is not surprising to find that things break down if we push the $p$ value calculation below $10^{-4}$. Empirically, we find that the `adtest` implementation of Marsaglia and Marsaglia (2004) breaks at $p \approx 6 \times 10^{-4}/N$. For illustration, Fig. 3.5 shows the dependence of the $p$ value from `adtest` on $A_N^2$ for $N = 50$. The curve levels out at $p = 1.2 \times 10^{-5} = 6 \times 10^{-4}/50$.

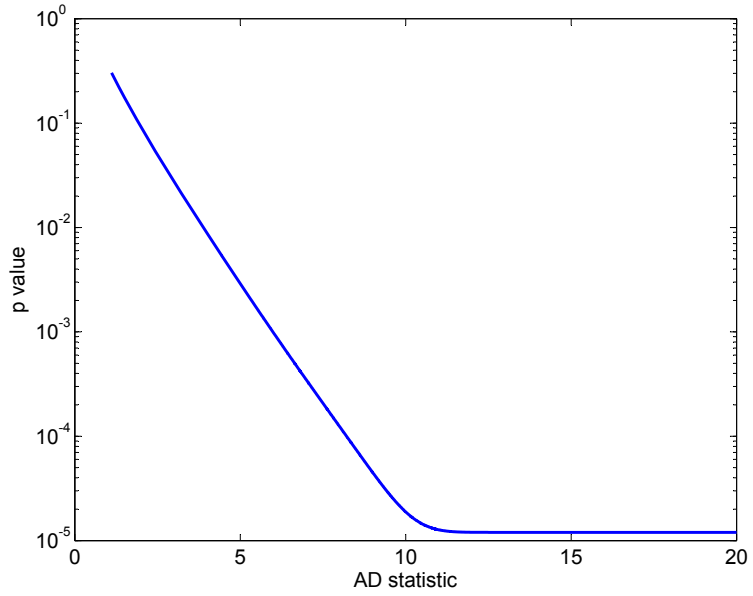


**Figure 3.5** The $p$ value vs $A_N^2$ for $N = 50$, as computed in `adtest` (modified version).

It is, of course, possible to extend the validity to smaller probabilities, at least using Monte Carlo methods. If you think it is important to do this however, you should first ask yourself why.

**✂The fine print – Kolmogorov-Smirnov**

The reader may be concerned that the dependence of power on significance for the KS test is not smooth, especially for small samples, as seen in Fig. 3.3. The locations of the singular points depend on $N$ as demonstrated below. This structure may be understood in terms of the discreteness of samples, emphasized by the substantial tails of the Cauchy distribution.

First, it can be remarked that the KS test doesn't look directly at probabilities, only at differences. For example, even if an observation occurs where the null hypothesis says the probability is zero, the null hypothesis may still be accepted.

Next, it is perhaps easier to think in terms of the critical values for the test rather than the significances, although these are monotonically related. Consider for example the negative region of the sampling space. The normal cdf quickly approaches zero as we go more negative. The Cauchy, however, approaches zero much more slowly, so there is an important probability to sample observations from the Cauchy in the region where the normal cdf is approximately zero. These observations come in discrete quanta, each such observation will contribute approximately $1/N$ to the difference in cdfs. If the critical value is near a multiple of $1/N$, singular behavior may be seen.

We can check this hypothesis by checking the location of the singular point in the top left graph in Fig. 3.3. It occurs at a significance of approximately 0.06. For $N = 10$ a signifcance level of $\alpha = 0.06$ corresponds to a KS test critical value of 0.40. This is $4/N$, that is, this singular point corresponds to the threshold for four events in the low tail (or four events in the high tail) to make the test fail. The upper range of the plot is too low for the 0.30 critical value, but at least one more singular point is visible on the log plot at the bottom.

To see this effect more clearly, we make the corresponding plots for the smaller sample sizes $N = 5$ and $N = 6$ in Fig. 3.6. For $N = 5$ there is a singular point at significance level $\alpha = 0.03$. For $N = 5$ and $\alpha = 0.03$ the KS test critical value is 0.60, that is $3/N$. For $N = 6$ there is a singular point at $\alpha = 0.07$. For $N = 6$ and $\alpha = 0.07$ the KS test critical value is 0.50, that is again at $3/N$.

