UiT The Arctic University of Norway

# FSK-2053 Data science & bioinformatics for fisheries and aquaculture

## Lecture 4: Statistics for Big Data

*Basic Statistics and Modeling*

Daniel Kumazawa Morais
*daniel.morais@uit.no*

# Learning objectives: Lecture 5

**4.1 Major fields and current "philosophical schools" in statistical analysis**

- Understand the different goals of the two major fields of statistical analysis: descriptive statistics vs inferential statistics.

- Understand the basic differences between the two major philosophical currents in inferential statistics: the frequentist paradigm vs the Bayesian paradigm.

**4.2. Frequentist methods for hypothesis testing**

- Become familiar with basic methods for hypothesis testing and learn how to decide which of them to use, depending on the types of the variables to be tested and their distributions.

- Understand the differences between parametric and non-parametric tests and how to check the assumptions to determine if a parametric test can be used.

**4.3. Linear models in R**

- Learn how to apply basic and multiple linear models in R to test the effects of categorical factors and/or numerical variables on a dependent variable.

- Introduce the concept of generalized linear models and explore their possible applications.

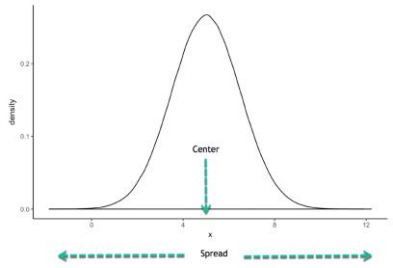
# Two major fields of statistical analysis

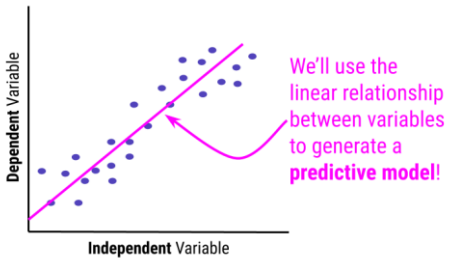
## Descriptive statistics

Deals with how variables, samples and populations are structured and distributed. It is focused on describing the **structure** of the variables. Which summarizing **parameters of central tendency and dispersion** (variability) can be used to describe a sample? What families of **distribution functions** can be used to describe a sample or a population? Descriptive methods are crucial to generate plots for **Exploratory Data Analysis**.

## Inferential statistics

Deals with how variables, samples and populations work and are related to each other. It is focused on analyzing the **function** of the variables. Inferential statistics uses **hypothesis testing** to find insights into how two or more variables are associated or related to each other. This allows to establish **statistical models** to explain how the studied systems have behaved in the past and/or to **make predictions** (inferences) about their future behaviour.




**Free online references for this session:**

Steve Midwey's **Data Analysis in R:**
https://bookdown.org/steve_midway/DAR/

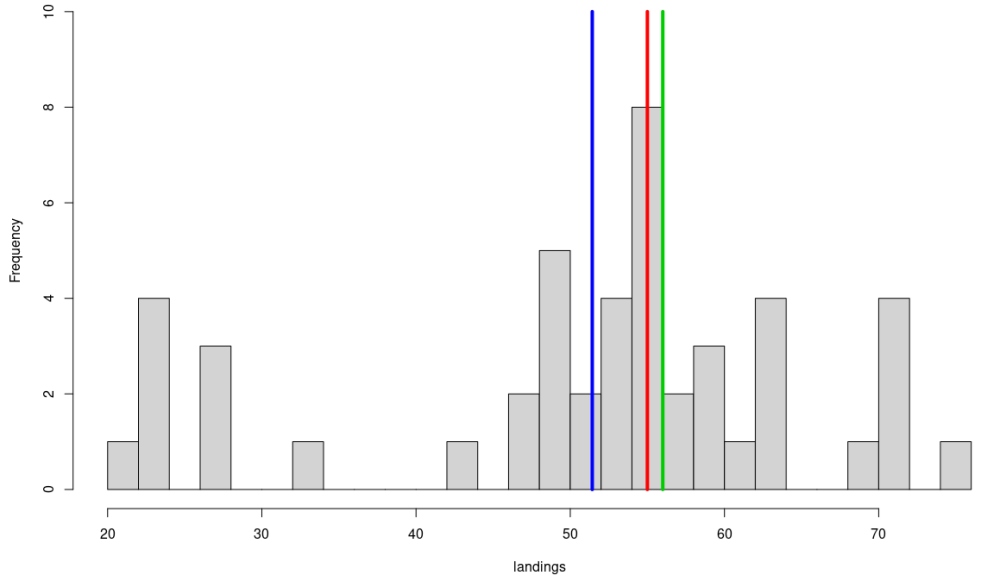
Carrie Wright et al. **Tidyverse Skills for Data Science:**
https://jhudatascience.org/tidyversecourse/model.html

# Descriptive statistics

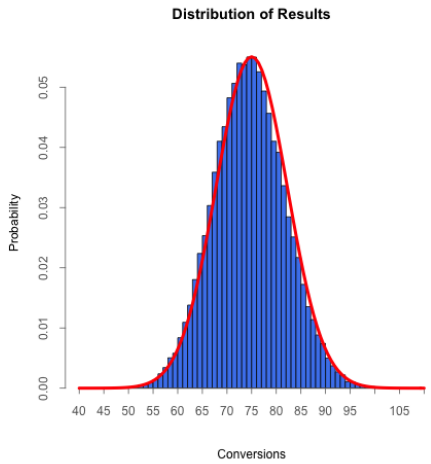
- **Measures of central tendency**
  - **mean**
  - **median**
  - **mode**

