4.1 Correlation testing

Figure 1 shows the early Hubble diagram from the book (Figure 4.3).

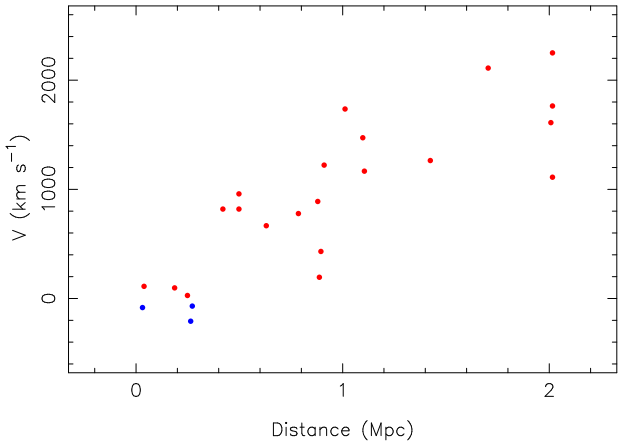

Figure 1: The early Hubble diagram for 24 galaxies.

(a) **The (Bayesian) Jeffreys test.** From the data provided for the example and plotted above, it is simple to calculate the probability of the correlation coefficient ρ via equations 4.3 and 4.8, the Jeffreys test, and we get the following picture:

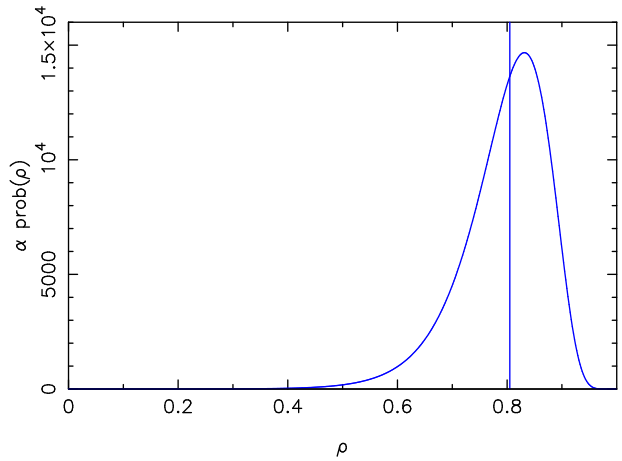

Figure 2: The Jeffreys test for correlation, assuming the bivariate Gaussian model of equation 4.1 for the data. The ordinate is proportional to the probability of the data being related by a correlation coefficient ρ. The vertical line shows the mean value of 0.807.

The mean value of ρ is 0.807 while the peak value is 0.831.

The correlation coefficient is clearly not zero, the value implying no correlation, or in more truthful terms, the value implying that the ellipses characterizing the point density contours in the $x-y$ plane are in fact circles so that x and y are independent. It shows us a probability distribution, so that we can answer the question: what possibility is there that given these data, they are NOT related, ie that $\rho = 0$? Figure 2 suggests very little chance indeed.

(b) **The classical Fisher test.** On the way through the above, we used equation 4.3 to calculate r, the famous Pearson product moment coefficient, and an estimator of ρ. The value obtained is $r = 0.837 \pm 0.062$ (with the standard deviation obtained from equation 4.9). $r = 0$ is the anticipated value for no correlation; and again it looks like it is safe to conclude we have a correlation here, as zero is many multiples of 0.062 away from 0.837. The classical test invites us to transform r (via equation 4.10) to a statistic $t' = r\sqrt{(N-2)}/\sqrt{(1-r^2)}$ which obeys the 'Students' t statistic with $N-2$ degrees of freedom. This yields a value of $t' = 7.177$. We rush off to Table A2.3, look for the line with $\nu = 24-2 = 22$ degrees of freedom, and find that our value of t' much exceeds that required for rejecting the null hypothesis, no correlation, at a significance level of 0.001. There is much less than 1 chance in 1000 that the correlation we observe could have arisen by chance from a random distribution of x and y.

We can do a bootstrap test to see if we believe the results of (a), or our result in (b) from looking up something in a table. The bootstrap test, as emphasized in Section 6.6, is particularly easy - all we do is draw at random (with replacement) 24 pairs of (x, y) from the tabulated 24 data pairs, and calculate a new value of r each time. The distribution of these r-values gives us the probability distribution of r, given these data. The calculation of r is not cpu-intensive so that we can achieve any accuracy for the distribution we like. Figure 3 shows the distribution for 10^6 trials.

