

Data set 4:

Butterfly Fish Pairs

Background

Ernie Reese (emeritus professor in the Zoology Department) and some of his students have been studying the social and mating behavior of *Chaetodon multivinctus* (the multibanded or pebbled butterflyfish). These fish, which are endemic to Hawai'i, typically are seen in male-female pairs. One question of interest was whether there is size-assortative mating, that is, whether the fish in a pair tend to be similar to each other in size.



The question:

Are paired fish similar in size (i.e. large males with large females, small with small)?

The data

Ernie collected 16 mated pairs of fish, and sexed and measured each individual. I do not know where they were from or how they were selected, but most likely they were haphazardly selected from a study site in Kaneohe Bay (O'ahu) or at Puako (Hawai'i Island). By this I mean that probably divers simply collected what pairs they could find and capture, in one area of reef.

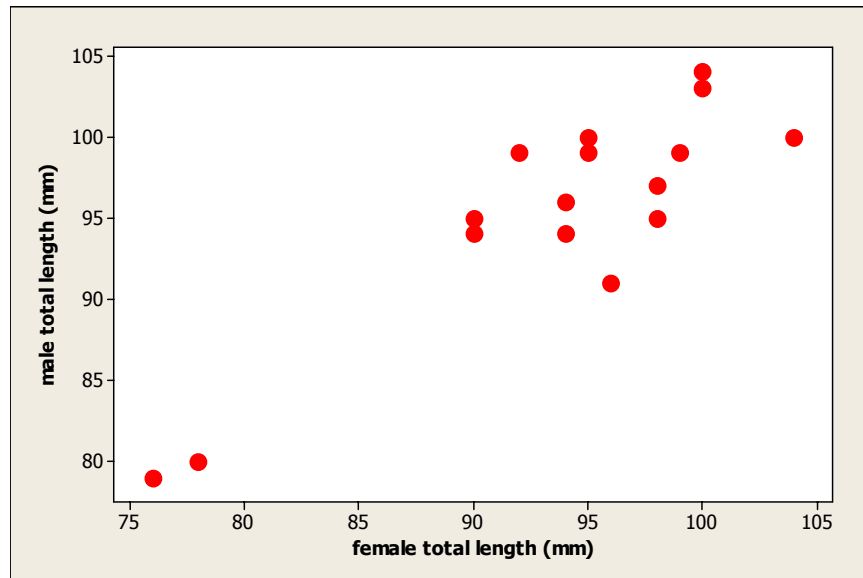
Observations are total lengths in mm:

| pair | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|--------|-----|----|-----|----|----|----|----|----|----|----|----|----|-----|----|-----|----|
| male | 103 | 97 | 100 | 95 | 99 | 80 | 79 | 94 | 94 | 96 | 99 | 91 | 104 | 95 | 100 | 99 |
| female | 100 | 98 | 95 | 90 | 95 | 78 | 76 | 90 | 94 | 94 | 92 | 96 | 100 | 98 | 104 | 99 |

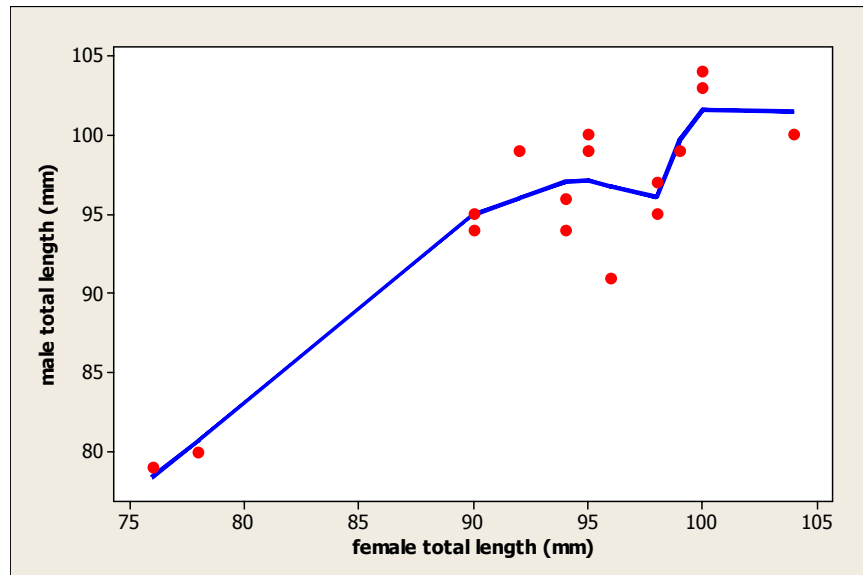
Data exploration

Displays of the relationship

Scatterplot



Scatterplot with LOWESS smoother



These plots show a positive association overall. Two pairs of fish are much smaller (female and male both) than the others. The positive association is still present if these outliers are ignored, but is considerably weaker looking.

