Statistics in Medicine

Calculating confidence intervals for regression and correlation

DOUGLAS G ALTMAN, MARTIN J GARDNER

Introduction

The most common statistical analyses are those that examine one or two groups of individuals with respect to a single variable, and methods of calculating confidence intervals for means or proportions and their differences have been described previously.[1] Also common are those analyses that consider the relation between two variables in one group of subjects. We use regression analysis to predict one variable from another, and correlation analysis to see if the values of two variables are associated. The purposes of these two analyses are distinct, and usually one only should be used.

This paper outlines the calculation of the linear regression equation for predicting one variable from another and shows how to calculate confidence intervals for the population value of the slope and intercept of the line, for the line itself, and for predictions made using the regression equation. It explains how to obtain a confidence interval for the population value of the difference between the slopes of regression lines in two groups of subjects and how to calculate a confidence interval for the vertical distance between two parallel regression lines. The calculations of confidence intervals for Pearson's correlation coefficient and Spearman's rank correlation coefficient are described.

Worked examples are included to illustrate each method. The calculations have been carried out to full arithmetical precision, as is recommended practice,[2] but intermediate steps are shown as rounded results. Methods of calculating confidence intervals for different aspects of regression and correlation are demonstrated, but the appropriate ones to use depend on the particular problem being studied.

The interpretation of confidence intervals has been discussed earlier.[1] Confidence intervals convey only the effects of sampling variation on the estimated statistics and cannot control for other errors such as biases in design, conduct, or analysis.

General form of confidence intervals

The basic method for constructing the confidence intervals is as previously described.[1] Each confidence interval is obtained by subtracting from, and adding to, the estimated statistic (or a transformation) a multiple of its standard error (SE). The multiple is determined by the theoretical distribution of the statistic: the t distribution for regression, or the Normal distribution for correlation. The multiple is taken as the value that corresponds to including the central $100(1-\alpha)\%$ of the theoretical distribution. So, for example, a 95% confidence interval is described by finding the value that cuts off 2½% from each tail of the distribution. Tables of the t and Normal distributions are available in most statistics books and *Geigy Scientific Tables*.[3] We denote the relevant value as either $t_{1-\alpha/2}$ or $N_{1-\alpha/2}$. For the t distribution the degrees of freedom, which depend on the sample size, must be known.

Section of Medical Statistics, MRC Clinical Research Centre, Harrow, Middlesex HA1 3UJ
DOUGLAS G ALTMAN, BSC, medical statistician

MRC Environmental Epidemiology Unit (University of Southampton), Southampton General Hospital, Southampton SO9 4XY
MARTIN J GARDNER, PHD, professor of medical statistics

Correspondence to: Mr D Altman, Medical Statistics Laboratory, Imperial Cancer Research Fund, PO Box 123, Lincoln's Inn Fields, London WC2A 3PX.

Regression analysis

For two variables x and y we wish to calculate the regression equation for predicting y from x. We call y the dependent variable and x the independent variable. The equation for the population regression line is:

$$y=A+Bx$$

where A is the intercept on the y axis (the value of y when $x=0$) and B is the slope of the line. In standard regression analysis it is assumed that the distribution of the y variable at each value of x is Normal with the same variance, but no assumptions are made about the distribution of the x variable. Sample estimates a (of A) and b (of B) are needed and also the means of the two variables ($\bar{x}$ and $\bar{y}$), the standard deviations of the two variables (s_x and s_y), and the residual standard deviation of y about the regression line (s_{res}). The formulas for deriving a, b, and s_{res} are given in the appendix.

All the following confidence intervals associated with a single regression line use the quantity $t_{1-\alpha/2}$, the appropriate value from the t distribution with $n-2$ degrees of freedom where n is the sample size.

