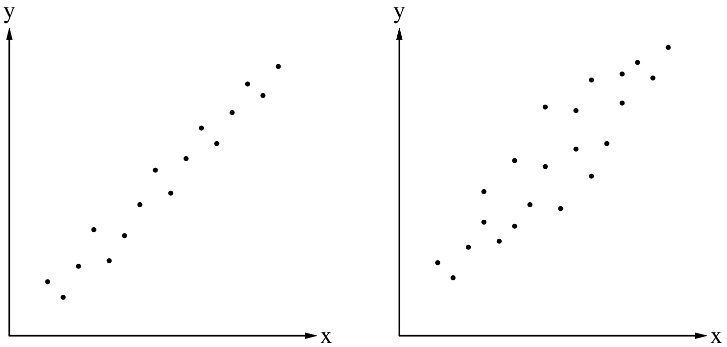
University of Waterloo W. H. Cherry

CORRELATION: Its Properties and Its Interpretation

This Highlight #66 describes correlation – what it is, how it works, how to calculate it, its properties and how to assess it.

1. What is Correlation?

A useful way of *looking* at *bi*variate data is with a scatter diagram – Cartesian axes with dots whose coordinates are the x and y values of each bivariate observation. The two scatter diagrams at the right show that two data sets with roughly the *same* average and *same* standard deviation for x and for y can visually *differ* from each other – the points in the diagram on the left are *more tightly clustered* than those in the diagram on the right. 'Tightness of clustering' on a scatter diagram is a convenient way to introduce correlation.



NOTES: 1. The *magnitude* of correlation *in*creases with tightness of clustering and its value ranges from –1 to +1.

- Tightness of clustering implies most obviously clustering about a *straight line*; thus, from the start, we associate correlation with a *straight-line* (or *linear*) relationship between two variates.
- The *sign* of correlation is that of the *slope* of the straight-line relationship between the two variates – correlation is positive when this slope is positive and negative when it is negative; correlation is zero if the slope is zero (*exactly* zero is a rare occurrence in practice).

2. For *uni*variate data sets, the sum of the observations and the sum of their squares are familiar initial numerical data summaries, but the two scatter diagrams above imply that four such summaries – two for x and two for y – are *not* enough for a bivariate data set; a *fifth* numerical summary is needed – it is the *sum of products.*

- We learn overleaf in Section 3 on page HL66.2 how these *five* numerical summaries of bivariate data – Σx_j, Σx_j^2, Σy_j, Σy_j^2 and $\Sigma x_j y_j$ – are involved in calculating the *value* of correlation.

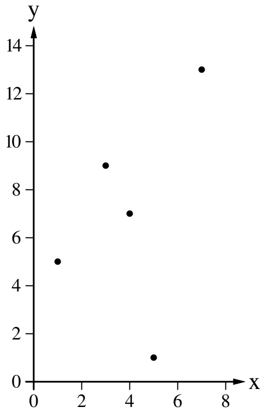
2. How Does Correlation Work Numerically?

To understand the value of correlation (denoted r) and how its sign follows that of the slope of the straight-line relationship between the two variates, we first look at equation (HL66.1) at the right – correlation is the '*average*' *product of the* n *standardized* x *and* y *variate values.* For the data set [numbered (HL66.2)] of five bivariate observations given at the right, the table in Example HL66.1 below shows how equation (HL66.1) yields a value of 0.4 for r. For interest, a scatter diagram of these data is also shown at the lower right.

$$r = \frac{1}{n-1}\sum_{j=1}^{n}\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right) \qquad \text{-----(HL66.1)}$$

x	1	3	4	5	7
y	5	9	7	1	13

-----(HL66.2)



Example HL66.1:

	x	y	$\frac{x-\bar{x}}{s_x}$	$\frac{y-\bar{y}}{s_y}$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
	1	5	$-3/\sqrt{5}$	$-2/\sqrt{20}$	6/10 = 0.6
	3	9	$-1/\sqrt{5}$	$-2/\sqrt{20}$	−2/10 = −0.2
	4	7	$0/\sqrt{5}$	$-2/\sqrt{20}$	0/10 = 0
	5	1	$1/\sqrt{5}$	$-2/\sqrt{20}$	−6/10 = −0.6
	7	13	$3/\sqrt{5}$	$-2/\sqrt{20}$	18/10 = 1.8
Average:	*4*	*7*		*Sum*:	$1.6 \div 4 \Rightarrow r = 0.4$
Standard deviation:	$\sqrt{5}$	$\sqrt{20}$			

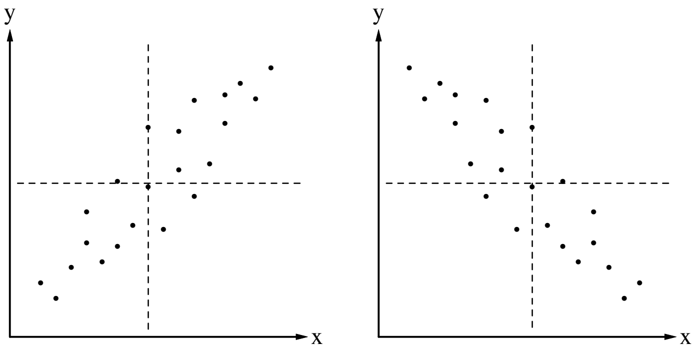
Next, we look at the two scatter diagrams given overleaf on page HL66.2, in which the additional sets of dashed axes have their origin at the point of averages ($\bar{x}$, $\bar{y}$) in both diagrams.

∗ In the left-hand diagram where the points cluster around a line with a *positive* slope, we see the points lie predominantly in the first and third quadrants of the set of dashed axes; the standardized x and y values of most of the points will therefore either both be *positive* (points in the first quadrant) or both be *negative* (points in the third quadrant), in either case yielding mainly *positive* products and, hence, a positive value of r.

∗ In the right-hand diagram where the points cluster around a line with a *negative* slope, the points lie predominantly in the second and fourth quadrants of the set of dashed axes; the standardized x and y values of most of the points will therefore be of *opposite* sign (*e.g.*, in the third quadrant, x is below its average while y is above its average), yielding mainly *negative* products and, hence, a negative value of r.

(*continued overleaf*)

It follows from this discussion that if the points of a scatter diagram show *no* relationship between x and y (*i.e.*, the points shown no obvious pattern, sometimes referred to as being 'randomly scattered'), the points will fall in roughly *equal* numbers in all four quadrants of axes with their origin at the points of averages ($\overline{x}$, $\overline{y}$); the average of the products of the standardized x and y values, and hence the value of r, will then be zero or close to it.



It is a useful statistical skill to be able to assess the approximate value of r from *looking* at a scatter diagram of bivariate data; to assist with acquiring this skill, pages HL66.10 to HL66.15 of this Highlight #66 show scatter diagrams for a variety of values of r. The axes in these diagrams are labelled u and v, instead of x and y, to avoid implying the variable on the horizontal axis is necessarily an explanatory variate and the variable on the vertical axis is a response variate.

3. Calculating Correlation in Practice

Equation (HL66.1) is useful for *understanding* correlation but is not a convenient way to *calculate* the value of r from data; instead, we use equation (HL66.3) given at the right.

$$r = \frac{\sum_{j=1}^{n} x_j y_j - (\sum_{j=1}^{n} x_j)(\sum_{j=1}^{n} y_j)/n}{\sqrt{[\sum_{j=1}^{n} x_j^2 - (\sum_{j=1}^{n} x_j)^2/n][\sum_{j=1}^{n} y_j^2 - (\sum_{j=1}^{n} y_j)^2/n]}} \quad \text{-----(HL66.3)}$$

For the data set [numbered (HL66.2)] of five bivariate observations given overleaf on page HL66.1, the values of the five numerical summaries are: $\Sigma x_j = 20$, $\Sigma x_j^2 = 100$, $\Sigma y_j = 35$, $\Sigma y_j^2 = 325$, $\Sigma x_j y_j = 156$; substituted into equation (HL66.3), these values give $r = 0.4$. Equation (HL66.3) can also be written in the form (HL66.4), where $SS_{xy} = 16$ is the expression in the numerator of equation (HL66.3) and $SS_x = 20$ and $SS_y = 80$ are the respective expressions in square brackets under the square root in the denominator of equation (HL66.3).

$$r = \frac{SS_{xy}}{\sqrt{SS_x \times SS_y}} \quad \text{-----(HL66.4)}$$

NOTES: 3. Having two forms of the symbolic expression for r – equation (HL66.1) to understand the idea and equation (HL66.3) to calculate the value of r from data – is reminiscent of (data) standard deviation.

- Recall that SS_y also occurs in the expression for calculating the (data) standard deviation which estimates σ in the response model: $Y_j = \mu + R_j$, $j = 1, 2, ..., n$, $R_j \sim N(0, \sigma)$, independent, EPS.

4. The divisor of n−1 in equation (HL66.1), which produces the 'average' of the products of the standardized x and y values, is *absent* from equation (HL66.3); in fact, this divisor is present in equation (HL66.1) *only* because this same divisor occurs (under a square root) in both s_x and s_y.

5. Using the symbols n, x, y and s in the foregoing discussion denotes the context of a *sample*; however, if the symbols were changed to **N**, **X**, **Y** and **S** and i instead of j used as the index of sigma signs [for instance, in equations (HL66.1) and (HL66.3) and the data table (HL66.2) overleaf on page HL66.1], the discussion would apply to the *respondent population*, whose correlation we denote **R** – see Appendix 4 on page HL66.8 for a notation summary.

6. **R** and r are also referred to as correlation *coefficients* – we use only 'correlation' in this Statistical Highlight #66.

4. Properties of Correlation

In addition to thinking of correlation as quantifying tightness of clustering about the straight-line trend of points on a scatter diagram and the sign of correlation being that of the slope of this straight line, equation (HL66.1) overleaf on page HL66.1], which defines correlation as the 'average' product of the standardized x and y values, gives r the following properties:

✻ although it can be calculated only for *quantitative* (not categorical) variates, r is unitless (or dimensionless);

✻ the value of r, which lies in the interval [–1, 1], is *un*affected by the units in which x and y are expressed – that is, the value of r is unaffected by the *scaling* of x or y;

✻ the value of r is *un*affected if a constant is added to or subtracted from all values of x and/or y – recognizing this property may be important when x and y are the *same* quantity measured by different processes or at different times;

✻ when r is 1 or –1 (a rare occurence in practice), the points of the scatter diagram lie *exactly* on a straight line with, respectively, positive or negative slope;

✻ the expression for r is *symmetrical* in x and y so the value of correlation is *un*affected by the designation of one variate as the response and the other as the explanatory variate;

✻ when r is positive, it means the values of x and y tend to increase or decrease together but, when r is negative, one variate tends to increase as the other decreases.

