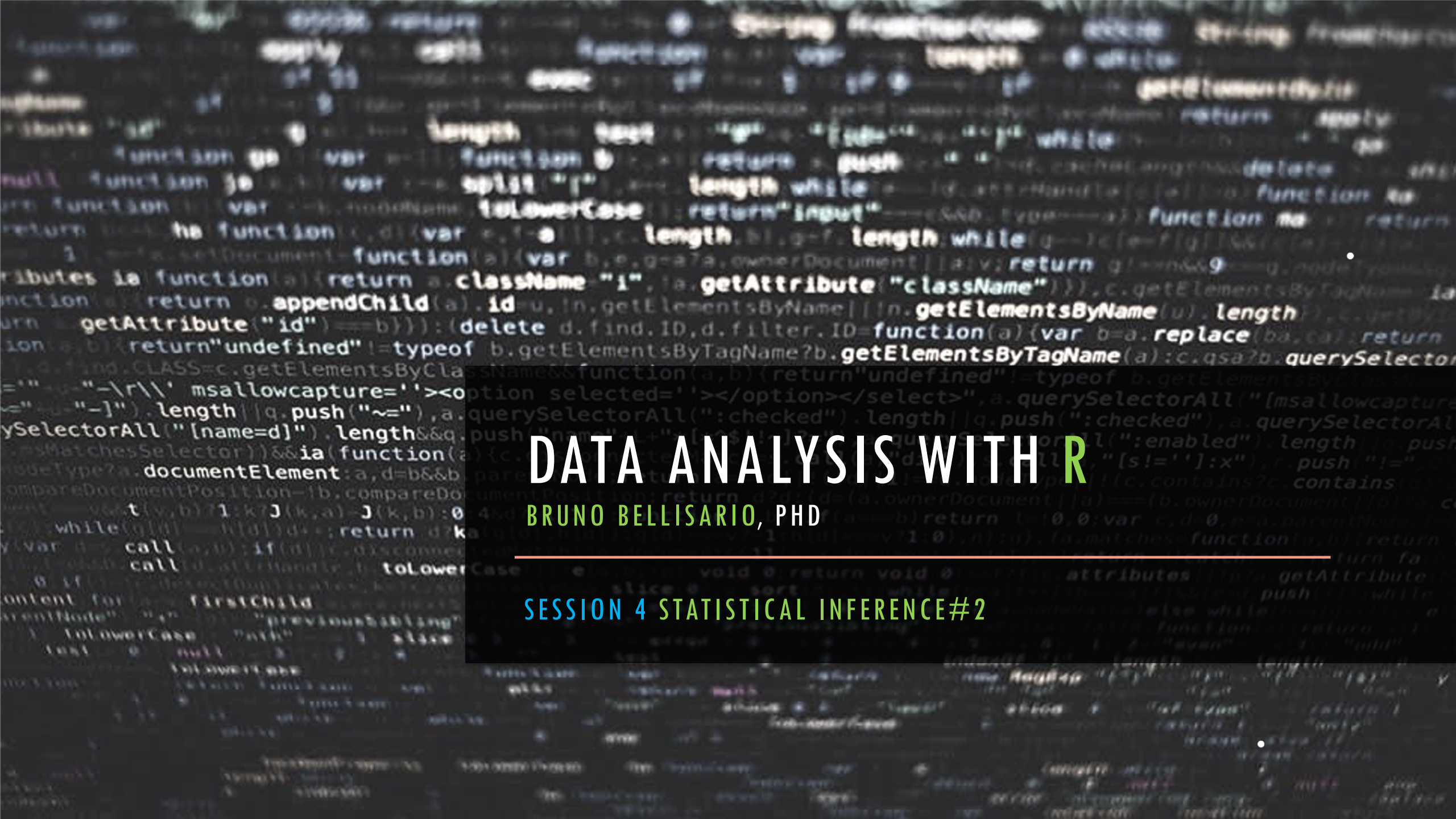
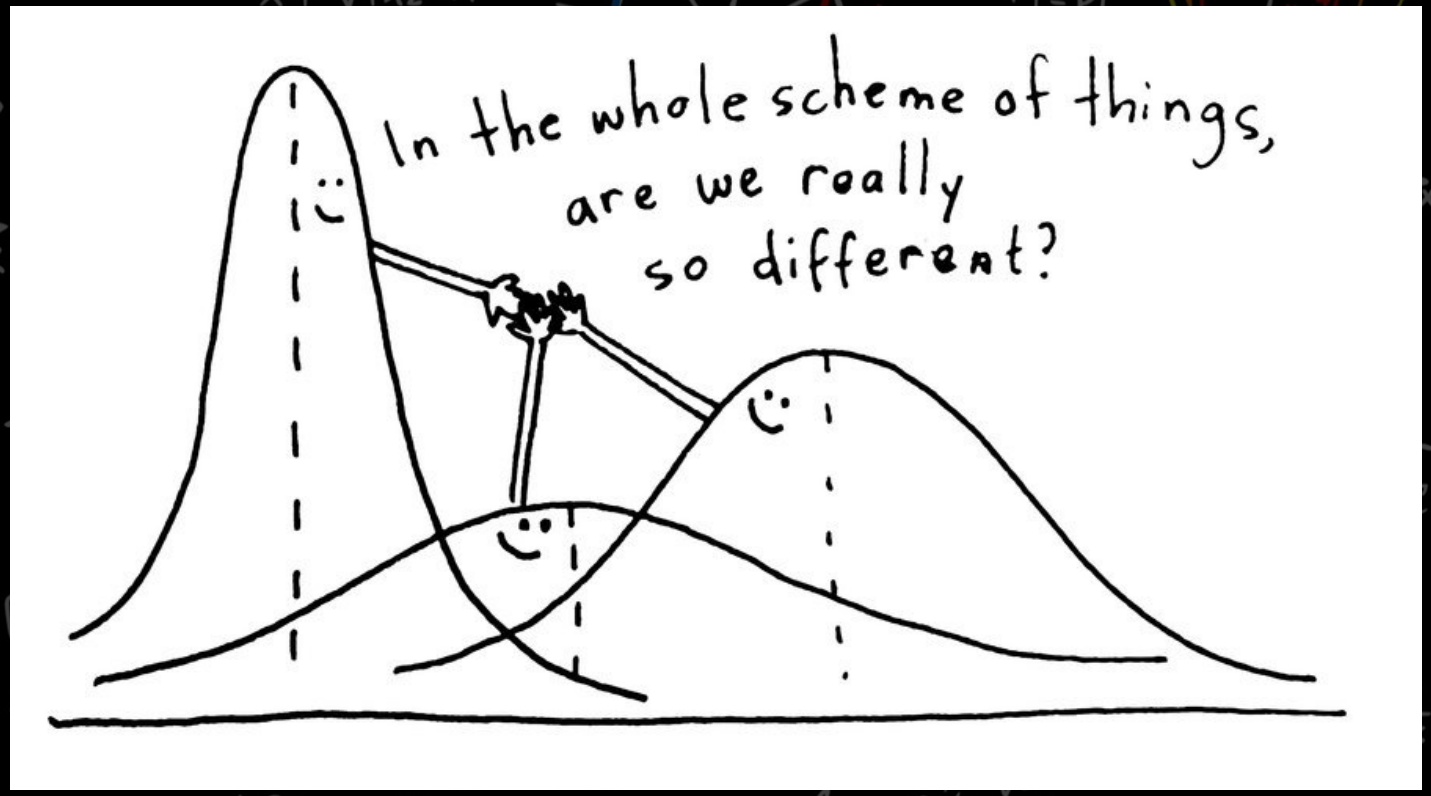