Statistical summary of the relationship

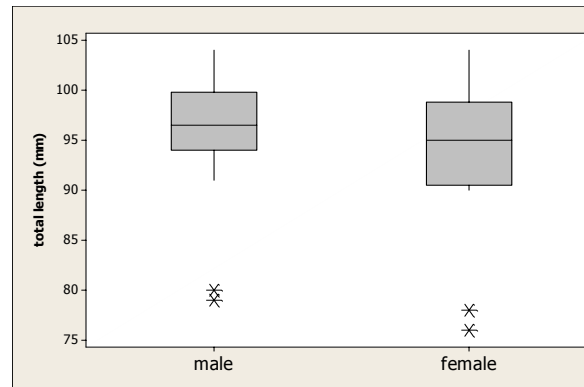
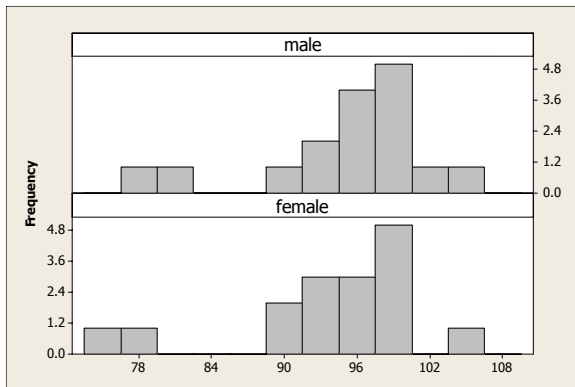
Correlation coefficient

all pairs: 0.888

without outliers: 0.536

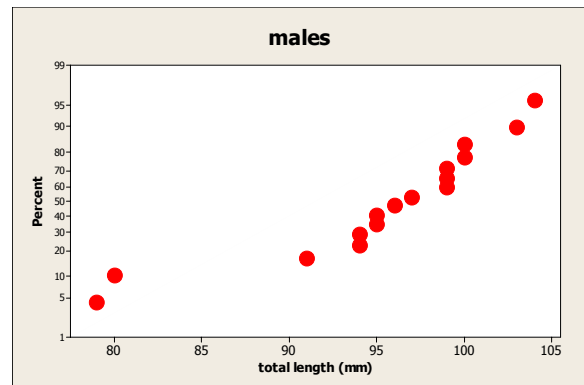
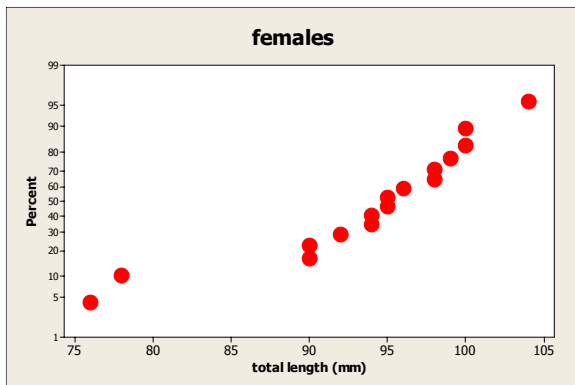
These statistics show there is a very strong positive relationship when all pairs are included, and a moderately strong positive relationship without the outliers.

Marginal distributions



Both distributions have two low outliers. The remainder of the observations show mildly skewed distributions, not with extended tails but with more observations below the peak than above it; in the boxplots this is shown by the shortness of the lower whiskers, especially for females.