(*continued*)

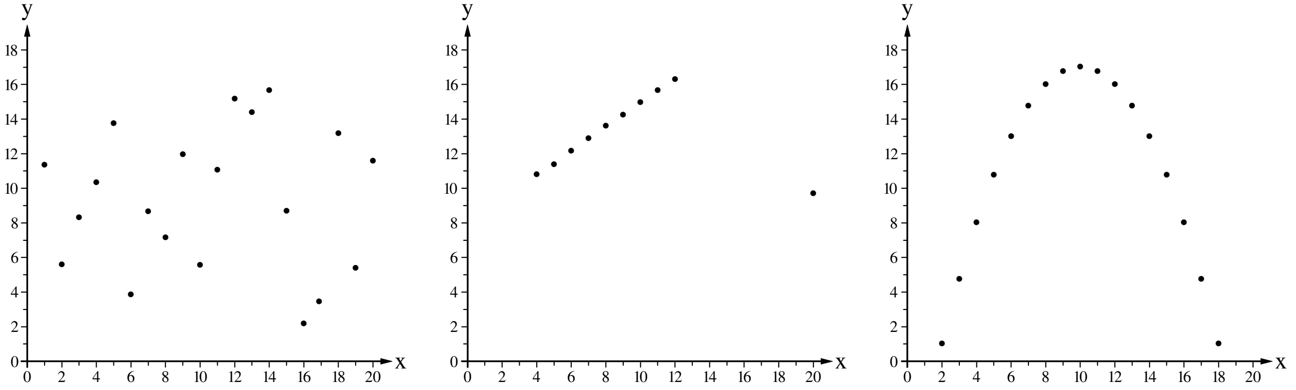
University of Waterloo W. H. Cherry

CORRELATION: Its Properties and Its Interpretation (continued 1)

5. Interpreting Correlation Correctly: Influential Observations, Outliers and Non-linear Relationships

The discussion in Sections 1-4 is concerned with starting with a scatter diagram and quantifying *one* characteristic of what we see by calculating the correlation. Difficulties may arise when this process is *reversed* to try to answer a Question about the appearance of a scatter diagram (*i.e.*, to answer a Question about an x-y relationship) from the value of the correlation; one value of note in this regard is a correlation of zero or close to it. The potential for a wrong answer is illustrated by the three diagrams below, all of which have (exactly) zero correlation but show very different x-y relationships. [The corresponding data sets and their five numerical summaries are given below the diagrams.] We note that:

- ∗ the left-hand diagram, with no obvious pattern, is what comes most naturally to mind when r is zero or close to it, making it easier to overlook possibilities like the other two diagrams;
- ∗ in the middle diagram, the correlation for the first 9 points is $r = 0.999\,769$, as we would anticipate for points we *see* lie almost *on* a straight line with positive slope – the overall correlation of 0 is, of course, a consequence of the influential observation at $x = 20$, and this overall value of $r = 0$ conveys little useful information about the appearance of the scatter diagram or the x-y relationship(s);
- ∗ the 17 points in the right-hand diagram lie on the parabola: $4y = -x^2 + 20x - 32$, reminding us that correlation quantifies tightness of clustering about a *straight line* – the respective correlations for the first and the last 9 points in this diagram are $\pm 0.962\,158$.



x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
y	11.34	5.58	8.31	10.32	13.74	3.84	8.64	7.14	11.94	5.55	11.04	15.15	14.37	15.66	8.55	2.16	3.45	13.17	5.37	11.58

$\Sigma x_j = 210$, $\Sigma x_j^2 = 2{,}870$,
$\Sigma y_j = 186.9$, $\Sigma y_j^2 = 2{,}067.0084$,
$\Sigma x_j y_j = 1{,}962.45$;

x	4	5	6	7	8	9	10	11	12	20
y	10.78	11.38	12.16	12.88	13.60	14.24	14.94	15.66	16.28	9.68

$\Sigma x_j = 92$, $\Sigma x_j^2 = 1{,}036$, $\Sigma y_j = 131.6$, $\Sigma y_j^2 = 1{,}774.3904$, $\Sigma x_j y_j = 1{,}210.72$;

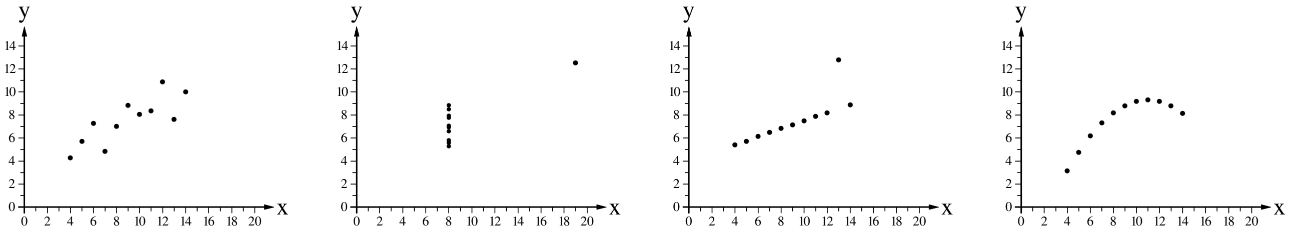
x	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
y	1	4¾	8	10¾	13	14¾	16	16¾	17	16¾	16	14¾	13	10¾	8	4¾	1

$\Sigma x_j = 170$, $\Sigma x_j^2 = 2{,}108$, $\Sigma y_j = 187$, $\Sigma y_j^2 = 2{,}541.5$, $\Sigma x_j y_j = 1{,}870$.

The same ideas are illustrated by the four diagrams below, where:

- ∗ r is about 0.66665 for each diagram (instead of 0 as in the three diagrams above);
- ∗ the second diagram is another instance of correlation providing little useful information about the appearance of a scatter diagram, because the value of r is determined largely by the influential observation at $x = 19$;
- ∗ the deviant observation in the third diagram below at $x = 13$ is an *outlier*, rather than an *influential observation* as in the second diagram at $x = 19$ or in the middle diagram above at $x = 20$ – the correlation for the *other* 10 points in the third diagram below is $r = 0.999\,607$, as we would again anticipate for points we *see* lie almost *on* a straight line with positive slope.

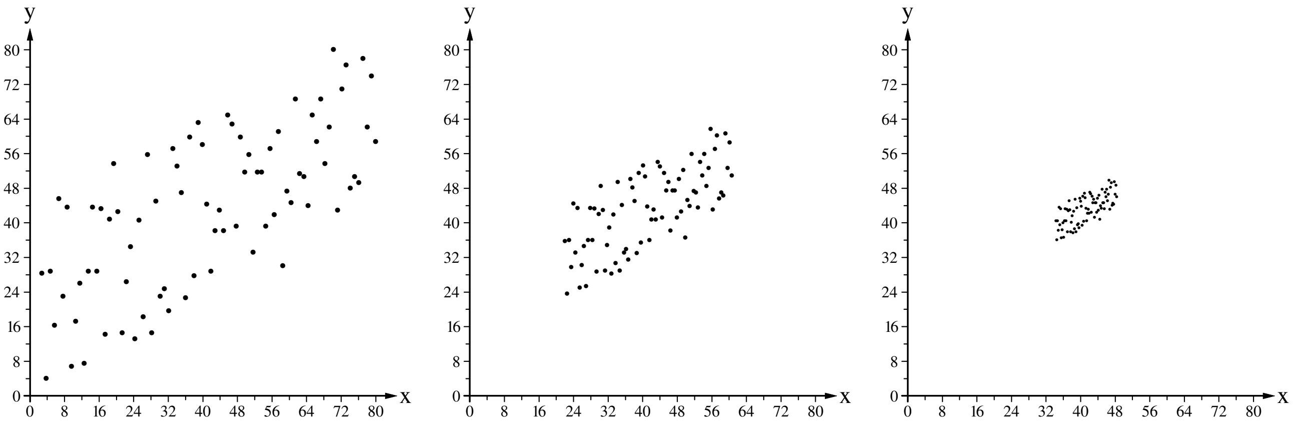
[The data sets for the points in these four diagrams, together with other matters they illustrate, are given in Statistical Highlight #53.]



(*continued overleaf*)

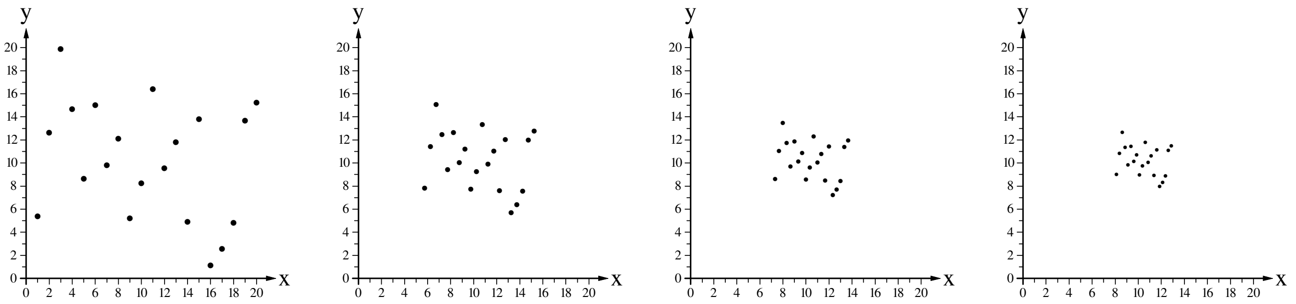
6. Assessing Correlation Correctly: Number of Observations and Scaling

Another use of correlation is in visually comparing the tightness of clustering of two (or more) scatter diagrams to answer, for example, a Question about which exhibits the stronger x-y relationship. As the diagrams below illustrate, when making such comparisons, we must be careful about the number of observations, the scaling of the diagrams and the size of the points.

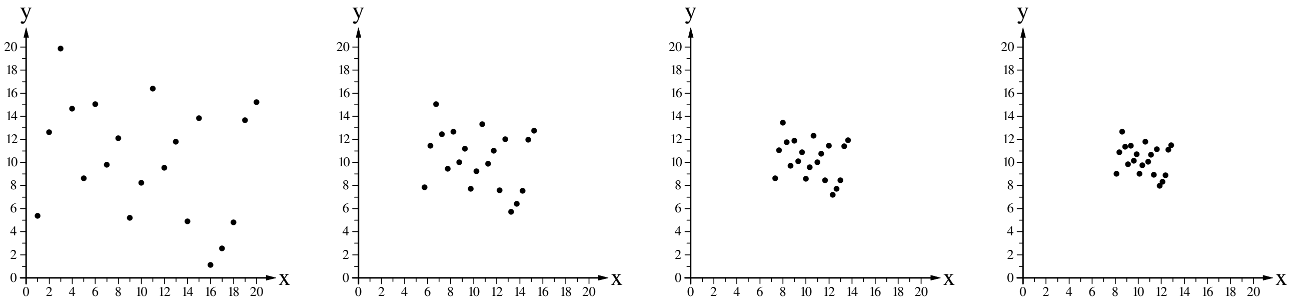


The three scatter diagrams above each contain 80 points with correlation 0.7 but, in the middle and right-hand diagrams, the point cloud had been shrunk in both x and y directions by respective factors of 2 and 5.5. The misperception of *in*creasing tightness of clustering across the three diagrams is obvious, warning us that scatter diagrams need to be of (roughly) the *same* size for reliable comparison of tightness of clustering and correlation. [A reducing/enlarging photocopier can be useful in achieving comparability of size.]

The four scatter diagrams immediately below each contain 20 points with correlation −0.25 and with the point cloud of the left-hand diagram shrunk successively in both x and y directions by factors of 2, 3 and 4 in the three diagrams to its right; misperception of increasing magnitude of correlation with decreasing cloud size is again apparent.



