1.2 Finding relationships among variables

- Descriptive methods (describe attributes of a data set)
- Inferential methods (draw conclusions about a population from samples)

An association means that knowing the value of one variable gives us some information about the possible values of the second variable.

Associations between pairs of variables are called bivariate associations.

1.2.1 Numerical variables

Pearson's product-moment correlation coefficient

Data: (1,2), (0,4), (3,2), (1,6), (5,1)

- Pearson's correlation coefficient takes a value of 0 if two variables are uncorrelated, and a value of +1 or -1 if we can predict the exact value of one variable given knowledge of the other using linear equation.
- A positive value indicates that higher values in one variable is associated with higher values of the second. A negative value indicates that higher values of one variable is associated with lower values of the second.
- Pearson's correlation coefficient will be misleading when relationship is curved or scattered and not linear.

Kendall’s τ

Measures of rank correlation are based on a comparison of the resulting ranks.

Any pair of observations i and j , where i < j, are said to be concordant if the sort order of i and j agrees.

$$\frac{\mathit{concordant\ pairs} - \mathit{disconcordant\ pairs}}{C(n,2)}$$

Data: (1,2), (0,4), (3,2), (1,6), (5,1)

1.2.2 Categorical variables

Numerically exploring associations between pairs of categorical variables is not as simple as the numeric variable case. We need to find out which combinations are common and which are rare.

Contingency table

Display frequencies in the form of a table

```
##                       month
## type                   Jun Jul Aug Sep Oct Nov Dec
##   Extratropical        27  38  23 149 129  42   4
##   Hurricane             3  31 300 383 152  25   2
##   Tropical Depression  22  59 150 156  84  42   0
##   Tropical Storm       31 123 247 259 204  61   1
```

Which months are most affected?

Which storms are most common?

Both variables are nominal so Pearson's coefficient is not possible.

Spearman's ρ rank correlation coefficient

It is valid for both numeric and categorical data.

$$\rho = 1 - \frac{6\sum(d_i^2)}{n(n^2-1)}$$

Data: (1,2), (0,4), (3,2), (1,6), (5,1)

Sample proportion

The sample proportion is used to estimate the value of a population proportion (the corresponding proportion of times something happens in the whole population).

1.2.3 Categorical variables and numerical variables

Such associations are studied visually using box plots and histogram that will be studied later.

Magic: Simpon's Paradox

	Men		**Women**	
	Selected	**Applied**	**Selected**	**Applied**
Computer Sc.	70	100	40	50
Electronics	10	50	30	100