THE SLOPE OF THE REGRESSION LINE

The slope of the sample regression line estimates the mean change in y for a unit change in x. The standard error of the slope, b, is calculated as:

$$SE(b)=\frac{s_{res}}{s_x\sqrt{n-1}}.$$

The $100(1-\alpha)\%$ confidence interval for the population value of the slope, B, is then given by

$$b-(t_{1-\alpha/2}\times SE(b)) \quad \text{to} \quad b+(t_{1-\alpha/2}\times SE(b)).$$

THE MEAN VALUE OF Y FOR A GIVEN VALUE OF X (AND FOR THE REGRESSION LINE)

The estimated mean value of y for any chosen value of x, say x_0, is obtained from the fitted regression line as:

$$y_{fit}=a+bx_0.$$

The standard error of y_{fit} is given by:

$$SE(y_{fit})=s_{res}\times\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}}.$$

The $100(1-\alpha)\%$ confidence interval for the population mean value of y at $x=x_0$ is then:

$$y_{fit}-(t_{1-\alpha/2}\times SE(y_{fit})) \quad \text{to} \quad y_{fit}+(t_{1-\alpha/2}\times SE(y_{fit})).$$

When this calculation is made for all values of x in the observed range a $100(1-\alpha)\%$ confidence interval for the position of the population regression line is obtained. Because of the last term in the formula for $SE(y_{fit})$ the confidence interval becomes wider with increasing distance of x_0 from $\bar{x}$.

THE INTERCEPT OF THE REGRESSION LINE

The intercept of the regression line on the y axis is generally of less interest than the slope of the line and does not usually have any obvious interpretation. It can be seen that the intercept is the fitted value of y when x is zero. Thus a $100(1-\alpha)\%$ confidence interval for the intercept, A, can be obtained using the formula from the preceding section with $x_0=0$ and $y_{fit}=a$. The confidence interval is thus given by:

$$a-(t_{1-\alpha/2}\times SE(a)) \quad \text{to} \quad a+(t_{1-\alpha/2}\times SE(a)).$$

PREDICTION FOR AN INDIVIDUAL (AND ALL INDIVIDUALS)

It is useful to calculate the uncertainty in y_{fit} as a predictor of y for an individual subject. The range of uncertainty is called a prediction (or tolerance) interval. A prediction interval is wider than the associated confidence interval for the mean value of y because the scatter of data about the regression line is more important. For an individual whose value of x is x_0 the predicted value of y is y_{fit}, given by:

$$y_{fit}=a+bx_0.$$

To calculate the prediction interval we use the estimated standard deviation of individual values of y when x equals x_0 (s_{pred}):

$$s_{pred}=s_{res}\times\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}}.$$

The $100(1-\alpha)\%$ prediction interval is then:

$$y_{fit}-(t_{1-\alpha/2}\times s_{pred}) \quad \text{to} \quad y_{fit}+(t_{1-\alpha/2}\times s_{pred}).$$

When this calculation is made for all values of x in the observed range the estimated prediction interval should include the values of y for $100(1-\alpha)\%$ of subjects in the population.

COMPARISON OF TWO REGRESSION LINES

Regression lines fitted to observations from two independent groups of subjects can be analysed to see if they come from populations with regression lines that are parallel or even coincident.[4] If we have fitted regression lines to two different sets of data on the same two variables we can construct a confidence interval for the difference between the population regression slopes using a similar approach to that for a single regression line. The standard error of the difference between the slopes is given by first calculating $s^{\star}_{res}$ as:

$$s^{\star}_{res}=\sqrt{\frac{(n_1-2)s^2_{res1}+(n_2-2)s^2_{res2}}{n_1+n_2-4}}$$

and then

$$SE(b_1-b_2)=s^{\star}_{res}\times\sqrt{\frac{1}{(n_1-1)s^2_{x1}}+\frac{1}{(n_2-1)s^2_{x2}}}$$

where the suffixes 1 and 2 indicate values derived from the two separate sets of data. The $100(1-\alpha)\%$ confidence interval for the population difference between the slopes is given by:

$$b_1-b_2-(t_{1-\alpha/2}\times SE(b_1-b_2)) \quad \text{to} \quad b_1-b_2+(t_{1-\alpha/2}\times SE(b_1-b_2)),$$

where $t_{1-\alpha/2}$ is the appropriate value from the t distribution with n_1+n_2-4 degrees of freedom.