To illustrate how *point* size impacts *cloud* size, these four diagrams are shown again below but with the *same* point size in each diagram, instead of the decreasing point size used across the diagrams above.



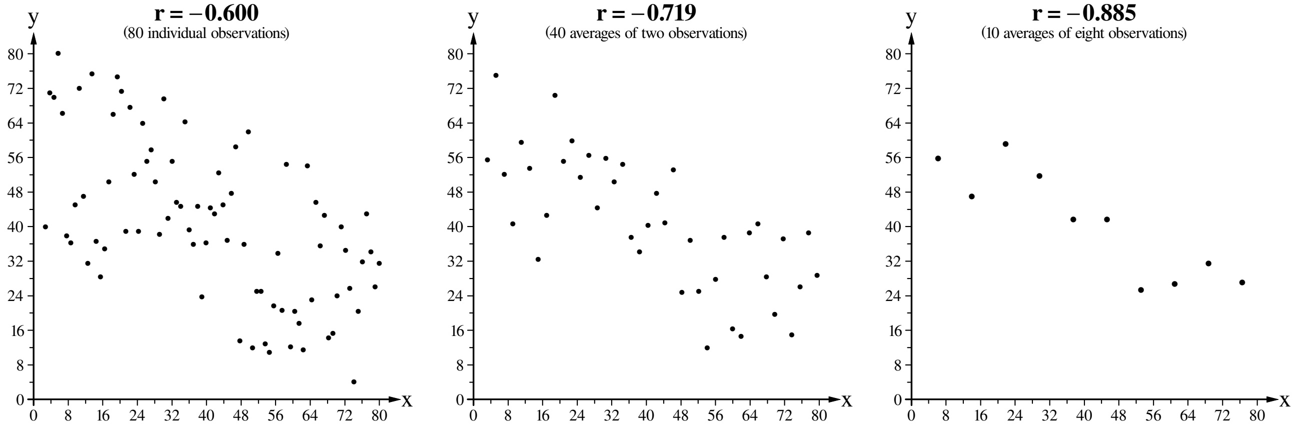
Visual perception of tightness of clustering and, hence, of the value of correlation, can also be affected by the *slope* of the line about which the points are clustered; this matter is discussed in Section 8, starting at the top of page HL66.6 of this Highlight #66.

7. Assessing Correlation Correctly: Scatter Diagrams of *Averages*

Scatter diagrams in which the points are *averages* typically show tighter clustering than diagrams where the points are for the corresponding *individuals* – that is, averaging usually *increases the magnitude* of correlation. This phenomenon is illustrated in fifteen diagrams on the facing page HL66.5.

(*continued*)

CORRELATION: Its Properties and Its Interpretation (continued 2)



r = −0.600
(80 individual observations)

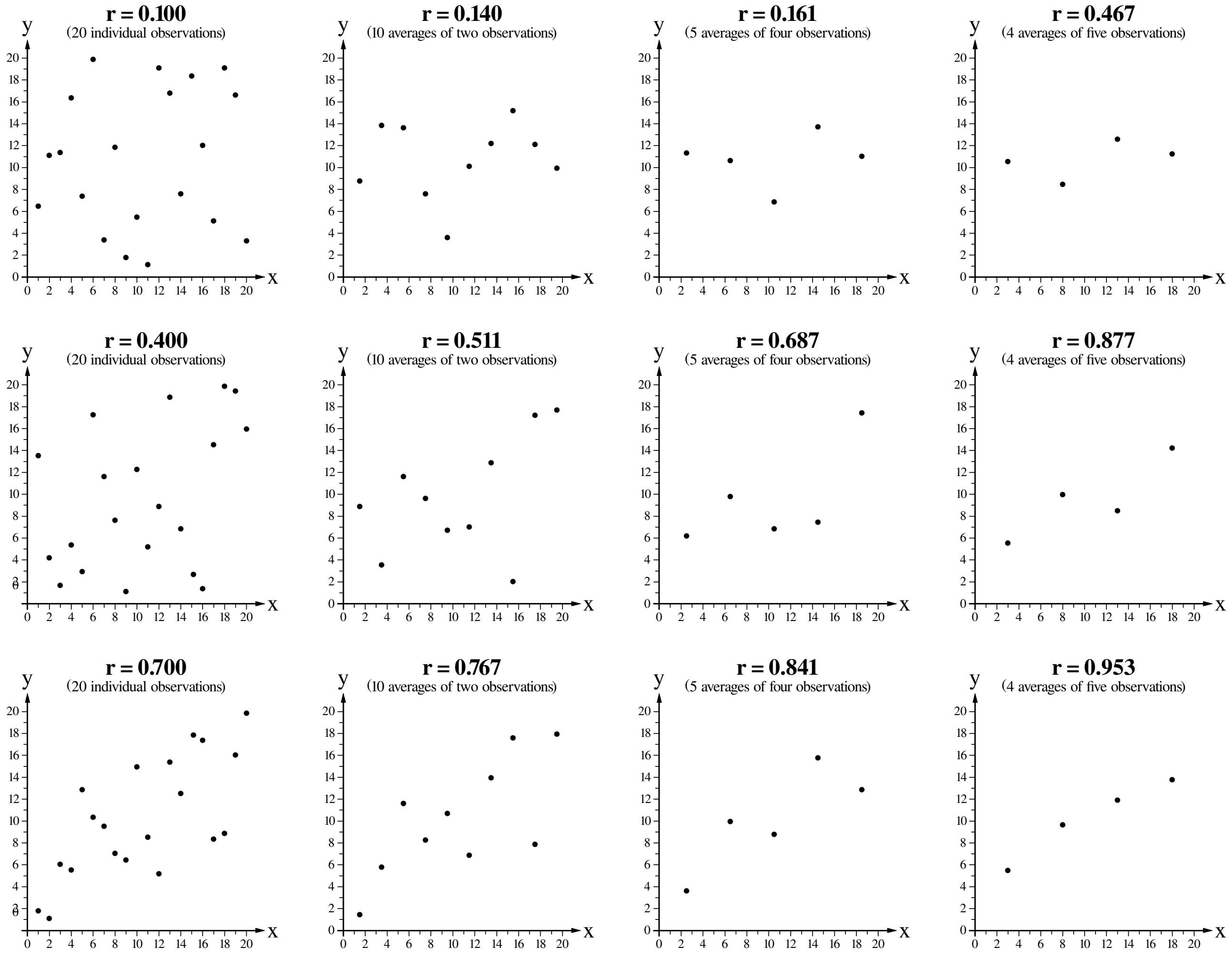
r = −0.719
(40 averages of two observations)

r = −0.885
(10 averages of eight observations)

The left-hand diagram above contains 80 points with correlation −0.6 and then the observations for successive sets of two and of eight points are averaged to produce the middle and right-hand diagrams of 40 and 10 points, respectively; these point clouds have correlations of about −0.72 and −0.88.

The three left-hand diagrams below contain 20 points with respective correlations of 0.1, 0.4 and 0.7, and then successive sets of two, four and five points are averaged to produce the points in the other sets of three diagrams with 10, 5 and 4 points; these have respective correlations of about 0.14, 0.16 and 0.47 in the upper set of diagrams, 0.51, 0.69 and 0.88 in the middle set, and 0.77, 0.84 and 0.95 in the bottom set. These are particular instances, for correlations of 0.1, 0.4 and 0.7 for the *individual* observations, of the increase in correlation magnitude (*i.e.*, increase in tightness of clustering) produced by averaging.

In sociology and related human sciences, the **ecological fallacy** refers to using inappropriately an exaggerated correlation based on averages to answer a Question of interest.



r = 0.100
(20 individual observations)

r = 0.140
(10 averages of two observations)

r = 0.161
(5 averages of four observations)

r = 0.467
(4 averages of five observations)

r = 0.400
(20 individual observations)

r = 0.511
(10 averages of two observations)

r = 0.687
(5 averages of four observations)

r = 0.877
(4 averages of five observations)

r = 0.700
(20 individual observations)

r = 0.767
(10 averages of two observations)

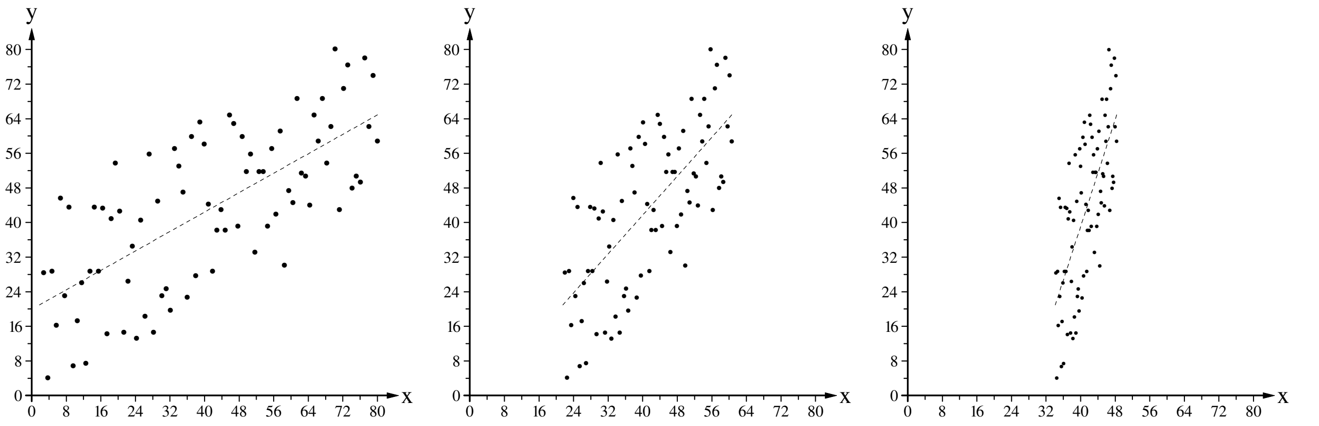
r = 0.841
(5 averages of four observations)

r = 0.953
(4 averages of five observations)

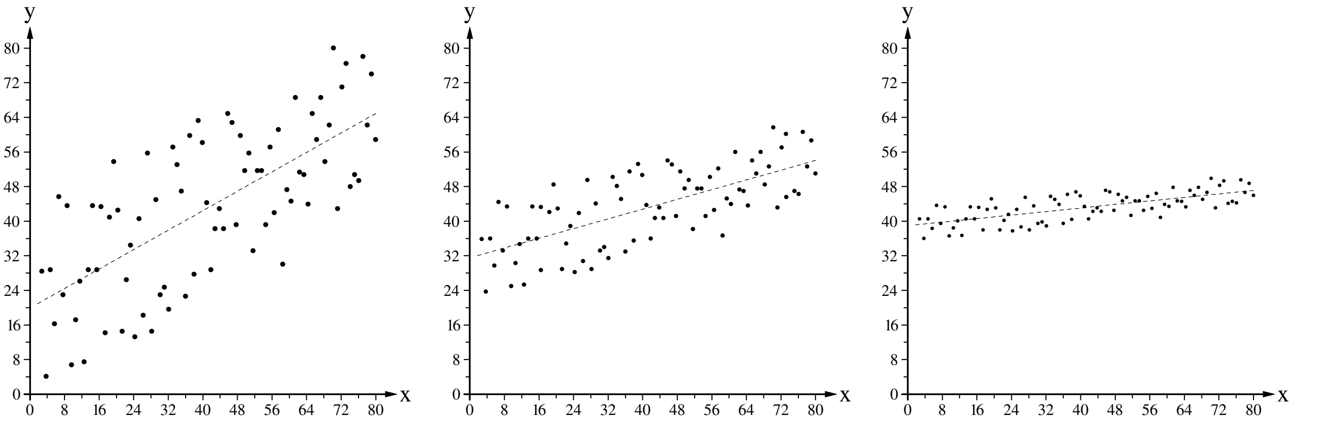
(*continued overleaf*)

