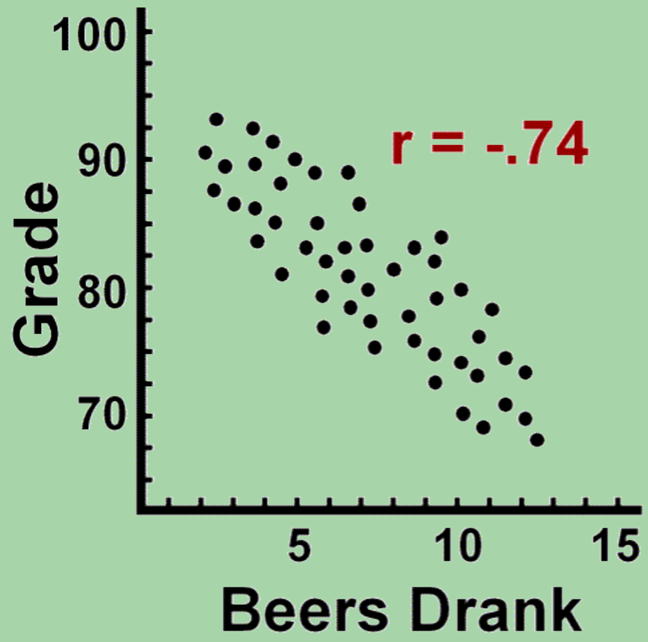
Statistics by Hand

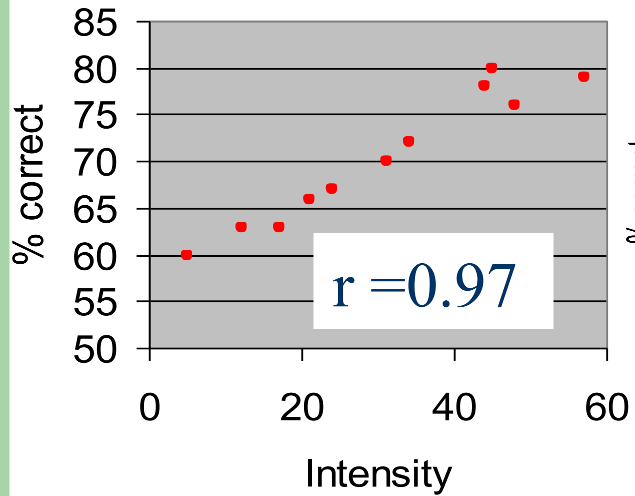
An Introductory Course for Psychologists

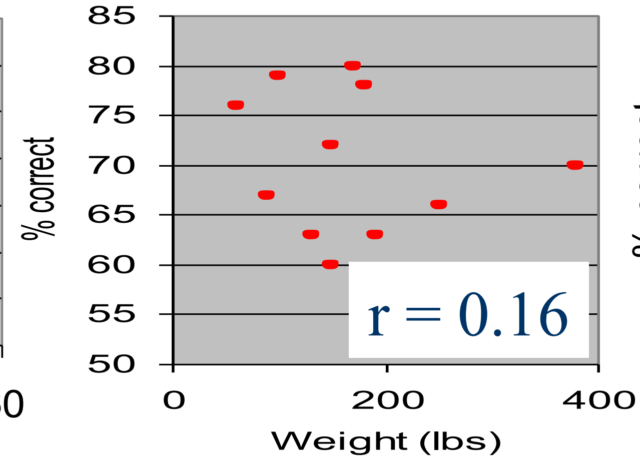


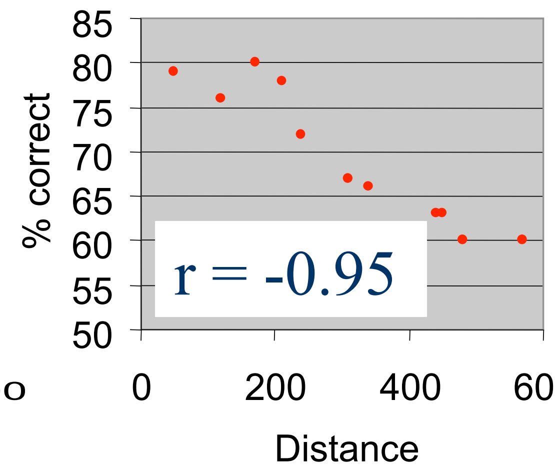
Correlation

Version 3.0

Correlation







- Degree of relationship between two *continuous* variables.
- Quantified by *r* - a correlation co-efficient.
- *r* ranges from -1 to +1.