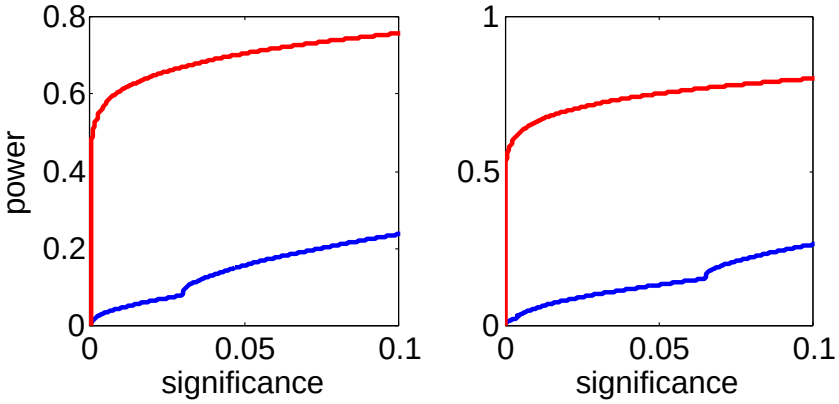


**Figure 3.6** Power vs significance for the KS test (blue) and the AD test (red). Left: $N = 5$. Right: $N = 6$.

**Discussion:**

The immediate conclusion from this study is that, for the situation considered, the AD test is much better than the KS test. Since the KS test is widely used in particle physics (and not the AD test), this should be taken as a warning that if you care about power, you should think about what test to use (not necessarily either of those considered here). This should not, however, be generalized to conclude that the KS test is always inferior, it depends on the situation. An example of where it performs well is provided in Freedman (1979) where possible seasonal variation is of interest.

It is easy to see why the AD test outperforms the KS test in this example. The AD test places special emphasis on the tails of the distribution. The Cauchy distribution has very long tails compared with the normal distribution, so the AD test works well. The KS test, on the other hand, looks for the maximum difference in cdfs. Since counting fluctuations tend to be largest near the peak of the distribution, the KS test emphasizes this region. A KS test would be more powerful, for example, in a situation where there is a shift in location between the sampled and hypothetical distributions.

The broad view is that different test statistics may be more optimal for different situations. There is no "one size fits all". The less you know about what you want to test, the less powerful you can expect your test to be. The more you know, the more knowledge you can put into your test and greater power can be obtained.

Note that, if the alternative hypothesis had been specified as the Cauchy with the given parameters, the situation would be different. We no longer test for "normal" against "not normal", but rather for "normal" against "Cauchy". This means our possibilities are more constrained, and we should be able to construct more powerful tests. In particular, both hypotheses are now simple hypotheses, and we can apply the likelihood ratio test, which is known to be uniformly most powerful for a simple test (section 2.5 of Narsky and Porter (2014a)).

Finally, we have seen that there may be numerical issues in the translation from theory to practice. If you use a "canned" package, it may not have been designed to handle your use case. To avoid mistakes, it is important to check for misbehavior.

## Bibliography

Freedman, L.S. (1979) The use of a Kolmogorov-Smirnov type statistic in testing hypotheses about seasonal variation. *Journal of Epidemiology and Community Health*, **33**, 223–228.

Marsaglia, G. and Marsaglia, J. (2004) Evaluating the Anderson-Darling distribution. *Journal of Statistical Software*, **9**, 1–5.

Narsky, I. and Porter, F.C. (2014a) *Statistical Analysis Techniques in Particle Physics*, Wiley-VCH.

Narsky, I. and Porter, F.C. (2014b) *Statistical Analysis Techniques in Particle Physics, online supplements*, Wiley-VCH. URL `http://www.hep.caltech.edu/~NarskyPorter/Solutions/NPsolution3p8.m`.

Narsky, I. and Porter, F.C. (2014c) *Statistical Analysis Techniques in Particle Physics, online supplements*, Wiley-VCH. URL `http://www.hep.caltech.edu/~NarskyPorter/Solutions/adstatnormcdf.m`.

Narsky, I. and Porter, F.C. (2014d) *Statistical Analysis Techniques in Particle Physics, online supplements*, Wiley-VCH. URL `http://www.hep.caltech.edu/~NarskyPorter/Solutions/lognormcdf.m`.