Chapter 4
Covariance, Regression, and Correlation

> "Co-relation or correlation of structure" is a phrase much used in biology, and not least in that branch of it which refers to heredity, and the idea is even more frequently present than the phrase; but I am not aware of any previous attempt to define it clearly, to trace its mode of action in detail, or to show how to measure its degree.(Galton, 1888, p 135)

A fundamental question in science is how to measure the relationship between two variables. The answer, developed in the late 19th century, in the the form of the *correlation coefficient* is arguably the most important contribution to psychological theory and methodology in the past two centuries. Whether we are examining the effect of education upon later income, of parental height upon the height of offspring, or the likelihood of graduating from college as a function of SAT score, the question remains the same: what is the strength of the relationship? This chapter examines measures of relationship between two variables. Generalizations to the problem of how to measure the relationships between sets of variables (multiple correlation and multiple regression) are left to Chapter 5.

In the mid 19th century, the British polymath, Sir Francis Galton, became interested in the intergenerational similarity of physical and psychological traits. In his original study developing the correlation coefficient Galton (1877) examined how the size of a sweet pea depended upon the size of the parent seed. These data are available in the **psych** package as `peas`. In subsequent studies he examined the relationship between the average height of mothers and fathers with those of their offspring Galton (1886) as well as the relationship between the length of various body parts and height Galton (1888). Galton's data are available in the **psych** packages as `galton` and `cubits` (Table 4.1)[1]. To order the `table` to match the appearance in Galton (1886), we need to `order` the rows in decreasing order. Because the `rownames` are characters, we first convert them to `ranks`.

Examining the table it is clear that as the average height of the parents increases, there is a corresponding increase in the heigh of the child. But how to summarize this relationship? The immediate solution is graphic (Figure 4.1). This figure differs from the original data in that the data are randomly *jittered* a small amount using `jitter` to separate points at the same location. Using the `interp.qplot.by` function to show the *interpolated medians* as well as the first and third quartiles, the medians of child heights are plotted against the middle of their parent's heights. Using a smoothing technique he had developed to plot meterological data Galton (1886) proceeded to estimate error ellipses as well as slopes through the smoothed

[1] For `galton`, see also **UsingR**.

Table 4.1 The relationship between the average of both parents (mid parent) and the height of their children. The basic data table is from Galton (1886) who used these data to introduce reversion to the mean (and thus, linear regression). The data are available as part of the **UsingR** or **psych** packages. See also Figures 4.1 and 4.2.

```
> library(psych)
> data(galton)
> galton.tab <- table(galton)
> galton.tab[order(rank(rownames(galton.tab)),decreasing=TRUE),] #sort it by decreasing row values

       child
parent 61.7 62.2 63.2 64.2 65.2 66.2 67.2 68.2 69.2 70.2 71.2 72.2 73.2 73.7
  73      0    0    0    0    0    0    0    0    0    0    0    1    3    0
  72.5    0    0    0    0    0    0    0    1    2    1    2    7    2    4
  71.5    0    0    0    0    1    3    4    3    5   10    4    9    2    2
  70.5    1    0    1    0    1    1    3   12   18   14    7    4    3    3
  69.5    0    0    1   16    4   17   27   20   33   25   20   11    4    5
  68.5    1    0    7   11   16   25   31   34   48   21   18    4    3    0
  67.5    0    3    5   14   15   36   38   28   38   19   11    4    0    0
  66.5    0    3    3    5    2   17   17   14   13    4    0    0    0    0
  65.5    1    0    9    5    7   11   11    7    7    5    2    1    0    0
  64.5    1    1    4    4    1    5    5    0    2    0    0    0    0    0
  64      1    0    2    4    1    2    2    1    1    0    0    0    0    0
```

medians. When this is done, it is quite clear that a line goes through most of the medians, with the exception of the two highest values.[2]

A finding that is quite clear is that there is a "*reversion to mediocrity*" Galton (1877, 1886). That is, parents above or below the median tend to have children who are closer to the median (reverting to mediocrity) than they. But this reversion is true in either direction, for children who are exceptionally tall tend to have parents who are closer to the median than they. Now known as *regression to the mean*, misunderstanding this basic statistical phenomena has continued to lead to confusion for the past century Stigler (1999). To show that regression works in both directions Galton's data are also plotted for child regressed on mid parent (left hand panel) or the middle parent height regressed on the child heights (right hand panel of Figure 4.2.

Galton's solution for finding the slope of the line was graphical although his measure of *reversion*, r, was expressed as a reduction in variation. Karl Pearson, who referred to *Galton's function* later gave Galton credit as developing the equation we now know as the *Pearson Product Moment Correlation Coefficient* Pearson (1895, 1920).

Galton recognized that the prediction equation for the best estimate of Y, $\hat{Y}$, is merely the solution to the linear equation

$$\hat{Y} = b_{y.x}X + c \tag{4.1}$$

which, when expressed in deviations from the mean of X and Y, becomes

$$\hat{y} = b_{y.x}x. \tag{4.2}$$

[2] As discussed by Wachsmuth et al. (2003), this bend in the plot is probably due to the way Galton combined male and female heights.

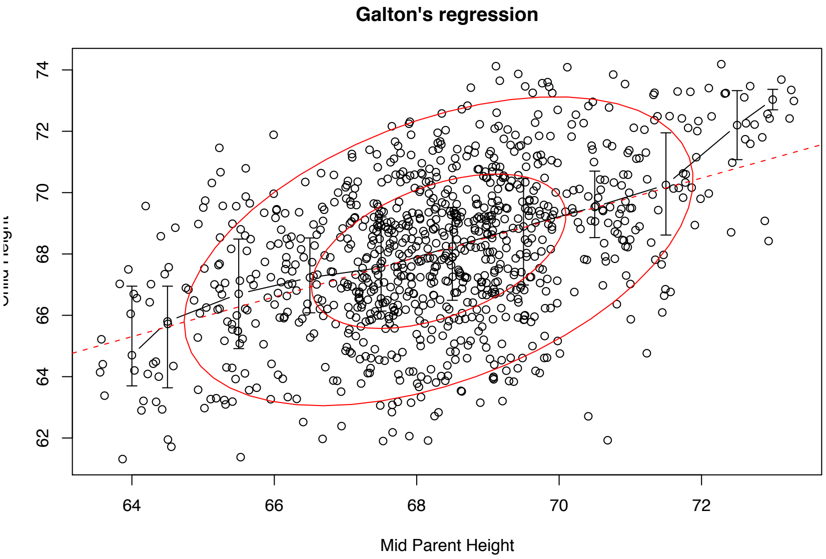


Fig. 4.1 The data in Table 4.1 can be plotted to show the relationships between mid parent and child heights. Because the original data are grouped, the data points have been *jittered* to emphasize the density of points along the median. The bars connect the first, 2nd (median) and third quartiles. The dashed line is the best fitting linear fit, the ellipses represent one and two standard deviations from the mean.

The question becomes one of what slope best predicts Y or y. If we let the residual of prediction be $e = y - \hat{y}$, then V_e, the average squared residual $\sum_{i=1}^{n} e^2/n$, will be a quadratic function of $b_{y.x}$:

$$V_e = \sum_{i=1}^{n} e^2/n = \sum_{i=1}^{n} (y - \hat{y})^2/n = \sum_{i=1}^{n} (y - b_{y.x}x)^2/n = \sum_{i=1}^{n} (y^2 - 2b_{y.x}xy + b_{y.x}^2 x^2)/n \qquad (4.3)$$

V_e is minimized when the first derivative with respect to b of equation 4.3 is set to 0.

$$\frac{d(V_e)}{d(b)} = \sum_{i=1}^{n} (2xy - 2b_{y.x}x^2)/n = 2Cov_{xy} - 2b\sigma_x^2 = 0 \qquad (4.4)$$

which implies that

$$b_{y.x} = \frac{Cov_{xy}}{\sigma_x^2}. \qquad (4.5)$$

That is, $b_{y.x}$, the slope of the line predicting y given x that minimizes the squared residual (also known as the squared error of prediction) is the ratio of the *Covariance* between x and y and the *Variance* of X. Similarly, the slope of the line that best predicts x given values of y will be

$$b_{x.y} = \frac{Cov_{xy}}{\sigma_y^2}. \qquad (4.6)$$

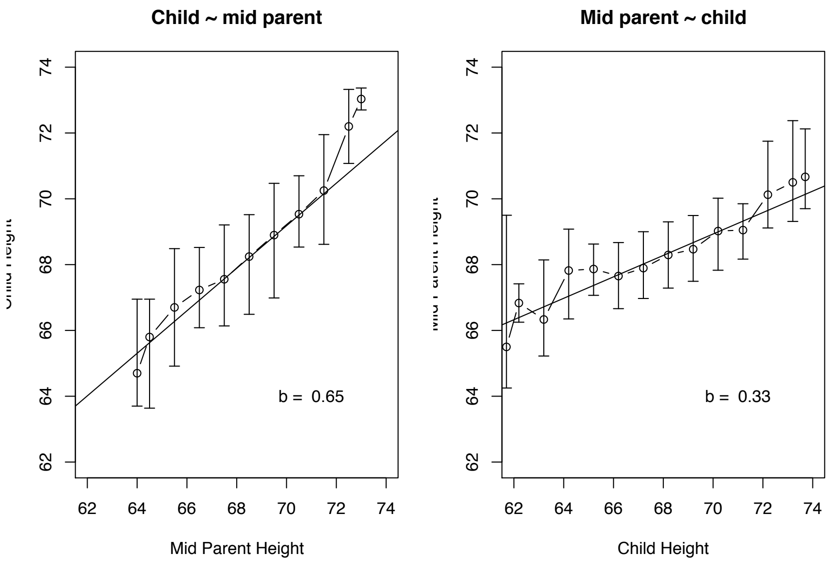


Fig. 4.2 Galton (1886) examined the relationship between the average height of parents and their children. He corrected for sex differences in height by multiplying the female scores by 1.08, and then found the average of the parents (the mid parent). Two plots are shown. The left hand panel shows child height varying as the mid parent height. The right hand panel shows mid parent height varying as child height. For both panels, the vertical lines and bars show the first, second (the median), and third interpolated quartiles. The slopes of the best fitting lines are given (see Table 4.2). Galton was aware of this difference in slopes and suggested that one should convert the variability of both variables to standard units by dividing the deviation scores by the inter-quartile range. The non-linearity in the medians for heights about 72 inches is discussed by Wachsmuth et al. (2003)

As an example, consider the `galton` data set, where the variances and covariances are found by the `cov` function and the slopes may be found by using the *linear model* function `lm` (Table 4.2). There are, of course two slopes: one for the best fitting line predicting the height of the children given the average (mid) of the two parents and the other is for predicting the average height of the parents given the height of their children. As reported by Galton, the first has a slope of .65, the second a slope of .33. Figure 4.2 shows these two regressions and plots the median and first and third quartiles for each category of height for either the parents (the left hand panel) or the children (the right hand panel). It should be noted how well the linear regression fits the median plots, except for the two highest values. This non-linearity is probably due to the way that Galton pooled the heights of his male and female subjects (Wachsmuth et al., 2003).

4.1 Correlation as the geometric mean of regressions

Galton's insight was that if both x and y were on the same scale with equal variability, then the slope of the line was the same for both predictors and was measure of the strength of their relationship. Galton (1886) converted all deviations to the same metric by dividing through

Table 4.2 The variance/covariance matrix of a data matrix or data frame may be found by using the `cov` function. The diagonal elements are variances, the off diagonal elements are covariances. Linear modeling using the `lm` function finds the best fitting straight line and `cor` finds the correlation. All three functions are applied to the Galton dataset `galton` of mid parent and child heights. As was expected by Galton (1877), the variance of the mid parents is about half the variance of the children, the slope predicting child as a function of mid parent is much steeper than that of predicting mid parent from child. The `cor` function finds the covariance for standardized scores.

```
> data(galton)
> cov(galton)
> lm(child~parent,data=galton)
> lm(parent~child,data=galton)
> round(cor(galton),2)


          parent    child
parent 3.194561 2.064614
child  2.064614 6.340029


Call:
lm(formula = child ~ parent, data = galton)

Coefficients:
(Intercept)       parent
    23.9415       0.6463

Call:
lm(formula = parent ~ child, data = galton)

Coefficients:
(Intercept)        child
    46.1353       0.3256



           parent child
parent   1.00  0.46
child    0.46  1.00
```

by half the interquartile range, and Pearson (1896) modified this by converting the numbers to standard scores (i.e., dividing the deviations by the standard deviation). Alternatively, the geometric mean of the two slopes (b_xy and b_yx) leads to the same outcome:

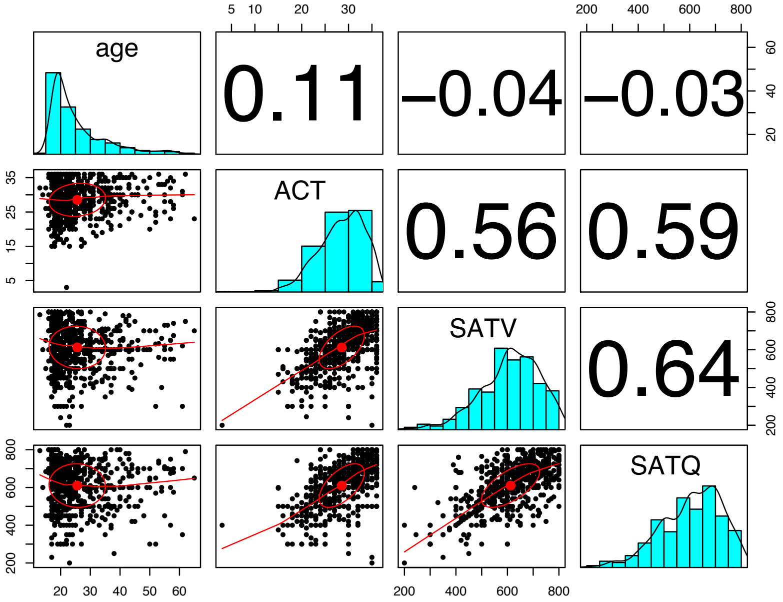
$$r_{xy} = \sqrt{b_{xy}b_{yx}} = \sqrt{\frac{(Cov_{xy}Cov_{yx}}{\sigma_x^2\sigma_y^2}} = \frac{Cov_{xy}}{\sqrt{\sigma_x^2\sigma_y^2}} = \frac{Cov_{xy}}{\sigma_x\sigma_y} \tag{4.7}$$

which is the same as the covariance of the standardized scores of X and Y.

$$r_{xy} = Cov_{z_xz_y} = Cov_{\frac{x}{\sigma_x}\frac{y}{\sigma_y}} = \frac{Cov_{xy}}{\sigma_x\sigma_y} \tag{4.8}$$

In honor of Karl Pearson (1896), equation 4.8, which expresses the correlation as the product of the two standardized deviation scores, or the ratio of the moment of dynamics to the square root of the product of the moments of inertia, is known as the *Pearson Product Moment Correlation Coefficient*. Pearson (1895, 1920), however, gave credit for the correlation coefficient to Galton (1877) and used r as the symbol for correlation in honor of *Galton's function* or the *coefficient of reversion*. Correlation is done in R using the `cor` function, as well as `rcorr` in the **Hmisc** package. Tests of significance (see section 4.4.1) are done using `cor.test`. Graphic representations of correlations that include locally smoothed linear fits (*lowess* regressions) are shown in the `pairs` or in the `pairs.panels` functions. For the `galton` data set, the correlation is .46 (Table 4.2).