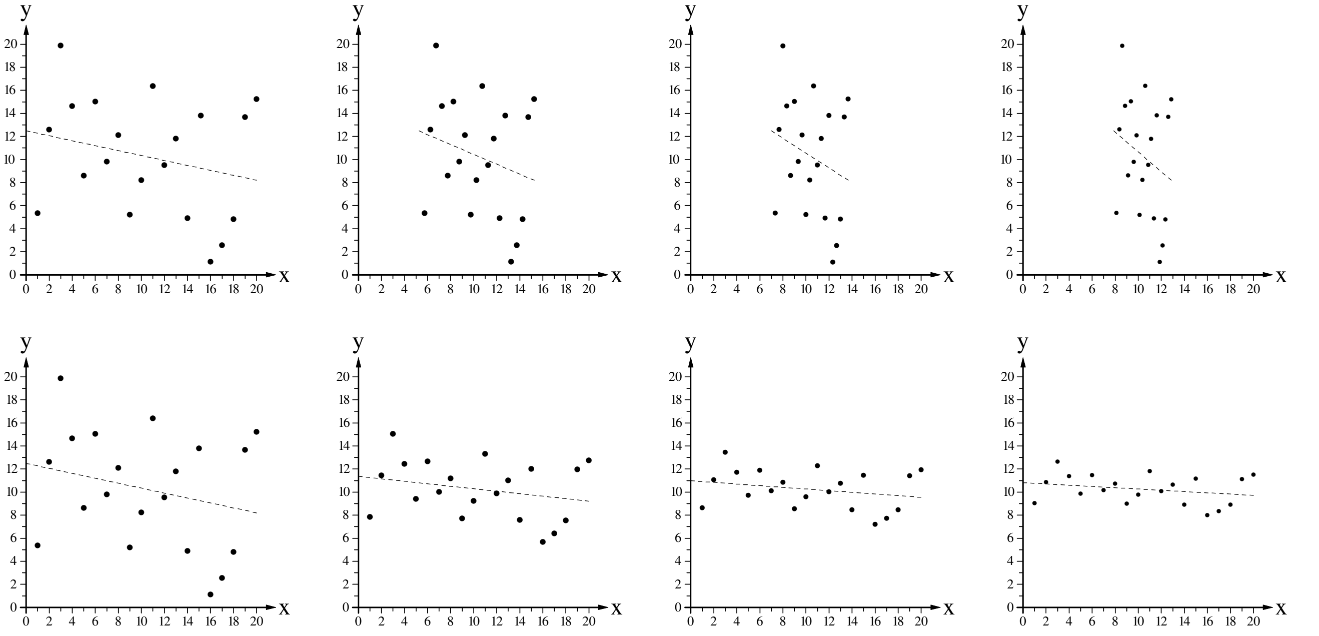
8. Assessing Correlation Correctly: Correlation and Slope

Correlation is tightness of clustering about a straight line so matters which involve correlation often also involve this line – the line of ‘best’ fit to bivariate data – which can be displayed by superimposing it on the corresponding scatter diagram.



The three scatter diagrams above each contain 80 points with correlation 0.7 but, in the middle and right-hand diagrams, the point cloud has been shrunk in the x direction by respective factors of 2 and 5.5; the same is true of the three scatter diagrams immediately below, except the point cloud has been shrunk in the y direction in the middle and right-hand diagrams. Thus, these six scatter diagrams show the *same* point cloud – the *difference* is the slope of the line of ‘best’ fit to the points, which is shown as a dashed line in each diagram. These matters are illustrated again by the (*continued on the facing page HL66.7*)





(*continued*)

CORRELATION: Its Properties and Its Interpretation (continued 3)

seven versions of ths 20-point cloud in the two sets of four smaller scatter diagrams on the lower half of the facing page HL66.6. The changes from the larger upper diagrams are that the correlation is −0.25 (instead of 0.7) and the shrinkage factors in the x direction and the y direction in the three right-hand diagrams of each set are 2, 3 and 4 respectively (instead of 2 and 5.5).

The fourteen diagrams on the facing page HL66.6 warn us that visual assessment of the tightness of clustering of a point cloud is affected by *differing dimensions* of the cloud (*i.e.*, *differing amounts of variation*) in the x and y directions, which distorts its 'true' slope – as the difference in dimensions (or variation) increases, it becomes easier to overestimate the magnitude of the correlation. Accurate assessment is easiest when the variation in x and in y occupy a *similar* distance on each axis.

$$\hat{\beta}_1 = r \times \frac{s_y}{s_x} \qquad \text{-----(HL66.5)}$$

$$r \simeq \hat{\beta}_1 \text{ when } s_x \simeq s_y \qquad \text{-----(HL66.6)}$$

An explanation of this matter is that (the estimate of) the slope of the line of 'best' fit (here denoted $\hat{\beta}_1$, which is *roughly* what the eye interprets as the 'slope' of the point cloud), is related to the correlation (r) as in equation (HL66.5) at the right; that is, the relationship between slope and correlation involves the quotient of the standard deviations of y and x as a multiplicative factor. Visual assessment of a scatter diagram is more straight-forward if this ratio is close to one; we can accomplish this state of affairs by making the variation in x and the variation in y occupy similar distances on their respective axes of the scatter diagram. When we look at such a diagram, where equation (HL66.6) applies, the extremes of what we may see are:

- a point cloud that is roughly square and the eye sees little evidence of clustering (*i.e.*, a correlation close to zero) between x and y, reinforced by an approximately horizontal (imagined) line of 'best' fit to the points; **OR**
- a point cloud that shows strong clustering about an (imagined) line of 'best' fit with a slope of approaching one (either positive or negative, depending on whether the upper left and lower right, or the upper right and lower left, regions of the scatter diagram have no points) and the eye sees clear evidence of strong correlation between x and y.

Most cases we encounter in practice lie between these extremes.

NOTE: 7. A matter of incidental interest is that, when estimating from *sample* data, *equiprobable* selecting is needed for correlation but data concentrated at the *ends* of the interval of x values increase precision for estimating slope.

9. Appendix 1: Sections 6 and 8 Data Sets

The two data sets used to construct the scatter diagrams in Section 6 (on page HL66.4 of this Highlight #66) and Section 8 (on the facing page HL66.6), are given below although, in the diagrams, the y-axis labels have been rescaled and r has changed sign.

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
y	1.99	2.71	1.98	2.35	1.48	2.15	1.54	2.63	2.32	2.06	2.61	1.98	1.54	1.98	1.55	2.41	1.62	1.24	1.57	2.40	2.05	1.81	2.44	1.63	2.29	1.18	2.40	1.50	2.15	2.10

x	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
y	2.25	1.14	1.26	1.44	2.16	1.06	2.01	0.96	1.11	1.52	1.98	1.70	1.56	1.70	0.91	0.97	1.67	1.06	1.30	1.18	1.85	1.30	1.30	1.67	1.14	1.59	1.02	1.94	1.43	1.51

x	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
y	0.80	1.31	1.33	1.53	0.91	1.09	0.80	1.24	0.99	0.46	1.56	0.73	0.55	1.41	1.33	1.37	0.52	0.99	0.64	1.09

$\Sigma x_j = 3{,}240$, $\Sigma x_j^2 = 173{,}880$, $SS_{xy} = -693.83$,
$\Sigma y_j = 124$, $\Sigma y_j^2 = 215.2298$, $SS_x = 42{,}660$,
$\Sigma x_j y_j = 4{,}328.17$, $r = -0.699\,999$, $SS_y = 23.0298$;

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
y	2.1	1.33	0.46	1.086	1.81	1.04	1.666	1.392	2.22	1.856	0.878	1.7	1.428	2.256	1.186	2.71	2.538	2.266	1.202	1.016

$\Sigma x_j = 210$, $\Sigma x_j^2 = 2{,}870$, $SS_{xy} = 17.14$,
$\Sigma y_j = 32.24$, $\Sigma y_j^2 = 59.037\,512$, $SS_x = 665$,
$\Sigma x_j y_j = 355.66$, $r = 0.250\,031$, $SS_y = 7.066\,632$.

10. Appendix 2: Section 7 Data Sets

The four data sets used to construct the scatter diagrams in Section 7 (on pages HL66.4 and HL66.5 of this Highlight #66) are given below although, in the diagrams, the y-axis labels have been rescaled and r has changed sign.

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
y	1.65	0.73	0.76	0.46	0.87	1.71	1.76	1.50	0.70	1.44	1.90	0.60	1.75	1.99	1.80	1.34	0.88	0.62	0.72	1.68	0.83	1.29	1.68	0.94	1.20	1.12	1.34	1.70	0.77	1.59

x	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
y	1.20	1.48	1.51	0.93	1.67	1.77	1.51	2.13	1.76	1.52	1.56	1.28	1.50	1.74	1.42	1.10	2.43	1.77	1.00	2.48	2.09	2.09	2.45	2.51	2.19	1.83	2.22	1.22	2.47	2.23

x	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
y	2.31	2.49	1.23	2.15	1.48	1.78	1.57	2.41	2.38	2.12	1.65	1.81	2.07	2.71	2.23	1.89	1.56	1.82	2.06	1.90

$\Sigma x_j = 3{,}240$, $\Sigma x_j^2 = 173{,}880$, $SS_{xy} = 595.23$,
$\Sigma y_j = 130$, $\Sigma y_j^2 = 234.3198$, $SS_x = 42{,}660$,
$\Sigma x_j y_j = 5{,}860.23$, $r = 0.600\,002$, $SS_y = 23.0698$;

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
y	2.07	1.51	1.48	0.88	1.96	0.46	2.44	1.42	2.63	2.19	2.71	0.55	0.83	1.93	0.64	1.40	2.23	0.55	0.85	2.45

$\Sigma x_j = 210$, $\Sigma x_j^2 = 2{,}870$, $SS_{xy} = -8.59$,
$\Sigma y_j = 31.18$, $\Sigma y_j^2 = 59.6964$, $SS_x = 665$,
$\Sigma x_j y_j = 318.8$, $r = -0.100\,041$, $SS_y = 11.086\,78$.

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
y	1.22	2.34	2.64	2.20	2.49	0.77	1.45	1.93	2.71	1.37	2.22	1.78	0.58	2.02	2.52	2.68	1.10	0.46	0.51	0.93

$\Sigma x_j = 210$, $\Sigma x_j^2 = 2{,}870$, $SS_{xy} = -35.24$,
$\Sigma y_j = 33.92$, $\Sigma y_j^2 = 69.208$, $SS_x = 665$,
$\Sigma x_j y_j = 320.92$, $r = -0.399\,862$, $SS_y = 11.679\,68$.