DATA ANALYSIS WITH R

BRUNO BELLISARIO, PHD

SESSION 4 STATISTICAL INFERENCE#2

ONE-WAY ANOVA



ONE-WAY ANOVA

A one-way analysis of variance (ANOVA) is like an independent t-test, except that it can compare more than two groups.

Appropriate data

→One-way data. That is, one measurement variable in two or more groups

→Dependent variable is interval/ratio, and is continuous

→Independent variable is a factor with two or more levels (two or more groups)

→Residuals are normally distributed (moderate deviation from normally-distributed residuals is permissible)

→Groups have the same variance (homoscedasticity)

→Observations among groups are independent.

Hypotheses

→Null hypothesis: The means of the measurement variable for each group are equal.

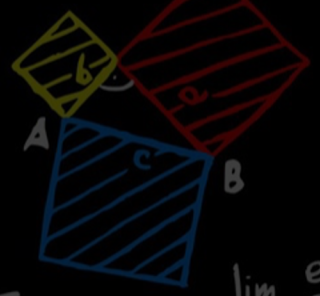
→Alternative hypothesis (two-sided): The means of the measurement variable for among groups are not equal.

ONE-WAY ANOVA W/BLOCKS

Blocks are used in an analysis of variance or similar models in order to account for suspected variation from factors other than the treatments or main independent variables being investigated.

Traditionally, in agricultural experiments, plots would be arranged into blocks according to factors in the field that could not be controlled.

The experiment might be designed in a randomized complete block design in which each block had a plot with each treatment.



ONE-WAY ANOVA W/BLOCKS

Appropriate data

→One-way data, with blocks. That is, one measurement variable in two or more groups, where each group is also distributed among at least two blocks

→Dependent variable is interval/ratio, and is continuous

→Independent variable is a factor with two or more levels.

→A second independent variable is a blocking factor variable with two or more levels

→Residuals are normally distributed (moderate deviation from normally-distributed residuals is permissible)

→Groups have the same variance (homoscedasticity)

→Observations among groups are independent.