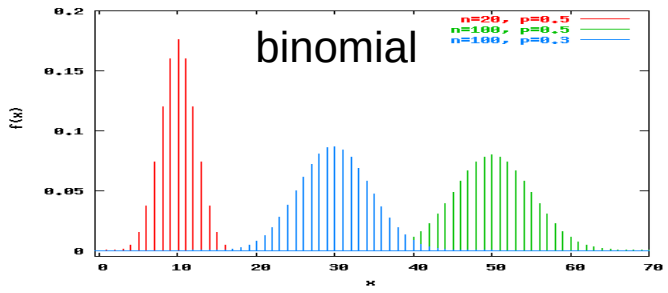
- **Measures of dispersion**
  - standard deviation
  - variance
  - interquartile range
  - range = max - min

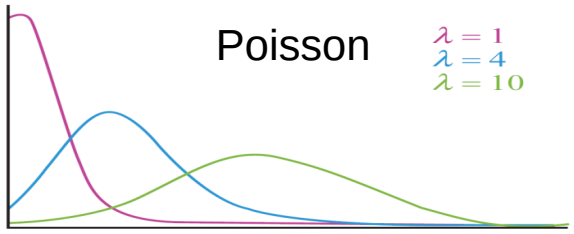
- **Probability distributions**
  - histograms
  - density functions
  - distribution families




normal



# Inferential statistics

Currently, two contrasting philosophical schools are "fighting" to dominate the field of statistical inference:

- **Frequentist statistics**
- **Bayesian statistics**

## The frequentist paradigm:

- Assume that data are random variables acquired from sampling.
- Assume that parameters (e.g. population mean, population variance) do not change and are often referred to as fixed and unknowable.
- All experiments are independent in the sense that no prior knowledge can be (directly) provided to a parameter estimate or model.
- Knowledge is driven by point estimates and ultimately is hypothesis-driven in the sense that we **accept or reject hypotheses** and outcomes.
- **p-values** are a key outcome in frequentist estimation.

## In contrast, Bayesians:

- Assume that data are fixed; data are the "real" things that are knowable, and parameters are random variables that we are seeking to estimate (based on the known data).
- Adopt a degree-of-belief from probability.
- Can update beliefs in the sense that prior information can be used to directly modify the estimate of certain parameters.
- Bayesian estimation is driven by probability distributions and uncertainty (as opposed to point estimates).
- Outcomes are **not required to accept or reject anything** based on a critical value; instead, **probabilistic interpretations of outcomes** are generated that may not distill to a simple yes or no but may in fact be more realistic to the question at hand.

# Bayes' theorem

prior probability

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}$$

posterior probability

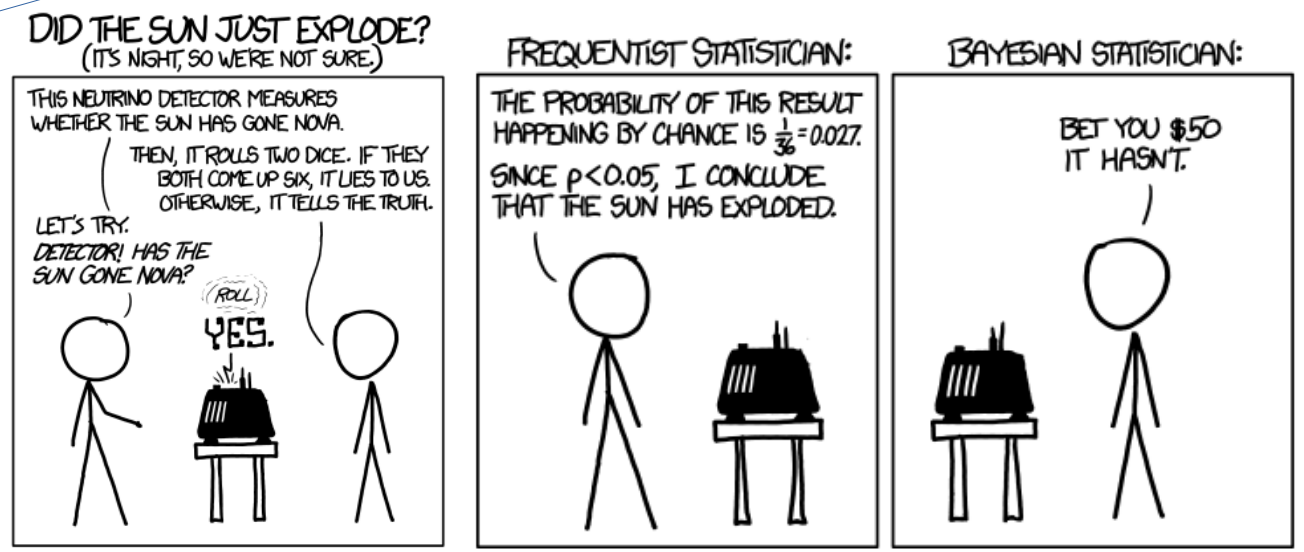
- $H$ stands for any *hypothesis* whose probability may be affected by data (called *evidence* below). Often there are competing hypotheses, and the task is to determine which is the most probable.
- $E$, the *evidence*, corresponds to new data that were not used in computing the prior probability.
- $P(H \mid E)$, the *posterior probability*, is the probability of $H$ *given* $E$, i.e., *after* $E$ is observed. This is what we want to know: the probability of a hypothesis *given* the observed evidence.
- $P(E \mid H)$ is the probability of observing $E$ *given* $H$, and is called the *likelihood*. As a function of $E$ with $H$ fixed, it indicates the compatibility of the evidence with the given hypothesis. The likelihood function is a function of the evidence, $E$, while the posterior probability is a function of the hypothesis, $H$.
- $P(H)$, the *prior probability*, is the estimate of the probability of the hypothesis $H$ *before* the data $E$, the current evidence, is observed.
- $P(E)$ is sometimes termed the marginal likelihood or "model evidence". This factor is the same for all possible hypotheses being considered (as is evident from the fact that the hypothesis $H$ does not appear anywhere in the symbol, unlike for all the other factors), so this factor does not enter into determining the relative probabilities of different hypotheses.