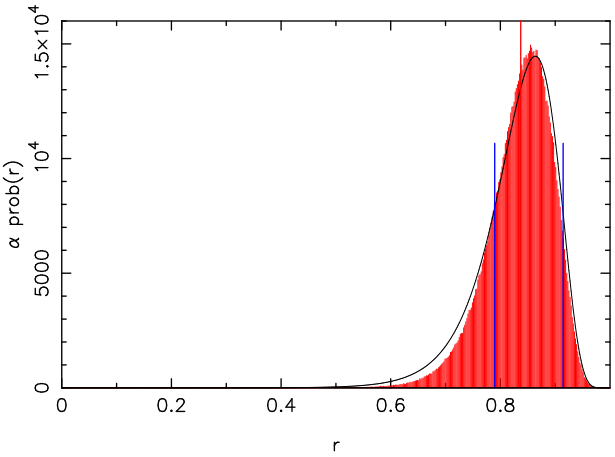

Figure 3: The bootstrap test to determine the distribution in r, the Pearson product moment coefficient, for the 24 data-pairs of the Hubble diagram in Figure 1. The histogram shows the r-values determined in 1,000,000 trials. The value of r determined from the data, 0.837, is shown as the extended red line, while the distance between the blue lines indicates $\pm 2\sigma$, with σ calculated as in equation 4.10. These $\pm\sigma$ lines have been placed approximately about the peak value of r in the distribution. The black curve represents the Fisher curve (equation 4.9), the probability of getting a value of r given that the value of ρ is 0.837.

We see that the distribution resembles that of the Bayesian calculation shown in Figure 2; we see that the calculated value for σ closely indicates the spread of the distribution; and we see that the distribution matches the Fisher distribution of equation 4.9 quite closely. This all seems self-consistent. In fact no value of r in 1,000,000 trials

has got anywhere near zero, so we can state that our significance level is at least 10^6. And of course we could carry out any further extension we like until we find a value of zero, and thus determine exactly the value of the significance of our result. Maybe it is 10^9, which surely must convince anybody....

Or so we might think. It is important to consider further what we have done and there are two serious caveats.

(1) First, look at the bootstrap in the following light - make the transformation to the t-plane and rebuild the histogram. The result is shown in Figure 4.

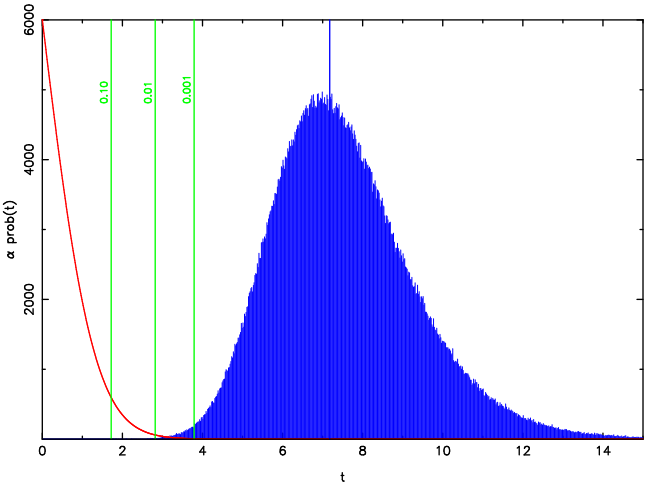

Figure 4: The bootstrap test for the 24 data-pairs of the Hubble diagram in Figure 4.3. The histogram shows the t-values calculated from the individual r values (with 22 degrees of freedom) determined in 1,000,000 trials; the extended vertical line shows the t-value (=7.18) corresponding to the r-value (0.837) of the original data. The red curve is the integrated t distribution, and the green vertical lines indicate the significance levels, or proportion of area in the tail of the distribution to the right of each vertical line. The t-values at these green lines correspond to the critical values given in Table A2.3 for 22 degrees of freedom.

The formal test then consists of comparing the position of the vertical blue line with the positions of the green lines representing proportions of the (integral) t-distribution. Our bootstrap (non-observed) values of t' are being compared with a distribution of non-observed t which we know from the start is incorrect (because it applies to values of r from a zero-correlation - and we *knew* we had some signal.) We produce a numerical, quantified result from this procedure and convince ourselves in this process that we have proved a correlation exists.

Tradition says that this is satisfactory. But really the probability distributions of Figures 2 and 3 tell us what we want to know. Indeed the bootstrap result of Figure 3 seems to settle the argument by providing the simplest and most graphic of tests. There is no need for comparison with fictitious distributions - the bootstrap *tells* you that r (and therefore ρ) is not zero in this case and never (well hardly ever) could be.