THE VERTICAL DISTANCE BETWEEN PARALLEL REGRESSION LINES

If the 95% confidence interval for the difference between population values of the slopes includes zero it is reasonable to fit two parallel regression lines with the same slope and calculate a confidence interval for their common slope. We can also calculate a confidence interval for the vertical distance between the parallel lines, which is the difference between the mean values of y calculated from the two lines at any value of x. This is equivalent to adjusting the observed mean values of y for the mean values of x and is known as analysis of covariance.[4]

All the calculation can be done using the results obtained by fitting separate regression lines to the two groups and the standard deviations of the x and y values in the two groups: s_{x1}, s_{x2}, s_{y1}, and s_{y2}. First define the quantity w as:

$$w=(n_1-1)s^2_{x1}+(n_2-1)s^2_{x2}.$$

The common slope of the parallel lines (b_{par}) is estimated as:

$$b_{par}=\frac{b_1(n_1-1)s^2_{x1}+b_2(n_2-1)s^2_{x2}}{w}.$$

The residual standard deviation of y around the parallel lines (s_{par}) is given by:

$$s_{par}=\sqrt{\frac{(n_1-1)s^2_{y1}+(n_2-1)s^2_{y2}-b^2_{par}\times w}{n_1+n_2-3}},$$

and the standard error of the slope by:

$$SE(b_{par})=\frac{s_{par}}{\sqrt{w}}.$$

The $100(1-\alpha)\%$ confidence interval for the population value of the common slope is then:

$$b_{par}-(t_{1-\alpha/2}\times SE(b_{par})) \quad \text{to} \quad b_{par}+(t_{1-\alpha/2}\times SE(b_{par}))$$

where $t_{1-\alpha/2}$ is the appropriate value from the t distribution with n_1+n_2-3 degrees of freedom.

The intercepts of the two parallel lines with the y axis are given by:

$$\bar{y}_1-b_{par}\bar{x}_1 \quad \text{and} \quad \bar{y}_2-b_{par}\bar{x}_2.$$

We are usually more interested in the difference between the intercepts, which is the vertical distance between the parallel lines. This is the same as the difference between the fitted y values for the two groups at the same value of x. The adjusted mean difference (y_{diff}) is calculated as:

$$y_{diff}=\bar{y}_1-\bar{y}_2-b_{par}(\bar{x}_1-\bar{x}_2)$$

and the standard error of y_{diff} is:

$$SE(y_{diff})=s_{par}\times\sqrt{\frac{\frac{1}{n_1}+\frac{1}{n_2}+(\bar{x}_1-\bar{x}_2)^2}{w}}.$$

The $100(1-\alpha)\%$ confidence interval for the population value of y_{diff} is:

$$y_{diff}-(t_{1-\alpha/2}\times SE\ (y_{diff})) \quad \text{to} \quad y_{diff}+(t_{1-\alpha/2}\times SE(y_{diff}))$$

where $t_{1-\alpha/2}$ is the appropriate value from the t distribution with n_1+n_2-3 degrees of freedom.

WORKED EXAMPLE

Table I shows data from a clinical trial of enalaparil versus placebo in diabetic patients.[5] The variables studied are mean arterial blood pressure (mm Hg) and total glycosylated haemoglobin concentration (%). The analyses presented here are illustrative and do not relate directly to the clinical trial. Most of the methods for calculating confidence intervals are demonstrated using only the data from the 10 subjects who received enalaparil.