**The frequentist** asks:

"Given a certain set of fixed parameters that this variable has, how likely am I to observe that data value? Depending on this probability, I will accept or reject my hypotheses about the values of those parameters, and my view of the world will be enhanced."

While **the Bayesian** asks:

"The only thing I can know is what I observe; based on the data I observe, what are the most likely parameter values I could infer? I will generate a probabilistic model to inform me about which are the most likely values for those parameters, which will provisionally enhance my view of the world, until I get better information."



# A practical example of Bayes' theorem: Do I have cancer?

- 1% of women have breast cancer (and therefore 99% do not).
- 80% of mammograms detect breast cancer when it is there (and therefore 20% miss it).
- 9.6% of mammograms detect breast cancer when it's **not** there (and therefore 90.4% correctly return a negative result).

| | Cancer (1%) | No Cancer (99%) |
|---|---|---|
| Test Pos | 80% | 9.6% |
| Test Neg | 20% | 90.4% |

$$\Pr(\mathrm{H}|\mathrm{E}) = \frac{\Pr(\mathrm{E}|\mathrm{H})\,\Pr(\mathrm{H})}{\Pr(\mathrm{E}|\mathrm{H})\,\Pr(\mathrm{H}) + \Pr(\mathrm{E}|\text{not H})\,\Pr(\text{not H})}$$

The chance evidence is real (supports a hypothesis) is the chance of a true positive among all positives (true or false)

$$0.078 = \frac{0.80 \times 0.01}{0.80 \times 0.01 + 0.096 \times 0.99}$$

Given a positive result in the mammogram, a simplistic frequentist could think that the probabilities of you having cancer are 80%, since this is the probability of a true positive result.

But when you consider the prevalence of cancer in the population (1%), the numbers change:

- $\Pr(\mathrm{H}|\mathrm{E})$ = Chance of having cancer (H) given a positive test (E). This is what we want to know: How likely is it to have cancer with a positive result? In our case it was 7.8%.
- $\Pr(\mathrm{E}|\mathrm{H})$ = Chance of a positive test (E) given that you had cancer (H). This is the chance of a true positive, 80% in our case.
- $\Pr(\mathrm{H})$ = Chance of having cancer (1%).
- $\Pr(\text{not H})$ = Chance of not having cancer (99%).
- $\Pr(\mathrm{E}|\text{not H})$ = Chance of a positive test (E) given that you didn't have cancer (not H). This is a false positive, 9.6% in our case.

The odds of really having cancer after you get a positive test result are just of 7.8%.

## Why frequentist methods are still so popular?

Why have frequentist approaches dominated and why might this be the first time you are hearing about Bayesian estimation?

Frequentist approaches were popular for much of the 20th century, and this is largely due to several reasons:

- Frequentist approaches are often operationalized within simple point and click routines and they do not require much computational power. As you may learn **Bayesian approaches can often require a lot of computational power** making them challenging to implement (until the advent of modern computers).

- Frequentist approaches were largely developed by R.A. Fisher who also popularized the **analysis of variance or ANOVA**. It does help your cause when a powerful and easy to use model is developed in parallel to an estimation routine.

- Finally, **p-values** have also helped the popularity of frequentist approaches. Although you may have questioned p-values yourself or read articles about concerns with using p-values, the reality for many people is that **an accept/reject framework remains simplistically attractive** for interpreting outcomes.

# Frequentist inference: hypothesis testing
# Parametric vs non-parametric tests

A frequentist **hypothesis test** involve the following elements:

- Model assumptions
- Null and alternative hypothesis
- A test statistic. This needs to have the property that extreme values of the test statistic cast doubt on the null hypothesis.
- A mathematical theorem saying, "If the model assumptions and the null hypothesis are both true, then the sampling distribution of the test statistic has this particular form."

Many commonly-used tests require that the tested variables follow normal distributions, or that variances are equal. These tests are usually called **Parametric tests**. The outcome of a parametric test is strictly valid only if all model assumptions are met. Examples of parametric tests are **Student's t test**, **ANOVA**, and **Pearson's correlation (r)**.