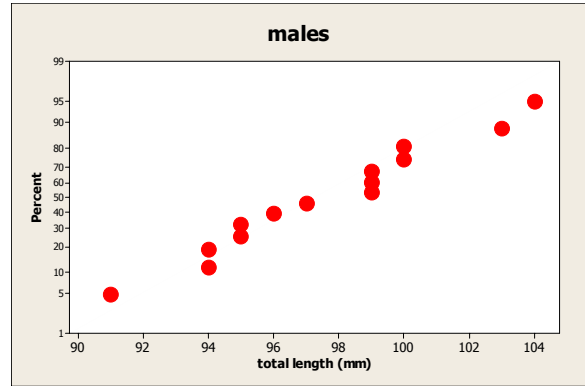
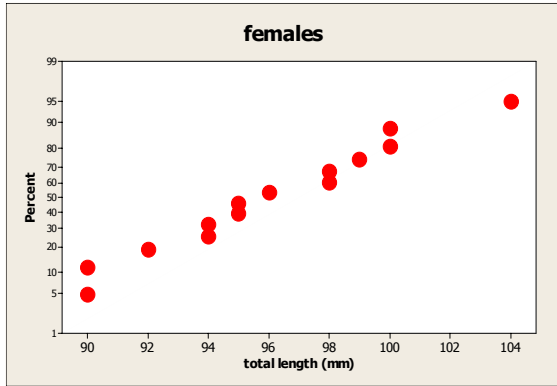
normal quantile-quantile plots



With all observations included, the distributions clearly are non-normal, with two low outliers in each. With the outliers excluded (next page) the distributions are fairly close to normal.

Conclusions (descriptive)

Yes, there is a positive association: larger than average females paired with larger than average males, and smaller than average females with smaller than average males. Two pairs of



very small fish contribute substantially to this association, though it is present even if these outliers are excluded.

Because of the outliers the full data set does not have a bivariate-normal distribution. If the outliers are excluded the distribution is roughly bivariate normal.

Inference

The scope of inference

As noted on the first page of this handout, I do not know exactly how the fish pairs were selected, but they certainly were not any sort of formal random sample from the entire species (e.g. randomized geographic multistage sampling). Rather, they probably were haphazardly sampled from some fairly localized population.

Any conclusions from these data therefore could be considered to apply to fish pairs in the locality where these were collected, if we are willing to assume that they are representative of that population, i.e. that whether a pair was collected was independent of how similar or dissimilar the male and female sizes were. I am willing to believe that this assumption is at least roughly true, and thus that the sampling of that locality was effectively unbiased.

To extend the conclusions to the species as a whole would require the assumption that fish in the locality these pairs came from are typical of the species, and there is no way this study can support (or refute) this assumption.

Assuming bivariate normality

Since the assumption of bivariate normality is not reasonable for all 16 observations, this test will be conducted only for the 14 larger pairs. Since we expect a positive association, the test is one-sided.

$$H_0: \rho = 0 \qquad H_a: \rho > 0$$
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.536\sqrt{12}}{\sqrt{1-0.287}} = 2.20 \quad df = 12 \quad P = 0.024$$

There is fairly strong evidence of a positive association between female and male sizes. The 2 small pairs, omitted from this analysis, further strengthen this conclusion, though they are not consistent with a bivariate-normal model.

Nonparametric analyses

To allow comparison with the preceding normal-theory test, the distribution-free procedures will be applied to both the full ($n = 16$) and reduced ($n = 14$) data sets.

Kendall's τ

$$H_0: \tau = 0 \qquad H_a: \tau > 0$$

$n = 16$, i.e. with outliers:

$$S = P - Q = 89 - 20 = 69$$
$$\hat{\tau} = \frac{S}{n(n-1)/2} = 0.575 \quad n = 16 \quad P < 0.005$$

A more extensive table in Hollander & Wolfe gives $P < 0.001$. SAS, using a large-sample approximation with correction for ties, gives $\hat{\tau} = 0.603$, $P = 0.00085$.

$n = 14$, i.e. without outliers:

$$S = P - Q = 60 - 20 = 40$$

$$\hat{\tau} = \frac{S}{n(n-1)/2} = 0.440 \quad n = 14 \quad P < 0.025$$

Spearman's rank correlation:

H_0 : no association

H_a : positive association

$n = 16$, i.e. with outliers:

$$r_s = 0.738 \quad n = 16 \quad P < 0.001$$

$n = 14$, i.e. without outliers:

$$r_s = 0.607 \quad n = 14 \quad P < 0.025$$

Resampling methods

Randomization test

H_0 : $\rho = 0$

H_a : $\rho > 0$

$n = 16$, i.e. with outliers:

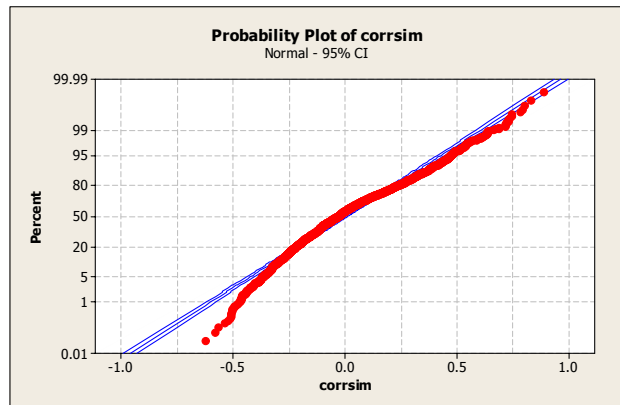
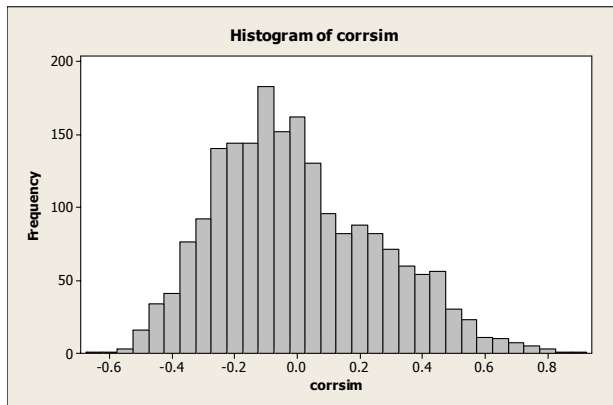
Correlation coefficient 0.888

Number of randomizations 1999

One sided randomization p-value, H_1 : +ve correlation 0.0005

This test again is highly significant; note that the P-value is the lowest value possible with 1999 randomizations.

The distribution of correlation coefficients from the randomizations is somewhat skewed, with more negative values but a longer positive tail. This shows the effect of the outlier pairs:



when the data are randomized with no relationship between the two variables, the very small fish most often are paired with normal-sized fish, which produces a negative correlation, while occasionally the small fish are paired with other small fish, producing large positive correlations.

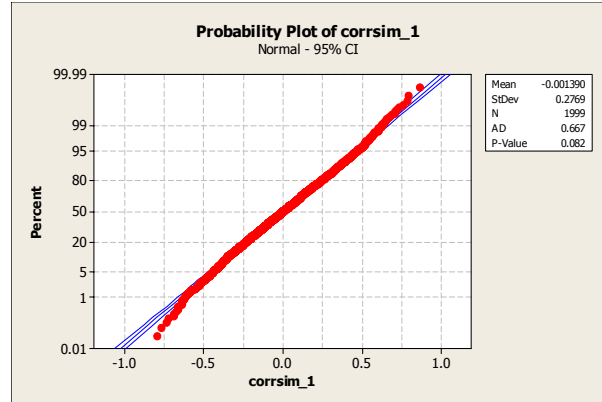
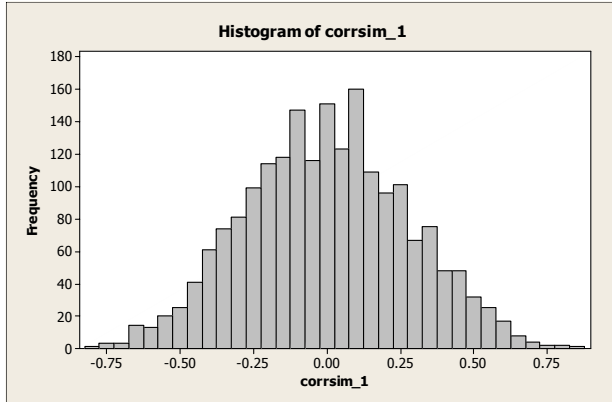
$n = 14$, i.e. without outliers:

Correlation coefficient 0.536

Number of randomizations 1999
 One sided randomization p-value, H1: +ve correlation **0.0275**

This result is very similar to those from all the previous tests, for this subset of the data: significant evidence of a positive correlation, but not nearly so strong or significant as when the two outlier pairs are included.

The distribution of correlation coefficients from the randomizations is much more symmetric and close to normal than was the case for the full data set (see graphs on next page).