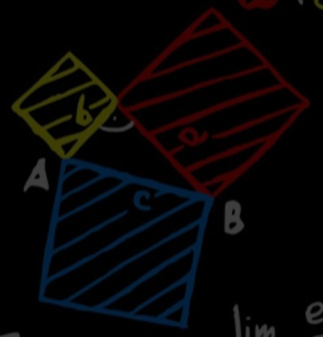
Hypotheses

→Null hypothesis: The means of the measurement variable for each group are equal

→Alternative hypothesis (two-sided): The means of the measurement variable for among groups are not equal

→An additional null hypothesis is tested for the effect of blocks: The means of the measurement variable for each block are equal

ONE-WAY ANOVA W/RANDOM BLOCKS

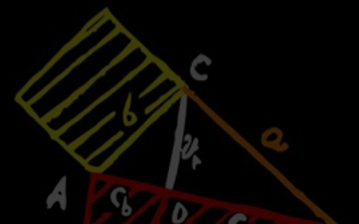


Here, the analysis is done with a mixed effects model, with the treatments treated as a fixed effect and the blocks treated as a random effect.

In analysis of variance, blocking variables are often treated as random variables.

This is because the blocking variable represents a random selection of levels of that variable.

The analyst wants to take the effects of the blocking variable into account, but the identity of the specific levels of the blocks are not of interest.



TWO-WAY ANOVA

A two-way ANOVA can investigate the main effects of each of two independent factor variables, as well as the effect of the interaction of these variables.

Appropriate data

→Two-way data

→Dependent variable is interval/ratio, and is continuous

→Independent variables are a factor with two or more levels

→Residuals are normally distributed (moderate deviation from normally-distributed residuals is permissible)

→Groups have the same variance (homoscedasticity)

→Observations among groups are independent

Hypotheses

→Null hypothesis 1: The means of the dependent variable for each group in the first independent variable are equal

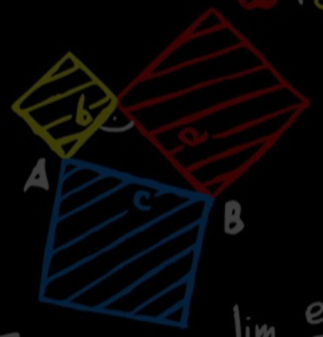
→Alternative hypothesis 1 (two-sided): The means of the dependent variable for each group in the first independent variable are not equal

REPEATED MEASURE ANOVA

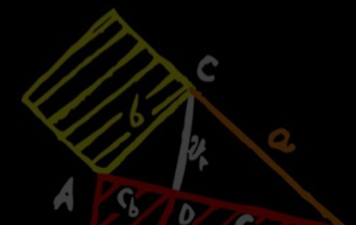
When an experimental design takes measurements on the same experimental unit over time, the analysis of the data must take into account the probability that measurements for a given experimental unit will be correlated in some way.

For example, if we were measuring calorie intake for students, we would expect that if one student had a higher intake at Time 1, that that student would have a higher intake others at Time 2, and so on.

CORRELATION & LINEAR REGRESSION



- Correlation and linear regression each explore the relationship between two quantitative variables.
- Correlation determines if one variable varies systematically as another variable changes. It does not specify that one variable is the dependent variable and the other is the independent variable.
- The three forms of correlation presented here are Pearson, Kendall, and Spearman.
- The test determining the p-value for Pearson correlation is a parametric test that assumes that data are bivariate normal. Kendall and Spearman correlation use nonparametric tests.



CORRELATION & LINEAR REGRESSION

- Linear regression specifies one variable as the independent variable and another as the dependent variable.
- The resultant model relates the variables with a linear relationship.
- The tests associated with linear regression are parametric and assume normality, homoscedasticity, and independence of residuals, as well as a linear relationship between the two variables.

CORRELATION & LINEAR REGRESSION

There are different methods to perform correlation analysis:

Pearson correlation (r), which measures a linear dependence between two variables (x and y). It's also known as a parametric correlation test because it depends to the distribution of the data. It can be used only when x and y are from normal distribution.

The plot of $y = f(x)$ is named the linear regression curve.

Kendall tau and Spearman rho are rank-based correlation coefficients (non-parametric).

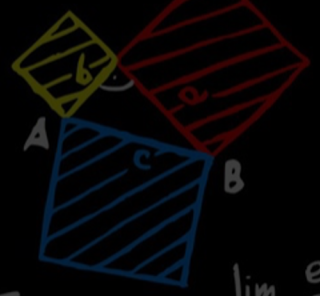
CORRELATION & LINEAR REGRESSION

In the formulas below, x and y are two vectors of length n mx and my corresponds to the means of x and y, respectively.

x' and y' correspond to the rank order of x and y

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

$$rho = \frac{\sum (x' - m_{x'})(y'_i - m_{y'})}{\sqrt{\sum (x' - m_{x'})^2 \sum (y' - m_{y'})^2}}$$



CORRELATION & LINEAR REGRESSION

The Kendall correlation method measures the correspondence between the ranking of x and y variables in a different way.