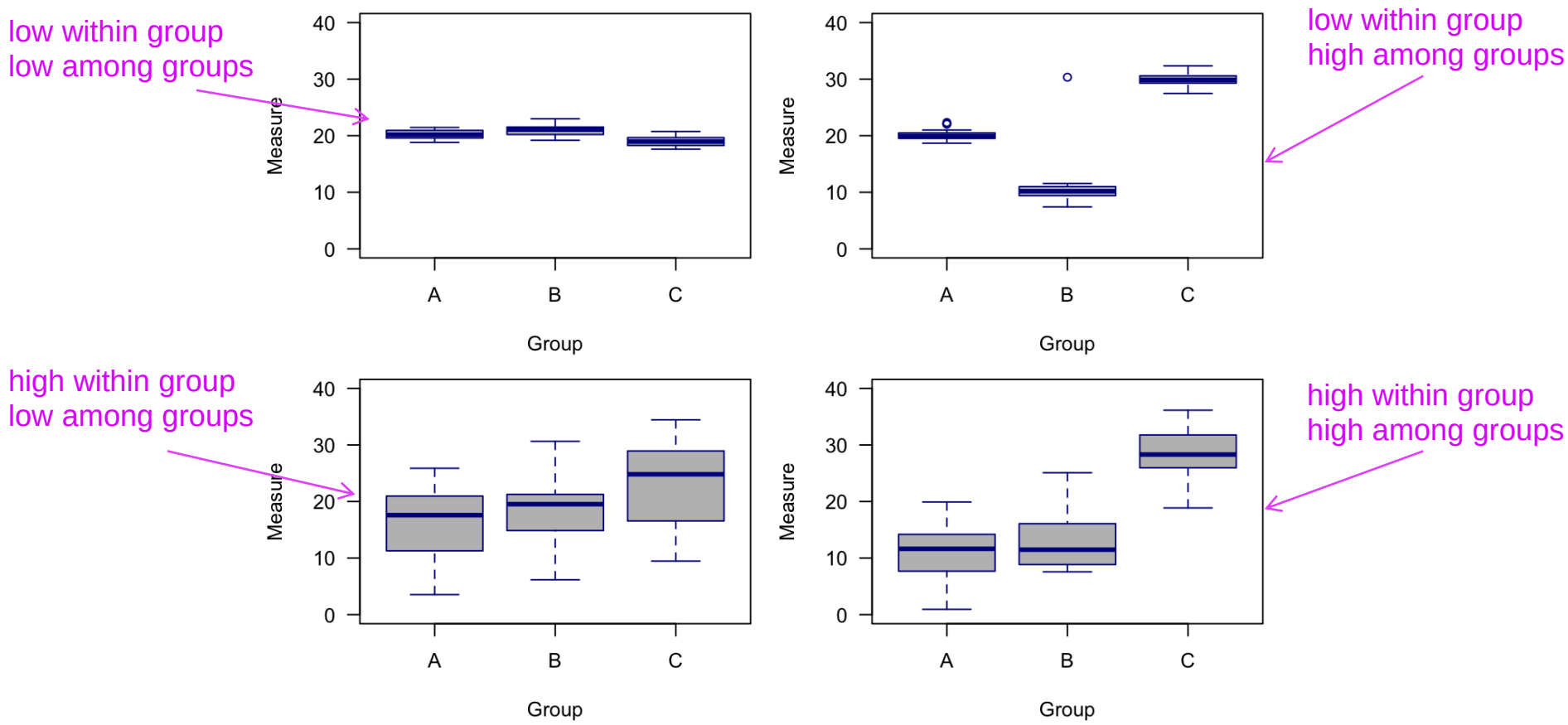
Other statistical tests do not require so strict assumptions and are widely applicable, although they often have lower statistical power to detect differences between samples. These are called **Non-parametric tests**. Many of these tests are not based on the values of the data, but in their ranks (how the values are ordered). Examples of non-parametric tests are **Mann-Whitney-Wilcoxon U-test**, **Kruskal-Wallis** test or **Spearman's correlation (ρ)**

**Some statisticians argue that the central limit theorem implies that large random samples automatically approximate normal distributions, so parametric tests would be generally valid for large samples (n>30).**

# Most used parametric & non-parametric tests

| | Parametric test | R function | Non-parametric test | R function |
|---|---|---|---|---|
| Comparing the means of two groups of a numerical variable | **Student's t** | t.test() | **Mann-Whitney-Wilcoxon** | wilcox.test() |
| Comparing the means of three or more groups of a numerical variable as a function of a categorical factor | **ANOVA** | aov() | **Kruskal-Wallis** | kruskal.test() |
| Post-hoc tests for testing pairwise differences between groups | **Tukey HSD** | TukeyHSD() | **Dunn's test (pairwise Mann-Whitney-Wilcoxon)** | pairwise.wilcox() |
| Comparing the association between two numerical variables | **Pearson's correlation** | cor.test() | **Spearman's correlation** | cor.test(,method="spearman") |
| Assessing normality of a distribution | | | **Shapiro-Wilk's test** | shapiro.test() |
| Assessing equality of variances | | | **Bartlett's test Levene test** | bartlett.test() car::leveneTest() |
| Comparing the means of a numerical variable as a function of many factors (both categorical and/or numerical) | **linear models** | lm( A ~ B) | **generalized linear models** | glm(A ~ B) |