We want to describe the way total glycosylated haemoglobin concentration (TGH) changes with mean arterial blood pressure (MAP). The regression line of total glycosylated haemoglobin concentration on mean arterial blood pressure for the 10 subjects receiving enalaparil is found to be:

$$TGH=20{\cdot}19-0{\cdot}1168\times MAP.$$

The estimated slope of the line is negative, indicating lower total glycosylated haemoglobin concentrations for subjects with higher mean arterial blood pressure.

TABLE I—*Mean arterial blood pressure and total glycosylated haemoglobin concentration in two groups of 10 diabetics on entry to a clinical trial of enalaparil versus placebo*[5]

Enalaparil group		Placebo group	
Mean arterial blood pressure (mm Hg) x	Total glycosylated haemoglobin (%) y	Mean arterial blood pressure (mm Hg) x	Total glycosylated haemoglobin (%) y
91	9·8	98	9·5
104	7·4	105	6·7
107	7·9	100	7·0
107	8·3	101	8·6
106	8·3	99	6·7
100	9·0	87	9·5
92	9·7	98	9·0
92	8·8	104	7·6
105	7·6	106	8·5
108	6·9	90	8·6
Means:			
$\bar{x}=101·2$	$\bar{y}=8·37$	$\bar{x}=98·8$	$\bar{y}=8·17$
Standard deviations:			
$s_x=6·941$	$s_y=0·9615$	$s_x=6·161$	$s_y=1·0914$
Standard deviations about the fitted regression lines:			
	$s_{res}=0·5485$		$s_{res}=0·9866$

The other quantities needed to obtain the various confidence intervals are shown in table I. The calculations use 95% confidence intervals. For this we need the value of $t_{0·975}$ with 8 degrees of freedom, and the appropriate table shows this to be 2·306.

Confidence interval for the slope of the regression line

The standard error of the slope is:

$$SE(b)=\frac{0·5485}{6·941\times\sqrt{9}}=0·02634\%\text{ per mm Hg.}$$

The 95% confidence interval for the population value of the slope is:

$$-0·1168-(2·306\times0·02634)\quad\text{to}\quad-0·1168+(2·306\times0·02634)$$

that is, from −0·178 to −0·056% per mm Hg.

Confidence interval for the mean total glycosylated haemoglobin concentration for a given mean arterial blood pressure (and for the regression line)

The confidence interval for the mean total glycosylated haemoglobin concentration can be calculated for any specified value of mean arterial blood pressure. If the mean arterial blood pressure of interest is 100 mm Hg the estimated total glycosylated haemoglobin concentration (y_{fit}) is 20·19−(0·1168×100)=8·51%. The standard error of this estimated value is:

$$SE(y_{fit})=0·5485\times\sqrt{\frac{1}{10}+\frac{(100-101·2)^2}{9\times6·941^2}}=0·1763\%.$$

The 95% confidence interval for the mean total glycosylated haemoglobin concentration for the population of diabetic subjects with a mean arterial blood pressure of 100 mm Hg is thus:

$$8·51-(2·306\times0·1763)\quad\text{to}\quad8·51+(2·306\times0·1763)$$

that is, from 8·10 to 8·92%.

By calculating the 95% confidence interval for the mean total glycosylated haemoglobin concentration for all values of mean arterial blood pressure within the range of observations we get a 95% confidence interval for the regression line. This is shown in figure 1. The confidence interval becomes wider moving away from the mean mean arterial blood pressure of 101·2 mm Hg.