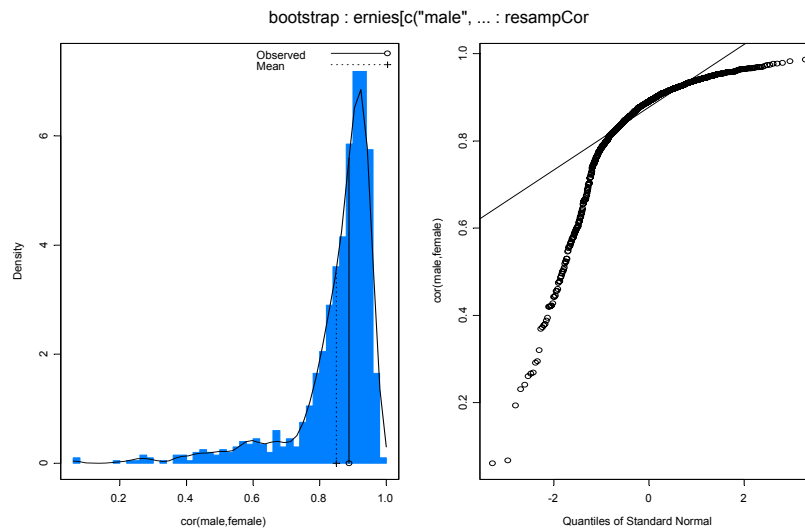


Bootstrap CIs for ρ

$n = 16$, i.e. with outliers:

Percentiles: 0.4457634 0.9643852
 BCa: 0.4922651 0.9663264
 Tilting: 0.6500244 0.9581673

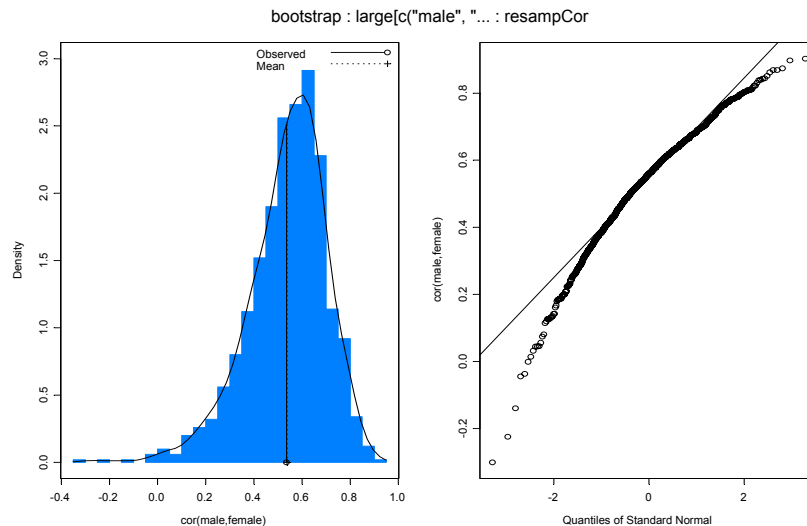
The upper bounds by all three methods are similar (near 0.96). The lower bounds, though, are fairly different. This is the result of the bootstrap distribution being very skewed, so that modifying exactly where in the distribution (i.e. what rank of the randomized values) a bound will be has a greater effect on the left side than the right.



$n = 14$, i.e. without outliers:

Percentiles: 0.1610173 0.7996115
BCa: 0.04445097 0.7674373
Tilting: 0.16346 0.7354139

In the absence of the outliers the confidence intervals are lower and wider than with all 16 observations, selecting the considerably smaller correlation coefficient without the outlier pairs. The upper bounds again are more similar among the three methods than are the lower bounds, though this difference is not as great as for the full data set because the bootstrap distribution is not as skewed.



Conclusions

Among the pairs of fish in this sample there was strong positive association of male and female sizes. If we assume these pairs to be representative of the locality from which they were collected, we can conclude that this positive association is present in that population; $\hat{\tau} = 0.575$, $P < 0.001$. The strongest evidence of this association comes from the two pairs of very small fish, but even if these are excluded there is a statistically significant association.

Which procedure to use?

Since all the tests gave similar results for the same data (i.e. for all 16 observations or for the 14 without the outliers), it doesn't matter much which procedure is adopted, except that it would be preferable to be able to use all the data. I would report Kendall's τ and test, for all 16 observations. This pertains to an understandable measure of association (unlike Spearman's), and uses the same logic as does the randomization test but is easier.