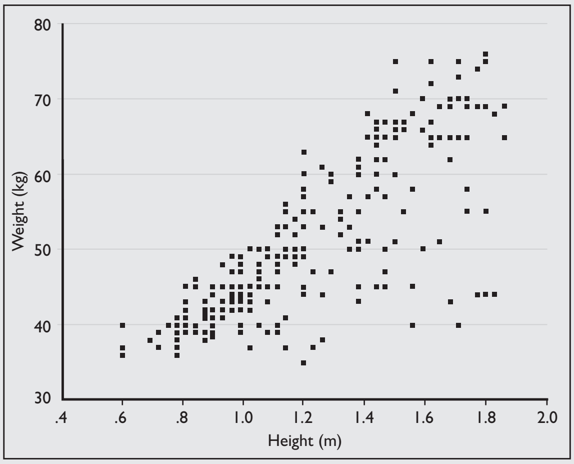
Singapore Med J 2003 Vol 44(12) : 614-619

Biostatistics 104: Correlational Analysis

Y H Chan

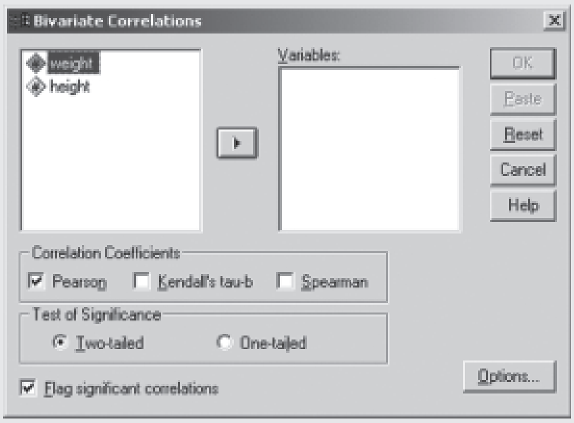
In this article, we shall discuss the analysis of the relationship between two quantitative outcomes, say, height and weight, which is shown graphically by a scatter plot (see Fig. 1).

Fig. 1 Relationship between height and weight.



From the plot it is obvious that as height increases, so does weight and we use the **correlation coefficient (r)** to describe the **degree of linear relationship** between the two variables. In SPSS, go to ***Analyse, Correlate, Bivariate*** to get template I.

Template I



If both height and weight are normally distributed, we use the **Pearson's** correlation; otherwise Spearman's correlation coefficient will be presented. Spearman's correlation is also applicable for two categorical ordinal variables like pain intensity (no, mild, moderate, severe) with burn-surface-area (< 10%, 10% - 19%, 20% - 29%, ≥ 30%).

Table I shows the correlation coefficient between height and weight (both variables are normally distributed): Pearson's r = 0.807 (p<0.001).

Table I. Correlation between Height and Weight.

		Weight (kg)	Height (m)
Weight (kg)	Pearson Correlation	1	.807**
	Sig. (2-tailed)	.	.000
	N	277	277
Height (m)	Pearson Correlation	.807**	1
	SIG. (2-TAILED)	.000	.
	N	277	277

**. Correlation is significant at the 0.01 level (2-tailed).

How to interpret the value of r? r lies between -1 and 1. Values near 0 means no (linear) correlation and values near ± 1 means very strong correlation. The negative sign means that the two variables are inversely related, that is, as one variable increases the other variable decreases. Table II gives a guideline on the strength of the linear relationship corresponding to the correlation coefficient value.

Table II. Strength of linear relationship.

Correlation Coefficient value	Strength of linear relationship
At least 0.8	Very strong
0.6 up to 0.8	Moderately strong
0.3 to 0.5	Fair
Less than 0.3	Poor

Thus for the above height and weight example, we have a strong linear relationship (as expected) and the p-value (affected by sample size) tells us that this relationship is unlikely to happen by chance (it is important to understand that we can have a poor

Clinical Trials and Epidemiology Research Unit
226 Outram Road
Blk B #02-02
Singapore 169039

Y H Chan, PhD
Head of Biostatistics

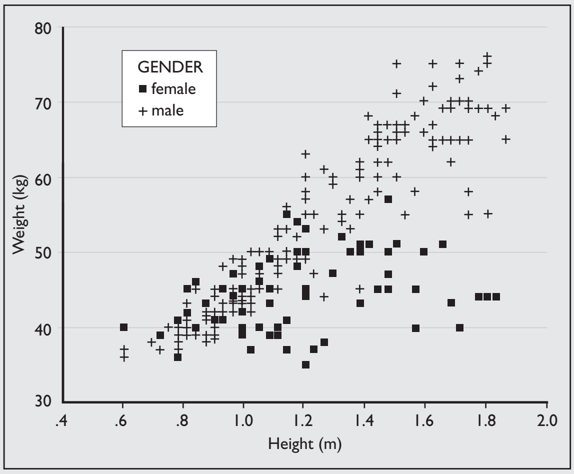
Correspondence to:
Y H Chan
Tel: (65) 6325 7070
Fax: (65) 6324 2700
Email: chanyh@cteru.com.sg

correlation and yet the p-value is statistically significant for a large study).