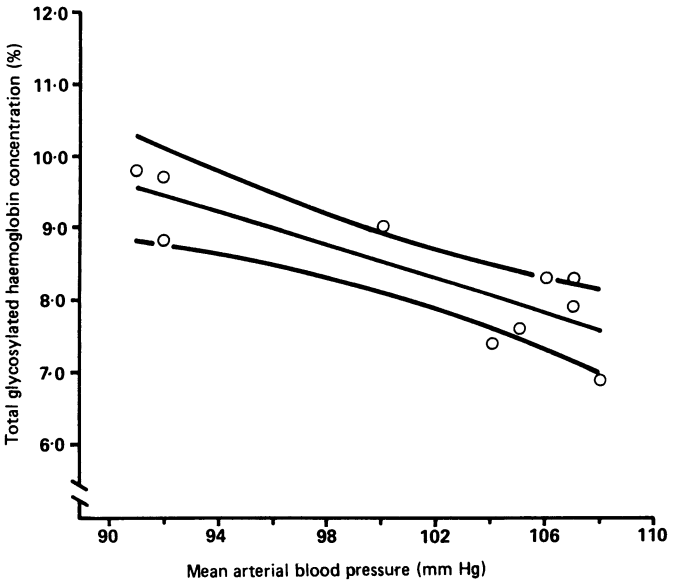

FIG 1—Regression line of total glycosylated haemoglobin concentration on mean arterial blood pressure, with 95% confidence interval for the mean total glycosylated haemoglobin concentration.

Confidence interval for the intercept of the regression line

The confidence interval for the population value of the intercept is the confidence interval for y_{fit} when x=0, and is calculated as before. In this case the intercept is 20·19%, with a 95% confidence interval from 14·03 to 26·35%.

Prediction interval for the total glycosylated haemoglobin concentration of an individual

The 95% prediction interval for the total glycosylated haemoglobin concentration of an individual subject with a mean arterial blood pressure of 100 mm Hg is obtained by first calculating s_{pred}:

$$s_{pred}=0·5485\times\sqrt{1+\frac{1}{10}+\frac{(100-101·2)^2}{9\times6·941^2}}=0·5761\%.$$

The 95% prediction interval is then given by:

$$8·51-(2·306\times0·5761)\quad\text{to}\quad8·51+(2·306\times0·5761)$$

that is, from 7·18 to 9·84%.

The contrast with the narrower 95% confidence interval for the mean total glycosylated haemoglobin concentration for a mean arterial blood pressure of 100 mm Hg calculated above is noticeable. The 95% prediction intervals for the range of observed levels of mean arterial blood pressure are shown in figure 2 and again these widen on moving away from the mean arterial blood pressure of 101·2 mm Hg.

Confidence interval for the difference between the slopes of two regression lines

The regression line for the placebo group from the data in table I is:

$$TGH=17·33-0·09268\times MAP.$$

The difference between the estimated slopes of the two lines is −0·1168−(−0·09268)=−0·02412% per mm Hg. The standard error of this difference is found either directly or through s^*_{res} as:

$$SE(b_1-b_2)=\sqrt{\frac{(8\times0·5485^2+8\times0·9866^2)}{16}\times\left[\frac{1}{9\times6·941^2}+\frac{1}{9\times6·161^2}\right]}$$

$$=0·05774\%\text{ per mm Hg.}$$

The value of $t_{0\cdot975}$ with 16 degrees of freedom is 2·120, so the 95% confidence interval for the population difference between the slopes is:

$$-0\cdot02412-(2\cdot120\times0\cdot05774) \quad \text{to} \quad -0\cdot02412+(2\cdot120\times0\cdot05774)$$

that is, from −0·147 to 0·098% per mm Hg.

Since a zero difference between slopes is near the middle of this confidence interval there is no evidence that the two population regression lines have different slopes. This is not surprising in this example as the subjects were allocated at random to the treatment groups.