Fig. 4.3 Scatter plots of matrices (SPLOMs) are very useful ways of showing the strength of relationships graphically. Combined with locally smoothed regression lines (lowess), histograms and density curves, and the correlation coefficient, SPLOMs are very useful exploratory summaries. The data are from the `sat.act` data set in **psych**.



4.2 Regression and prediction

The slope $b_{y.x}$ was found so that it minimizes the sum of the squared residual, but what is it? That is, how big is the variance of the residual? Substituting the value of $b_{y.x}$ found in Eq 4.6 into Eq 4.3 leads to

$$V_r = \sum_{i=1}^{n} r^2/n = \sum_{i=1}^{n} (y-\hat{y})^2/n = \sum_{i=1}^{n} (y-b_{y.x}x)^2/n = \sum_{i=1}^{n} (y^2 + b_{y.x}^2 x^2 - 2b_{y.x}xy)/n$$

$$V_r = V_y + b_{y.x}^2 V_x - 2b_{y.x}Cov_{xy} = V_y + \frac{Cov_{xy}^2}{V_x^2}V_x - 2\frac{Cov_{xy}}{V_x}Cov_{xy}$$

$$V_r = V_y + \frac{Cov_{xy}^2}{V_x} - 2\frac{Cov_{xy}^2}{V_x} = V_y - \frac{Cov_{xy}^2}{V_x}$$

$$V_r = V_y - r_{xy}^2 V_y = V_y(1-r_{xy}^2) \tag{4.9}$$

That is, the *variance of the residual* in Y or the variance of the error of prediction of Y is the product of the original variance of Y and one minus the squared correlation between X and Y. This leads to the following table of relationships:

Table 4.3 The basic relationships between Variance, Covariance, Correlation and Residuals

	Variance	Covariance with X	Covariance with Y	Correlation with X	Correlation with Y
X	V_x	V_x	C_{xy}	1	r_{xy}
Y	V_y	C_{xy}	V_y	r_{xy}	1
$\hat{Y}$	$r_{xy}^2 V_y$	$C_{xy} = r_{xy}\sigma_x\sigma_y$	$r_{xy}V_y$	1	r_{xy}
$Y_r = Y - \hat{Y}$	$(1-r_{xy}^2)V_y$	0	$(1-r_{xy}^2)V_y$	0	$\sqrt{1-r^2}$

4.3 A geometric interpretation of covariance and correlation

Because X and Y are vectors in the space defined by the observations, the covariance between them may be thought of in terms of the average squared distance between the two vectors in that same space (see Equation 3.14). That is, following Pythagorus, the *distance*, d, is simply the square root of the sum of the squared distances in each dimension (for each pair of observations), or, if we find the average distance, we can find the square root of the sum of the squared distances divided by n:

$$d_{xy} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$$

or

$$d_{xy}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2.$$

which is the same as

$$d_{xy}^2 = V_x + V_y - 2C_{xy}$$

but because

$$r_{xy} = \frac{C_{xy}}{\sigma_x\sigma_y}$$

$$d_{xy}^2 = \sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y r_{xy} \tag{4.10}$$

or

$$d_{xy} = \sqrt{2 * (1 - r_{xy})}. \tag{4.11}$$

Compare this to the trigonometric *law of cosines*,

$$c^2 = a^2 + b^2 - 2ab \cdot cos(ab),$$

and we see that the distance between two vectors is the sum of their variances minus twice the product of their standard deviations times the cosine of the angle between them. That is, the correlation is the cosine of the angle between the two vectors. Figure 4.4 shows these relationships for two Y vectors. The correlation, r_1, of X with Y_1 is the cosine of $\theta_1 =$ the ratio of the projection of Y_1 onto X. From the *Pythagorean Theorem*, the length of the residual Y with X removed $(Y.x)$ is $\sigma_y \sqrt{1 - r^2}$.

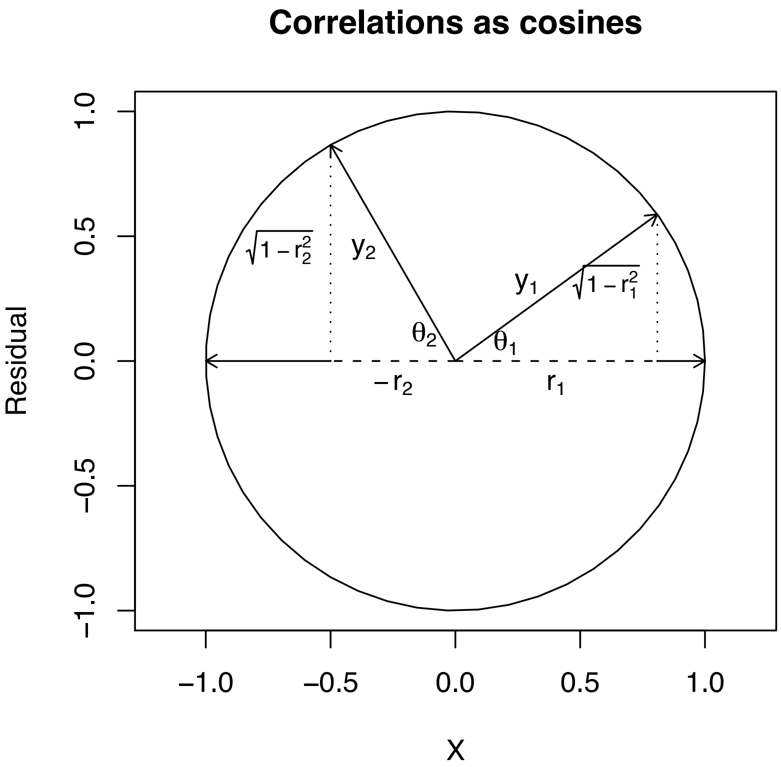


Fig. 4.4 Correlations may be expressed as the cosines of angles between two vectors or, alternatively, the length of the projection of a vector of length one upon another. Here the correlation between X and $Y_1 = r_1 = cos(\theta_1)$ and the correlation between X and $Y_2 = r_2 = cos(\theta_2)$. That Y_2 has a negative correlation with X means that unit change in X lead to negative changes in Y. The vertical dotted lines represent the amount of residual in Y, the horizontal dashed lines represent the amount that a unit change in X results in a change in Y.

Linear regression is a way of decomposing Y into that which is predicable by X and that which is not predictable (the residual). The variance of Y is merely the sum of the variances of bX and residual Y. If the standard deviation of X, Y, and the residual Y are thought of as the length of their respective vectors, then the sin of the angle between X and Y is $\sqrt{\frac{V_r}{V_y}}$ and the vector of length $1-V_r$ is the projection of Y onto X. (Refer to Table 4.3).

4.4 The bivariate normal distribution

If x and y are both continuous normally distributed variables with mean 0 and standard deviation 1, then the *bivariate normal distribution* is

$$f(x) = \frac{1}{\sqrt{(2\pi(1-r^2)}} e^{-\frac{x_1^2+2x_1x_2+x_2^2}{2(1-r^2)}}. \tag{4.12}$$

The **mvtnorm** and **MASS** packages provides functions to find the cumulative density function, probability density function, or to generate random elements from the *bivariate normal* and *multivariate normal* and t distributions (e.g., `rmvnorm` and `mvrnorm`).

4.4.1 *Confidence intervals of correlations*

For a given correlation, r_{xy}, estimated from a sample size of n observations and with the assumption of bivariate normality, the t statistic with degrees of freedom, $df = n-2$ may be used to test for deviations from 0 (Fisher, 1921).

$$t_{df} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{4.13}$$

Although Pearson (1895) showed that for large samples, that the standard error of r was

$$\frac{(1-r^2)}{\sqrt{n(1+r^2)}}$$

Fisher (1921) used the geometric interpretation of a correlation and showed that for population value ρ, by transforming the observed correlation r into a z using the arc tangent transformation:

$$z = 1/2\ log(\frac{1+r}{1-r}) \tag{4.14}$$

then z will have a mean

$$\bar{z} = 1/2\ log(\frac{1+\rho}{1-\rho})$$

with a standard error of

$$\sigma_z = 1/\sqrt{(n-3)} \tag{4.15}$$

Confidence intervals for r are thus found by using the *r to z transformation* (Equation 4.14), the standard error of z (Equation 4.15), and then back transforming to the r metric (`fisherz` and `fisherz2r`). The `cor.test` function will find the t value and associated probability value and the confidence intervals for a pair of variables. The `rcorr` function from Frank Harrell's **Hmisc** package will find Pearson or Spearman correlations for the columns of matrices and handles missing values by pairwise deletion. Associated sample sizes and p-values are reported for each correlation. The `r.con` function from the **psych** package will find the confidence intervals for a specified correlation and sample size (Table 4.4).

Table 4.4 Because of the non-linearity of the r to z transformations, and particularly for large values of the estimated correlation, the confidence interval of a correlation coefficient is not symmetric around the estimated value. The two-tailed p values in the following table are based upon the t-test for a difference from 0 with a sample size of 30 and are found using the `pt` function. The t values are found directly from equation 4.13 by the `r.con` function.

```
> n <- 30
> r <- seq(0,.9,.1)
> rc <- matrix(r.con(r,n),ncol=2)
> t <- r*sqrt(n-2)/sqrt(1-r^2)
> p <- (1-pt(t,n-2))/2
>  r.rc <- data.frame(r=r,z=fisherz(r),lower=rc[,1],upper=rc[,2],t=t,p=p)
> round(r.rc,2)
```

```
     r    z lower upper     t    p
1  0.0 0.00 -0.36  0.36  0.00 0.25
2  0.1 0.10 -0.27  0.44  0.53 0.15
3  0.2 0.20 -0.17  0.52  1.08 0.07
4  0.3 0.31 -0.07  0.60  1.66 0.03
5  0.4 0.42  0.05  0.66  2.31 0.01
6  0.5 0.55  0.17  0.73  3.06 0.00
7  0.6 0.69  0.31  0.79  3.97 0.00
8  0.7 0.87  0.45  0.85  5.19 0.00
9  0.8 1.10  0.62  0.90  7.06 0.00
10 0.9 1.47  0.80  0.95 10.93 0.00
```

add z to the table

4.4.2 Testing whether correlations differ from zero

Null Hypothesis Significance Tests, *NHST*, examine the likelihood of observing a particular correlation given the null hypothesis of no correlation. This is may be found by using Fisher's test (equation 4.13) and finding the probability of the resulting t statistic using the `pt` function or, alternatively, directly by using the `corr.test` function. Simultaneous testing of sets of correlations may be done by the `rcorr` function in the **Hmisc** package.

The problem of whether a matrix of correlations, **R**, differs from those that would be expected if sampling from a population of all zero correlations was addressed by Bartlett (1950, 1951) and Box (1949). Bartlett showed that a function of the natural logarithm of the determinant of a correlation matrix, the sample size (N), and the number of variables (p) is asymptotically distributed as χ^2:

$$\chi^2 = -ln|\mathbf{R}| * (N - 1 - (2p + 5)/6) \tag{4.16}$$

with degrees of freedom, $df = p * (p - 1)/2$. The determinant of an identity matrix is one, and as the correlations differ from zero, the determinant will tend towards zero. Thus Bartlett's test is a function of how much the determinant is less than one. This may be found by the `cortest.bartlett` function.

Given that the standard error of a z transformed correlation is $1/\sqrt{n-3}$, and that a squared z scores is χ^2, a very reasonable alternative is to consider whether the sum of the squared correlations differs from zero. When multiplying this sum by n-3, this is distributed as χ^2 with p*(p-1)/2 degrees of freedom. This is a direct test of whether the correlations differ from zero Steiger (1980c). This test is available as the `cortest` function.