Squaring r gives the **Coefficient of determination** which tells us the proportion of variance that the two variables have in common. For the height-weight example, r = 0.807 and squaring r gives 0.6512, which means that the height of a person explains 65% of the person's weight; the other 35% could probably be explained by other factors, perhaps nature and nurture (for example).

Precautions have to be taken when we suspect that the overall correlation coefficient is affected by sub-populations. From the above height-weight example, is the relationship of 0.807 applicable to both gender? Fig. 2 shows the scatter plot (again) with the males and females specified.

Fig. 2 Height-Weight by Gender.



Here, we observe that the height-weight relationship for both gender is very different! The strong relationship is only applicable for males but not for females (as most of them tend to be watchful of their weight regardless of their height). Table III shows the correlations by gender.

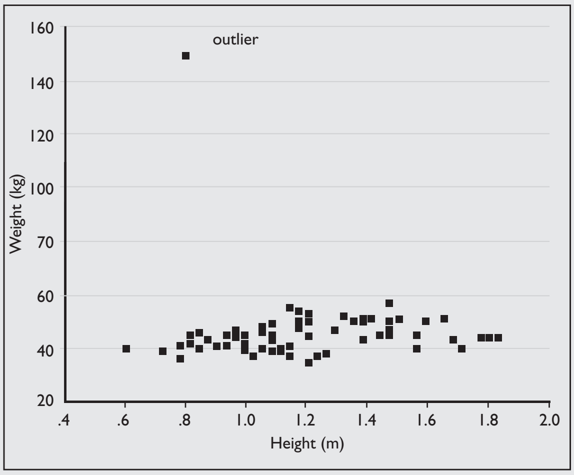
Table III. Pearson's Correlations for Height-Weight by Gender.

GENDER			Weight (kg)	Height (m)
Male	Weight (kg)	Pearson Correlation	1	.930**
		Sig. (2-tailed)	.	.000
		N	207	207
	Height (m)	Pearson Correlation	.930**	1
		Sig. (2-tailed)	.000	.
		N	207	207
Female	Weight (kg)	Pearson Correlation	1	.383**
		Sig. (2-tailed)	.	.001
		N	70	70
	Height (m)	Pearson Correlation	.383**	1
		Sig. (2-tailed)	.001	.
		N	70	70

**. Correlation is significant at the 0.01 level (2-tailed).

Another important precaution to note is the existence of outliers that may affect the correlation. Fig. 3 shows the scatter plot for the females with an outlier, perhaps a wrong entry (or a baby hippo!!), which shows a false negative relationship (r = -0.006).

Fig. 3 Effect of outliers on the correlation coefficient.



To handle the above two pitfalls, a graphical presentation should be performed whenever we want to determine the correlation between two quantitative variables.

PARTIAL CORRELATION

Table IV shows that the age of a subject is highly correlated with both the height and weight.

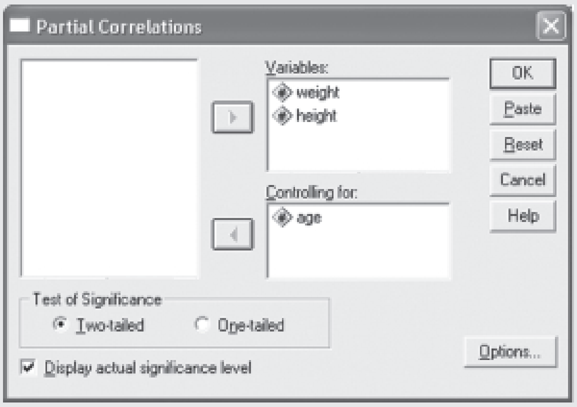
Table IV. Bivariate correlations between Height, Weight and Age.

		Weight (kg)	Height (m)	AGE
Weight (kg)	Pearson Correlation	1	.807**	.889**
	Sig. (2-tailed)	.	.000	.000
	N	277	277	277
Height (m)	Pearson Correlation	.807**	1	.984**
	Sig. (2-tailed)	.000	.	.000
	N	277	277	277
AGE	Pearson Correlation	.889**	.984**	1
	Sig. (2-tailed)	.000	.000	.
	N	277	277	277

**. Correlation is significant at the 0.01 level (2-tailed).

If we want to determine the correlation between height and weight without the effect of age, a partial correlation analysis to control for age is carried out. In SPSS, go to ***Analyse, Correlate, Partial*** to get template II.