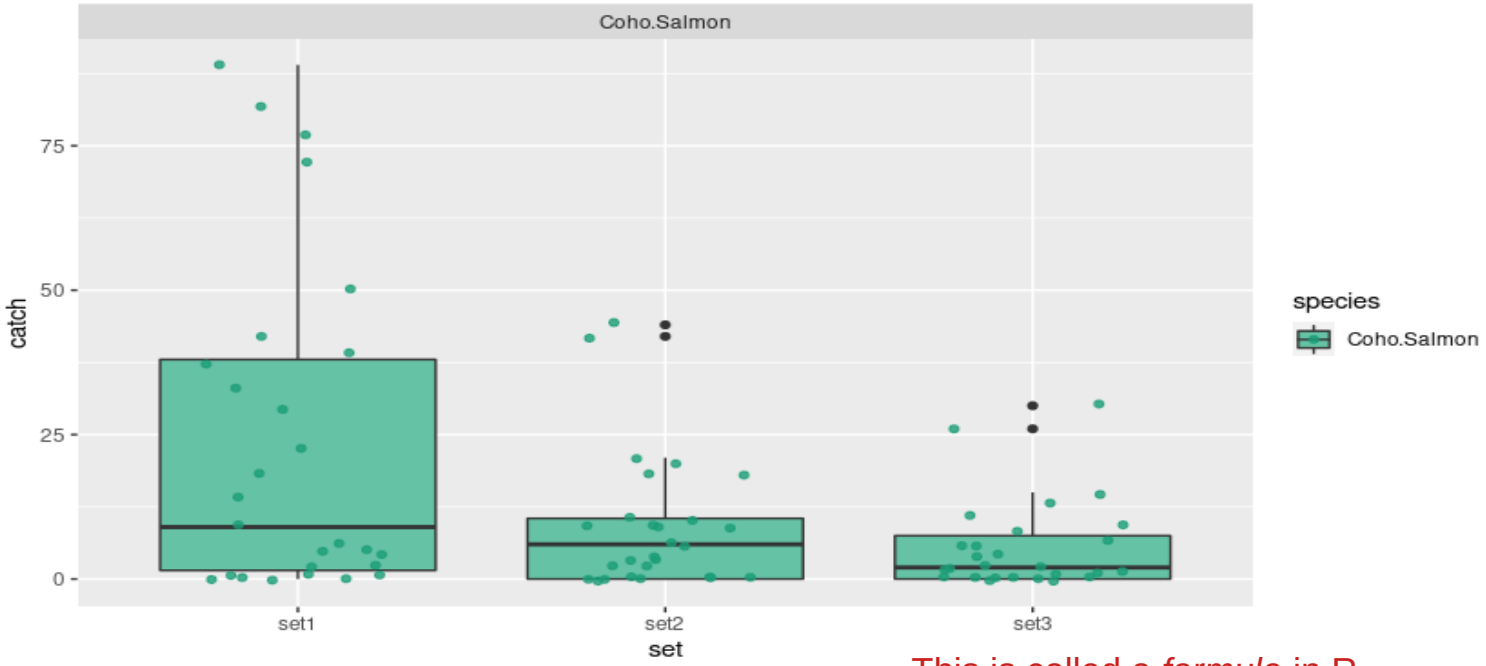
# 2.1. Relationship between a numerical variable and a categorical factor: ANOVA




ANOVA is based on comparing the variability among groups with the variability within groups.
If among-groups variability is higher than within groups, then the test will detect significant differences between the average values of the groups.

A post-hoc test must be run to determine which groups are different to which ones.