Calculation of *r*

- Calculation of *r* is based on the concept of *co-variance*.
- *Co-variance:* The extent to which changes in one variable are reflected in changes of the other.

Co-variance

- If X increases as Y increases, co-variance will be positive.
- If X increases as Y decreases, co-variance will be negative.
- If X and Y are independent, co-variance will the close to zero.

Co-variance: Worked e.g.

											Mean
X: Intensity	5	12	17	21	24	31	34	44	45	48	28.1
Y: Detection accuracy (%)	60	63	63	66	67	70	72	78	80	76	69.5
(X- mean of X)	-23.1	-16.1	-11.1	-7.1	-4.1	2.9	5.9	15.9	16.9	19.9	
(Y - mean of Y)	-9.5	-6.5	-6.5	-3.5	-2.5	0.5	2.5	8.5	10.5	6.5	Sum
Product	219	105	72.2	24.9	10.3	1.45	14.8	135	177	129	889.5

- $cov_{XY} = 889.5 / 9 = 98.8$

$$\mathrm{cov}_{XY} = \frac{\sum\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)}{N-1}$$

Pearson Product-Moment

$$r = \frac{\text{cov}_{XY}}{s_X s_Y}$$

- Co-variance then needs to be scaled by the total amount of variability in the data.
- Pearson showed that doing this meant *r* always ranged from -1 to +1.

Significance & assumptions

- Significance testing: Use an *r* table.
- If you're testing $r \neq 0$, then it's two-tailed.
- X and Y should be normally distributed (quite important unless *N* is large).

Uses

- Use Pearson correlation when you want to assess the relationship between two variables, and...
  - …both variables are continuously valued, and
  - …both variables are approximately normal
- Otherwise, use
  - contingency chi-square (if categorical) or
  - Spearman’s (see later) if non-normal

Method summary

											Mean
X: Intensity	5	12	17	21	24	31	34	44	45	48	28.1
Y: Detection accuracy (%)	60	63	63	66	67	70	72	78	80	76	69.5
(X- mean of X)	-23.1	-16.1	-11.1	-7.1	-4.1	2.9	5.9	15.9	16.9	19.9	
(Y - mean of Y)	-9.5	-6.5	-6.5	-3.5	-2.5	0.5	2.5	8.5	10.5	6.5	Sum
Product	219.45	104.65	72.15	24.85	10.25	1.45	14.75	135.15	177.45	129.35	889.5
(X- mean of X) squared	533.61	259.21	123.21	50.41	16.81	8.41	34.81	252.81	285.61	396.01	1961
(Y - mean of Y) squared	90.25	42.25	42.25	12.25	6.25	0.25	6.25	72.25	110.25	42.25	424.5

- $s_x = \sqrt{}\ (1961 / 9) = 14.76$
- $s_y = \sqrt{}\ (424.5 / 9) = 6.87$
- $cov_{xy} = 889.5 / 9 = 98.83$
- $r = 98.83 / (14.76 \text{ x } 6.87) = 0.97$

Spearman's r

- If only ranks are available, the same equations can be applied. The test is then called Spearman's *r* or r_s
- Where $N > 9$ the critical value of r_s is numerically close to the values in a Pearson's *r* table.
- Even if actual data are available, we sometimes use just the ranks. The main advantage is that it avoids the assumption that X and Y are normally distributed.

Problems

- *CLR-1 – Assume it is a Pearson’s.*
- *CLR-2 – Use histograms to determine whether it is a Pearson’s or a Spearman’s.* ***Do not attempt last sentence.***