The total number of possible pairings of x with y observations is $n(n-1)/2$, where n is the size of x and y.

In the formula below, n_c is total number of concordant pairs n_d is total number of discordant pairs and n is the size of x and y

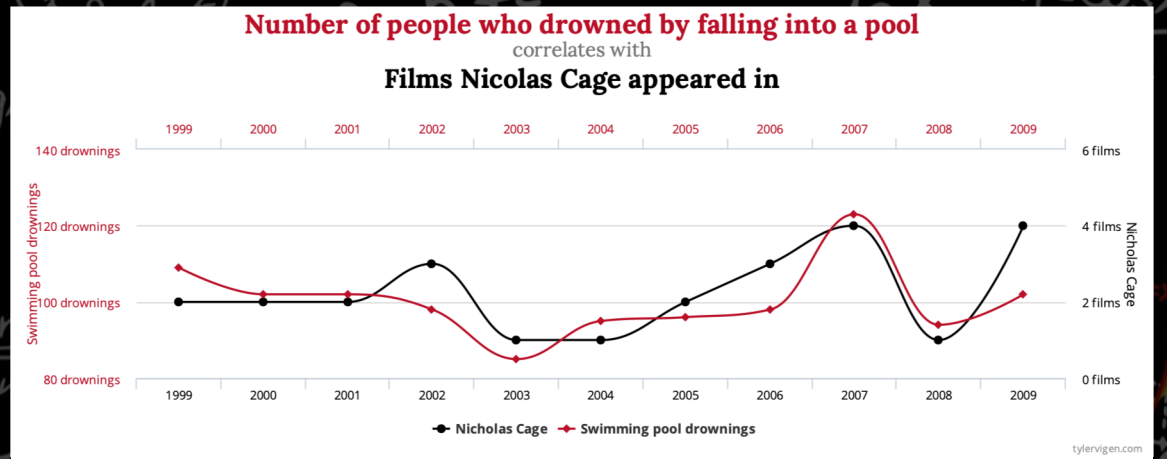
$$tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

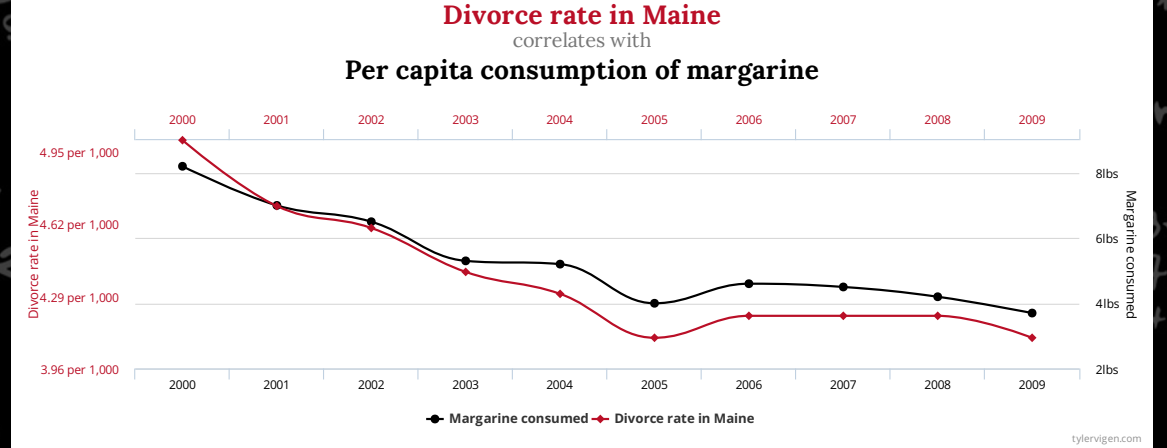
CORRELATION & LINEAR REGRESSION

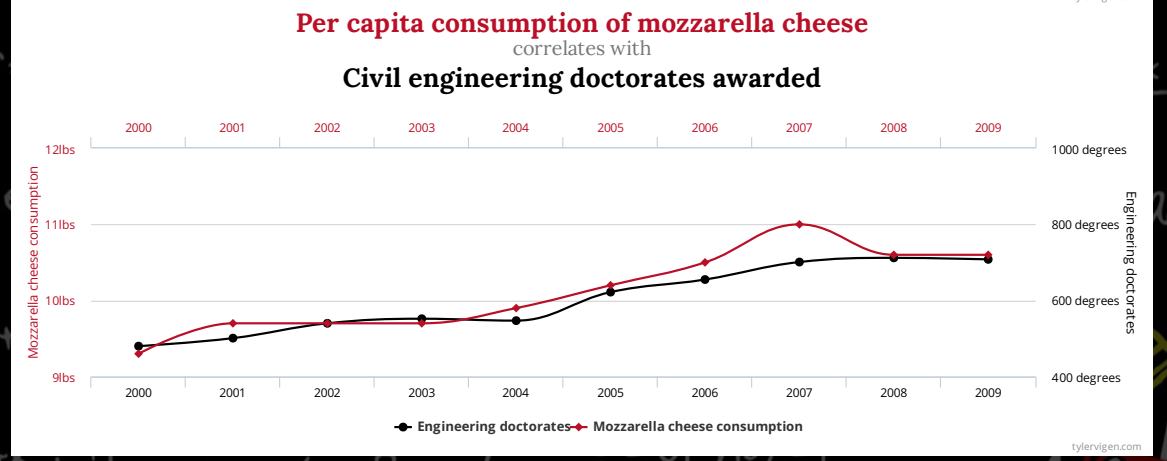
Beware of spurious correlations!