# ANOVA table




This is called a *formula* in R

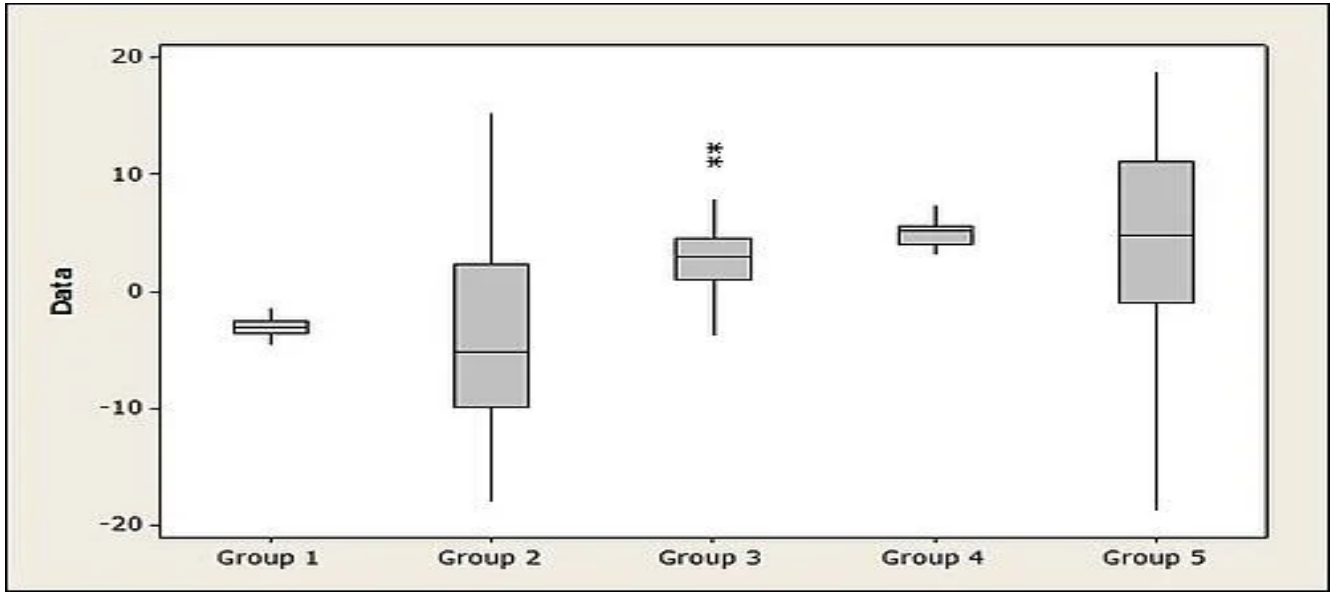
```
anova_test_coho <- aov(catch ~ set, data=dat_coho)

summary(anova_test_coho)

             Df Sum Sq Mean Sq F value  Pr(>F)
set           2   5016    2508    7.51 0.00104 **
Residuals    78  26049     334
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

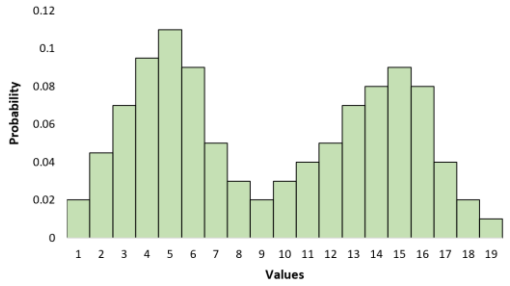
*F* is the ratio of among-groups variance divided by within-groups variance

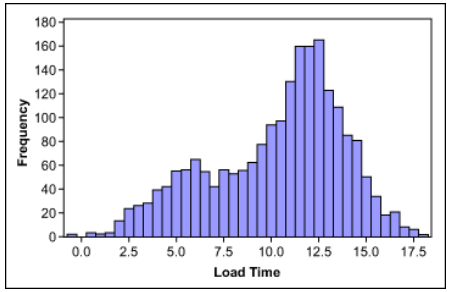
# Assumptions of ANOVA: homoscedasticity & normality of residuals

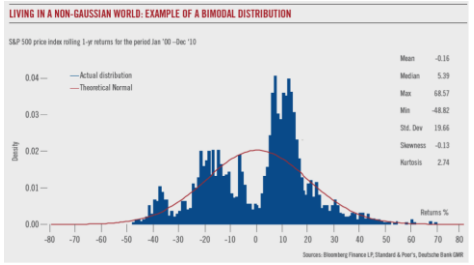



Groups with significantly different variances cannot be strictly compared by a parametric ANOVA