Template II



Put Height and Weight in the *Variables* option and Age in the *Controlling for* option.

Table V. Partial correlations between Height and Weight: controlling for Age.

Partial Correlation Coefficients			
Controlling for ..Age			
		Height	Weight
Height	Pearson Correlation	1.0000	0.6126
		(0)	(274)
		p= .	p= .000
Weight	Pearson Correlation	0.6126	1.0000
		(274)	(0)
		p= .000	p= .

The influence of age has been removed from the correlation between height and weight which is reduced from 0.807 to 0.6126, see table V. If age has no effect on height and weight, then there will be not much change in the original correlation even after controlling for age. Qualitative variables like gender could also be used as a *controlling for* variable and more than one controlling variables could be factored for.

CORRELATION DOES NOT MEAN CAUSATION

A high correlation **does not** give us the evidence to make a cause-and-effect statement. A common example given is the high correlation between the cost of damage in a fire and the number of firemen helping to put out the fire. Does it mean that to cut down the cost of damage, the fire department should dispatch less firemen for a fire rescue! We know that there is this intensity of the fire that is highly correlated with the cost of damage and the number of firemen dispatched.

Another example is the high correlation between smoking and lung cancer. However, one may argue that both could be caused by stress; and smoking does not cause lung cancer. In this case, a correlation between lung cancer and smoking may be a result of a cause-and-effect relationship (by clinical experience + common sense?). To establish this cause-and-effect relationship, controlled experiments should be performed (see table VI for the required sample size for different correlation values at power of 80% and 90% with a 2-sided 5%).

Table VI. Sample sizes for different correlation values.

		Pearson's Correlation								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Power	80%	780	190	82	44	26	17	11	7	5
(2-sided 5%)	90%	1,045	255	110	58	34	21	14	9	6

AGREEMENT BETWEEN TWO QUANTITATIVE OUTCOMES

Firstly the paired t-test is definitely not appropriate to show agreement between two quantitative measurements (for example, two instruments measuring temperature). Does it mean that we want the p-value to be greater than 0.05 to imply agreement? Surely by now we know that the p-value is affected by the sample size and thus there's no way to comment on the agreement whether the paired t-test gives a statistical or non-statistical result.

On the other hand, using correlation to describe agreement between two quantitative variables needs caution. Definitely, a high correlation is required but that does not imply agreement (see Fig. 4a). The "line of agreement" should be a 45 degrees (x = y) line (see Fig. 4b)

Fig. 4a

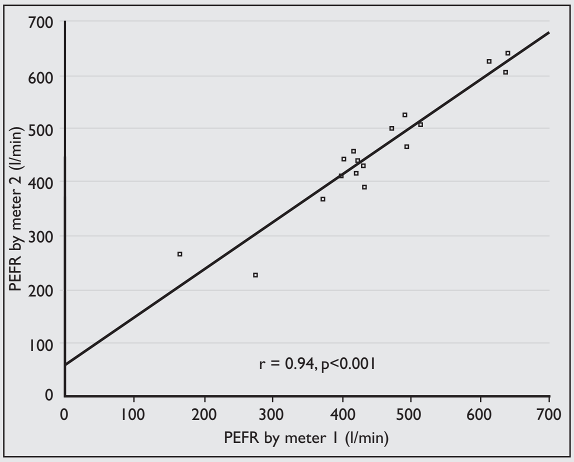
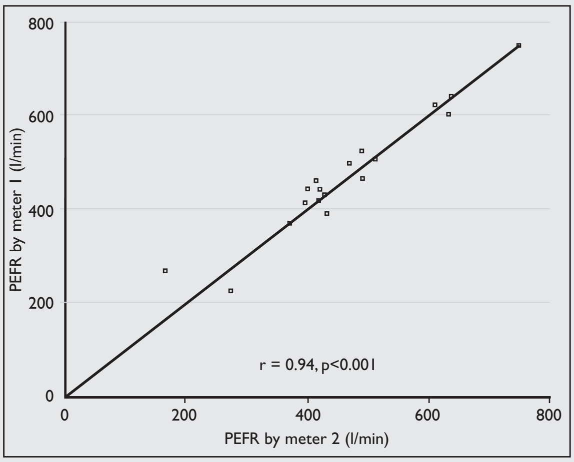