The vertical distance between two parallel lines

First calculate the quantity w as:

$$w=9\times6\cdot941^2+9\times6\cdot161^2$$

$$=775\cdot22.$$

The common slope of the parallel lines is then found as:

$$b_{par}=\frac{-0\cdot1168\times9\times6\cdot941^2+(-0\cdot09268)\times9\times6\cdot161^2}{775\cdot22}$$

$$=-0\cdot1062\% \text{ per mm Hg.}$$

The residual standard deviation of y around the parallel lines is:

$$s_{par}=\sqrt{\frac{9\times0\cdot9615^2+9\times1\cdot0914^2-(-0\cdot1062)^2\times775\cdot2}{10+10-3}}$$

$$=0\cdot7786\%$$

and the standard error of the common slope is thus:

$$SE(b_{par})=\frac{0\cdot7786}{\sqrt{775\cdot22}}=0\cdot02796\% \text{ per mm Hg.}$$

The value of $t_{0\cdot975}$ with 17 degrees of freedom is 2·110, so that the 95% confidence interval for the population value of b_{par} is therefore:

$$-0\cdot1062-(2\cdot110\times0\cdot02796) \quad \text{to} \quad -0\cdot1062+(2\cdot110\times0\cdot02796)$$

that is, from −0·165 to −0·047% per mm Hg.

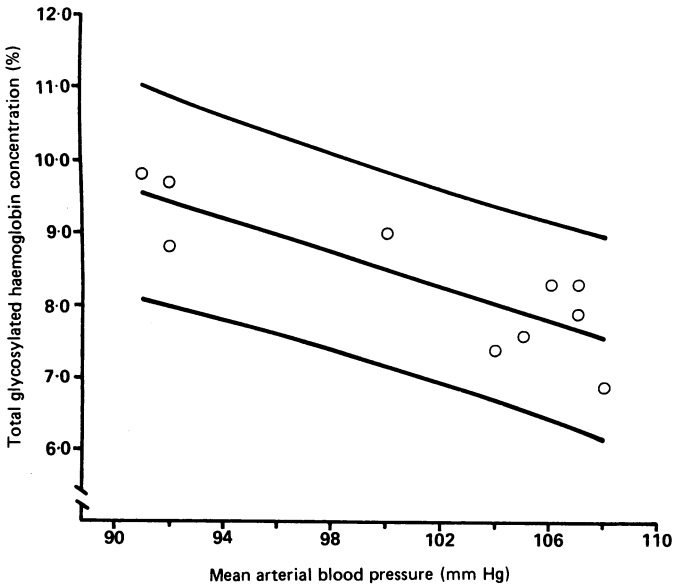

FIG 2—Regression line of total glycosylated haemoglobin concentration on mean arterial blood pressure, with 95% prediction interval for an individual total glycosylated haemoglobin concentration.

Using the calculated value for the common slope the adjusted difference between the mean total glycosylated haemoglobin concentration in the two groups is:

$$y_{diff}=(8\cdot37-8\cdot17)-(-0\cdot1062)\times(101\cdot2-98\cdot8)=0\cdot4548\%,$$

and its standard error is:

$$SE(y_{diff})=0\cdot7786\times\sqrt{\frac{1}{9}+\frac{1}{9}+\frac{(101\cdot2-98\cdot8)^2}{775\cdot22}}$$

$$=0\cdot3731\%.$$

The 95% confidence interval for the population value of y_{diff} is then given by:

$$0\cdot4548-(2\cdot110\times0\cdot3731) \quad \text{to} \quad 0\cdot4548+(2\cdot110\times0\cdot3731)$$

that is, from −0·33 to 1·24%.

EXTENSIONS

The ideas introduced in this section can be extended to studies with more than two groups and where there are more than two variables by using analysis of covariance and multiple regression.[4] Confidence intervals are constructed in much the same way, and are based on the relevant standard error. These more complex situations are beyond the scope of this paper.

Correlation analysis

PEARSON'S PRODUCT MOMENT CORRELATION COEFFICIENT

The correlation coefficient usually calculated is the "product moment correlation coefficient" or "Pearson's r." This measures the degree of linear co-relation between two variables x and y. The formula for calculating r for a sample of observations is given in the appendix.