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
y	2.63	2.71	2.12	2.18	1.30	1.60	1.70	2.00	2.07	1.05	1.82	2.22	1.00	1.34	0.70	0.76	1.84	1.78	0.92	0.46

$\Sigma x_j = 210$, $\Sigma x_j^2 = 2{,}870$, $SS_{xy} = -50.76$,
$\Sigma y_j = 32.2$, $\Sigma y_j^2 = 59.7512$, $SS_x = 665$,
$\Sigma x_j y_j = 287.34$, $r = -0.699\,914$, $SS_y = 7.9092$.

(*continued overleaf*)

11. Appendix 3: Correlation in Probability Theory

As with standard deviation, learning the ideas can be easier if we distinguish *data* correlation, the topic of the foregoing discussion in this Highlight #66, and *probabilistic* correlation (discussed, for example, in STAT 230); an overview of the latter follows.

The (probabilistic) correlation (here denoted *cor* but possibly ρ elsewhere) of the random variables U and V is given by equation (HL66.7) at the right; the three probabilistic measures of variation on the right-hand side of this equation are defined in terms of probabilistic expectation in equations (HL66.8) to (HL66.10). [*Expectation* is the average of the values of a random variable, each value being weighted by its probability in calculating the average; this operation is denoted E here. This average is called the *mean* of the random variable.]

$$cor(U,V) = \frac{cov(U,V)}{s.d.(U) \times s.d.(V)} \quad \text{-----(HL66.7)}$$

$$cov(U,V) = E(UV) - E(U)E(V) \quad \text{-----(HL66.8)}$$

$$s.d.(U) = \sqrt{E[U-E(U)]^2} \quad \text{-----(HL66.9)}$$

$$s.d.(V) = \sqrt{E[V-E(V)]^2} \quad \text{----(HL66.10)}$$

If the random variables U and V are *probabilistically independent*, $cor(U,V)$ is *zero* because $cov(U,V) = 0$; the converse is *not* true – zero correlation does *not* imply probabilistic independence. This distinction is often given prominence in introductory probability teaching; the idea behind it is illustrated by the first and third larger scatter diagrams in Section 5 near the middle of the third side (page HL66.3) of this Highlight #66: r = 0 can mean no relationship *or* a non-linear relationship. To remind us of the distinction, when all we know is $cor(U,V) = 0$, we refer to the random variables as *uncorrelated*, a (much) *weaker* term than probabilistically independent.

12. Appendix 4: Notation Summary

For those using this Highligh #66 on a stand-alone basis, a summary of notation is given in Table HL66.1 at the right; the following conventions are involved although only the first and third occur in the table:

- lower-case roman letters for sample quantities and data;
- lower-case *italic* letters for values of random variables;
- upper-case **bold** letters for respondent population quantities; the line (called 'cross') through these symbols is to distinguish them (in handwriting) from:
- upper-case *italic* letters for random variables (see Appendix 3 above).

Table HL66.1

SYMBOL		**DESCRIPTION**
Sample	**Respondent Population**	
r	**R̵**	(Data) correlation (coefficient)
n	**N̵**	Number of elements (or 'size')
y	**Y̵**	Response variate
x	**X̵**	Explanatory variate
$\overline{\mathrm{x}}$, $\overline{\mathrm{y}}$	$\overline{\textbf{X̵}}$, $\overline{\textbf{Y̵}}$	Average (sum ÷ number)
s, $\hat{\sigma}$	**S̵**	(Data) standard deviation
j	i	Summation index
$\hat{\beta}_1$	**B̵**$_1$	Slope of line of 'best' fit to points

NOTE: 8. It is unfortunate the symbol r is now widely used for (data) correlation; c would have been a more evocative choice originally and would have left r for a ratio (*e.g.*, in survey sampling).

13. Appendix 5: Exercising the Ideas

The reader can work on the following problems as part of the process of learning about correlation; the data sets in Appendix 1 and Appendix 2 given overleaf on page HL66.7 can be used to practise *calculating* r values.

Ex. HL66.1: (a) The correlation between the ages of husbands and wives in Canada in 2002 was:

exactly −1 ; close to −1 ; around − 0.5 ; close to 0 ; around +0.5 ; close to +1 ; exactly +1.

Circle the correct option and briefly justify your choice.

(b) If men *always* married women who were three years younger, what would the correlation be between the ages of husbands and wives? Explain briefly.

Ex. HL66.2: Suppose a psychologist speaking at a meeting of the Canadian Association of University Teachers said: "The evidence suggests there is nearly zero correlation between teaching ability of a faculty member and her or his research productivity." A student newspaper reported this statement as: "Professor McDaniel said good teachers tend to be poor researchers and good researchers tend to be poor teachers". Explain whether you consider the newspaper report to be accurate; if not, write a succinct version of an accurate report, using language suitable for a newspaper.

Ex. HL66.3: If the response variate takes on values that are usually smaller than those of the explanatory variate, the correlation of the two variates over a large data set will tend to be negative. True or False? Explain briefly.

Ex. HL66.4: If the correlation between the values of a response variate and an explanatory variate is − 0.8, below-average values of the response variate tend to be associated with above-average values of the explanatory variate. True or False? Explain briefly.

Ex. HL66.5: For the population of North American automobiles, would the correlation between vehicle weight and fuel consumption be positive or negative if fuel consumption is measured in: km per litre ; litres per 100 km?
Explain briefly in each case.

Ex. HL66.6: Give the likely order of increasing magnitude of the three correlations calculated from large data sets collected on:

(*continued*)

CORRELATION: Its Properties and Its Interpretation (continued 4)

Ex. HL66.6: (cont.) university student cumulative averages in first and second years;
university student cumulative averages in first and fourth years;
length and weight of off-cuts from two-by-four lumber.

Ex. HL66.7: For data collected from a large group of people of the same sex, which of the two correlations is likely to be of greater magnitude in each of the following three cases? Justify each answer briefly.

(a) Height at age 4 and age 18, height at age 16 and age 18.
(b) Height at age 4 and age 18, weight at age 4 and age 18.
(c) Height and weight at age 4, height and weight at age 18.

Would your answers be altered if the group of people were roughly half females and half males? Explain briefly.

Ex. HL66.8: For a large group of married couples, arrange in increasing order the likely magnitude of the husband-wife correlations for the following four factors. Justify your order briefly.

Height ; income ; obesity ; weight.

Ex. HL66.9: Over a large group of homes heated by natural gas, gas consumption and outside temperature are correlated; this correlation is *positive* when outside temperature is quantified in terms of 'degree days'. Suppose data on these two variates are available for the eight months of a winter heating season for homes in Toronto; what would happen to the correlation (sign and magnitude) as the data are aggregated progressively from daily to weekly to fortnightly to monthly? Explain briefly.

Ex. HL66.10: In 1910, Hiram Johnson entered the California gubnatorial primaries. For each county, data are available on the percentage of native-born Americans in the county and the percentage of the vote for Johnson. If these data show a correlation of 0.5 over the counties of California, is this value a fair measure of the extent to which *Johnson received native, as opposed to immigrant, support*? Explain briefly.

Ex. HL66.11: Measuring inaccuracy occurs when a measuring process yields values that differ from the true value by a constant amount. In a calibration investigation, 10 samples are measured by each of two measuring processes and a correlation of 0.996 is found between the two sets of measurements. Explain briefly what information the high positive correlation provides about inaccuracy in either or both of the measuring processes.

Ex. HL66.12: On a multiple-choice examination with 100 questions written by a large class of students, the average number of right answers was 69 and the standard deviation was 12; for wrong answers, the average number was 31 and the standard deviation was 12. The correlation between right and wrong answers over the class of students was:

−1 ; −0.69 ; −0.31 ; 0 ; +0.31 ; +0.69 ; +1 ; can't tell without the data.

Circle the correct option and briefly justify your choice.

Ex. HL66.13: Participants in Cycle II of the U.S. Health Examination Survey were children aged 6 to 11 years. For *each* of these six years of age, the correlation between height and weight was around 0.5. For all the children *combined* across the six years of age, the correlation between height and weight would have been:

quite a bit below 0.5 ; around 0.5 ; quite a bit above 0.5.

Circle the correct option and briefly justify your choice.

Ex. HL66.14: There is a high *positive* correlation between years of education and income for enonomists who work in business; in particular, economists with doctorates earn substantially more than economists with baccalaureates. There is also a high correlation between years of education and income for economists who work in post-secondary education. However, when *all* such economists are considered together, there is a *negative* correlation between years of education and income. The explanation is that business tends to pay higher salaries and employs mainly economists with baccalaureates whereas post-secondary education tends to pay lower salaries and employs economists with doctorates. Sketch a scatter diagram with appropriate labels which illustrates these matters.

SOURCES: Freedman, D., Pisani, R. and R. Purves: *Statistics.* First Edition, W. W. Norton & Company, New York, 1980, pages 130-133 and 143.

Moore, D.S. and G.P. McCabe: *Introduction to the Practice of Statistics.* Second Edition, W. H. Freeman and Company, New York, 1993, pages 175 and 183.

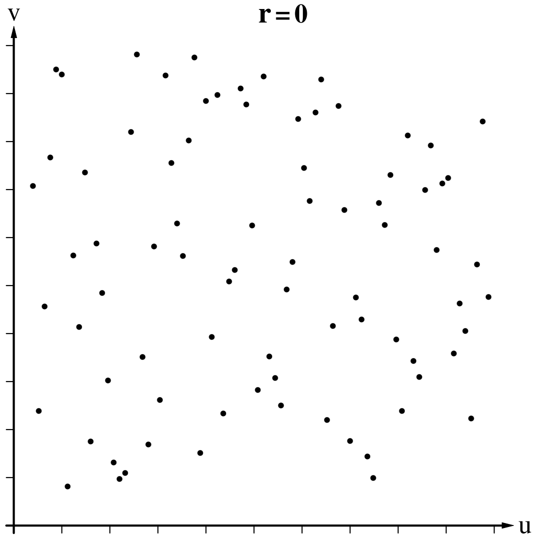
□ The diagrams on the following pages HL66.10 to HL66.15 have axes with scale divisions but *no numerical labels.*

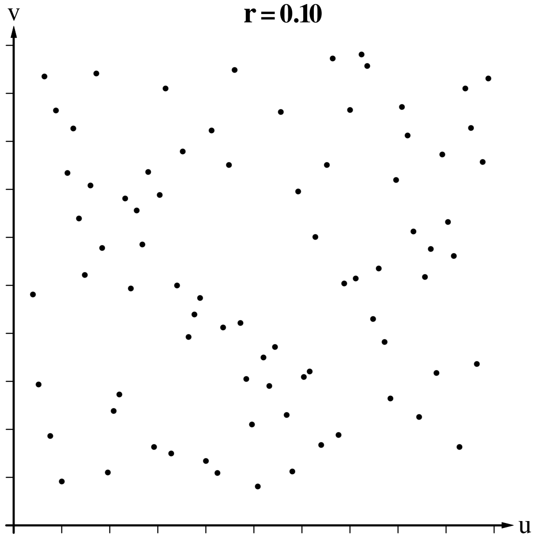
- Indicate briefly the purpose(s) served by the *inclusion* of the scale divisions.
- Indicate briefly the reason(s) for the *omission* of numerical labels; your explanation should include explicit discussion of the *form* of the relevant algebraic expression for r.
  - Outline any *practical* advantage(s) or disadvantage(s) that are a consequence of this property of correlation.