In statistics, a spurious relationship (or spurious correlation) is a mathematical relationship in which two or more events or variables are associated but not causally related, due to either coincidence or the presence of a certain third, unseen factor (referred to as a "common response variable", "confounding factor", or "lurking variable").

CORRELATION & LINEAR REGRESSION







CORRELATION & LINEAR REGRESSION

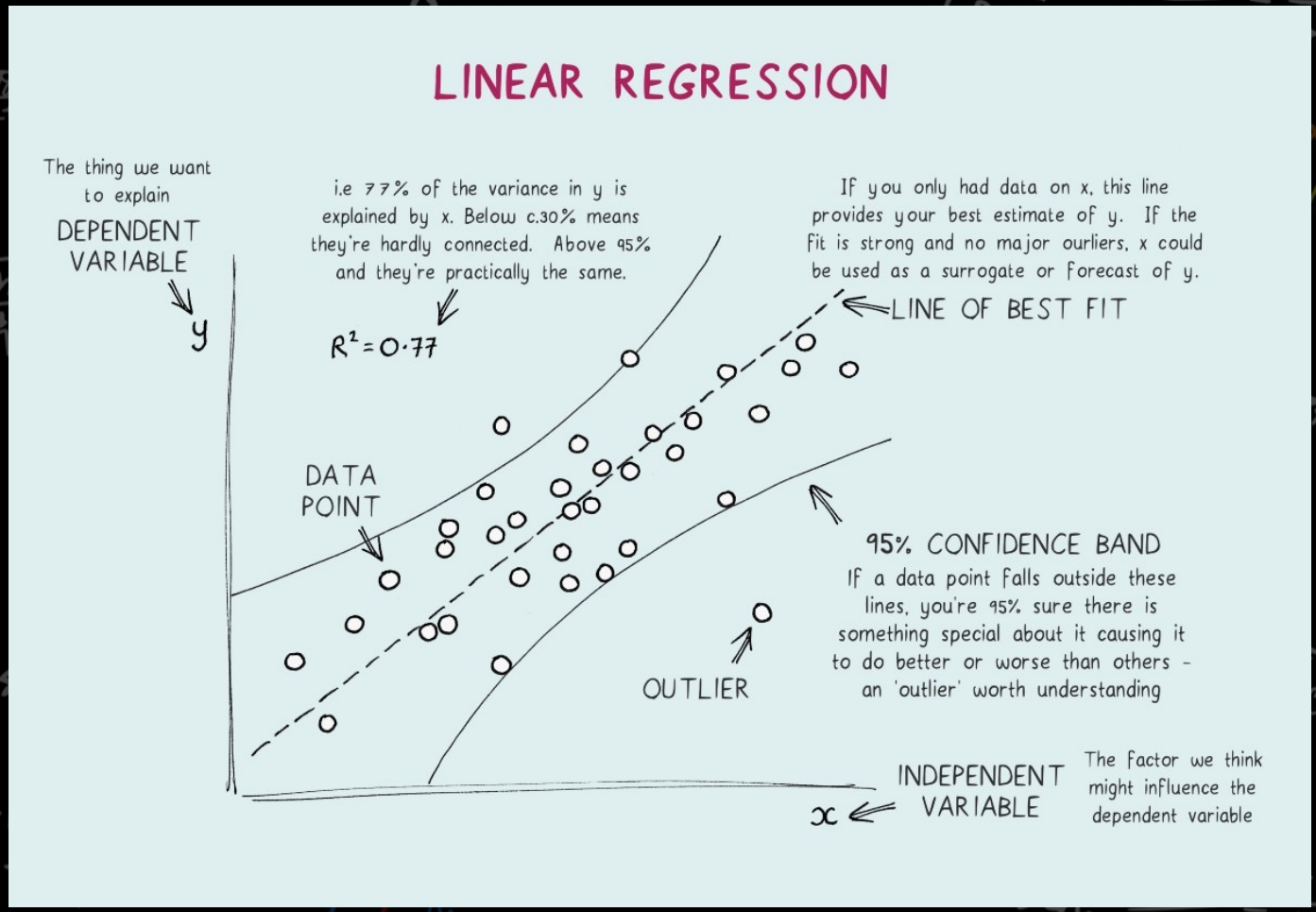
Linear regression is a very common approach to model the relationship between two interval/ratio variables.