4.4.3 Testing the difference between correlations

There are four different tests of correlations and the *differences between correlations* that are typically done: 1) is a particular correlation different from zero, 2) does a set of of correlations differ from zero, 3) do two correlations (taken from different samples) differ from each other, and 4) do two correlations taken from the same sample differ from each other. The first question, does a correlation differ from zero was addressed by Fisher (1921) and answered using a *t-test* of the observed value versus 0 (Equation 4.13) and looking up the probability of observing that size t or larger with degrees of freedom of n-2 with a call to `pt` (see Table 4.4 for an example). The second is answered by applying a χ^2 test (equation 4.16) using either the `cortest.bartlett` or `cortest` functions. The last two of these questions are more complicated and have two different sub questions associated with them.

Olkin and Finn (1990, 1995) and Steiger (1980a) provide very complete discussion and examples of tests of the *differences between correlations* as well as the confidence intervals for correlations and their differences. Three of the Steiger (1980a) tests are implemented in the `r.test` function in the **psych** package. Olkin and Finn (1995) emphasize confidence intervals for differences between correlations and address the problem of what variable to choose when adding to a multiple regression.

4.4.3.1 Testing *independent correlations*: r_{12} is different from r_{34}

To test two whether two correlations are different involves a z-test and depends upon whether the correlations are from different samples or from the same sample (the dependent or correlated case). In the first case, where the correlations are independent, the correlations are transformed to zs and the test is just the ratio of the differences (in z units) compared to the standard error of a difference. The standard error is merely the square root of the sum of the squared standard errors of the two individual correlations:

$$z_{r_{12}-r_{34}} = \frac{z_{r_{12}} - z_{r_{34}}}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}} \tag{4.17}$$

which is the same as

$$z_{r_{12}-r_{34}} = \frac{1/2\log(\frac{1+r_{12}}{1-r_{12}}) - 1/2\log(\frac{1+r_{34}}{1-r_{34}})}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}}.$$

This seems more complicated than it really is and can be done using the `paired.r` or `r.test` functions (Table 4.5).

Table 4.5 Testing two *independent correlations* using Equation 4.17 and `r.test` results in a z-test.

```
> r.test(r12=.25,r34=.5,n=100)
$test
[1] "test of difference between two independent correlations"
$z
[1] 2.046730
$p
[1] 0.04068458
```

4.4.3.2 Testing *dependent correlations*: r_{12} is different from r_{23}

A more typical test would be to examine whether two variables differ in their correlation with a third variable. Thus with the variables X_1, X_2, and X_3 with correlations r_{12}, r_{13} and r_{23} the t of the difference of r_{12} minus r_{13} is

$$t_{r_{12}-r_{13}} = (r_{12} - r_{13}) * \sqrt{\frac{(n-1)*(1+r_{23})}{2\frac{(n-1)}{n-3}|R| + \left(\frac{r_{12}+r_{13}}{2}\right)^2 (1-r_{23})^3}} \tag{4.18}$$

where $|R|$ is the *determinant* of the 3 *3 matrix of correlations and is

$$|R| = 1 - r_{12}^2 - r_{13}^2 - r_{12}^2 + 2 * r_{12} r_{13} r_{23} \tag{4.19}$$

(Steiger, 1980a). Consider the case of Extraversion, Positive Affect, and Energetic Arousal with PA and EA assessed at two time points with 203 participants (Table 4.6). The t for the difference between the correlations of Extraversion and Positive Affect at time 1 (.6) and Extraversion and Energetic Arousal at time 1 (.5) is found from equation 4.18 using the `paired.r` function and is 1.96 with a p value $< .05$. Steiger (1980a) and Dunn and Clark (1971) argue that Equation 4.18, *Williams' Test* (Williams, 1959), is preferred to an alternative test for dependent correlations, the *Hotelling T*, which although frequently recommended, should not be used.

4.4.3.3 Testing *dependent correlations*: r_{12} is different from r_{34}

Yet one more case is the test of equality of two correlations both taken from the same sample but for different variables (Steiger, 1980a). An example of this would be whether the correlations for Positive Affect and Energetic Arousal at times 1 and 2 are the same. For four variables ($X_1 ... X_4$) with correlations $r_{12} ... r_{34}$, the z of the difference of r_{12} minus r_{34} is

$$z_{r_{12}-r_{34}} = \frac{(z_{12} - z_{34})\sqrt{n-3}}{\sqrt{2(1-r_{12,34})}} \tag{4.20}$$

Table 4.6 The difference between two dependent correlations, $r_{ext,PA1}$ and $r_{ext,EA1}$ is found using Equation 4.18 which is implement in the `paired.r` and `r.test` functions. Because these two correlations share a common element (Extraversion), the appropriate test is found in Equation 4.18.

	Ext 1	PA 1	EA 1	PA 2	EA 2
Extraversion	1				
Positive Affect 1	.6	1			
Energetic Arousal 1	.5	.6	1		
Positive Affect 2	.4	.8	.6	1	
Energetic Arousal 2	.3	.6	.8	.5	1

```
>  r.test(r12=.6,r13=.5,r23=.6,n=203)
Correlation tests
Call:r.test(n = 203, r12 = 0.6, r23 = 0.6, r13 = 0.5)
Test of difference between two correlated  correlations
 t value 2    with probability < 0.047
```

where

$$r_{12,34} = 1/2([(r_{13} - r_{12}r_{23})(r_{24} - r_{23}r_{34})] + [(r_{14} - r_{13}r_{34})(r_{23} - r_{12}r_{13})] \\ +[(r_{13} - r_{14}r_{34})(r_{24} - r_{12}r_{14})] + [(r_{14} - r_{12}r_{24})(r_{23} - r_{24}r_{34})])$$

reflects that the correlations themselves are correlated. Under the null hypothesis of equivalence, we can also assume that the correlations $r_{12} = r_{34}$ (Steiger, 1980a) and thus both of these values can be replaced by their average

$$\bar{r_{12}} = \bar{r_{34}} = \frac{r_{12} + r_{34}}{2}.$$

Calling `r.test` with the relevant correlations strung out as a vector shows that indeed, the two correlations (.6 and .5) do differ reliably with a probability of .04 (Table 4.7).

Table 4.7 Testing the difference between two correlations from the same sample but that do not overlap in the variables included. Because the correlations do not involve the same elements, but do involve the same subjects, the appropriate test is Equation 4.20.

```
> r.test(r12=.6,r34=.5,r13=.8,r14=.6,r23=.6,r24=.8,n=203)


Correlation tests
Call:r.test(n = 203, r12 = 0.6, r34 = 0.5, r23 = 0.6, r13 = 0.8, r14 = 0.6,
    r24 = 0.8)
Test of difference between two dependent correlations
 z value 2.05    with probability  0.04
```

4.4.3.4 Testing whether correlation matrices differ across groups

The previous tests were comparisons of single correlations. It is also possible to test whether the observed p * p correlation matrix, $\mathbf{R_1}$ for one group of subjects differs from $\mathbf{R_2}$ in a different group of subjects. This has been addressed by several tests, perhaps the easiest to understand is by Steiger (1980b). Given the null hypothesis of no difference, the sum of squared differences of the z transformed correlation should be distributed as χ^2 with p * (p-1) degrees of freedom. A somewhat more complicated derivation by Jennrich (1970) also leads to a χ^2 estimate:

$$\chi^2 = \frac{1}{2} tr(\mathbf{Z}^2) - diag(\mathbf{Z})\mathbf{S}^{-1} diag(\mathbf{Z}) \tag{4.21}$$

where $\mathbf{R} = (\mathbf{R_1} + \mathbf{R_2})/2$, and the elements of $\mathbf{S}$ are the squared elements of $\mathbf{R}$, c = n1 * n2 /(n1+n2), and $\mathbf{Z} = c^{.5}\mathbf{R}^{-1}(\mathbf{R_1} - \mathbf{R_2})$. Both of these tests are available in the **psych** package: the `normal.cortest` finds the sum of squared differences of either the raw or z transformed correlations, the `jennrich.cortest` finds χ^2 as estimated in equation 4.21. Monte Carlo simulations of these and an additional test `mat.cortest` suggest that all three, and in particular the z-transformed and Jennrich tests are very sensitive to differences between groups (Revelle and Wilt, 2008).

4.5 Other estimates of association

The *Pearson Product Moment Correlation Coefficient* (*PPMCC*) was developed for the case of two continuous variables with interval properties, which, with the assumption of bivariate normality, can lead to estimates of confidence intervals and statistical significance (see `cor.test`). As pointed out by Charles Spearman (1904b), the Pearson correlation may be most easily thought of as

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \tag{4.22}$$

Dividing the numerator and the two elements in the square root by either n or n-1, this is, of course, equivalent to Equation 4.8 for the Pearson Product Moment Correlation Coefficient. A calculating formula that is sometimes used when doing hand calculations (for those who are stuck without a working copy of R) and that is useful when finding PPMCC for special cases (see below) uses raw scores rather than deviation scores:

$$\frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{(n\sum X_i^2 - (\sum X_i)^2)(n\sum Y_i^2 - (\sum Y_i)^2)}}. \tag{4.23}$$

Generalizing this to formula to $\mathbf{R_x}$, the matrix of correlations between the columns of a matrix $\mathbf{x}$ where $\mathbf{x}$ has been zero centered, let

$$\mathbf{I_{sd}} = diag(\frac{1}{\sqrt{diag(\mathbf{x})}})$$

that is, where $\mathbf{I_{sd}}$ is a diagonal matrix of the reciprocals of the standard deviations of the columns of $\mathbf{x}$, then

$$\mathbf{R}_x = \mathbf{I_{sd}}\mathbf{x}\mathbf{x}'\mathbf{I_{sd}} \tag{4.24}$$

is the matrix of correlations between the columns of **x**.

There are a number of alternative measures of association, some of which appear very different but are merely the PPMCC for special cases, while there are other measures for cases where the data are clearly neither continuous nor at the interval level of measurement. Even more coefficients of association are used as estimates of effect sizes.

4.5.1 Pearson correlation equivalents

Using Spearman's formula for the correlation (Equation 4.22) allows a simple categorization of a variety of correlation coefficients that at first appear different but are functionally equivalent (Table 4.8).

Table 4.8 A number of correlations are Pearson r in different forms, or with particular assumptions. If $r = \frac{\Sigma x_i y_i}{\sqrt{\Sigma x_i^2 \Sigma y_i^2}}$, then depending upon the type of data being analyzed, a variety of correlations are found.

Coefficient	symbol	X	Y	Assumptions
Pearson	r	continuous	continuous	
Spearman	rho (ρ)	ranks	ranks	
Point bi-serial	r_{pb}	dichotomous	continuous	
Phi	ϕ	dichotomous	dichotomous	
Bi serial	r_{bis}	dichotomous	continuous	normality of latent X
Tetrachoric	r_{tet}	dichotomous	dichotomous	bivariate normality of latent X, Y
polychoric	r_{pc}	categorical	categorical	bivariate normality of latent X, Y
polyserial	r_{ps}	categorical	continuous	bivariate normality of latent X, Y

4.5.1.1 Spearman ρ: a Pearson correlation of ranks

In the first of two major papers published in the *American Journal of Psychology* in 1904, Spearman (1904b) reviewed for psychologists the efforts made to define the correlation coefficient by Galton (1888) and Pearson (1895). Not only did he consider the application of the Pearson correlation to ranked data, but he also developed corrections for attenuation and the partial correlation, two subjects that will be addressed later. The advantage of using ranked data rather than the raw data is that it is more robust to variations in the extreme scores. For whether a person has an 8,000 or a 6,000 on an exam, that he or she is the highest score makes no difference to the ranks. Consider Y as ten numbers sampled from 1 to 20 and then find the Pearson correlation with Y^2 and e^Y. Do the same things for the ranks of these numbers. That is, find the Spearman correlations. As is clear from Figure 4.5, the Spearman correlation is not affected by the large non-linear transformation applied to the data Spearman (1907).

> It should be observed, that in many cases the non-linear form is more apparent than real. Generally speaking, a mere tendency of two characteristics to vary concurrently must be taken,

> it seems to me, as the effect of some particular underlying strict law (or laws) partly neutralized by a multitude of 'casual' disturbing influences. The quantity of a correlation is neither more nor less than the relative influence of the underlying law in question as compared with the total of all the influences in play. Now, it may easily happen, that the underlying law is one of simple proportionality but the disturbing influences become greater when the correlated characteristics are larger (or smaller, as the case may be). Then the underlying simple proportionality will not appear on the surface; the correlation will seem non-linear. Under such circumstances, r cannot, it is true, express these variations in the quantity of correlation; it continues, however, to express completely the mean quantity of correlation.
>
> In the majority of the remaining cases of non-linearity, the latter is merely due to a wrong choice of the correlated terms. For instance, the correlation between the length of the skull and the weight of the brain must, obviously, be very far from linear. But linearity is at once restored (supposing all the skulls to belong to one type) if we change the second term from the brain's weight to the cube root of the weight.
>
> To conclude, even when the underlying law itself really has a special non-linear form, although r by itself reveals nothing of this form, it nevertheless still gives (except in a few extreme and readily noticeable cases) a fairly approximate measure of the correlation's quantity. Spearman, 1907 p 168-169

Although a somewhat different formula is reported in Spearman (1904b), the calculating formula is

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} \tag{4.25}$$

where d is the difference in ranks. Holmes (2001) presents a very understandable graphical proof of this derivation. Alternatively, just find the Pearson correlation of the ranked data.

4.5.1.2 Point biserial: A Pearson correlation of a continuous variable with a dichotomous variable

If one of two variables, X, is dichotomous, and the other, Y, is continuous, it is still possible to find a Pearson r, but this can also be done by using a short cut formula. An example of this problem would be to ask the correlation between gender and height. Done this way, the correlation is known as the *point biserial correlation* but it is in fact, just a Pearson r.

