Statistics – chapter 3

**The association between two categorical vars – the contingency table**

The aim is to explore the **relationship**, if any, between the variables

**An association** (aka, dependence) **between two vars exists if a particular value for one var is more likely to occur with certain values of the other var**

Such associations can be explored via: (not explored in this chapter)

* **cross-tabulation** and/or,
* **the chi-square test**

***Explanatory and Response variables***

An **explanatory variable** is what you manipulate or observe changes in (e.g., caffeine dose), this is aka **expected cause (or result explaining variable) - Independent**

A **response variable** is what changes as a result (e.g., reaction times), this is aka the **expected effect (or respond to explaining variable)** **- Dependent**

The words “explanatory variable” and “response variable” are often interchangeable with other terms used in research

| **Cause (what changes)** | **Effect (what’s measured)** |
| --- | --- |
| [Independent variable](https://www.scribbr.com/methodology/independent-and-dependent-variables/#independent) | [Dependent variable](https://www.scribbr.com/methodology/independent-and-dependent-variables/#dependent) |
| Predictor variable | Outcome/criterion variable |
| Explanatory variable | Response variable |

In general, two variables are independent if the conditional percentage distributions of the response (given the categories of the explanatory) are **equal**

The aim is to analyze the **conditional probabilities** in the contingency tables, we need to find, for each category of the **response** variable, under which category of the **explanatory** variable its percentage is greater than the corresponding marginal

Graphical user interface, text

Description automatically generated with medium confidence

**Text, letter

Description automatically generated**

Chart, scatter chart

Description automatically generated**The association between two quantitative vars – The scatterplot**

The interpretation depends on the correlation coefficient goes as follows: (seen in ML)

* **Positive correlation**, straight line visible with m > 0
* **Negative correlation**, straight line visible with m < 0 (anticorrelation)
* Text, letter

  Description automatically generated**No correlation**, no linear relationship is visible in the scatterplot

**Predicting the outcome of a variable – The regression line**

A regression line is a straight line that describes how the response variable (y) changes as the explanatory variable changes, predicting its value for a given explanatory variable (x):

[ŷ](https://www.compart.com/en/unicode/charsets/containing/U+0177) = a + bx

The y -intercept is the predicted value of y when x = 0

The true formula, accounting for the **error** is:

y= a + bx + ε

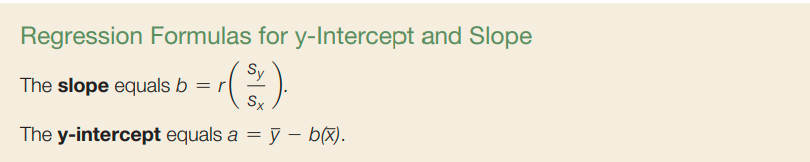
Text, letter

Description automatically generatedText, letter

Description automatically generatedThe size of the error is the **residual** = y - [ŷ](https://www.compart.com/en/unicode/charsets/containing/U+0177) > 0, that is, the distance between a point and the regression line

Where s is the sample standard deviation

**The square of the correlation coefficient is used to measure if the residuals are small enough**

****The evaluation of regression lines is done via **least square method**

**REM : Correlation does not imply causation**

**https://www.math.utah.edu/~morris/Courses/1070/notes/l4.pdf**