(2) The second caveat is the most important. *All of the foregoing is modelled on the bivariate Gaussian premise.* Modern data provides very good evidence that the

Hubble diagram is not well represented by a bivariate Gaussian! Even this version of the Hubble diagram does not really resemble a bivariate Gaussian - the scatter is too even. The scatter in Figure 1 at the low end is boosted by Hubble's inclusion of Local Group objects, objects which do not represent the Hubble flow, while the scatter at the 'distant' end is boosted by measurement error - and these combine to give a fairly uniform scatter, unlike a bivariate Gaussian. However, the fact that the bootstrap result of Figure 3 is similar to the Fisher distribution suggests that the deviation from Gaussian is not great enough in this case to invalidate the Fisher test.

Nevertheless, to play safe a non-parametric test should be used.

(c) **The Spearman rank correlation coefficient.** Ranking tests for correlation are excellent in that they don't care about the form of the relation. They simply get you to find the rank of each of the two variables and then examine statistically how much the ranks of the individual (X_i, Y_i) pairs differ. If on average they don't differ by much, then a strong correlation is present. The best-known of these tests comes from calculating the Spearman rank correlation coefficient (equation 4.12)

$$r_s = 1 - 6\frac{\sum^N (X_i - Y_i)^2}{N^3 - N}.$$

(Ranking can be done by eye for 24 variable-pairs. For larger data sets, resort to available indexing/ranking routines, such as the pair of routines *indexx.for* and *rank.for* of *Numerical Recipes*. Remember that the index of the variable is not the rank of the variable; at least with the *Numerical Recipes* routines getting the rank is a two-stage process.)

The result here is $r_s = 0.879$

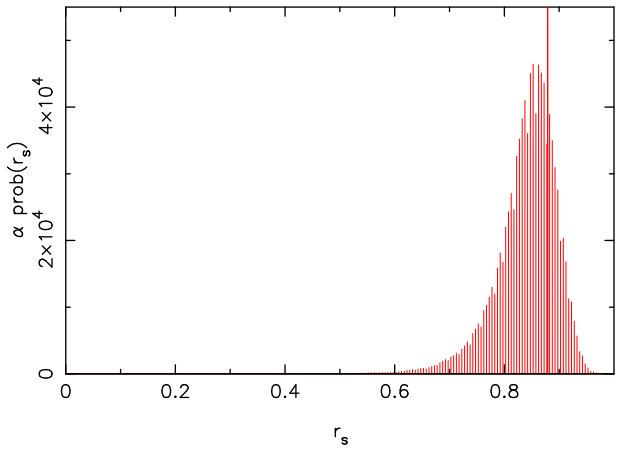

Figure 5: The bootstrap test to determine the distribution in r_s, the Spearman rank correlation coefficient, for the 24 data-pairs of the Hubble diagram in Figure 4.3. The tall red line indicates the value of $r_s = 0.879$ for the original 24 data-pairs.

The significance of the test is assessed in the traditional way; a high value of r_s indicates a correlation, the level of significance depending on the number of data-pairs.

Consulting Table A2.5 we see once again that the indicated level of significance is off the end of the 22-degrees-of-freedom line, i.e. much smaller than 0.001.

But the same criticism as described for Figure 4 prevails with this; what are we really saying when we make this statement? Now however we can truly call on the bootstrap to tell us what the probabilities are. The results of 10^6 trials appear in Figure 5.

Never mind the tables and comparison of mismatched distributions. The bootstrap trials in this non-parametric test tell us unequivocally that the chances of a value of $r_s = 0$ arising by chance from this data set is far less than 1 in 10^6.

This seems the best answer we can get. We have a result that

1. ignores the form or origin of the data;
2. gives us a view of the probabilities of correlation offered by the data;
3. enables us to make a quantitative (and in this case unequivocal) judgement about the presence of a correlation; and
4. avoids any need to consult a fictitious distribution table.

What is perhaps surprising is how robust the parametric method is, at least in terms of the Fisher coefficient r, its bootstrap distribution, and the Jeffreys test for correlation.

There are many other ways of looking at the problem. For example note the following:

Suppose you fit a line of zero intercept (ie just a slope) assuming errors only in y. The maximum likelihood estimate of the slope is

$$\sum(x_i y_i)/\sum(x_i^2)$$

which will be large if r is large. So in the line-fitting case r becomes an indicator of a non-zero slope.

In considering some additional and educational analysis, use formulations of Chapter 6 to produce some bivariate Gaussian distributions of differing ρ. Examine these with the different approaches above. Throw in some outliers; check out robustness.

Consider in particular $\rho = 0$. If you carry out bootstrap tests, do you recover the famous t distribution for no correlation, the basis of the significance tables for standard correlation testing?