If we code one of the two genders as 0 and the other as one, then Equation 4.23 becomes

$$r_{pb} = \frac{npq(\bar{Y}_2 - \bar{Y}_1)}{\sqrt{npq(n-1)\sigma_y^2}} = \frac{\bar{Y}_2 - \bar{Y}_1}{\sigma_y}\sqrt{\frac{npq}{(n-1)}} \tag{4.26}$$

where n is the sample size, p and q are the probabilities of being in group 1 or group 2, $\bar{Y}_1$ and $\bar{Y}_2$ are the mean of the two groups and σ_y^2 is the variance of the continuous variable. That is, the point biserial correlation is a direct function of the difference between the means and the relative frequencies of two groups. For a fixed sample size and difference between the group means, it will be maximized when the two groups are of equal size.

Thinking about correlations as reflecting the differences of means compared to the standard deviation of the dependent variable suggests a comparison to the t-test. And in fact, the point biserial is related to the t-test, for with $df = n-2$,

Table 4.9 The point biserial correlation is a Pearson r between a continuous variable (height) with a dichotomous variable (gender). It is equivalent to a t-test with pooled error variance.

```
> set.seed(42)
> n <- 12  #sample size
> gender <- sample(2,n,TRUE)  #create a random vector with two values
> height <- sample(10,n,TRUE)+ 58 + gender*3 #create a second vector with up to 10 values
> g.h  <- data.frame(gender,height) #put into a data frame
> g.h[order(g.h[,1]),]  #show the data frame
> cor(g.h)   #the Pearson correlation between height and gender
> t.test(height~gender,data=g.h,var.equal=TRUE)
> r <- cor(g.h)[1,2]  #get the value of the correlation
> r * sqrt((n-2)/(1-r^2)) #find the t- equivalent of the correlation, compare to the t-test.
```

```
   gender height
2       1     63
3       1     67
4       1     67
6       1     63
9       1     66
11      1     66
12      1     67
1       2     66
5       2     70
7       2     71
8       2     66
10      2     67

        gender  height
gender 1.00000 0.54035
height 0.54035 1.00000

        Two Sample t-test

data:  height by gender
t = -2.0307, df = 10, p-value = 0.06972
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.0932281  0.2360852
sample estimates:
mean in group 1 mean in group 2
       65.57143        68.00000

# t calculated from point biserial
[1] 2.030729
```

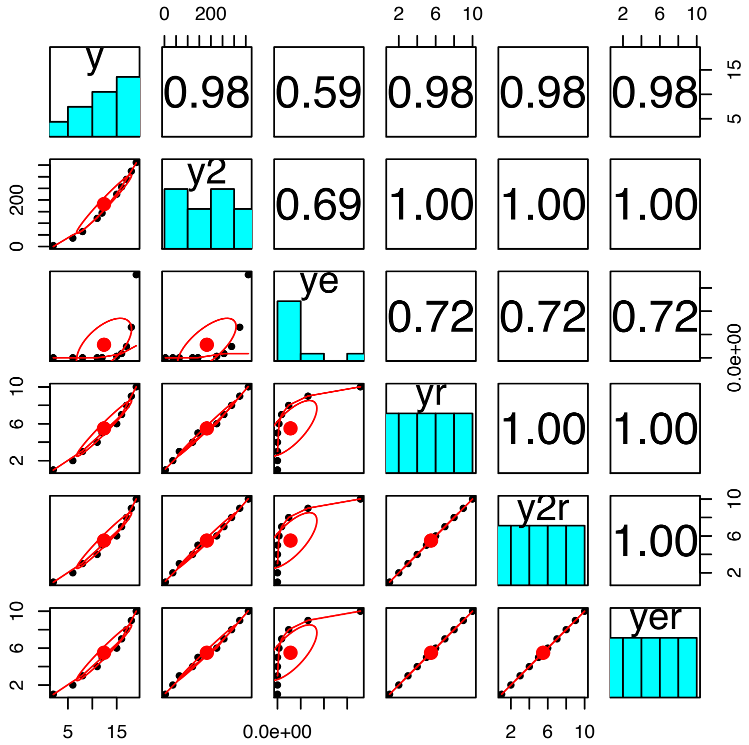


Fig. 4.5 Spearman correlations are merely Pearson correlations applied to ranked data. Here y is randomly sampled from the interval 1-20. y2 is Y^2, ye is e^Y, yr, y2r and yer are y, y2 and ye expressed as ranks. The correlations are found as Pearson r, but those between the second three variables are equivalent to Spearman ρ.

$$t_{df} = \frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\frac{\sigma_{y.x}}{n_1} + \frac{\sigma_{y.x}}{n_2}}} = \frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\frac{(n_1+n_2)\sigma^2_{y.x}}{n_1 n_2}}} = \frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\frac{\sigma^2_{y.x}}{npq}}} = \frac{\bar{Y}_2 - \bar{Y}_1}{\sigma_{y.x}}\sqrt{npq} \tag{4.27}$$

Comparing Equations 4.27 and 4.26 and recognizing that the within cell variance in the t-test is the residual variance in y after x is removed $\sigma^2_{y.x} = \sigma^2_y(1-r^2)$

$$t = r_{pb}\sqrt{\frac{(n-2)}{(1-r^2_{pb})}} = r_{pb}\sqrt{\frac{df}{(1-r^2_{pb})}} \tag{4.28}$$

Although the *t* and *point biserial correlation* are transforms of each other, it is incorrect to artificially dichotomize a continuous variable to express the relationship as a t value. If X and Y are both continuous, the appropriate measure of relationship is the Pearson correlation. By artificially dichotomizing one variable in order to express the effect as a t rather than a r, the strength of the relationship is reduced. Compare the four panels of Figure 4.6. The underlying scatter plot is shown for four values of a Pearson r (.9, .6, .3, and .0). Forming groups by setting values of Y < 0 to 0 and values greater than or equal to 0 to 1, results in the frequency distributions shown at the bottom of each panel. The corresponding point biserial correlations are reduced by 20%. That is, the point biserial for equal sized groups is .8 of the original Pearson r. In terms of power to detect a relationship, this is equivalent of throwing away 36% of the observations.

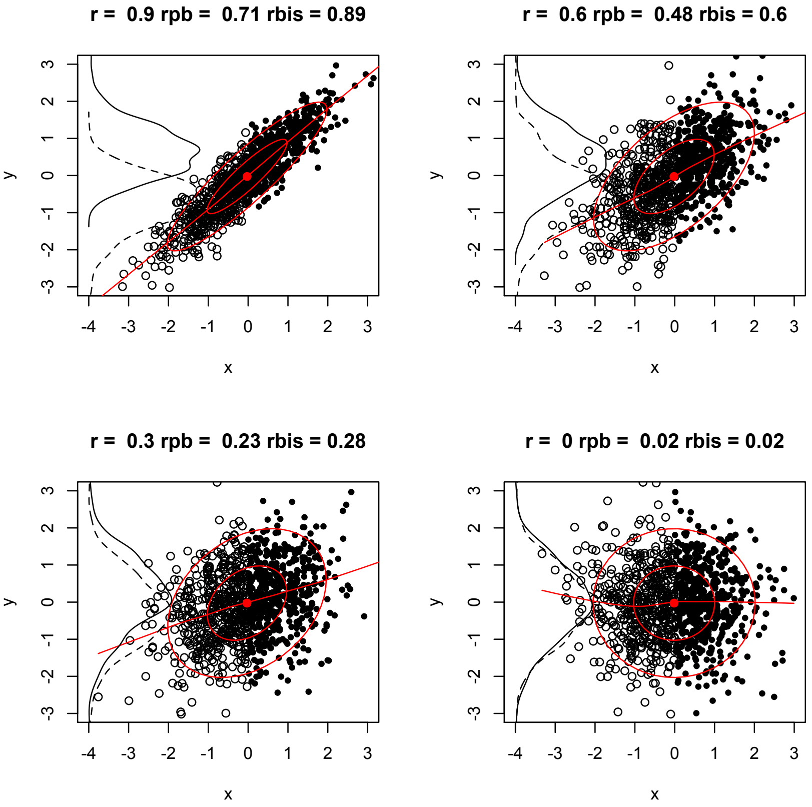


Fig. 4.6 The *point biserial correlation* is a Pearson r between a continuous variable and a dichotomous variable. If both X and Y are continuous and X is artificially dichotomized, the point biserial will be less than the original Pearson correlation. The *biserial correlation* is based upon the assumption of underlying normality for the dichotomized variable. It more closely approximates the "real" correlation. The estimated density curves for y are drawn for the groups formed from the dichotomized values of x.

4.5.1.3 Phi: A Pearson correlation of dichotomous data

In the case where both X and Y are naturally dichotomous, another short cut for the Pearson correlation is the phi (ϕ) coefficient. A typical example might be the success of predicting applicants to a graduate school. Two actions are taken, accept or reject, and two outcomes are observed, success or failure. This leads to the two by two table (Table 4.10) In terms of the raw data coded 0 or 1, the *phi coefficient* can be derived directly from Equation 4.23 by direct substitution, recognizing that the only non zero product is found in the A cell

Table 4.10 The basic table for a phi, ϕ coefficient, expressed in raw frequencies in a four fold table is taken from Pearson and Heron (1913)

	Success	Failure	Total
Accept	A	B	R_1 = A + B
Reject	C	D	R_2 = C + D
Total	C_1 = A + C	C_2=B + D	n = A + B + C + D

$$n\sum X_iY_i - \sum X_i \sum Y_i = nA - R_1C_1$$

$$\phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}. \tag{4.29}$$

Table 4.10 may be converted from frequency counts to proportions by dividing all entries by the total number of observations (A+B+C+D) to produce a more useful table (Table 4.11). In this table, the total counts (A, B, C, D) are expressed as their proportions (a, b, c, d) and the fraction of applicants accepted $\frac{R_1}{n} = \frac{A+B}{A+B+C+D}$ may be called the Selection Ratio, the fraction rejected is thus 1-SR. Similarly, the fraction of students who would have succeeded if accepted is $\frac{C_1}{n} = \frac{A+C}{A+B+C+D}$ may be called the Hit Rate, and the proportion who would fail is 1-HR. If being accepted or succeeding is given a score of 1, and rejected or failing, a score of 0, then the PPMCC of Table 4.11 may be found from Equation 4.23 as

Table 4.11 The basic table for a phi coefficient expressed in proportions

	Success	Failure	Total
Accept	Valid Positive	False Positive	R
Reject	False Negative	Valid Negative	$1-R$
Total	C	$1-C$	

$$\phi = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} = \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{\sqrt{(n\sum X_i^2 - (\sum X_i)^2)(n\sum Y_i^2 - (\sum Y_i)^2)}} = \frac{VP - RC}{\sqrt{R(1-R)C(1-C)}}. \tag{4.30}$$

The numerator is the number of valid positives (cell a or the percent of *valid positives*) less the number expected if there were no relationship ($R * C$). The denominator is the square root of the product of the row and column variances. As can be seen by equation 4.30, for fixed marginals the correlation is linear with the percentage of valid positives.

The ϕ *coefficient* is a PPMCC between two dichotomous variables. It is not, however, the equivalent of the PPMCC of continuous data that have been being artificially dichotomized. In addition, where the cuts are made greatly affects the correlation. Consider the case of two normally distributed variables (X and Y) that are correlated .6 in the population. If these variables are dichotomized at -1, 0, or 1 standard deviation from the mean, the correlations between them are attenuated, most so for the case of one variable being cut at lower value and the other being cut at the higher value. More importantly, the correlation of two dichotomized variables formed from the same underlying continuous variable is also seriously attenuated (Figure 4.7). That is, the correlations between the four measures of X (X, Xlow, Xmid, and Xhigh), although based upon exactly the same underlying numbers range from .19 to .43. Indeed, the *maximum value of phi* or *phi max* (ϕ_{max}) is a function of the marginal distributions and is

$$\phi_{max} = \sqrt{\frac{p_x q_y}{p_y q_x}} \tag{4.31}$$

where $p_x + q_x = p_y + q_y = 1$ and p_x, p_y represent the proportion of subjects "passing" an item.

The point bi-serial correlation is also affected by the distribution of the dichotomous variable. The first two rows in Figure 4.7 show how a continuous variable correlates between .65 to .79 with a dichotomous variable based upon that continuous variable.

4.5.1.4 Tetrachoric and polychoric correlations

If a two by two table is thought to represent an artificial dichotomization of two continuous variables with a bivariate normal distribution, then it is possible to estimate that correlation using the *tetrachoric correlation* (Pearson, 1900; Carroll, 1961). A generalization of the tetrachoric to more than two levels is the *polychoric correlation.* The `tetrachoric` function may be used to find the tetrachoric correlation as can the `polychor` function in the *polycor* package which also will find *polychoric correlations*.

Perhaps the major application of the tetrachoric correlation is when doing item analysis when each item is assumed to represent an underlying ability which is reflected as a probability of responding correctly to the item and the items are coded as correct or incorrect. In this case (discussed in more detail when considering *Item Response Theory* in Chapter 8), the difficulty of an item may be expressed as a function of the item *threshold*, τ, or the cumulative normal equivalent of the percent passing the item. The *tetrachoric correlation* is then estimated by comparing the number in each of the four cells with that expected from a bivariate normal distribution cut at τ_x and τ_y (see Figure 4.8 which was drawn using `draw.tetra`).