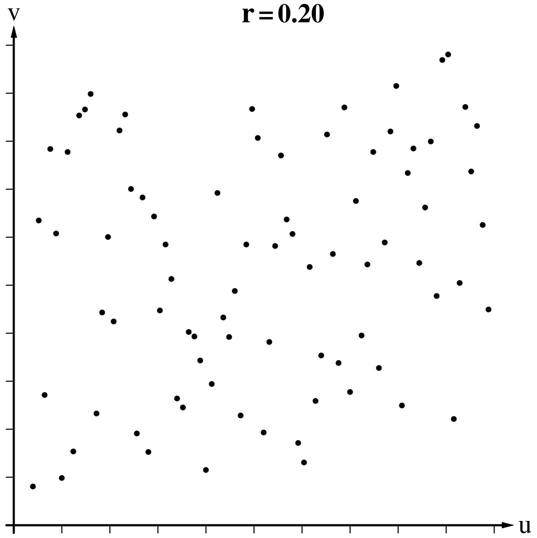
#HL66.10

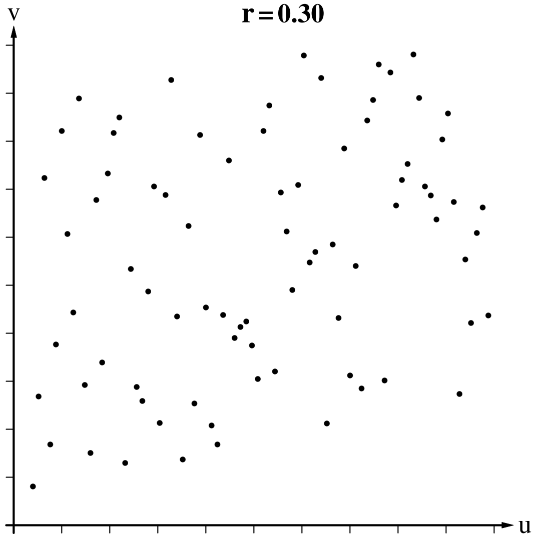
University of Waterloo G.W. Bennett

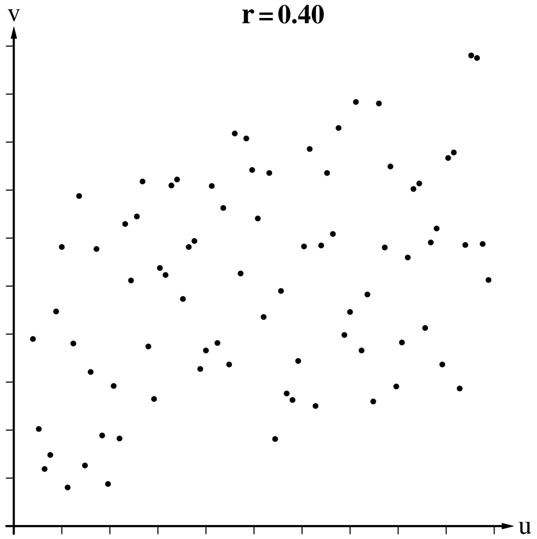
CORRELATION: Scatter Diagrams (n = 80) for Positive r-values

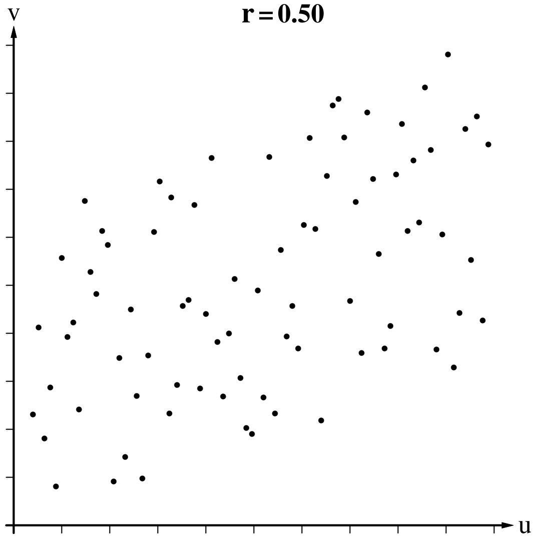

r = 0


r = 0.10


r = 0.20

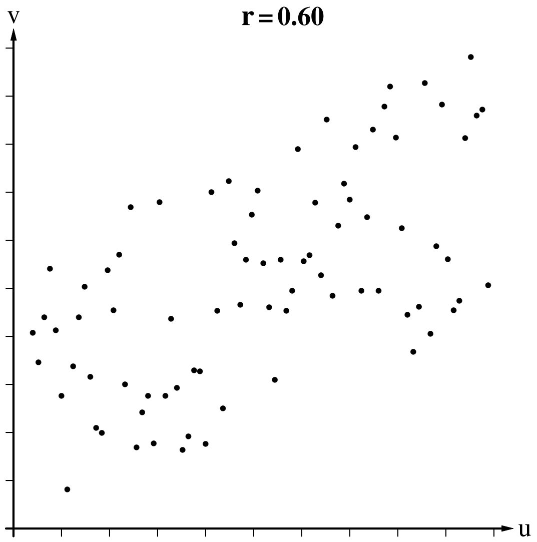

r = 0.30

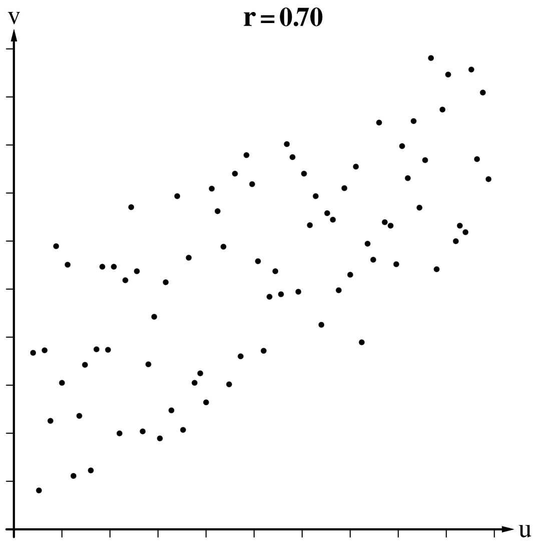

r = 0.40

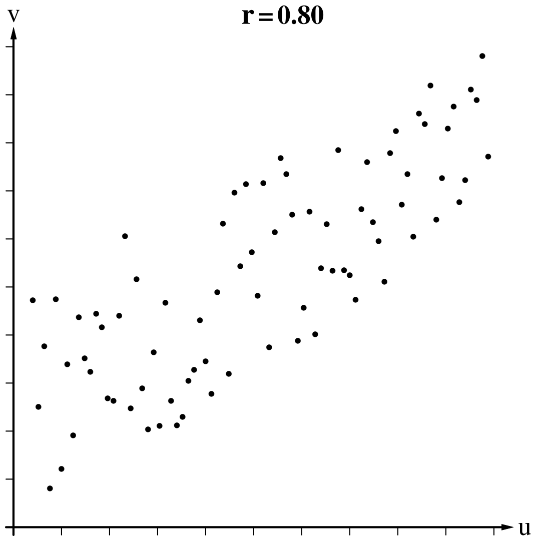

r = 0.50

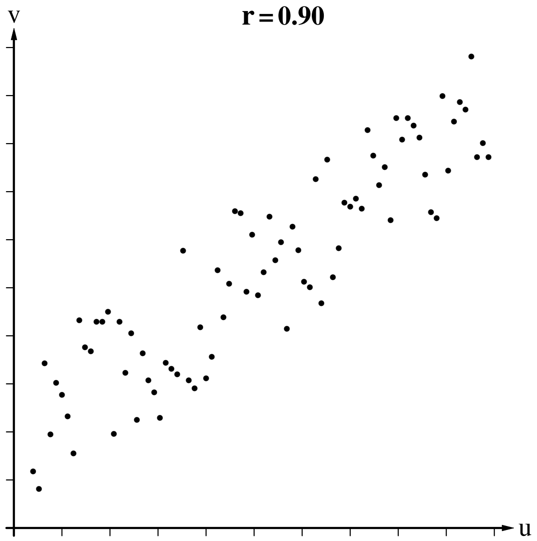
(*continued*)

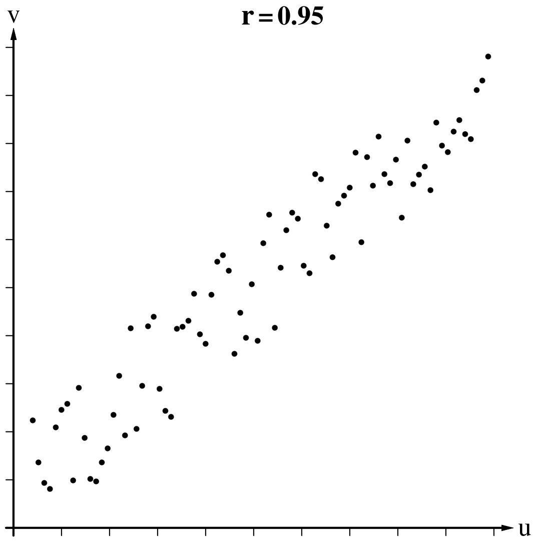
CORRELATION: Scatter Diagrams (n = 80) for Positive r-values (continued)