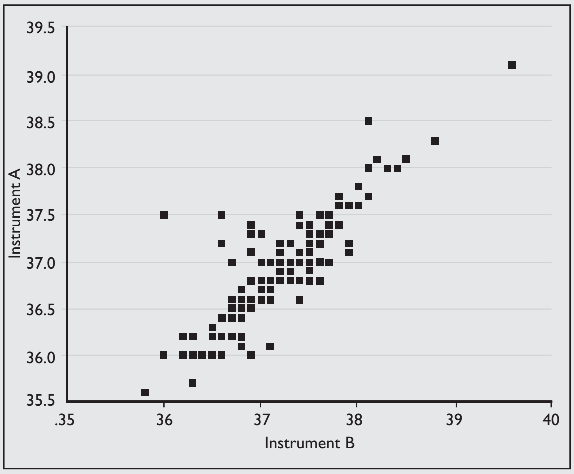
Fig. 4b



Bland & Altman introduced the Bland-Altman plot[1] to describe agreement between two quantitative measurements. There's no p-value available to describe this agreement but rather a "quality control" concept. The difference of the paired two measurements is plotted against the mean of the two measurements and they recommend that 95% of the data points should lie within the ± 2sd of the mean difference.

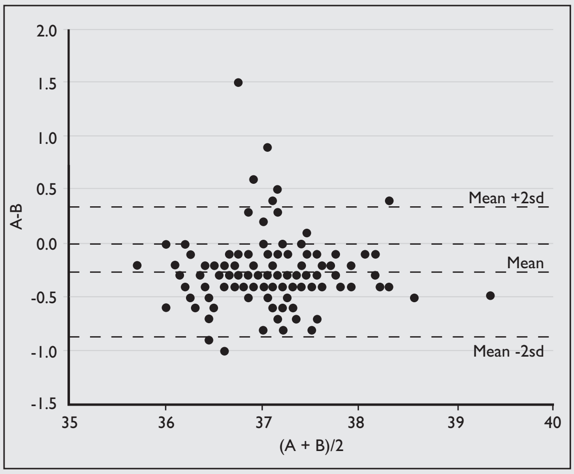
We shall use Bland Altman plot to assess the agreement of two temperature-measuring instruments. One hundred and fifty measurements were taken and Fig. 5 shows the scatter plot between instrument A vs instrument B, the correlation is 0.871, $p<0.001$.

Fig. 5 Scatter plot between Instruments A and B.



To do the Bland Altman plot (see Fig. 6), we have to compute the differences between the instruments and the mean of both instruments for all the paired values. The mean (sd) of the differences is -0.2665 (0.3022), thus the mean ± 2sd are (-0.8709, 0.3379).

Fig. 6 Bland Altman plot.



From the plot, we want the cluster of points to be around the difference = 0 line and any large deviated differences have to be checked: whether the "large" difference is due to wrong data-entry, operator dependent or the flaw of one (or both) of the instruments.

From the plot, we observe that 8/150 (5.3%) of the points are beyond the ± 2sd lines and instrument A seems to be measuring 'low' most of time but have a trend of high scores amidst the 37 degree value. The extreme difference is ± 1 degree. The question now is: are the two instruments agreeable? Well, if one does not mind a one-degree difference, then they are agreeable; otherwise the researcher has to make the judgement call.

Another way to assess the agreement is to set acceptable tolerances for the differences between the two instruments; for example, how many percent in the 0.1 degree difference, 0.2, etc. Table VII shows the "tolerance" table.

Table VII. "Tolerance" table: absolute differences between instruments A & B.

Tolerance (degrees)	n (%)
0.2	29 (19.3)
0.4	117 (78.0)
0.5	131 (87.3)
0.6	133 (88.7)
0.8	144 (96.0)
1.0	149 (99.3)

At least 87% of the measurements agree on the 0.5 degree tolerance or is it acceptable to have 13% being more than 0.5 degrees difference?

AGREEMENT BETWEEN TWO INTER-RATERS ON QUALITATIVE OUTCOMES

Frequently, we are interested to determine whether two raters agree in the rating of a subject on a qualitative scale (disease, no disease). Table VIII shows the responses of two raters on 100 subjects on their disease-status in a SPSS dataset.

Table VIII. Inter-raters' dataset in SPSS.

Rater_a	Rater_b	Count
Disease	Disease	40
Disease	No disease	5
No disease	Disease	5
No disease	No disease	50

To measure the agreement between the two raters, Kappa[2] estimate is used. Kappa lies between zero to one and has similar interpretation as the strength of correlation given in Table II. To obtain Kappa from

SPSS, firstly **weight the count**[3], then go to ***Analyse, Descriptive statistics, Crosstab - choose Statistics and tick on Kappa***, see template 3.