Unfortunately, for extreme differences in marginals, estimates of the tetrachoric do not provide a very precise estimate of the underlying correlation. Consider the data and ϕ correlations from Figure 4.7. Although the polychoric correlation does a very good job of estimating the correct value of the underlying correlation between X and Y (.60) for different values of dichotomization, and correctly finds a very high value for the correlation between the various sets of Xs and Ys (1.0), in some cases, if there are zero entries in one cell, the estimate is seriously wrong. One solution to this problem is to apply a *correction for continuity* which notes that a 0 case represents some where between 0 and .5 cases. `tetrachoric` automatically applies this correction but warns when this happens. In Table 4.12, this correction was

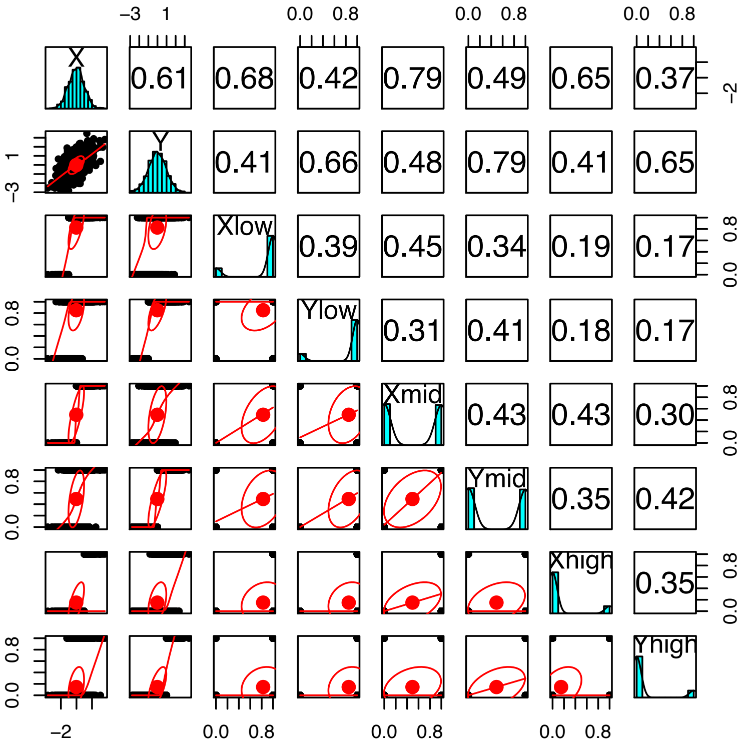


Fig. 4.7 The ϕ coefficient is a Pearson correlation applied to dichotomous data. It should not be used as a short cut for the relationship between two continuous variables. ϕ is sensitive to the marginal frequencies of the two variables. Shown is the SPLOM of two continuous variables, X and Y, and dichotomous variables formed from cutting X and Y at -1, 0 or 1 standard devations from the mean. Note how the ϕ coefficients underestimate the underlying correlation, particularly if the marginals differ. The first two rows of the correlations are point bi-serial correlations between continuous X and Y and dichotomized scores.

not applied. Redoing this analysis with the correction will yield somewhat different results. Günther and Höfler (2006) give an example from a comorbidity study where applying or not applying the correction makes a very large difference. Examples of the effect of the continuity correction are in the help for `tetrachoric`.

This problem of differences in endorsement frequency (differences in marginals) will addressed again when considering *factor analysis of items* (6.6) where the results will be much clearer when using tetrachoric correlations.

Table 4.12 The effect of various cut points upon *polychoric* and *phi* correlations. The original data matrix is created for X and Y using the `rmvnorm` function with a specified mean and covariance structure. Three dichotomous versions of X and Y are formed by cutting the data at -1, 0, or 1 standard deviations from the mean. The tetrachoric correlations are found using the `tetrachoric` function. These are shown below the diagonal by using `lower.tri`. Similarly, by using `upper.tri` the entries above the diagonal are *phi* correlations which are, of course, just the standard Pearson correlation applied to dichotomous data. The empirical thresholds, τ, are close to the -1, 0, and 1 cutpoints. The data are also tabled to show the effect of extreme cuts. Compare these correlations with those shown in Figure 4.7.

```
> library(psych)
> library(mvtnorm)  #needed for rmvorm
> set.seed(17) #to reproduce the results
> cut <- c(-1,-1,0,0,1,1)
> D <- rmvnorm(n=1000, sigma=Sigma)  #create the data
> D3 <- cbind(D,D,D)
> d <- D3
> D3[t(t(d) > cut)] <- 1
> D3[t(t(d) <= cut)] <- 0
> xy <- D3[,c(1,3,5,2,4,6)]
> colnames(xy) <- c("Xlow","Xmid","Xhigh","Ylow","Ymid","Yhigh")
> #describe(xy)
> tet.mat <- tetrachor(xy,FALSE)  #don't correct for continuity
> phi.mat <- cor(xy)
> both.mat <- tet.mat$rho * lower.tri(tet.mat$rho,TRUE) + phi.mat * upper.tri(phi.mat) #combine them
> round(both.mat,2)
> round(tet.mat$tau,2)  #thresholds

      Xlow Xmid Xhigh Ylow Ymid Yhigh
Xlow  1.00 0.44  0.19 0.35 0.30  0.18
Xmid  0.99 1.00  0.43 0.34 0.41  0.36
Xhigh 0.97 0.99  1.00 0.19 0.32  0.40
Ylow  0.60 0.64  0.66 1.00 0.44  0.20
Ymid  0.57 0.60  0.61 0.99 1.00  0.44
Yhigh 0.60 0.70  0.65 0.97 0.99  1.00

 Xlow  Xmid Xhigh  Ylow  Ymid Yhigh
-0.99 -0.02  0.99 -0.97  0.00  0.97
> table(xy[,2],xy[,5]) #both middle range

      0   1
  0 350 144
  1 150 356
> table(xy[,3],xy[,4]) #x high, y low

      0   1
  0 164 676
  1   1 159
> table(xy[,1],xy[,3]) #both low range

      0   1
  0 162   0
  1 678 160
> table(xy[,1],xy[,6]) #x low, y high

      0   1
  0 160   2
  1 675 163
```

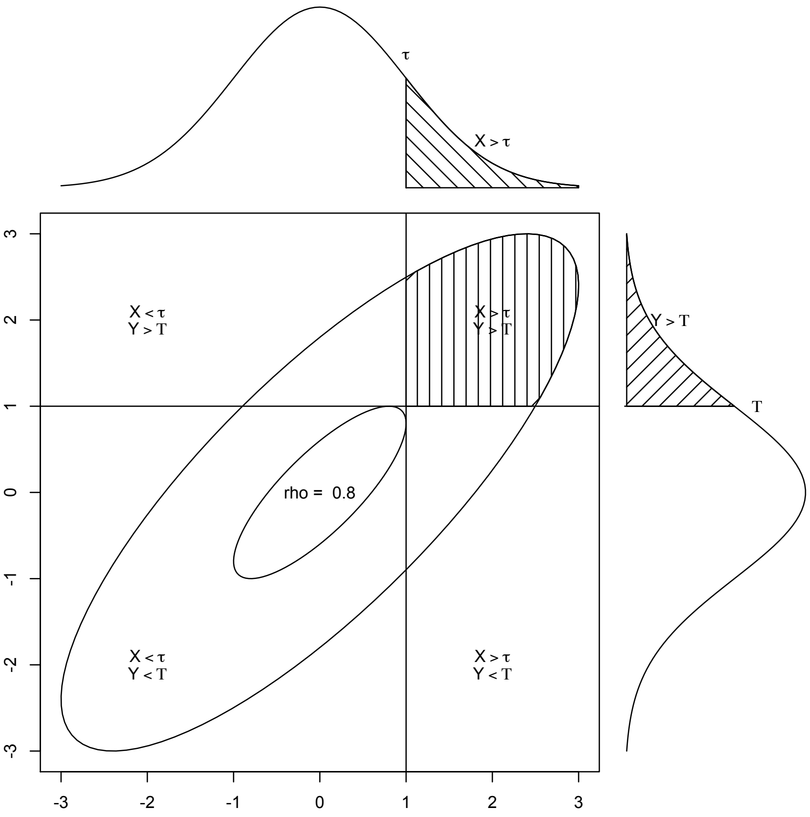


Fig. 4.8 The tetrachoric correlation is estimated from the marginal distributions of x and y and well as the joint frequency of x and y. The maximum likelihood estimate assumes bivariate normality.

4.5.1.5 Biserial and polyserial correlations: An estimated Pearson correlation of a continuous variable with a ordered categorical variable

While the *point biserial correlation* and the ϕ *coefficient* are equivalent to a Pearson r, the *biserial correlation* and *polyserial correlation* are not. The point biserial is just a short cut formula (Equation 4.26) for the Pearson r where one of the two variables, Y, is continuous and the other, X, is dichotomous. If, however, the dichotomous variable is *assumed* to be a dichotomy of a normally distributed variable divided at a particular cut point into two levels (0 and 1) with probabilities of q and p, then the *biserial correlation* (r_{bis}) is

$$r_{bis} = \frac{\bar{Y}_2 - \bar{Y}_1}{\sigma_y} \frac{pq}{z_p} \tag{4.32}$$

where z_p is the ordinate of the normal curve for the item *threshold*, τ, where τ is the cumulative normal equivalent of the probability, p. Thus,

$$r_{bis} = r_{pb}\frac{\sqrt{pq}}{z_p}. \tag{4.33}$$

The use of the biserial correlation is highly discouraged by some (e.g., Nunnally, 1967), and recommend to be used with extreme caution by others (Nunnally and Bernstein, 1984) but is probably appropriate in the case of modeling the underlying latent relationships between dichotomous items and continuous measures when the sample size is not too small. Biserial correlations may be found by using the `biserial` function. When applied to the data of Table 4.9, the biserial correlation is .70 compared to the observed Pearson (and thus point-biserial) correlation of .54. Examining Equation 4.33, it is clear that $r_{bis} > r_{pb}$ and that even in the best case (a 50-50 split), the point-biserial will be just 80% of the underlying correlation.

The biserial correlation is a special case of the *polyserial correlation*, r_{ps} which is the estimate of a Pearson correlation of two continuous variables when one is continuous and the other is an ordered categorical variable (say the four - six levels of a personality or mood item). For a continuous X and a ordered categorical Y, the simple "ad hoc" estimator of r_{ps} (Olsson et al., 1982) is a function of the observed point-polyserial correlation (which is just the Pearson r), the standard deviation of y, and the normal ordinates of the cumulative normal values of the probabilities of the alternatives:

$$r_{ps} = \frac{r_{xy}\sigma_y}{\sum z_{p_i}}. \tag{4.34}$$

This "ad hoc" estimator is simple to find and is a close approximation to that found by maximum likelihood. Just as with the biserial and point-biserial correlations, the polyserial correlation will be greater than the equivalent point-polyserial.

4.5.1.6 Correlation and comorbidity

In medicine and clinical psychology, diagnoses tend to be categorical (someone is depressed or not, someone has an anxiety disorder or not). Co-occurrence of both of these symptoms is called *comorbidity*. Diagnostic categories vary in their degree of comorbidity with other diagnostic categories. From the point of view of correlation, comorbidity is just a name applied to one cell in a four fold table. It is thus possible to analyze comorbidity rates by considering the probability of the separate diagnoses and the probability of the joint diagnosis. This gives the two by two table needed for a ϕ or r_{tet} correlation. Table 4.13 gives an example using the `comorbidity` function.

4.6 Other measures of association

Although most of psychometrics is concerned with combining and partitioning variances and covariances and the resulting correlations in the manner developed by Galton (1888), Pearson (1895) and Spearman (1904b), it is useful to consider other measures of association that are used in various applied settings. The first set of these are concerned with naturally occurring dichotomies while a second set has to do with measuring the association between categorical variables (e.g., diagnostic categories). A third set of correlations are those measuring asso-

Table 4.13 Given the base rates (proportions) of two diagnostic categories (e.g., .2 and .15) and their co-occurence (comorbidity, e.g. .1), it is straightforward to find the correlation between the two diagnoses. The tetrachoric coefficient is most appropriate for subsequent analysis.

```
> comorbidity(.2,.15,.1,c("Anxiety","Depression"))

Call: comorbidity(d1 = 0.2, d2 = 0.15, com = 0.1, labels = c("Anxiety",
    "Depression"))
Comorbidity table
            Anxiety -Anxiety
Depression      0.1     0.05
-Depression     0.1     0.75

implies phi =  0.49  with Yule =  0.87  and tetrachoric correlation of  0.75
```

ciations within classes of equivalent measures and uses an analysis of variance approach to find the appropriate coefficients.

4.6.1 Naturally dichotomous data

There are many variables, particularly those that reflect actions that are dichotomous (giving a vaccine, admitting to graduate school, diagnosing a disease). Similarly, there are many outcomes of these actions that are also dichotomous (surviving vs. dieing, finishing the Ph.D or not, having a disease or not). Although Pearson argued that the latent relationship was best descried as bivariate normal, and thus the appropriate statistic would be the tetrachoric correlation, Yule (1912) and others have examined measures of relationship that do not assume normality. Table 4.14, adapted from Yule (1912) provides the four cell entries that enter into multiple estimates of associations. Pearson and Heron (1913) responded to Yule (1912) and showed that even with extreme non-normality, the phi and tetrachoric correlations were superior to others that had been proposed.

Table 4.14 Two dichotomous variables produce four outcomes. Yule (1912) used the example of vaccinated and not vaccinated crossed with survived or dead. Similarly, colleges accept or reject applicants who either do or do not graduate.

	Action	non-action	total
Positive Outcome	a	b	a+b
Negative Outcome	c	d	c + d
total	a+c	b+d	a+b+c+d

Given such table, there are a number of measures of association that have been or are being used.