The method assumes that there is a linear relationship between the dependent variable and the independent variable, and finds a best fit model for this relationship.

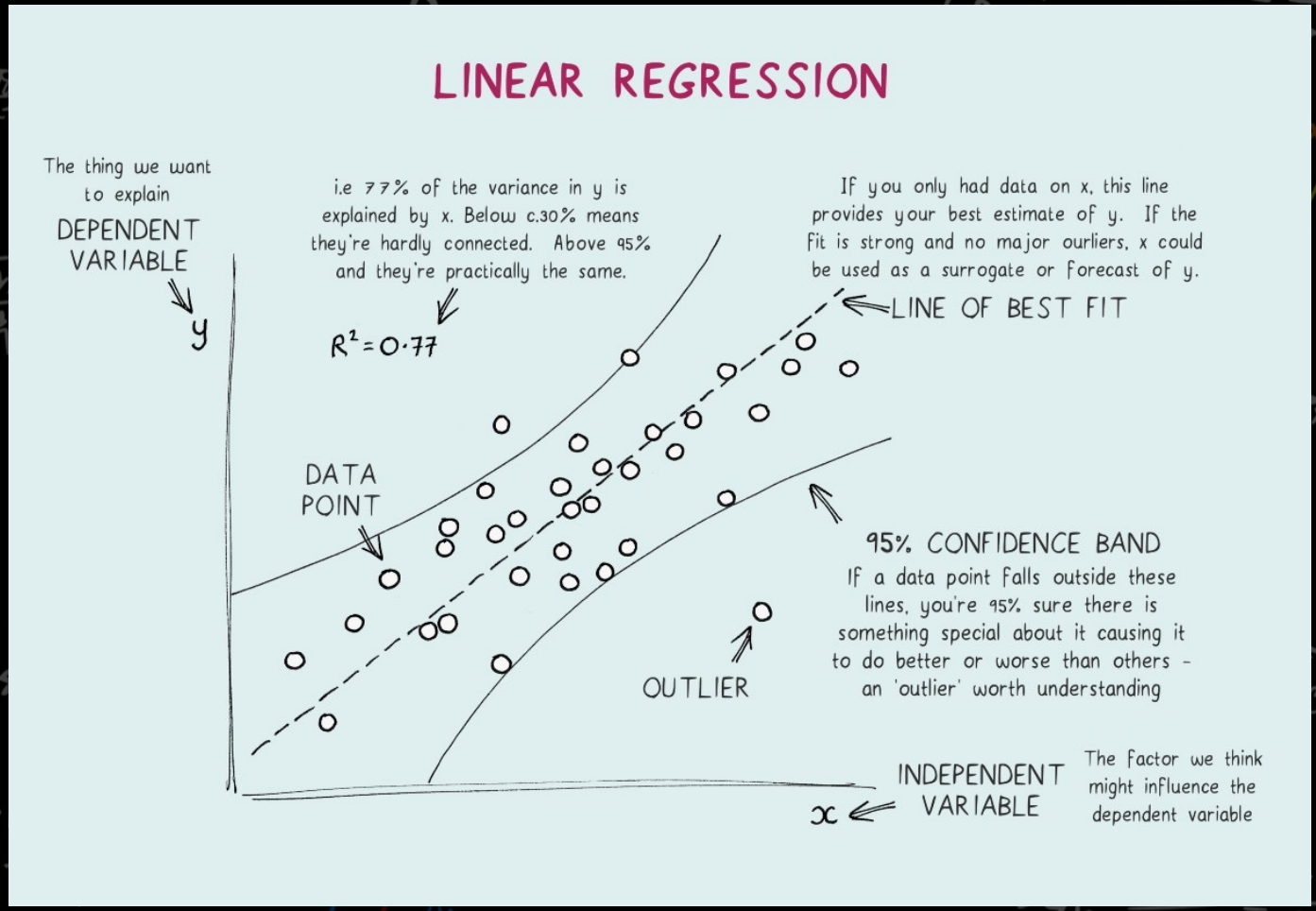
The outcome of linear regression includes estimating the intercept and the slope of the linear model.

Linear regression can then be used as a predictive model, whereby the model can be used to predict a y value for any given x. In practice, the model shouldn't be used to predict values beyond the range of the x values used to develop the model.

CORRELATION & LINEAR REGRESSION



CORRELATION & LINEAR REGRESSION



Mathematically, a regression line assumes the form of $y = ax + b$, with a the intercept (the location where it intersects an axis) and b the slope (the steepness of a line).

Both can be used to estimate an average rate of change. The greater the magnitude of the slope, the steeper the line and the greater the rate of change.

CORRELATION & LINEAR REGRESSION

If the independent variable were of nominal type, then the linear regression would become a one-way analysis of variance!!