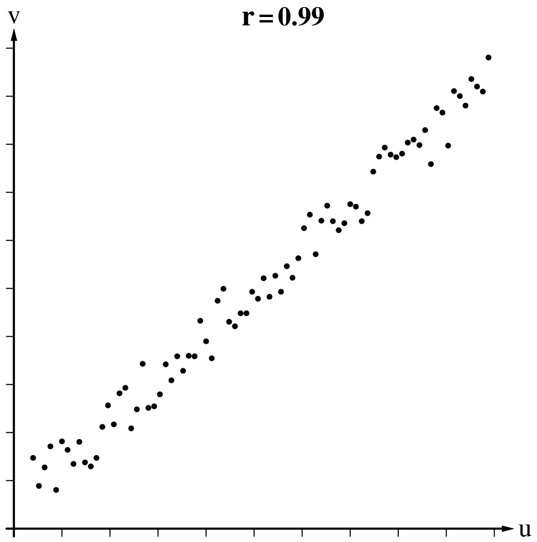




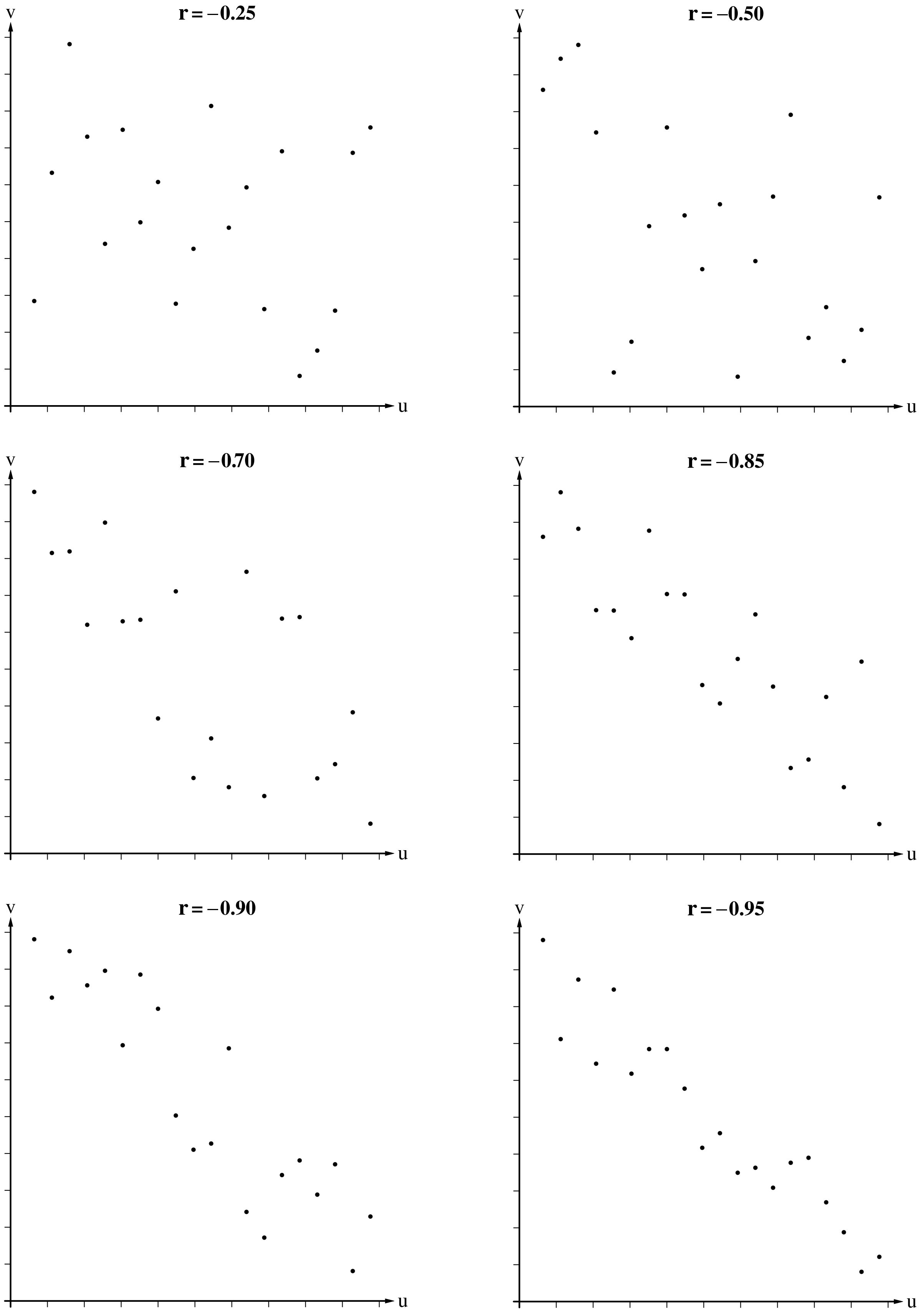








CORRELATION: Scatter Diagrams (n = 20) for Negative r-values



r = −0.25

r = −0.50

r = −0.70

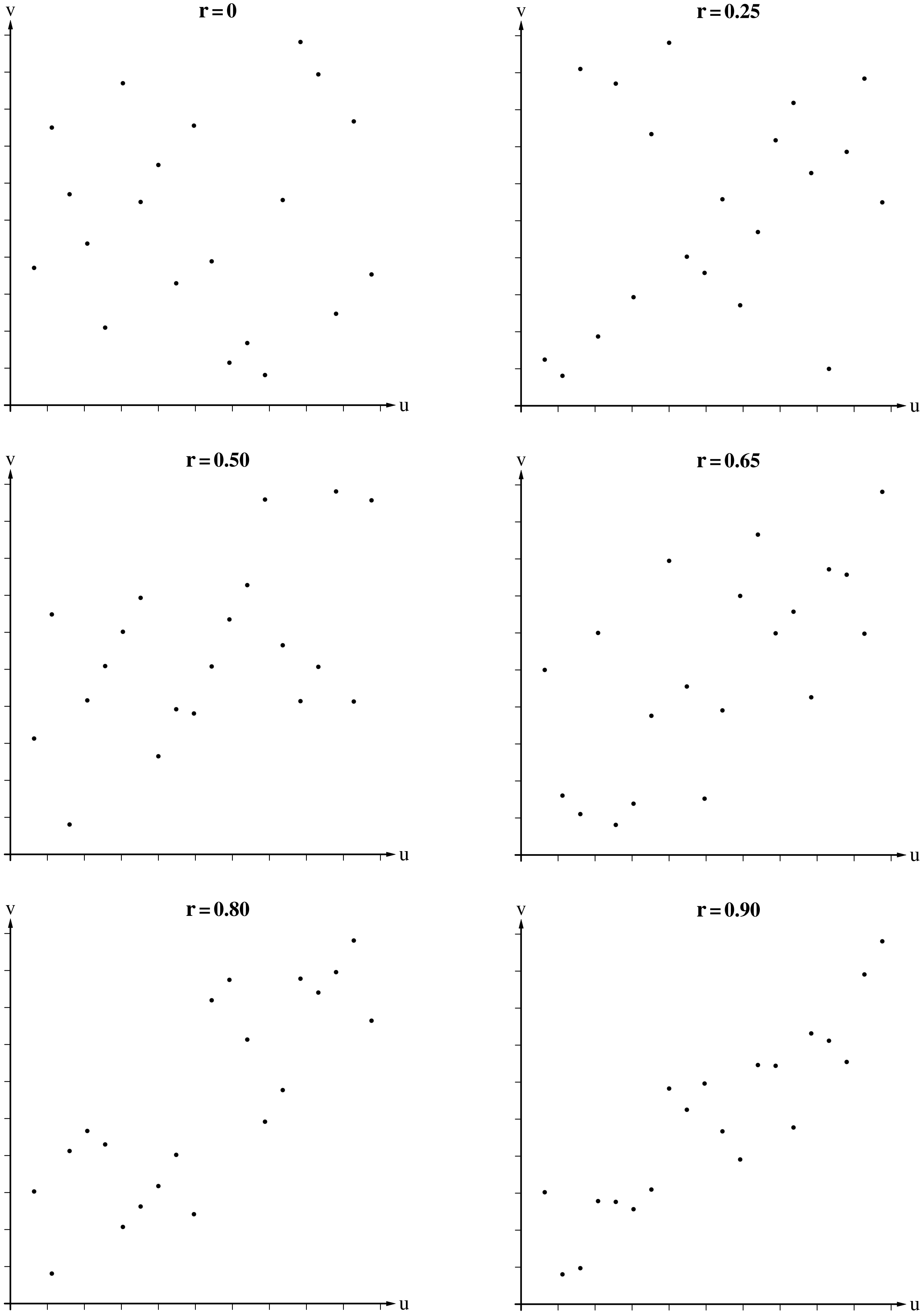
r = −0.85

r = −0.90

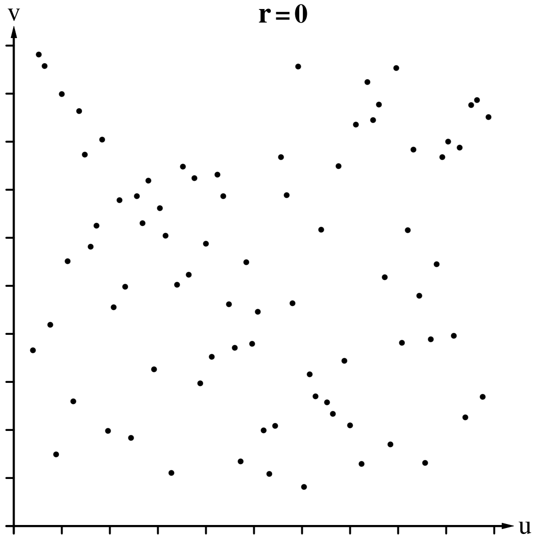
r = −0.95

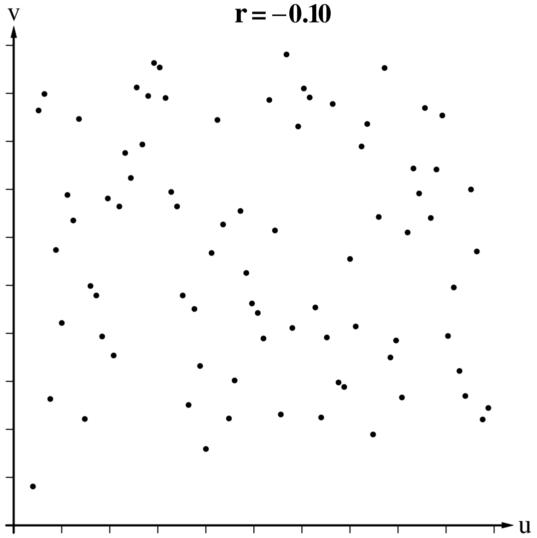
(*continued*)

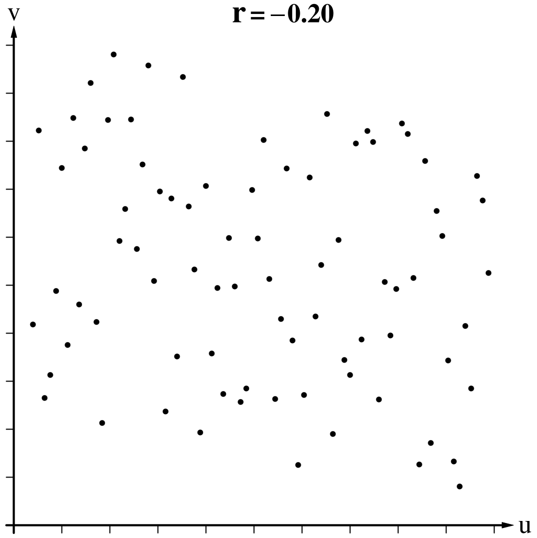
CORRELATION: Scatter Diagrams (n = 20) for Positive r-values (continued)

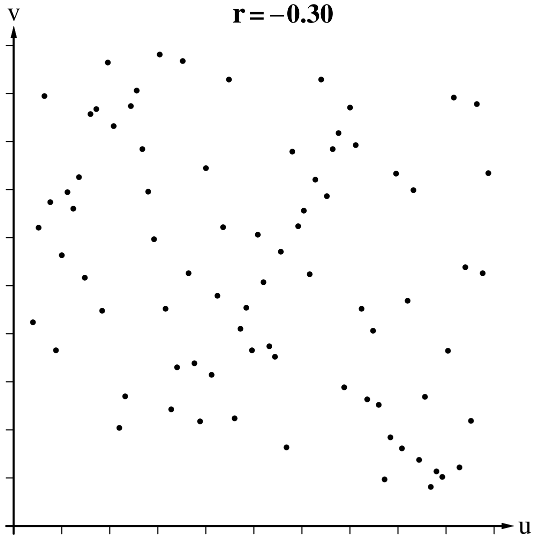


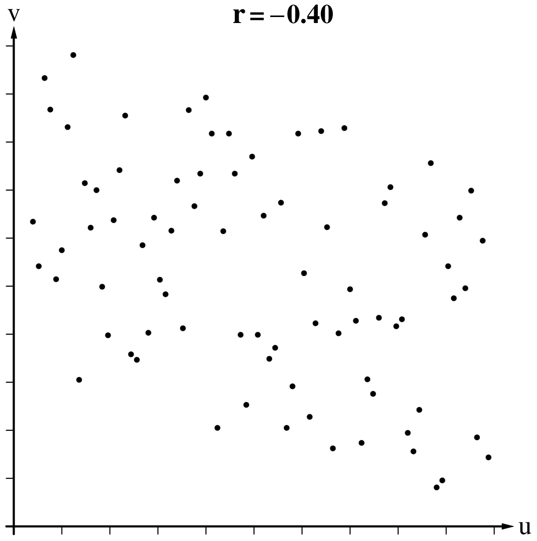
CORRELATION: Scatter Diagrams (n = 80) for Negative r-values

