COMPARATIVE STATISTICS

COM 221 - ML | CCDATSCL

VISULIZING RELATIONSHIP

It is usually more informative to explore the relationship between different continuous variables using a scatterplot. The next diagram illustrates three scatterplots.

a) , there is a positive relationship between the two variables and from inspection appears to be linear.

b) , there is a negative relationship between the variables and it also appears to be non linear.

c) , it is difficult to see any relationship between the two variables.

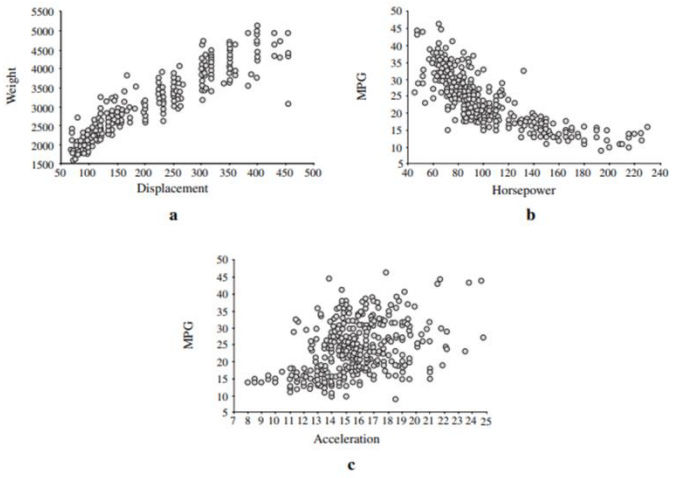


Figure 5.21. Illustrating different relationships using scatterplots

The Correlation Coefficient (r)

- Sometimes called **Pearson product-moment correlation**.
- For pairs of variables measured on an interval or ratio scale, a **correlation coefficient (r)** can be calculated. This value quantifies the linear relationship between the variables. It generates values ranging from **-1.0 to +1.0**.
- Positive numbers indicate a positive correlation and negative numbers indicate a negative correlation. Little or no correlation is centered around 0.
- The formula used to calculate **r** is shown below:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Table 5.12. Table showing the calculation of the correlation coefficient

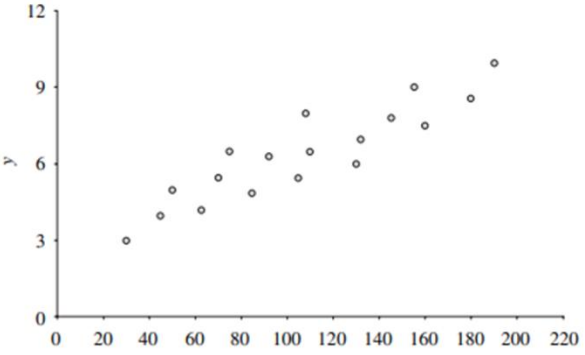
x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
92	6.3	−14.94	−0.11	1.58
145	7.8	38.06	1.39	53.07
30	3	−76.94	−3.41	262.04
70	5.5	−36.94	−0.91	33.46
75	6.5	−31.94	0.09	−3.02
105	5.5	−1.94	−0.91	1.76
110	6.5	3.06	0.094	0.29
108	8	1.06	1.59	1.68
45	4	−61.94	−2.41	149.01
50	5	−56.94	−1.41	80.04
160	7.5	53.06	1.09	58.07
155	9	48.06	2.59	124.68
180	8.6	73.06	2.19	160.32
190	10	83.06	3.59	298.54
63	4.2	−43.94	−2.21	96.92
85	4.9	−21.94	−1.51	33.04
130	6	23.06	−0.41	−9.35
132	7	25.06	0.59	14.89
$\bar{x}$ = **106.94**	$\bar{y}$ = **6.41**			**Sum = 1357.01**
s_x = **47.28**	s_y = **1.86**			

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

$$r = \frac{1357.01}{(18-1)(47.28)(1.86)}$$

$$r = 0.91$$



INTERPRETATION OF R

Size of correlation	Interpretation
0.90-1.00 (−0.90 to −1.00)	Very high positive (negative) correlation
0.70-0.90 (−0.70 to −0.90)	High positive (negative) correlation
0.50-0.70 (−0.50 to −0.70)	Moderate positive (negative) correlation
0.30-0.50 (−0.30 to −0.50)	Low positive (negative) correlation
0.00-0.30 (0.00 to −0.30)	Negligible correlation

When describing relationships

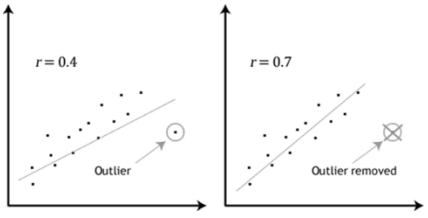
Four things must be reported to describe a relationship:

1. The strength of the relationship given by the correlation coefficient.
2. The direction of the relationship, which can be positive or negative based on the sign of the correlation coefficient.
3. The shape of the relationship, which must always be linear to computer a Pearson correlation coefficient.
4. Whether or not the relationship is statistically significant, which is based on the p-value.