A confidence interval for the population value of r, assuming that x and y have a joint bivariate Normal distribution, can be constructed by using a transformation of r to a quantity z, which has an approximately Normal distribution. This transformed value, z, is given by:

$$z=\frac{1}{2}\log_e\frac{1+r}{1-r},$$

which for all values of r has a standard error of $1/\sqrt{n-3}$ where n is the sample size.

For a $100(1-\alpha)\%$ confidence interval we then calculate the two quantities:

$$z_1=z-(N_{1-\alpha/2}/\sqrt{n-3})$$

and

$$z_2=z+(N_{1-\alpha/2}/\sqrt{n-3}),$$

where $N_{1-\alpha/2}$ is the appropriate value from the standard Normal distribution for the $100(1-\alpha/_2)$ percentile. This can be found in appropriate tables.

The values z_1 and z_2 need to be transformed back to the original scale to give a $100(1-\alpha)\%$ confidence interval for the population correlation coefficient as:

$$\frac{e^{2z_1}-1}{e^{2z_1}+1} \quad \text{to} \quad \frac{e^{2z_2}-1}{e^{2z_2}+1}.$$

Worked example

Table II shows the basal metabolic rate and total energy expenditure in 24 hours from a study of 13 non-obese women.[6] The data are ranked by increasing basal metabolic rate. Pearson's r for these data is 0·7283, and the transformed value z is:

$$z=\frac{1}{2}\log_e\frac{1+0\cdot7283}{1-0\cdot7283}=0\cdot9251.$$

The values of z_1 and z_2 for a 95% confidence interval are:

$$z_1=0\cdot9251-(1\cdot96/\sqrt{10})=0\cdot3053$$

and

$$z_2=0\cdot9251+(1\cdot96/\sqrt{10})=1\cdot545.$$

TABLE II—*Basal metabolic rate and isotopically measured 24 hour energy expenditure in 13 non-obese women*[6]

Basal metabolic rate (MJ/day)	24 Hour total energy expenditure (MJ)
4·67	7·05
5·06	6·13
5·31	8·09
5·37	8·08
5·54	7·53
5·65	7·58
5·76	8·40
5·85	7·48
5·86	7·48
5·90	8·11
5·91	7·90
6·19	10·88
6·40	10·15

From these values we derive the 95% confidence interval for the population correlation coefficient:

$$\frac{e^{2\times 0\cdot 3053}-1}{e^{2\times 0\cdot 3053}+1} \text{ to } \frac{e^{2\times 1\cdot 545}-1}{e^{2\times 1\cdot 545}+1}$$

that is, from 0·296 to 0·913.

SPEARMAN'S RANK COEFFICIENT

To calculate Spearman's rank correlation coefficient (r_s) the values of x and y for the n individuals have to be ranked separately in order of increasing size from 1 to n. Spearman's rank correlation coefficient is then obtained either by using the standard formula for Pearson's product moment correlation coefficient on the ranks of the two variables, or as shown in the appendix using the difference in their two ranks for each individual. The distribution of r_s is similar to that of Pearson's r, so that confidence intervals can be constructed as shown in the previous section.

Appendix: Formulas for regression and correlation analyses

REGRESSION

The slope of the regression line is given by:

$$b=\frac{\Sigma xy-n\bar{x}\bar{y}}{\Sigma x^2-n\bar{x}^2},$$

where Σ represents summation over the n sample points. The intercept is given by:

$$a=\bar{y}-b\bar{x}.$$

The residual standard deviation of y about the regression line is:

$$s_{res}=\sqrt{\frac{\Sigma(y-y_{fit})^2}{n-2}}$$

$$=\sqrt{\frac{\Sigma y^2-n\bar{y}^2-b^2(\Sigma x^2-n\bar{x}^2)}{n-2}}$$

$$=\sqrt{\frac{(n-1)(s_y^2-b^2s_x^2)}{n-2}}.$$

Most statistical computer programs give all the necessary quantities to derive confidence intervals, but you may find that the output gives a quantity called the "standard error of the estimate" (SEE), from which s_{res} can be calculated using:

$$s_{res}=SEE\sqrt{n}.$$

Some of the standard errors needed to calculate confidence intervals may be given directly by the computer program.