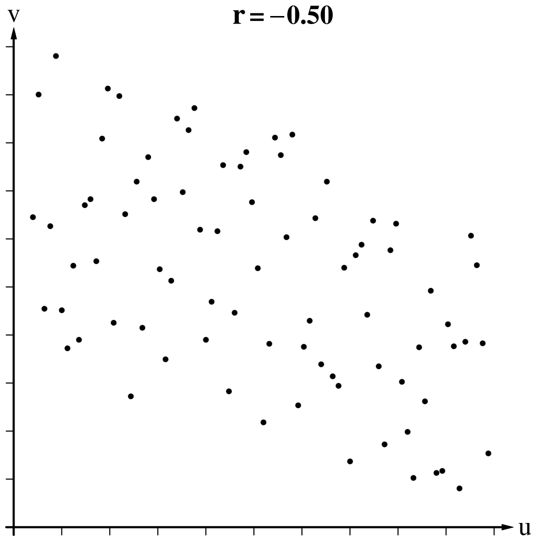

r = 0


r = −0.10


r = −0.20

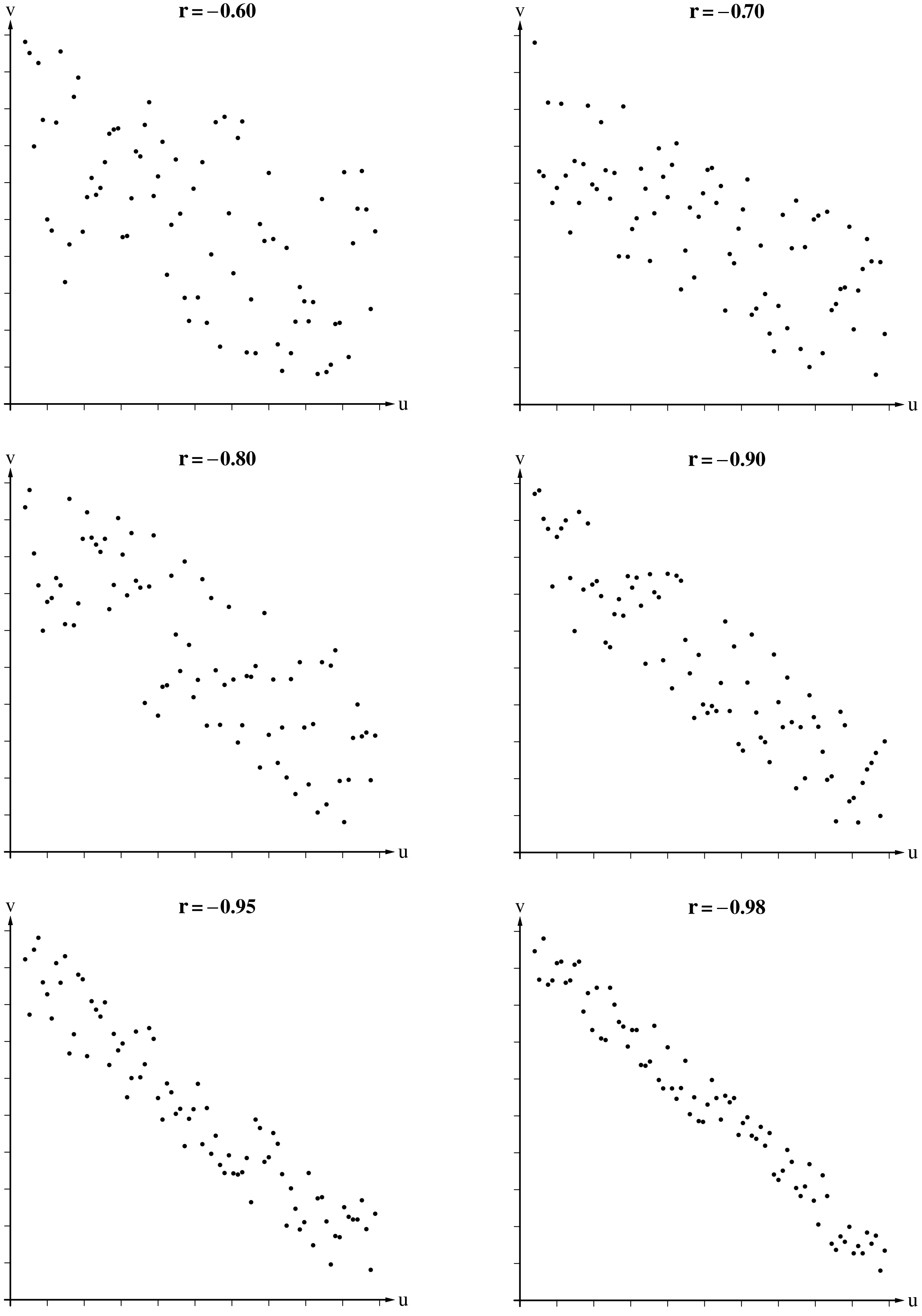

r = −0.30


r = −0.40


r = −0.50

(*continued*)

CORRELATION: Scatter Diagrams (n = 80) for Negative r-values (continued)



(*continued overleaf*)

EM9026: The Globe and Mail, March 6, 1990, page A8

Statistics: time to update our brains

THOMAS HURKA

Questions of Principle

CALGARY

OUR BRAINS aren't designed for life in the modern world. They support thought processes that were appropriate for our ancestors in the prehistoric forest, when the important information was vivid and particular, such as a sabre-toothed tiger about to pounce. But they go wonky when faced with statistics.

Some years ago, Maude Barlow, then special advisor to prime minister Pierre Trudeau on pornograpy policy, gave the following argument to show that pornography causes sexual crimes: "It's no accident that 48 per cent of rapists and 14 per cent of child molesters are pornography fans."

The argument seems impressive. But if we're persuaded by it we're persuaded by something that doesn't do half the work it has to.

To prove that pornography causes sexual crimes, you first have to show that the percentage of pornography fans who commit sexual crimes is higher than the percentage of non-fans who do. By this comparison you establish a correlation between sexual crimes and pornography. Then you have to rule out the other possible explanations of this correlation: that sexual crimes cause the viewing of pornography (unlikely) or that some third factor like upbringing causes both.

Quite apart from not attempting the second task, Ms. Barlow's argument doesn't complete the first.

Take her first statistic: that 48 per cent of rapists are pornography fans. On its own this statistic is useless. We have to know how the figure for rapists compares with that for non-rapists, and we aren't given the second figure.

This point is even clearer for the second statistic: that 14 per cent of child molesters are pornography fans.

If pornography caused child molesting, the comparable figure for non-child molesters would have to be lower than 14 per cent. But this is implausible. Pornography is a billion-dollar industry, its products consumed by millions of men. Surely, more than 14 per cent of men count as pornography fans. If so, child molesters as a group view less pornography, and, if anything, pornography may reduce child molesting.

What's going on? Why are we persuaded by statistics that are incomplete or even support an opposed conclusion?

Our brains evolved in a simpler world and lead us constantly astray when we reason about statistics. This is shown by the many psychological experiments summarized in *Human Inference Strategies and Shortcomings of Social Judgement* by Richard Nisbett and Lee Ross.

Just one thing we're bad at is judging correlations. Deciding whether B is correlated with A requires comparing the percentage of As that are B with the percentage of non-As that are B. But we conclude there's a correlation on the basis just of non-comparative information about some As that are B, for example, some child molesters who view pornography. We reach a statistical conclusion using less than all the evidence.

There are many other examples.

We've all seen the athlete, interviewed on TV after his game-winning home run, who says he dreamed of hitting it the night before.

He evidently thinks such dreams are significant; so probably does the interviewer. But they're deciding on incomplete evidence. We need to know not only how many players dream of home runs and hit them, but how many dream of them and don't, how many don't dream of them and do, and even how many don't dream and don't hit.

(We rarely get this extra evidence. Sluggers don't say, "Fergy, it's amazing, I didn't dream of hitting a home run last night." And people who don't get hits aren't interviewed.)

Our weakness also contributes, more alarmingly, to the persistence of false racial and gender stereotypes.

Stereotypes are really claims about correlation. If you believe that redheads are hot-tempered, you don't believe that all redheads are hot-tempered, nor do you believe that only redheads are hot-tempered. But you believe the percentage of hot tempers among redheads is higher than among blondes, brunettes, and so on.

Imagine that this stereotype is false, and your evidence shows it is false. You may still believe it because you attend more to the evidence in its favour than to what refutes it.

When you see a redhead getting angry, you're reminded of the stereotype and take it to be confirmed. ("Another hot-tempered redhead.") But when you see one being placid you don't take this as evidence against it. Why remember a stereotype about hot temper when no one's showing temper? Likewise when you see a blonde throwing a fit. Why connect that to any ideas about redheads? By attending more to evidence about As that are B you keep believing As are specifically B when they're not.

Stereotypes about redheads may be harmless, but ones like "Jews are clannish" and "women are emotional" are not. If they persist partly because of fallacious statistical reasoning, we should be worried.

As statistical information becomes more important, you would expect basic training about it to become more common, and even to be seen as part of essential literacy. But who is giving this training, either in our schools or, outside specialist courses, in university? Who is trying to bring our brains into the modern world?

Prof. Hurka teaches philosophy at the University of Calgary.

2 In the first column of the article, Prof. Hurka points out two *logical* inaccuracies associated with Ms. Barlow's interpretation of percentages of 48 and 14. Indicate briefly *other* type(s) of inaccuracies that may be involved in these two values.

3 Suppose the two additional percentages for 'non-fans' of pornography, discussed by Prof. Hurka in the first column of the article, were available and that they *were* appreciably lower than 48 and 14 respectively. Would the Answer (in the second paragraph) that pornography *causes* sexual crimes then be justified? Explain briefly.

- ● Explain briefly what *additional* line(s) of evidence would be needed to *reduce the limitations* on the Answer of a causal connection between pornography and sexual crimes?

4 Comment briefly on the plausibility of the argument, at the top of the middle column of the article, that the percentage of 'non-child molesters' who are 'pornography fans' is *greater* than 14 per cent.

5 In the last paragraph of the middle column, Professor Hurka describes *four* proportions involving dreams and home runs. Describe the equivalent four proportions in the context of assessing the usefulness of a medical diagnostic test for a disease.