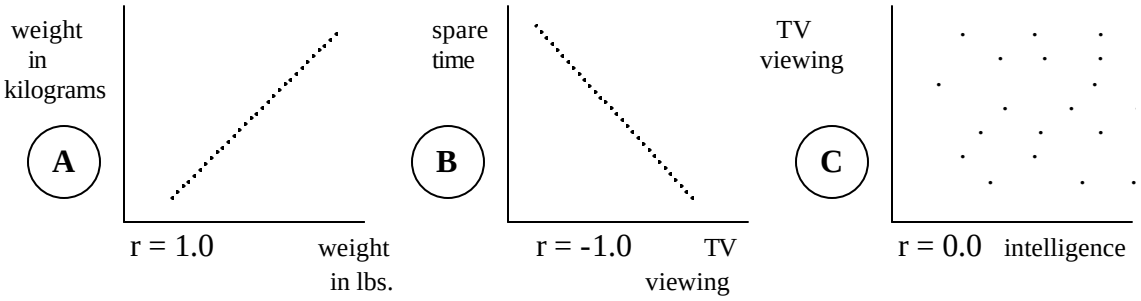
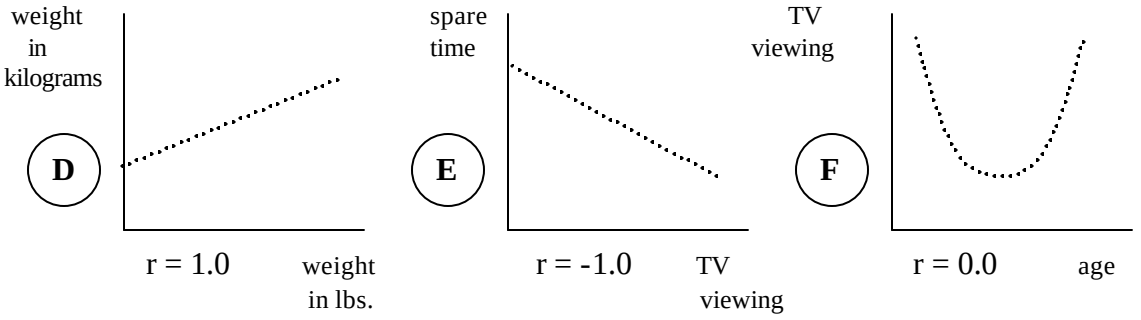
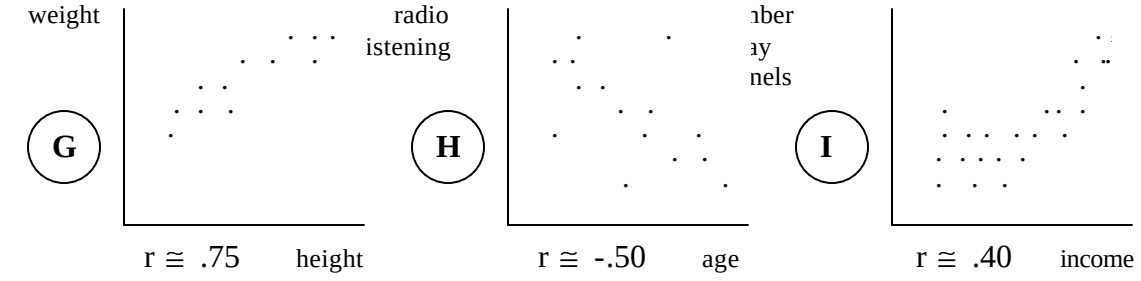
Neuendorf

Bivariate Correlation

The Pearson correlation coefficient (r) is a measure of how closely related two variables are, both of which must be measured at the interval/ratio level. This relationship is assumed to be linear, and the correlation is a measure of how tightly clustered data points are about a correlation line. Correlation ranges from –1.0 (perfect negative relationship) to 1.0 (perfect positive relationship). A correlation of 0.0 indicates no discernable linear relationship. Below are some graphical representations of sample correlations:







Note that the slope of the line does not indicate the correlation; the closeness of the data points to the line is what determines the size of the r. The greater the absolute value of r, the more closely related the two variables.

Also note that a very strong nonlinear relationship (e.g., above, age and TV viewing) will not usually show a very strong correlation! That doesn't mean there is not a relationship, just not a linear one. This inability of the Pearson correlation coefficient to discern nonlinear relationships is perhaps its greatest shortcoming, and should be remembered when interpreting correlational results.

A correlation may be tested for its statistical significance by consulting the r table. For example, suppose we found a relationship between grade point average and hours per week of partying of r = .55. If our n was 25, the critical value in the table at df = 23 and p = .05 is .396. Since our r exceeds this, it is statistically significant at the .05 level. We are 95% certain that grade point average and partying are positively related in the population (the more one parties, the greater one's grade point average; or, the greater one's grade point average, the more one parties).

Another bit of information that can be derived from the correlation is the amount of shared variance. r^2 is called the coefficient of determination, and represents the proportion of the variance of x that is shared by y, and correspondingly the proportion of the variance of y that is shared by x. Thus, if we have an r of .30, $r^2 = .09$ – that is, 9% of the variance is shared. If we have an r of .90, $r^2 = .81$ – 81% of the variance is shared.

For the record, one formula for r (there are several) is:

$$r_{xy} = \frac{\sum xy}{\sqrt{\sum(x^2)\sum(y^2)}}$$

x = individual deviation score

$(X - \text{mean}_x)$

y = individual deviation score

$(Y - \text{mean}_y)$

df = n-2

Appendix B Values of r_{xy} beyond which 5 percent or 1 percent of the area falls

Number of pairs -2	05	01	Number of pairs -2	05	01
1	.997	1.000	24	.388	.496
2	.950	.990	25	.381	.487
3	.878	.959	26	.374	.478
4	.811	.917	27	.367	.470
5	.754	.874	28	.361	.463
6	.707	.834	29	.355	.456
7	.666	.798	30	.349	.449
8	.632	.765	35	.325	.418
9	.602	.735	40	.304	.393
10	.576	.708	45	.288	.372
11	.553	.684	50	.273	.354
12	.532	.661	60	.250	.325
13	.514	.641	70	.232	.302
14	.497	.623	80	.217	.283
15	.482	.606	90	.205	.267
16	.468	.590	100	.195	.254
17	.456	.575	125	.174	.228
18	.444	.561	150	.159	.208
19	.433	.549	200	.138	.181
20	.423	.537	300	.113	.148
21	.413	.526	400	.098	.128
22	.404	.515	500	.088	.115
23	.396	.505	1000	.062	.081