4.6.1.1 Odds ratios, risk ratios, and the problem of base rates

From a patient's point of view, it would seem informative to know the ratio of how many survived versus how many died given a vaccine (a/c). But this ratio is only meaningful in contrast to the ratio of how many survive who are not vaccinated (b/d). The *Odds Ratio* compares these two odds

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc} \tag{4.35}$$

Unfortunately, if the number of cases is small, and if either the b or c cells is empty, the OR is infinite. A standard solution in this case is to add .5 to all cells. For cell sizes of $n_a ... n_d$, the standard error of the logarithm of the odds ratio is (Fleiss, 1993)

$$SE(ln(OR)) = \sqrt{\frac{1}{n_a} + \frac{1}{n_b} + \frac{1}{n_c} + \frac{1}{n_d}}.$$

An alternative to the *Odds Ratio* is the *Risk Ratio*. For instance, what fraction of patients given a medication survive $\frac{a}{a+c}$, or if a test is given for diagnostic reasons, what percentage of patients with the disease (a+c) test positive for the disease (a). This is known as the *sensitivity* of the test. But it is also important to know what percentage of patients without the disease test negative (the *specficity* of the test). More informative is the *Relative Risk Ratio* (comparing the risk given the action to the risk not given the action). That is, what is the ratio of patients who survive given treatment to those who survive not given treatment.

$$RRR = \frac{a/(a+c)}{b/(b+d)} = \frac{sensitivity}{1 - specificity} = \frac{a(b+d)}{b(a+c)} \tag{4.36}$$

Odds ratios and relative risk ratios along with confidence intervals may be found using the epidemiological packages **Epi** and **epitools**. Just as the regressions of **Y** on **X** and **X** on **Y** yield different slopes, so does the odds ratio depend upon the direction of prediction. That is, the odds of a positive outcome given an action (a/c) is not the same as the odds of an action having happened given a positive outcome (a/b). This difference depending upon direction can lead to serious confusion, for many phenomena that seem highly related in one direction have only small odds ratios in the opposite direction. This is important to realize, and somewhat surprising, that the frequency of observing an event associated with a particular condition (having lung cancer and having been a smoker, having an auto accident and having been drinking, being pregnant and having had recent sexual intercourse) is much higher than the frequency in the reverse direction (the percentage of smokers who have lung cancer, the fraction of drivers who have been drinking who have accidents, the percentage of women who have recently had sexual intercourse who are pregnant). Consider the case of the relationship between sexual intercourse and pregnancy (Table 4.17). In this artificial example, the odds of becoming pregnant given intercourse (a/(a+c)) are .0019, while the odds of having had intercourse given that one is pregnant is 1.0. The odds ratio $\frac{ad}{bc}$ is undefined, although adding .5 to all cells to correct for zero values, yields an odds ratio of 12.51. The relative risk is also undefined unless .5 is added, in which case it becomes 12.49. The ϕ coefficient is .04 while the tetrachoric correlation (found using the `tetrachoric` function) is .938. But the latter makes the assumption of bivariate normality with extreme cut points. It is not at all clear that this is appropriate for the data of 4.17.

4.6.1.2 Yule's Q and Y

Yule (1900, 1912) developed two measures of association, one of which, Q (for Quetelet), may be seen as a transform of the Odds Ratio into a metric ranging from -1 to 1. *Yule's Q* statistic is

$$Q = \frac{ad - bc}{ad + bc} = \frac{ad/bc - 1}{ad/bc + 1} = \frac{OR - 1}{OR + 1}. \tag{4.37}$$

A related measure of association, the *coefficient of colligation*, called by Yule (1912) ω but also known as the Yule's Y is

$$\omega = Y = \frac{\sqrt{OR} - 1}{\sqrt{OR} + 1}.$$

Yule's coefficient has the advantage over the odds ratio that is defined even if one of the off diagonal elements (b or c) is zero in which case Q = 1. However, if b or c is 0, then no matter what the other cells are, Q will still be one. As a consequence, Yule's Q was not uniformly appreciated, and was strongly attacked by Pearson's student Heron (1911); MacKenzie (1978) as not consistent with ϕ or the tetrachoric correlation. In a very long article Pearson and Heron (1913) gives many examples of the problems with Yule's coefficient and why ϕ gives more meaningful results. In their paper, Pearson and Heron also consider the problem with ϕ, which is that it is limited by differences in the marginal frequencies. Both the Q and Y statistics are found in the `Yule` function.

4.6.1.3 Even more measures of association for dichotomous data

In psychometrics, the preferred measures of association for dichotomous data is either the *tetrachoric* correlation (based upon normal theory) or a Pearson correlation, which when applied to dichotomous data is the *phi* coefficient. However, in other fields a number of measures of similarity have been developed. Jackson et al. (1989) distinguishes between measures of co-occurence and measures of association. The *phi*, *tetrachoric*, and *Yule* coefficients are measures of association, while there are at least 8 measures of similarity. The oldest, Jaccard's *coefficient of community* was developed to measure the frequency with which two species co-occured across various regions of the Jura mountains and is just the number of co-occurences (a) divided by the number of species in one or the other or both districts (a + b + c).

When these measures are rescaled so that their maximum for perfect association is 1, and the index when there is no association is zero, most of these indices are equivalent to *Loevinger's H* statistic (Loevinger, 1948; Warrens, 2008).

complete

These different measures of similarity typically are used in fields where the clustering of objects (not variables) is important. That most of them are just transforms of *Loevinger's H* suggests that there is less need to consider each one separately (Warrens, 2008).

4.6.1.4 Base rates and inference

In addition to the problem of direction of inference is the problem of base rates. Even for tests with high *sensitivities* $(\frac{a}{a+c})$ and *specificities* $(\frac{d}{b+d})$, if the *base rates* are extreme, misdiagnosis is common. Consider the well known example of a very sensitive and specific test for HIV/AIDS. Let these two values be .99. That is, out of 100 people with HIV/AIDS,

Table 4.15 It is important to realize, and somewhat surprising, that the frequency of observing an event associated with a particular condition (e.g., lung cancer and smoking, auto accidents and drinking, pregnancy and sexual intercourse) are very different from the inverse (e.g. smoking and lung cancer, drinking and auto accidents, sexual intercourse and pregnancy). In this hypothetical example, a couple is assumed to have had sexual intercourse twice a week for ten years and to have had two children. From the two x two table, it is possible calculate the tetrachoric correlation using the `polychor` or `tetrachoric` functions, Yule's Q using `Yule`, the ϕ correlation using the `phi`, and Cohen's kappa using `cohen.kappa` function as well .

	Intercourse	No intercourse	total
Pregnant	2	0	2
Not pregnant	1038	2598	3638
total	1040	2600	3640

```
> pregnant = matrix(c(2,0,1038,2598),ncol=2)
> colnames(pregnant) <- c("sex","nosex")
> rownames(pregnant) <- c("yes","no")
> pregnant

   sex nosex
yes   2  1038
no    0  2598
> polychor(pregnant)
[1] 0.9387769
> Yule(pregnant)
[1] 1

> phi(pregnant)

[1] 0.0371

> wkappa(pregnant)

$kappa
[1] 0.002744388
```

Table 4.16 Alternative measures of co-occurence for binary data (adapted from Jackson et al. (1989); Warrens (2008)).

Coefficient	Index	Reference
Jaccard	$\frac{a}{a+b+c}$	Jaccard (1901)
Sorenson-Dice	$\frac{2a}{2a+b+c}$	??
Russell-Rao	$\frac{a}{a+b+c+d}$	?
Sokal	$\frac{a+b}{2a+b+c}$	?
Ochiai	$\frac{a}{\sqrt{(a+b)(a+c)}}$	?
etc.		

the test correctly diagnoses 99 of them. Of 100 people for whom the test returns a negative result, 1 has the disease. Assume that 1% of a sample of 10,000 people are truly infected with HIV/AIDS. What percentage of the sample will test positive? What is the likelihood of having the disease if the test returns positive? The answer is that roughly 2% of the sample tests positive and half of those are false positives.

Table 4.17 High specificity and sensitivity do not necessarily imply a low rate of false positives or negatives if the base rates are extreme. Even with specificities and sensitivies of 99%, 50% of those diagnosed positive are false positives.

	HIV/AIDS Yes	HIV/AIDS No	total
Test Positive	99	99	198
Test Negative	1	2598	9802
total	100	9,900	10,000

4.6.2 Measures of association for categorical data

Some categorical judgments are made using more than two outcomes. For example, two diagnosticians might be asked to categorize patients three ways (e.g., Personality disorder, Neurosis, Psychosis) or to rate the severity of a tumor (not present, present but benign, serious, very serious). Just as base rates affect observed cell frequencies in a two by two table, they need to be considered in the n way table Cohen (1960). Consider Table 4.18 which is adapted from Cohen (1968). Let **O** be the matrix of observed frequencies and $\mathbf{E} = \mathbf{RC}$ where **R** and **C** are the row and column (marginal) frequencies. *Kappa* corrects the proportion of matches, p_o, (the number of times the judges agree) with what would be expected by chance, p_e (the sum of the diagonal of the product of the row and column frequencies). Thus

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{tr(\mathbf{O} - \mathbf{E})}{1 - tr\mathbf{E}}. \tag{4.38}$$

Table 4.18 *Cohen's kappa* measures agreement for n-way tables for two judges. It compares the observed frequencies on the main diagonal with those expected given the row marginals. The expected scores (shown in parentheses) are the products of the marginals. $\kappa = \frac{(.44+.20+.06)-(.30+.09+.02)}{1-(.30+.09+.02)} = .49$. Adapted from Cohen (1968).

		Judge 1			
		Personality disorder	Neurosis	Psychosis	$p_{i.}$
Judge 2	Personality disorder	.44 (.30)	.07 (.18)	.09 (.12)	.60
	Neurosis	.05 (.15)	.20 (.09)	.05 (.06)	.30
	Psychosis	.01 (.05)	.03 (.03)	.06 (.02)	.10
	$p_{.j}$	.50	.30	.20	1.00

As discussed by Hubert (1977) and Zwick (1988), *kappa* is one of a family of statistics that correct observed agreement with expected agreement. If raters are assumed to be random samples from a pool of raters, then the marginal probabilities for each rater may be averaged and the expected values will be the squared marginals (Scott, 1955). *kappa* does not assume equal marginal frequencies and follows the χ^2 logic of finding expectancies based upon the product of the marginal frequencies. However, *kappa* considers all disagreements to be equally important and just considers entries on the main diagonal.

If some disagreements are more important than others, then the appropriate measure is *weighted kappa* (Cohen, 1968):

$$\kappa_w = \frac{wp_o - wp_e}{1 - wp_e} \tag{4.39}$$

where $wp_o = \sum w_{ij} p_{oij}$ and similarly $wp_e = \sum w_{ij} p_{eij}$. With the addition of a weighting function, w_{ij} that weights the diagonal 1 and the off diagonal with weights depending upon the inverse of the squared distance between the categories, the *weighted kappa coefficient* is equivalent to one form of the *intraclass correlation coefficient* (see 4.6.3) (Fleiss and Cohen, 1973). *Weighted Kappa* is particularly appropriate when the categories are ordinal and a near miss is less important than a big miss (i.e., having one judge give a medical severity rating of not present and the other judge rating the same case as very serious shows less agreement than one judge giving a very serious and the other a serious rating). Another use of weighted kappa is even if the categories are not ordinal, some mistakes are more important than others.

The variance of *kappa* or *weighted kappa* for large samples may be found using formulas in Fleiss et al. (1969). By using the resulting standard errors, it is possible to find confidence intervals for *kappa* and *weighted kappa* (Hubert, 1977; Fleiss et al., 1969). Calculations of kappa and weighted kappa are done in several packages: `Kappa` in **vcd** (a very nice package for Visualizing Categorical Data), `wkappa` in **psy**, and `cohen.kappa` in **psych**.

4.6.3 Intraclass Correlation

The Pearson correlation coefficient measures similarity of patterns of two distinct variables across people. The variables are two measures (say height and weight) on the same set of people, and the two variables are logically distinct. But sometimes it is desired to measure how similar pairs (or more) of people are on one variable. Consider the problem of similarity of pairs twins on a measure of ability (Table 4.19). For five pairs of twins, they may be assigned to be the first or second twin based upon observed score (Twin 1 and Twin 2), or as they are sampled (Twin 1* and Twin 2*). The correlation between the twins in the first

Table 4.19 Hypothetical twin data. The Twin 1 and Twin columns have been ordered by the value of the lower scoring twin, the Twin 1* and Twin 2* columns suggest what happens if the twins are randomly assigned to twin number.

Pair	Twin 1	Twin 2	Twin 1*	Twin 2*
1	80	90	80	90
2	90	100	100	90
3	100	110	110	100
4	110	120	110	120
5	120	130	130	120

two columns is 1, but between the second two sets of columns it is .80. That the twins in the first two columns are not perfectly similar is obvious, in that their scores systematically differ by 10 points. The normal correlation, by removing the means of the scores, does not detect this effect. One solution, sometimes seen in the early behavior genetics literature was to double enter each twin, that is, to have each twin appear once in the first column and once in the second column. This effectively pools the mean of the two sets of twins and finds the correlation with respect to deviations from this pooled mean. The value in this case is .77. Why do these three correlations differ and what is is the correct value of the similarity of the twins?