Indeed, in R, the same syntax can be used to perform both ANOVA and linear regression!

ASSESSING MODEL ASSUMPTIONS: FORMAL TESTS TO ASSESS NORMALITY OF RESIDUALS

There are formal tests to assess the normality of residuals.

Common tests include Shapiro-Wilk, Anderson–Darling, Kolmogorov–Smirnov, and D'Agostino–Pearson.

Their results are dependent on sample size.

When the sample size is large, the tests may indicate a statistically significant departure from normality, even if that departure is small.

When sample sizes are small, they won't detect departures from normality!!

ASSESSING MODEL ASSUMPTIONS: FORMAL TESTS TO ASSESS NORMALITY OF RESIDUALS

Skewness and kurtosis

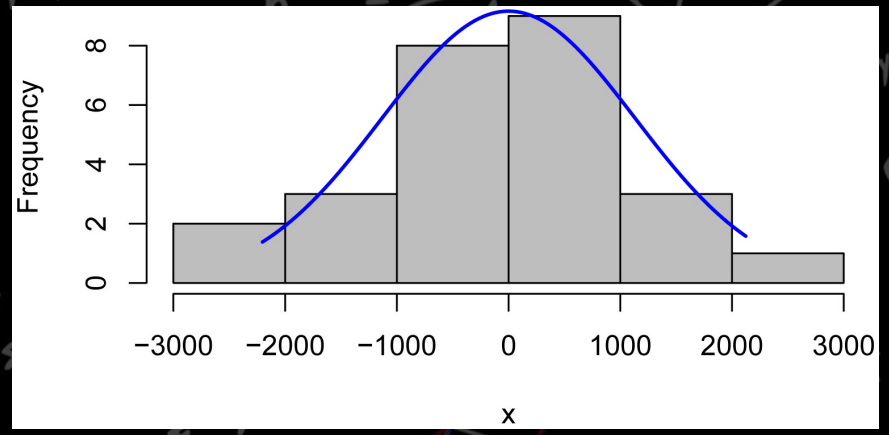
There are no definitive guidelines as to what range of skew or kurtosis are acceptable for considering residuals to be normally distributed.

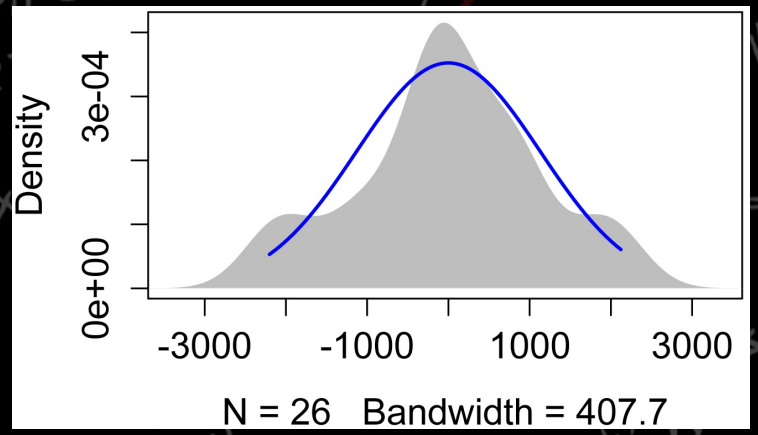
If we were forced to give advice for skewness calculations, be cautious if the absolute value is > 0.5, and consider it not normally distributed if the absolute value is > 1.

Some authors use 2 as a cut-off for normality, and others use a higher limit for kurtosis.

ASSESSING MODEL ASSUMPTIONS: FORMAL TESTS TO ASSESS NORMALITY OF RESIDUALS

Usually, the best method to see if model residuals meet the assumptions of normal distribution and homoscedasticity are to plot them and inspect the plots visually.