Template 3

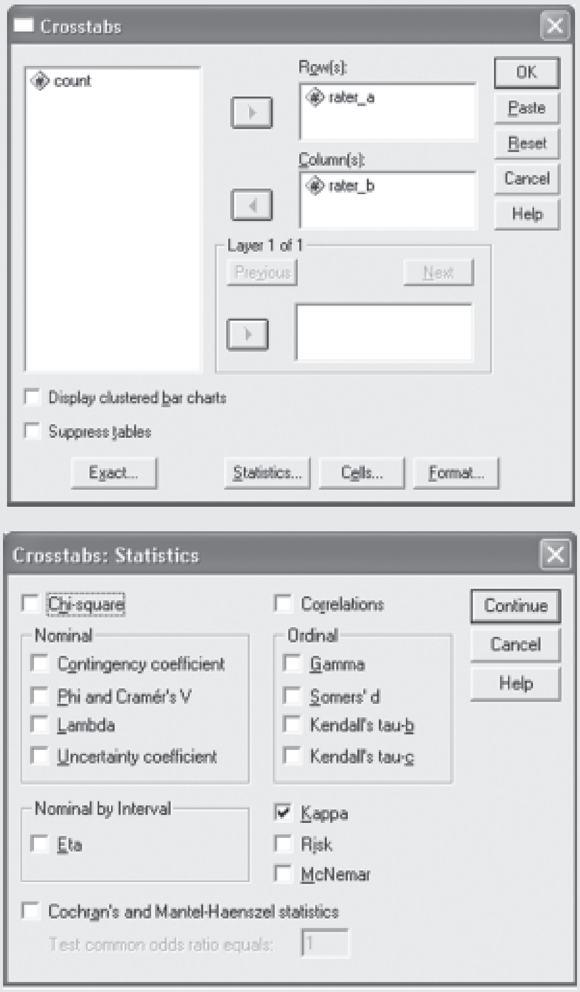


Table IX gives a Kappa value of 0.798 (p<0.001), a strong agreement between the two raters since they agree on 90% of the subjects.

Table IX. Kappa estimate.

Symmetric Measures

	Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement Kappa	.798	.061	7.980	.000
N of Valid Cases	100			

a. Not assuming the null hypothesis
b. Using the asymptotic standard error assuming the null hypothesis

LIMITATION OF KAPPA

Table X shows the same two raters agreeing on 90% of the subjects but this time 85% on no-disease and 5% disease but the Kappa value is only 0.444 (and sometimes could be negative!). See Table XI. This happens because one of the agreed-category has a small percentage.

Table X.

RATER_A * RATER_B Crosstabulation

Count

		RATER_B		
		no disease	disease	Total
RATER_A	no disease	85	5	90
	disease	5	5	10
Total		90	10	100

Table XI. Kappa estimate.

Symmetric Measures

	Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement Kappa	.444	.147	4.444	.000
N of Valid Cases	100			

a. Not assuming the null hypothesis
b. Using the asymptotic standard error assuming the null hypothesis

Table XII.

Rater B	Rater A		
	1	2	Total
1	A	B	B1
2	C	D	B2
Total	A1	A2	N

Gwet[4] proposed an alternative statistic referred as AC1-statistic which is more consistent with the percentage of agreement between raters in all situations and is given by (using Table XII)

$$AC1 = \frac{p - ø}{1 - ø} \text{ where } p = \frac{A + D}{N} \text{ and } ø = 2q(1-q),\ q = \frac{A1 + B1}{2N}$$

Using the values from Table X, the AC1-statistic for agreement is 0.8780 (which is more consistent with the two raters having the same rating for 90% of the subjects).

CONCLUSIONS

This article, in a way, ties up the univariate techniques of analyses for both quantitative and qualitative data. In such analyses, we have not taken into account the effect of other variables (confounders/covariates) except for the technique of partial correlation.

In the next article, we will discuss the multivariate technique of analysis for the regression model of quantitative outcomes: Biostatistics 201 – Linear Regression Analysis.

REFERENCES

1. Bland JM & Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement, Lancet, February, 1986; 307-10.
2. Cohen J. A coefficient of agreement for nominal scales, Educational and Psychological Measurement. 1960; 20:37-46.
3. Chan YH, Biostatistics 103: Qualitative Data —Tests of Independence. SMJ 2003; Vol 44 (10):498-503.
4. Gwet K. Handbook of inter-rater reliability. STATAXIS Publishing Company 2001.