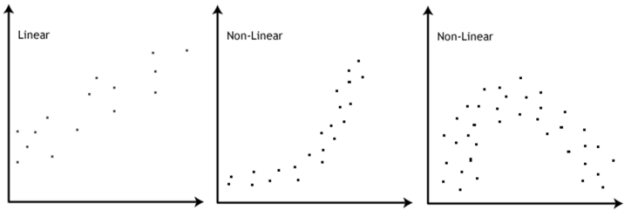
Assumptions for Pearson correlation

1. Level of measurement refers to each variable. For a Pearson correlation, **each variable should be continuous**.
2. Related pairs refers to the pairs of variables. **Each participant or observation should have a pair of values**. So if the correlation was between weight and height, then each observation used should have both a weight and a height value.
3. **Absence of outliers refers to not having outliers in either variable**. Having an outlier can skew the results of the correlation by pulling the line of best fit formed by the correlation too far in one direction or another. Typically, an outlier is defined as a value that is 3.29 standard deviations from the mean, or a standardized value of less than ±3.29.
4. **Your variables should be approximately normally distributed**. In order to assess the statistical significance of the Pearson correlation, you need to have bivariate normality, but this assumption is difficult to assess, so a simpler method is more commonly used.

Assumptions: No outliers



Assumption: There is a linear relationship



Spearman's rank-order correlation

- Also called Spearman's rho ρ
- The Spearman rank-order correlation coefficient (Spearman's correlation) is a **nonparametric measure of the strength and direction** of association that exists between two variables measured on at least an ordinal scale.
- Spearman's correlation can be used when your two variables are not normally distributed.
- It generates values ranging from **-1.0 to +1.0**.

With tied ranks

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)\cdot\left(R(y_i)-\overline{R(y)}\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)^2\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i)-\overline{R(y)}\right)^2\right)}}$$

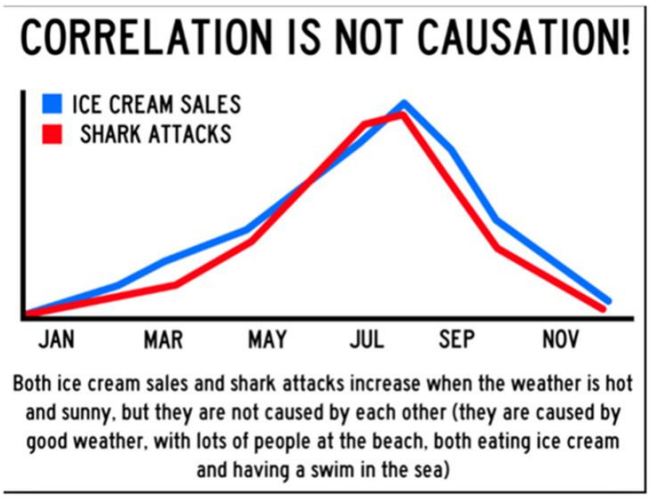
$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$ **No tied ranks**

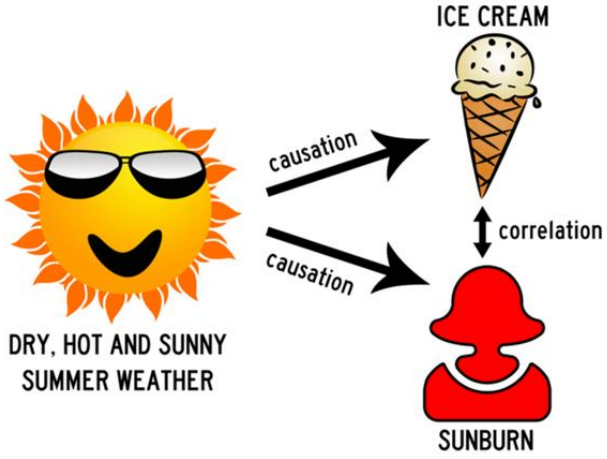
Assumptions for Spearman correlation

1. **Your two variables should be measured on an ordinal, interval or ratio scale**. Examples of ordinal variables include Likert scales (e.g., a 7-point scale from "strongly agree" through to "strongly disagree.
2. **Your two variables represent paired observations**. With 30 participants in the study, this means that there would be 30 paired observations.
3. **There is a monotonic relationship between the two variables.** A monotonic relationship exists when either the variables increase in value together, or as one variable value increases, the other variable value decreases.

Correlation does not imply causation

The phrase refers to the **inability to legitimately deduce a cause-and-effect relationship** between two events or variables solely on the basis of an observed association or correlation between them





Correlation is a statistical technique which tells us how strongly the pair of variables are linearly related and change together. **It does not tell us why and how behind the relationship but it just says the relationship exists.**