The answer comes from the *intraclass correlation coefficient*, or *ICC* and a consideration of the sources of variance going into the twin scores. Consider the traditional analysis of variance model:

$$x_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij}$$

where μ is the overall mean for all twins, a_i is the mean for the ith pair, b_j is the mean for the first or second column, ab_{ij} reflects the interaction of particular twin pair and being in column 1 or 2, and e_{ij} is residual error. In the case of twins, ab_{ij} and e_{ij} are indistinguishable and may be combined as w_{ij}. Then the total variance σ_t^2 may be decomposed

$$\sigma_t^2 = \sigma_i^2 + \sigma_j^2 + \sigma_w^2.$$

and the fraction of total variance (between + within pair variance) due to difference between the twin pairs is the intraclass correlation measure of similarity:

$$\rho = \frac{\sigma_i^2}{\sigma_t^2} = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_j^2 + \sigma_w^2}.$$

An equivalent problem is the problem of estimating the agreement in their ratings between two or more raters. Consider the ratings of six targets by four raters shown in Table 4.20. Although the raters have an average intercorrelation of .76, they differ drastically in their mean ratings and one (rater 4) has a much higher variance. As reviewed by Shrout and Fleiss (1979) there are at least six different intraclass correlations that are commonly used when considering the agreement between k different judges (raters) of n different targets:

1. Case 1: Targets are randomly assigned to different judges. (This would be equivalent to the twins case above).
2. Case 2: All targets are rated by the same set of randomly chosen judges.
3. Case 3: All targets are rated by the same set of fixed judges.
4. Case 1-b:The expected correlation of the average ratings across targets of the mean ratings of randomly assigned judges with another set of such measures.
5. Case 2-b: The expected correlation of the average ratings across targets from one set of randomly chosen judges with another set.
6. Case 3-b: The expected correlation of the average ratings across targets of fixed judges.

All six of these intraclass correlations may be estimated by standard analysis of variance implemented in the `ICC` function in **psych**. If the ratings are numerical rather than categorical, the ICC is to be preferred to κ or *weighted* κ which were discussed above (4.6.2).

Table 4.20 Example data from four raters for six targets demonstrate the use of the intraclass correlation coefficient. Adapted from Shrout and Fleiss (1979)

Subject	Rater 1	Rater 2	Rater 3	Rater 4
1	9	2	5	8
2	6	1	3	2
3	8	4	6	8
4	7	1	2	6
5	10	5	6	9
6	6	2	4	7

```
> round((sum(cor(SF79)) - 4)/12,2)

[1] 0.76

> describe(SF79)

   var n mean   sd median trimmed  mad min max range  skew kurtosis   se
V1   1 6 7.67 1.63    7.5    7.67 2.22   6  10     4  0.21    -1.86 0.67
V2   2 6 2.50 1.64    2.0    2.50 1.48   1   5     4  0.45    -1.76 0.67
V3   3 6 4.33 1.63    4.5    4.33 2.22   2   6     4 -0.21    -1.86 0.67
V4   4 6 6.67 2.50    7.5    6.67 1.48   2   9     7 -0.90    -0.83 1.02

> ICC(SF79)

     ICC1 ICC2 ICC3 ICC12 ICC22 ICC32
[1,] 0.17 0.29 0.71 0.44  0.62  0.91
```

4.6.4 Quantile Regression

Originally introduced by Galton (1889), regression of deviations from the median in terms of quantile units has been rediscovered in the past decade Gilchrist (2005). The package **quantreg** by Koenker (2007) implements these procedures.

4.6.5 Kendall's Tau

τ is a rank order correlation based on the number of concordant (same rank order) and disconcordant (different rank order) pairs (Dalgaard, 2002). If there are no ties in the ranks for the x_i and y_i

$$\tau = \frac{\sum_{i<j} sign(x_j - x_i) * sign(y_j - y_i)}{n(n-1)/2}.$$

τ counts the number of pairs that have the same rank orders and compares this to the number of pairs. If two vectors, **x** and **y**, are monotonically the same, τ will be one. Kendall is an option in the `cor` function in base R and is also available as the `Kendall` function in the **Kendall** package.

4.6.6 Circular-circular and circular-linear correlations

As discussed earlier (3.4.1), when data represent angles (such as the hours of peak alertness or peak tension during the day), we need to apply *circular statistics* rather than the more normal linear statistics (see Jammalamadaka and Lund (2006) for a very clear set of examples of circular statistics). The generalization of the Pearson correlation to circular statistics is straight forward and is implemented in `cor.circular` in the **circular** package and in `circadian.cor` in the **psych** package. Just as the Pearson r is a ratio of covariance to the square root of the product of two variances, so is the *circular correlation*. The *circular covariance* of two circular vectors is defined as the average product of the sines of the deviations from the *circular mean*. The variance is thus the average squared sine of the angular deviations from the circular mean.

Consider the data shown in Table 3.8. Although the Pearson r of these variables range from -.78 to .06, the circular correlations among all of them are exactly 1.0. (The separate columns are just phase shifted by 5 hours and thus the deviations from the circular means are identical.)

In addition to the *circular-circular correlation*, there is also the correlation between a circular variable and a linear variable (the *circular-linear correlation*). The *circular-linear covariance* is the product of the sine of the angular deviation from the *circular mean* times the deviation of the linear variable from its mean. It may be found by the `cor.circular` or the `circadian.linear.cor` functions. In the example in Table 4.21, the circular variable of hour of peak mood for Tense Arousal has a perfect positive circular-linear correlation with the linear variable, Extraversion, and a slight positive correlation with Neuroticism. By comparison, the traditional, Pearson correlations for these variables were -.78 and -.18.

4.7 Alternative estimates of effect size

There are a number of ways to summarize the importance of a relationship. The slope of the linear regression, $b_{y.x}$is a direct measure of how much one variable changes as a function of changes in the other variable. The regression is in the units of the measure. Thus, Galton could say that the height of children increased by .65 inches (or centimeters) for every increase in 1 inch (or centimeter) of the mid parent. As a measure of effect with meaningful units, the slope is probably the most interpretable.

But, for much of psychological data, the units are arbitrary, and discussing scores in terms of deviations from the mean with respect to the standard deviation (i.e., standard scores) is more appropriate. In this case, the correlation is the preferred unit of effect. Using correlations, Galton would have said that the relationship between mid parent and child height was .46. Experimentalists tend to think of the effects in terms of differences between the means of two groups, the two standard estimates of *effect size* of group differences are *Cohen's d* (Cohen, 1988) and *Hedges' g* (Hedges and Olkin, 1985), both of which compare the mean difference to estimates of the within cell standard deviation. Cohen's d uses the population estimate, Hedge's g the sample estimate. Useful reviews of the use of these and other ways of estimating *effect sizes* for *meta-analysis* include Rosnow et al. (2000) and the special issue of Psychological Methods devoted to effect sizes Becker (2003). Summaries of these various formulae are in Table 4.22.

Table 4.21 The Pearson correlation for circular data misrepresents the relationships between data that have a circular order (such as time of day, week, or year). The *circular correlation* considers the sine of deviations from the *circular mean*. The correlation between a linear variable (e.g., extraversion or neuroticism) with a circular variable is found using the *circular-linear correlation*.

```
> time.person  #the raw data (four circular variables, two linear variables)

 EA PA TA NegA extraversion neuroticism
1  9 14 19   24            1           3
2 11 16 21    2            2           6
3 13 18 23    4            3           1
4 15 20  1    6            4           4
5 17 22  3    8            5           5
6 19 24  5   10            6           2

> round(cor(time.person),2)  #the Pearson correlations

                EA    PA    TA  NegA extraversion neuroticism
EA            1.00  1.00 -0.78 -0.34         1.00       -0.14
PA            1.00  1.00 -0.78 -0.34         1.00       -0.14
TA           -0.78 -0.78  1.00  0.06        -0.78       -0.18
NegA         -0.34 -0.34  0.06  1.00        -0.34       -0.23
extraversion  1.00  1.00 -0.78 -0.34         1.00       -0.14
neuroticism  -0.14 -0.14 -0.18 -0.23        -0.14        1.00

> circadian.cor(time.person[1:4])  # the circular correlations
> round(circadian.linear.cor(time.person[1:4],time.person[5:6]),2)

     EA PA TA NegA
EA    1  1  1    1
PA    1  1  1    1
TA    1  1  1    1
NegA  1  1  1    1

     extraversion neuroticism
EA              1        0.18
PA              1        0.18
TA              1        0.18
NegA            1        0.18
```

4.8 Sources of confusion

The correlation coefficient, while an extremely useful measure of the relationship between two variables, can sometimes lead to improper conclusions. Several of these are discussed in more detail below. One of the most common problems is *restriction of range* of either one of the two variables. The use of sums, ratios, or differences can also lead to *spurious correlations* when none are truly present. Some investigators will *ipsatize* scores either intentionally or non-intentionally and discover that correlations of related constructs are seriously reduced. *Simpson's paradox* is a case of correlations between data measured at one level being reversed when pooling data across a grouping variable at a different level. The importance of a correlation for practical purposes is also frequently dismissed by reflexively squaring the correlation to understand the reduction in variance accounted for by the correlation. In practical decision making situations, the slope of the linear relationship between two variables is much more

Table 4.22 Alternative Estimates of effect size. Using the correlation as a scale free estimate of effect size allows for combining experimental and correlational data in a metric that is directly interpretable as the effect of a standardized unit change in x leads to r change in standardized y.

Regression	$b_{y.x} = \frac{Cxy}{\sigma_x^2}$	$b_{y.x}$	$b_{y.x} = r\frac{\sigma_x}{\sigma_y}$
Pearson correlation	$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}$	r_{xy}	
Cohen's d	$d = \frac{X_1 - X_2}{\sigma_x}$	$r = \frac{d}{\sqrt{d^2+4}}$	$d = \frac{2r}{\sqrt{1-r^2}}$
Hedge's g	$g = \frac{X_1 - X_2}{s_x}$	$r = \frac{g}{\sqrt{g^2+4(df/N)}}$	$g =$
t - test	$t = 2d\sqrt{df}$	$r = \sqrt{t^2/(t^2+df)}$	$t = \sqrt{\frac{r^2 df}{1-r^2}}$
F-test	$F = 4d^2 df$	$r = \sqrt{F/(F+df)}$	$F = \frac{r^2 df}{1-r^2}$
Chi Square		$r = \sqrt{\chi^2/n}$	$\chi^2 = r^2 n$
Odds ratio	$d = \frac{ln(OR)}{1.81}$	$r = \frac{ln(OR)}{1.81\sqrt{(ln(OR)/1.81)^2+4}}$	$ln(OR) = \frac{3.62r}{\sqrt{1-r^2}}$
$r_{equivalent}$	r with probability p	$r = r_{equivalent}$	

important than the squared correlation and it is more appropriate to consider the slope of the mean differences between groups (Lubinski and Humphreys, 1996). Finally, correlations can be seriously attenuated by differences in *skew* between different sets of variables.

4.8.1 Restriction of range

The correlation is a ratio of covariance to the square root of the product of two variances 4.8. As such, if the variance of the predictor is artificially constrained, the correlation will be reduced, even though the slope of the regression remains the same. Consider an example of 1,000 simulated students with GREV and GREQ scores with a population correlation of .6. If the sample is restricted in its variance (say only students with GREV > 600 are allowed to apply, the correlation drops by almost 1/2 from .61 to .34.(Table 4.23, Figure 4.9).

An even more serious problem occurs if the range is restricted based upon the sum of the two variables. This might be the case if an admissions committee based their decisions upon total GRE scores and then examined the correlation between their predictors. Consider the correlation within those applicants who had total scores of more than 1400. In this case, the correlation for those 11 hypothetical subjects has become -.34 even though the underlying correlation was .61! Similar problems will occur when choosing a high group based upon several measures of a related concept. Some researchers examine the relationship among measures of negative affecting with a group chosen to be extreme on the trait. That is, what is the correlation between measures of neuroticism, anxiety, and depression within a selected set of patients rather than the general population. Consider the data set `epi.bfi` which includes measures of Neuroticism using the *Eysenck Personality Inventory* (Eysenck and Eysenck, 1964), of Depression using the Beck Depression Inventory Beck et al. (1961) and Trait Anxiety using the State Trait Anxiety Inventory (Spielberger et al., 1970) for 231 undergraduates. For the total sample, these three measures have correlations of .53, .73 and .65, but if a broad trait of negative affectivity is defined as the sum of the three standardized scales, and an "at risk" group is defined as more than 1 s.d. on this composite is chosen, the correlations become -.08, -.11, and .17.

Table 4.23 Restricting the range of one variable will reduce the correlation, although not change the regression slope. Simulated data are generated using the `mvrnorm` function from the **MASS** package. To give the example some context, the variables may be taken to represent “GRE Verbal” and “GRE Quantitative”. The results are shown in Figure 4.9.

```
library(MASS)
set.seed(42)
GRE <- data.frame(mvrnorm(1000,c(500,500),matrix(c(10000,6000,6000,10000),ncol=2)))
colnames(GRE) <- c("GRE.V","GRE.Q")
op  <- par(mfrow = c(1,2))
plot(GRE,xlim=c(200,800),ylim=c(200,800),main="Unrestricted")
lmc <- lm(GRE.Q ~ GRE.V,data=GRE)
abline(lmc)
text(700,200,paste("r =",round(cor(GRE)[1,2],2)))
text(700,250,paste("b =",round(lmc$coefficients[2],2)))
GREs <- subset(GRE,GRE$GRE.V > 600)
plot(GREs,xlim=c(200,800),ylim=c(200,800),main="Range restricted")
lmc <- lm(GRE.Q ~ GRE.V,data=GREs)
abline(lmc)
text(700,200,paste("r =",round(cor(GREs)[1,2],2)))
text(700,250,paste("b =",round(lmc$coefficients[2],2)))
```

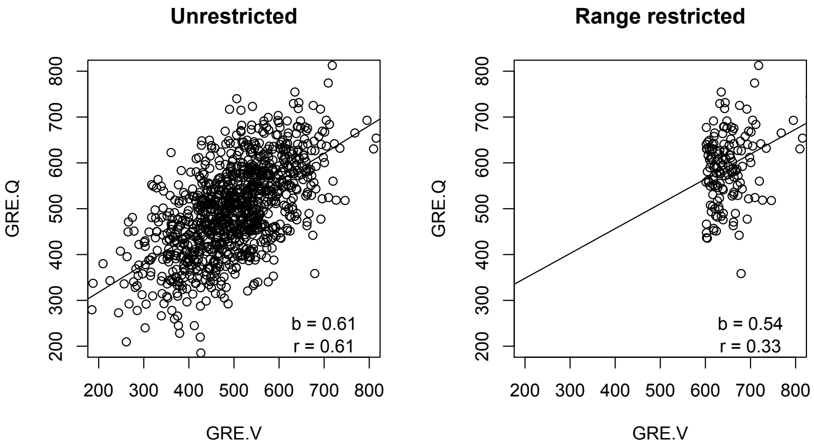


Fig. 4.9 The effect of restriction of range on correlation and regression. If the predictor (X) variable is restricted by selection, the regression slope does not change, but the correlation drops. Data generated using `mvrnorm` as shown in Table 4.23.