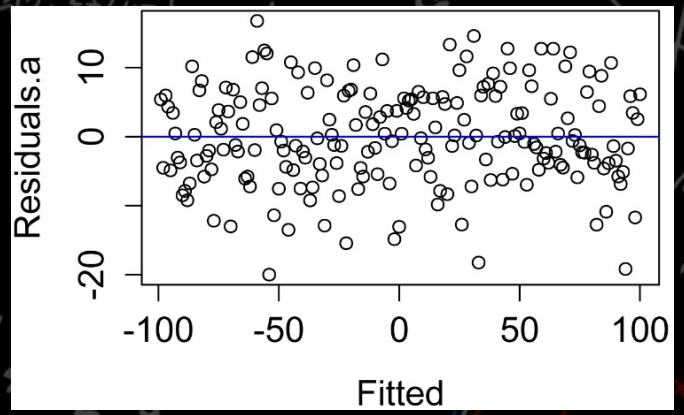




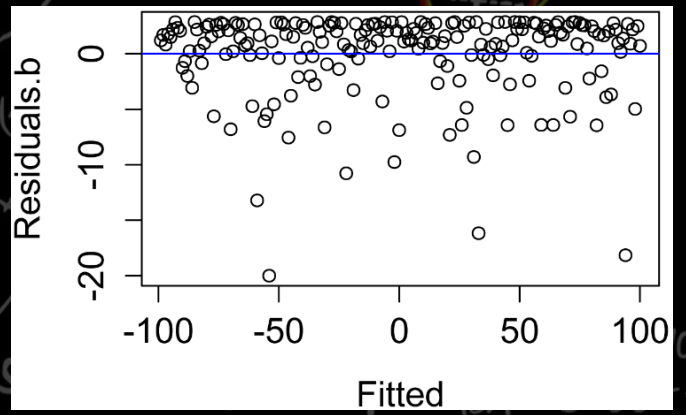
ASSESSING MODEL ASSUMPTIONS: FORMAL TESTS TO ASSESS NORMALITY OF RESIDUALS

Patterns in the plot of residuals versus fitted values can indicate a lack of homoscedasticity or that errors are not independent of fitted values.

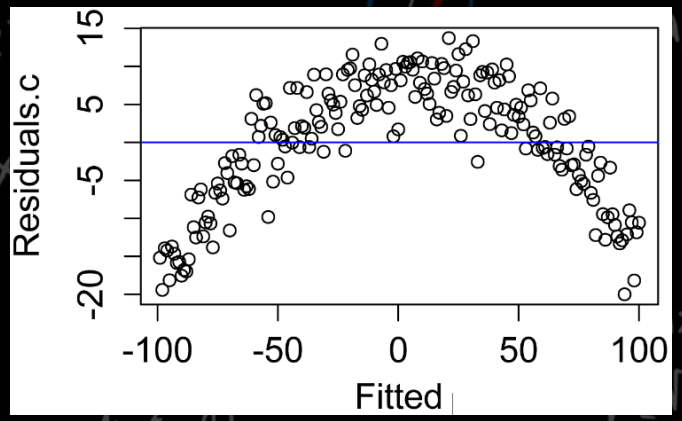
Normally distributed and homoscedastic



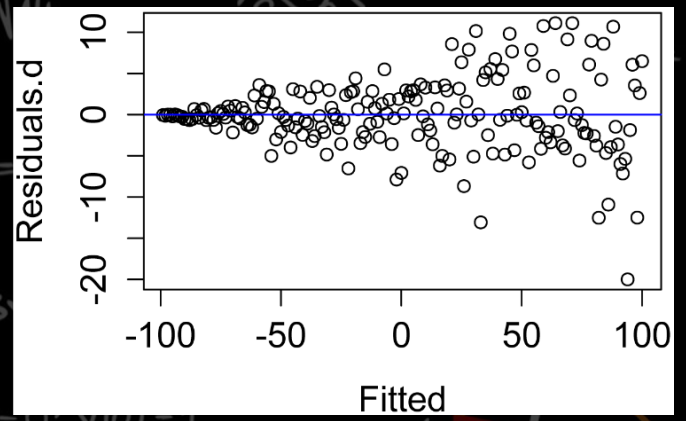
Non-normal distribution of residuals



Not independent of the fitted values



Heteroscedasticity



MODEL SELECTION: THE AKAIKE INFORMATION CRITERION

AIC

The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data. AIC is calculated from:

→The number of independent variables used to build the model.

→The maximum likelihood estimate of the model (how well the model reproduces the data).

→The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest possible independent variables.

MODEL SELECTION: THE AKAIKE INFORMATION CRITERION

AIC

AIC is most often used for model selection. By calculating and comparing the AIC scores of several possible models, you can choose the one that is the best fit for the data.

When testing a hypothesis, you might gather data on variables that you aren't certain about, especially if you are exploring a new idea. You want to know which of the independent variables you have measured explain the variation in your dependent variable.

A good way to find out is to create a set of models, each containing a different combination of the independent variables you have measured. These combinations should be based on:

→Your knowledge of the study system – avoid using parameters that are not logically connected, since you can find spurious correlations between almost anything!

→Your experimental design – for example, if you have split two treatments up among test subjects, then there is probably no reason to test for an interaction between the two treatments.

MODEL SELECTION: THE AKAIKE INFORMATION CRITERION

AIC

Once you've created several possible models, you can use AIC to compare them.

Lower AIC scores are better, and AIC penalizes models that use more parameters.

If two models explain the same amount of variation, the one with fewer parameters will have a lower AIC score and will be the better-fit model.

MODEL SELECTION: THE AKAIKE INFORMATION CRITERION

AIC

AIC determines the relative information value of the model using the maximum likelihood estimate and the number of parameters (independent variables) in the model.

AIC = 2K-2Ln(L)

K is the number of independent variables used and L is the log-likelihood estimate (a.k.a. the likelihood that the model could have produced your observed y-values).

To compare models using AIC, you need to calculate the AIC of each model. If a model is more than 2 AIC units lower than another, then it is considered significantly better than that model.