CORRELATION

The correlation coefficient (Pearson's r) is estimated by:

$$r=\frac{\Sigma xy-n\bar{x}\bar{y}}{\sqrt{[\Sigma x^2-n\bar{x}^2][\Sigma y^2-n\bar{y}^2]}}.$$

Spearman's rank correlation coefficient is given by:

$$r_s=1-\frac{6\Sigma d_i^2}{n^3-n},$$

where d_i is the difference in the ranks of the two variables for the ith individual. Alternatively, r_s can be obtained by applying the formula for Pearson's r to the ranks of the variables. The calculation of r_s should be modified when there are tied ranks in the data, but the effect is minimal unless there are many tied ranks.

We thank the referee for his helpful suggestions and Mrs Brigid Howells for her careful typing.

References

1 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986;292:746-50.
2 Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. *Br Med J* 1983;286:1489-93.
3 Lentner C, ed. *Geigy scientific tables*. Vol 2. 8th ed. Basle: Ciba-Geigy, 1982.
4 Armitage P, Berry G. *Statistical methods in medical research*. 2nd ed. Oxford: Blackwell, 1987.
5 Marre M, Leblanc H, Suarez L, Guyenne T-T, Ménard J, Passa P. Converting enzyme inhibition and kidney function in normotensive diabetic patients with persistent microalbuminuria. *Br Med J* 1987;294:1448-52.
6 Prentice AM, Black AE, Coward WA, *et al*. High levels of energy expenditure in obese women. *Br Med J* 1986;292:983-7.

(Accepted 4 February 1988)

WORDS HOMO—NOT THE SAME MAN. A reader enquires: "Nowadays the word homosexual is on everyone's lips. Does this word really mean what people generally intend? Surely the opposite of heterosexual is homeosexual. Does not homosexual refer to normal sexual practice as opposed to sex with other animals?"

The prefix HOMO- has two entirely unrelated origins and meanings. *Homo* (Latin) means "man," as distinct from "woman." *Homos* (Greek) means "same." HOMEO- (Gk *homoios*) means "like, similar," and is obviously related to *homos*. The confusion has arisen from mistakenly thinking that the HOMO- of homosexual refers solely to sexual relations between *men*, whereas it refers to the occurrence of this relationship between members of the *same* sex, and is equally applicable to women. It is likewise appropriate that heterosexual (Gk *heteros*, other) should refer to sexual relations with the *other* sex, and not with animals, as suggested, for which the epithet is bestiality.

There are not many words beginning with HOMO- (man). Here are a few: homicide, hominid, homunculus, and the naturalised Latin, *Homo sapiens*, *H erectus*, and *H neanderthalensis*. By contrast there are many words beginning with HOMO- (same): apart from homosexual, we have homogeneous and its derivatives—not to be confused with homogenous (stressed vowels italicised), which means "similar owing to common descent"; homologous and homograft; the chemical term homocyclic, together with those of numerous isomers. Also, in linguistics we have homophone to describe words that have the same sound but different meanings, and homonym and homograph, where the spelling as well as the sound is identical despite different meanings. HOMEO- appears in homeostasis and homeopathy. HOMEO- yields no advantages over HOMO-, except, perhaps, where caution suggests that two subjects are similar but not the same.

As to pronunciation, there appears to be some freedom of choice. The first "o" is usually short as in "hot," except in homunculus and *Homo sapiens* and his predecessors, where it is long as in "home." This is odd because the first "o" in the Latin *homo* is short and the second is long. But no one ever suggested that pronunciation in the English language should be dependent on origins. Pedants might wish for some means of distinguishing "man" HOMO- from "same" HOMO-. They might also object that "homosexual" is a Greek-Latin hybrid, but then so is "bicycle."—B J FREEDMAN.