4.8.2 Spurious correlations

Although viewing the correlation coefficient as perhaps his greatest accomplishment, Pearson (1910) listed a number of sources of *spurious correlations* (Aldrich, 1995). Among these was the problem of ratios and of sums, and of correlations induced by mixing different groups.

4.8.2.1 The misuse of ratios, sums and differences

It is not uncommon to convert observations to ratios of the amount X_i with respect to some common baseline, T. Consider the case of discretionary income spent on CDs, books, and wine with respect to the total discretionary income, T. Although these four variables might themselves be uncorrelated, by expressing them all as a ratio of a common variable, the ratios are spuriously correlated (Table 4.24).

Just as forming ratios can induce spurious correlations, so can addition or subtraction. Table 4.24 considers the difference between amount spent on CDs, Books, or Wine and total income. Even though the raw data are uncorrelated, the differences are correlated. The reason behind this is that 50% of the variance of the ratio or difference score is associated with the common variable. Thus, the expected amount of variance beween two such ratios or differences would be 25% for an expected correlation of .5.

Table 4.24 When expressing variables as a ratio or a sum or difference of two unrelated variable, the ratios and differences are correlated, even though the variables themselves are not. The amount of money spent on CDs, books or wine is unrelated to the other two and to total income. But the ratio of amount spent on CDs, books, or wine as fraction of total income (CDsratio , etc) as well as the differences in amount spent (CDs - Income) are correlated.

```
>   x <- matrix(rnorm(1000),ncol=4) + 4
>   colnames(x) <- c("CDs","Books","Wine","Income")
>  x.df <- data.frame(x,x/x[,4],(x-x[,4]))
>  colnames(x.df) <- c("CDs","Books","Wine","Income",
+    paste(c("CDs","Books","Wine","Income"),"ratio",sep=""),
+    paste(c("CDs","Books","Wine","Income"),"diff",sep=""))
>   round(cor(x.df[-c(8,12)]),2)
```

```
             CDs Books  Wine Income CDsratio Booksratio Wineratio CDsdiff Booksdiff Winediff
CDs         1.00  0.00 -0.04   0.02     0.64      -0.04     -0.06    0.71     -0.02    -0.05
Books       0.00  1.00 -0.01  -0.03     0.00       0.61      0.00    0.02      0.72     0.01
Wine       -0.04 -0.01  1.00   0.04    -0.07      -0.05      0.58   -0.06     -0.04     0.69
Income      0.02 -0.03  0.04   1.00    -0.68      -0.73     -0.72   -0.69     -0.71    -0.70
CDsratio    0.64  0.00 -0.07  -0.68     1.00       0.55      0.52    0.94      0.47     0.44
Booksratio -0.04  0.61 -0.05  -0.73     0.55       1.00      0.59    0.49      0.94     0.50
Wineratio  -0.06  0.00  0.58  -0.72     0.52       0.59      1.00    0.46      0.49     0.94
CDsdiff     0.71  0.02 -0.06  -0.69     0.94       0.49      0.46    1.00      0.49     0.46
Booksdiff  -0.02  0.72 -0.04  -0.71     0.47       0.94      0.49    0.49      1.00     0.49
Winediff   -0.05  0.01  0.69  -0.70     0.44       0.50      0.94    0.46      0.49     1.00
```

4.8.2.2 Correlation induced by ipsatization and other devices

When studying individual differences in values (e.g, Allport et al., 1960; Hinz et al., 2005), it is typical to *ipsatize* the scores (Cattell, 1945). That is, the total score of all the values is fixed at a constant for all participants and an increase in one necessarily implies a decrease in the others. Essentially this is zero centering the data for each participant. Psychologically this means everyone has the same total of value strength. Even for truly uncorrelated variables, ipsatization forces a correlation of -1/(k -1) for k variables and reduces the rank of the correlation matrix by 1 (Dunlap and Cornwell, 1994).

This problem can occur in more subtle ways than just fixing the sum to be a constant. Sometimes ratings are made on a forced choice basis (choose which behavior is being shown) and leads to the strange conclusion that e.g., being friendly is unrelated to being sociable. When allowing the ratings not to be forced choices but rather the amount of each behavior is rated separately, the normal structure is observed (Romer and Revelle, 1984). This problem also can be seen in cognitive psychology when raters are asked to choose which cognitive process is being used, rather than how much each process is being used.

4.8.2.3 Simpson's Paradox and the within versus between correlation problem

The confounding of group differences with within group relationships to produce spurious overall relationships has plagued the use of correlation since it was developed. Consider the classic example of inappropriately deciding that an antitoxin is effective even though in reality it has no effect (Yule, 1912). If women have a higher mortality from a disease than do men, but more men are given the antitoxin than the women, the pooled data would show a favorable effect of the antitoxin, even though it in fact had no effect. Similarly, between 1974 and 1978 the tax rate decreased within each of several income categories, although the overall tax rate increased Wagner (1982). So called *ecologicial correlations* Robinson (1950) are correlations of group means and either can or can not reflect relationships within groups.

One of the most well known examples of this effect, known as *Simpson's paradox*, where relationships within groups can be in the opposite direction of the relationships for the entire sample (Simpson, 1951) was found when studying graduate admissions to the University of California, Berkeley. In 1973, UCB had 2691 male applicants and 1198 females applicants. Of the males, about 44% were admitted, of the females, about 35%. What seems to be obvious sex discrimination in admissions became a paper in *Science* when it was discovered that the individual departments, if discriminating at all, discriminated in favor of women (Bickel et al., 1975). The women were applying to the departments which admitted fewer applicants as a percentage of applicants (i.e., two thirds of the applicants to English but only 2 percent to mechanical engineering were women). The correlation across departments of percent female applicants and difficulty of admission was .56. This data set `UCBAdmissions` is used as an example of various graphical displays.

Problems similar to the UCB case can arise when pooling within subject effects across subjects. For instance when examining the structure of affect the structure across subjects is very different from the structure within subjects. Across subjects, positive and negative affect are almost independent, while within subjects the correlation reliabily varies from highly positive to highly negative (Rafaeli et al., 2007).

4.8.2.4 Correlations of means $\neq$ correlations of observations

There are other cases, however, when the correlations of group means clarifies the importance of the underlying relationship (Lubinski and Humphreys, 1996). elaborate

4.8.2.5 Base rates and skew

A difficulty with the Pearson correlation (but not rank order correlations such as Spearman's ρ) is when the data differ in the amount of *skew*. This problem arises, for example when examining the structure of measures of positive and negative affect (Rafaeli and Revelle, 2006), or when looking for correlates of psychopathology. The Pearson r can be greatly attenuated with large differences in skew. Correlations of rank orders (i.e., Spearman's rho) do not suffer from this problem. Consider a simple example of two bivariate normal variables, **x** and **y** with a population correlation of .71. Consider also various transformations of these original data to have positive and negative skew. log.x and log y are negatively skewed, -log(-x) (log.nx) and -log(-y) have positive skew. Similarly the exponential of x (exp.x) and y (exp.y) have very large positive skews while their negative inverses (exp.nx= $-e^{-x}$) have large negative skews (Table 4.25, Figure 4.8.2.5).

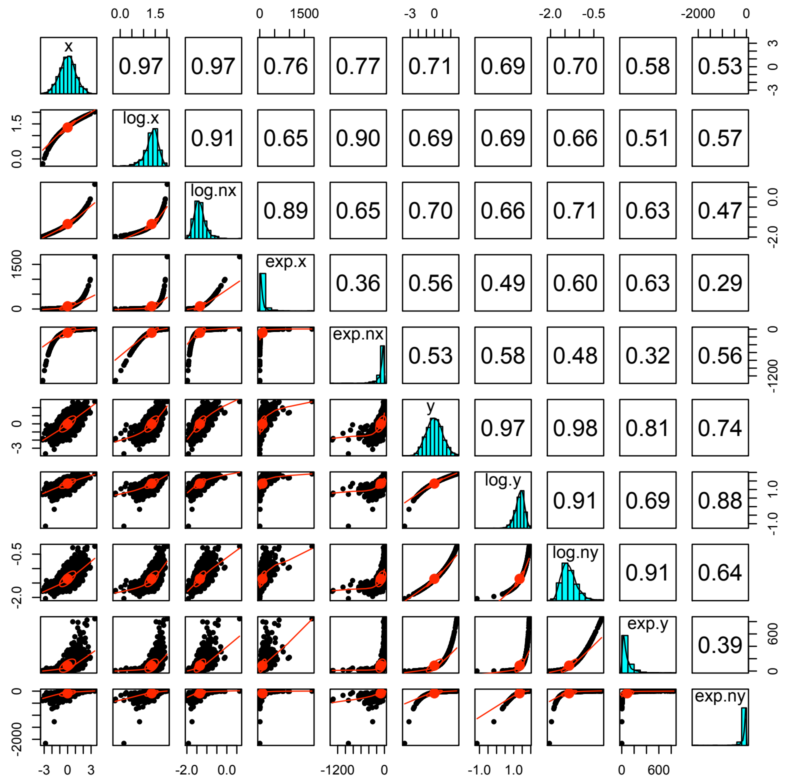


Fig. 4.10 Differences in skew can attenuate correlations. Two variables, x and y, are correlated .71 in the bivariate normal case. Four variations of each and x and y are generated by log and exponential transforms of original or reversed values to induce positive and negative skew.

Table 4.25 Descriptive statistics of skew example

> describe(skew.df)

```
> describe(skew.df)
        var    n   mean     sd median trimmed   mad      min     max   range  skew kurtosis   se
x         1 1000   0.00   1.03   0.01    0.00  0.96    -3.20    3.47    6.67 -0.05     0.07 0.03
log.x     2 1000   1.35   0.29   1.39    1.37  0.24    -0.22    2.01    2.23 -1.15     2.73 0.01
log.nx    3 1000  -1.35   0.29  -1.38   -1.37  0.24    -1.97    0.64    2.61  1.10     3.23 0.01
exp.x     4 1000  91.93 122.03  55.22   67.64 49.38     2.23 1756.20 1753.98  4.94    43.58 3.86
exp.nx    5 1000 -93.82 127.51 -53.98  -67.54 47.03 -1338.84   -1.70 1337.14 -4.34    27.74 4.03
y         6 1000  -0.02   0.99  -0.04   -0.03  1.02    -3.68    2.74    6.42  0.04    -0.08 0.03
log.y     7 1000   1.35   0.28   1.38    1.37  0.25    -1.14    1.91    3.05 -1.36     7.12 0.01
log.ny    8 1000  -1.36   0.27  -1.40   -1.38  0.25    -2.04   -0.23    1.81  0.91     1.40 0.01
exp.y     9 1000  87.51 106.14  52.40   65.79 45.79     1.38  844.76  843.38  3.39    15.45 3.36
exp.ny   10 1000 -90.05 119.15 -56.89  -69.21 49.25 -2167.17   -3.53 2163.64 -7.40   103.09 3.77
```

Examining the *SPLOM* it is clear that small differences in skew do not lead to a large attenuation, but that as the differences in skew go up, particularly if they are in opposite directions, the correlations are seriously attenuated. This is true not just with the set of transformations of each variable **x** with transformations of **x**, but even more serious when examining the correlations between the transformed values **x** and **y**. For this particular example, because the transformations were monotonic, the *Spearman rho* correlations correctly were 1s within the **x** and **y** set, and .69 between.

When working with only a few levels rather than the many shown in Figure 4.8.2.5, the problems of skew are also known as a problems of base rates. If the probability of success on a task is much greater than the probability of failure, and the probability of a predictor of success being positive is much than the probability of it being negative, then the dichotomous variable of success/failure can not have a high correlation with the predictor, even if the underlying relationship were perfect.

4.8.3 Non linearity, outliers and other problems: the importance of graphics

Although not all threats to inference can be detected graphically, one of the most powerful statistical tests for non-linearity and outliers is the well known but not often used "interoccular trauma test". A classic example of the need to examine one's data for the effect of nonlinearity and the effect of outliers is the data set of Anscombe (1973) which is included as the `data(anscombe)` data set. This data set is striking for it shows four patterns of results, with equal regressions and equal descriptive statistics. The graphs differ drastically in appearance for one actually has a curvilinear relationship, two have one extreme score, and one shows the expected pattern. Anscombe's discussion of the importance of graphs is just as timely now as it was 35 years ago:

> Graphs can have various purposes, such as (i) to help us perceive and appreciate some broad features of the data, (ii) to let us look behind these broad features and see what else is there. Most kinds of statistical calculaton rest on assumptions about the behavior of the data. Those assumptions may be false, and the calculations may be misleading. We ought always to try to check whether the assumptions are reasonably correct; and if they are wrong we ought to be able

> to perceive in what ways the are wrong. Graphs are very valuable for these purposes. (Anscombe, 1973, p 17).

The next chapter will generalize the correlation coefficient from the case of two variables to the case of multiple predictors (multiple R) and the problem of statistical control from one or more variables when considering the relationship between variables. The problems that arise in the two variable case are even more pronounced in the multiple variable case.