Non-normal distribution of variables also make the comparisons unreliable.

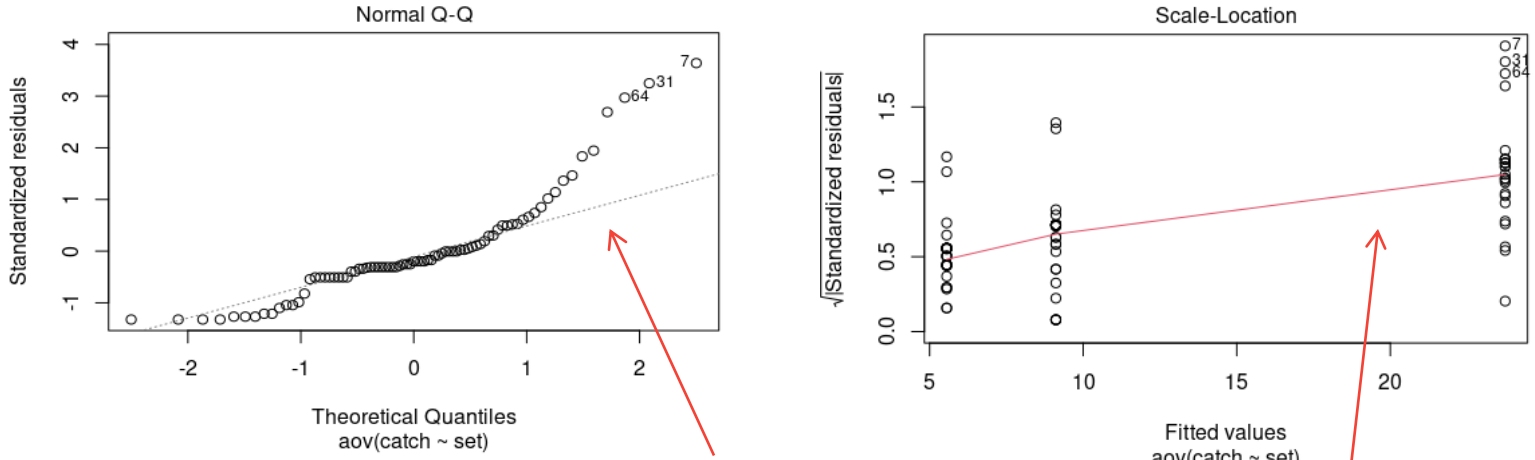


# Testing the assumptions of ANOVA:

After performing an ANOVA, we must check if:
- residuals are distributed normally
- variances of the different groups are equal (homoscedasticity)




Residuals are not well distributed along a straight line in the Q-Q plot => they are not normally distributed

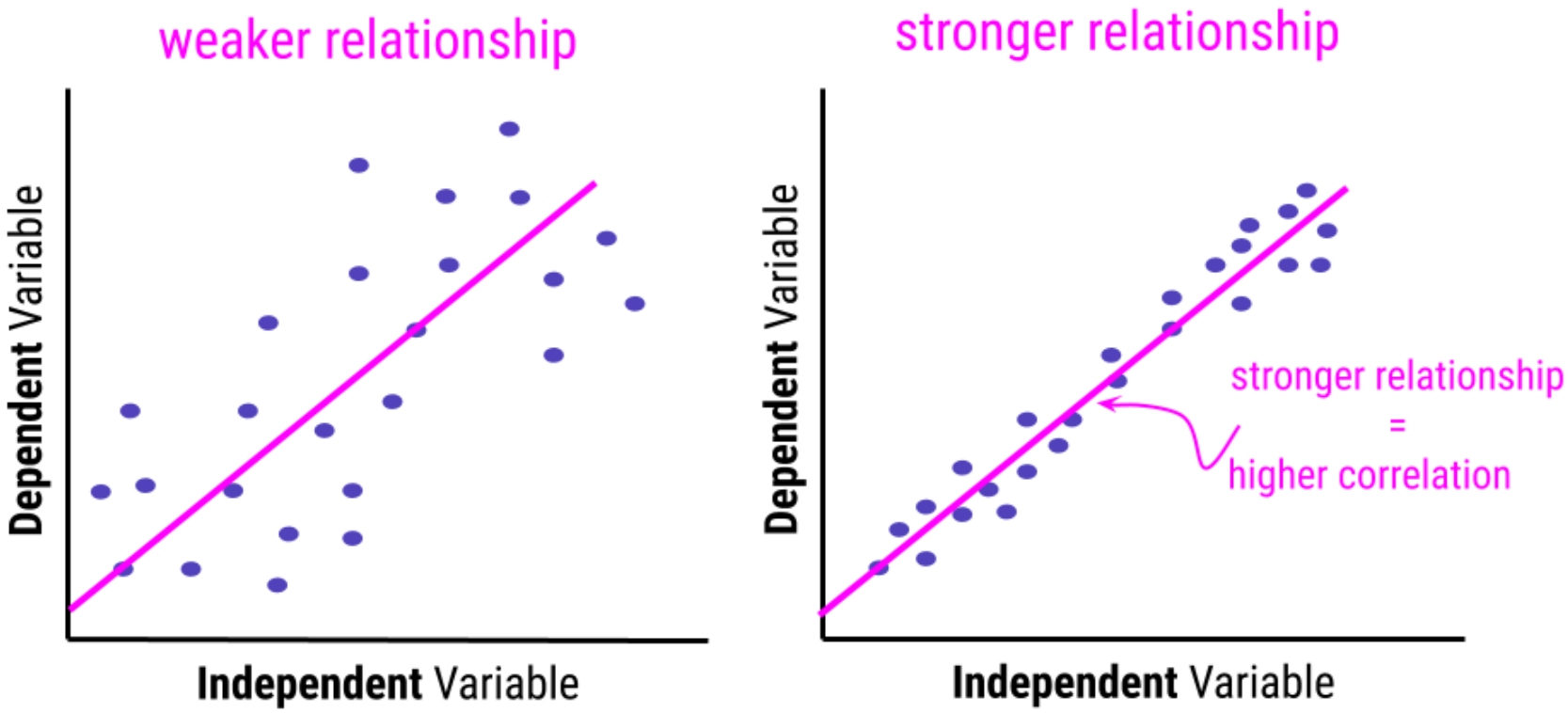
This line is not horizontal, so the variances of the three groups are not equal.

If any of these assumptions is not met, then the p-value estimation is not reliable.
Then we can:

- Try to transform the data, re-do the ANOVA and check that the assumptions are now met.

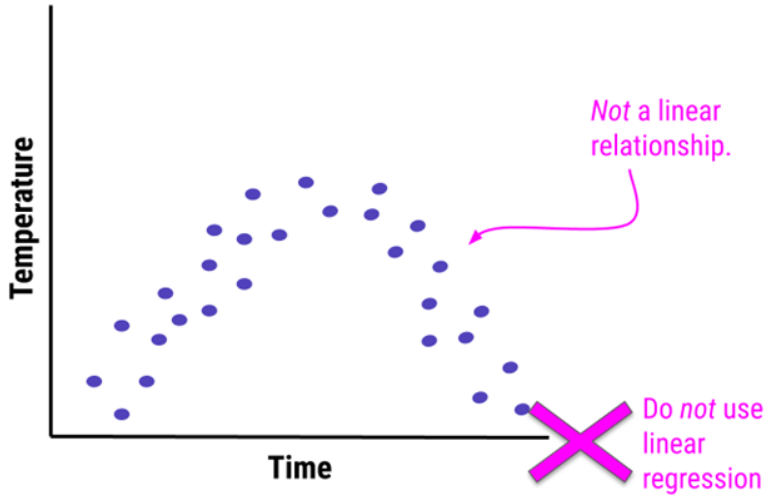
- Use a non-parametric alternative (Kruskal-Wallis test)

# 2.2. Relationship between 2 numerical variables: linear regression

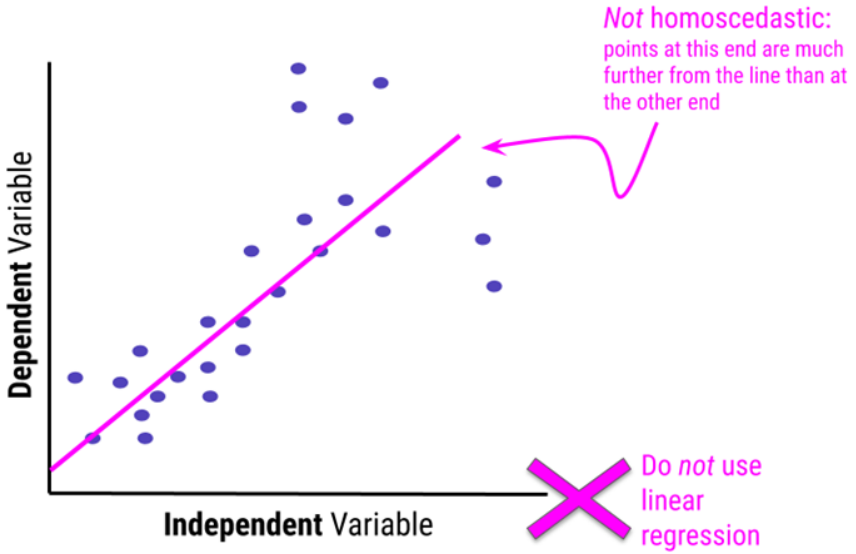


# Assumptions of linear regression: linearity, homoscedasticity & normality of residuals

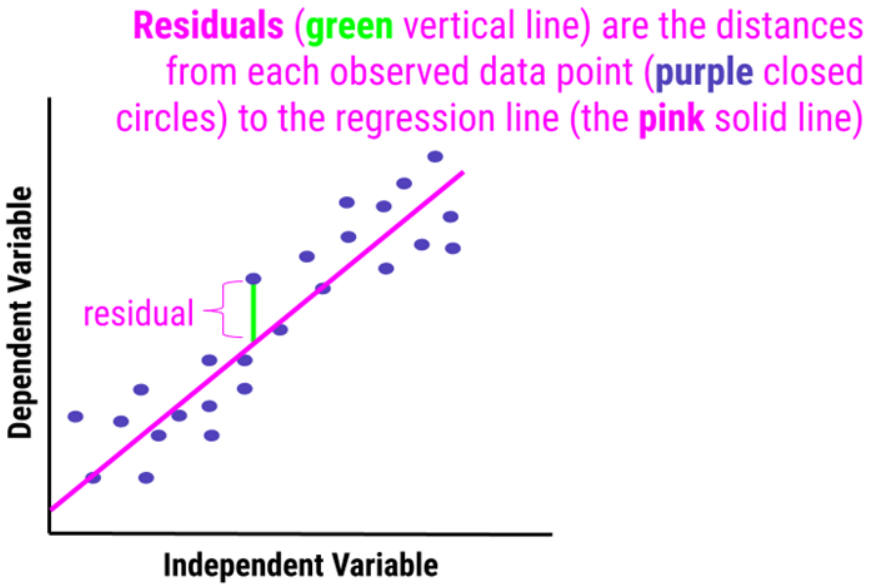
## linearity

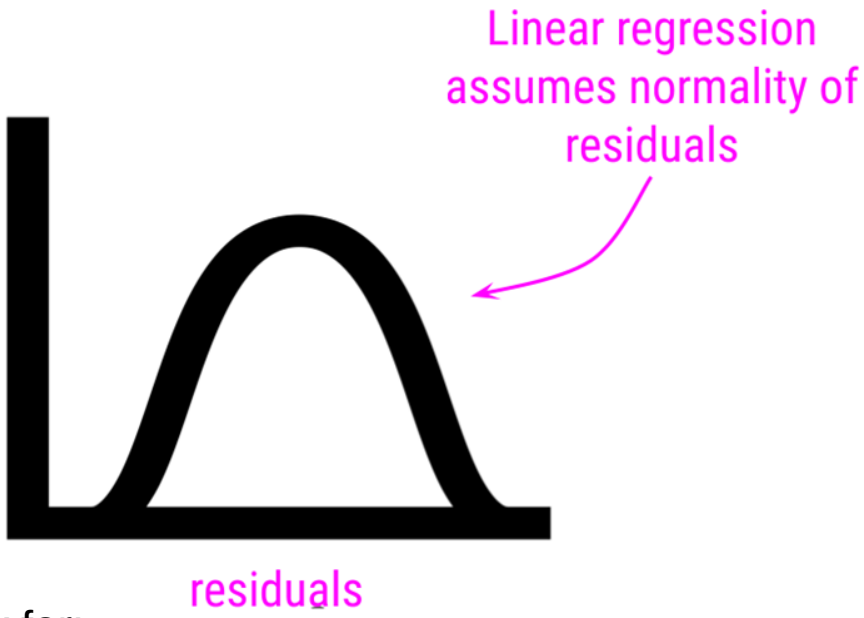



## homoscedasticity



# Normality of residuals




It is your job, when running linear regression to check for:

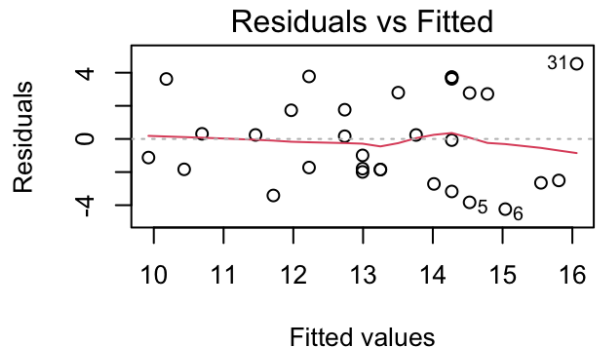
**- Non-linearity**

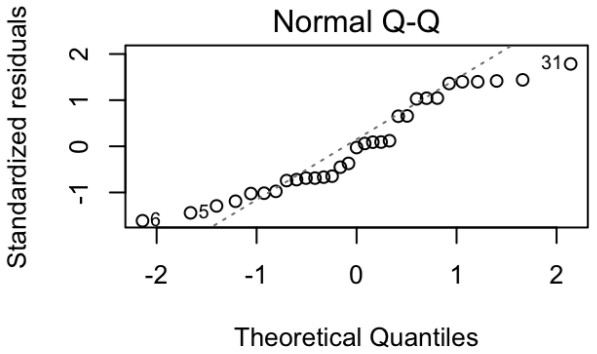
**- Heteroscedasticity**

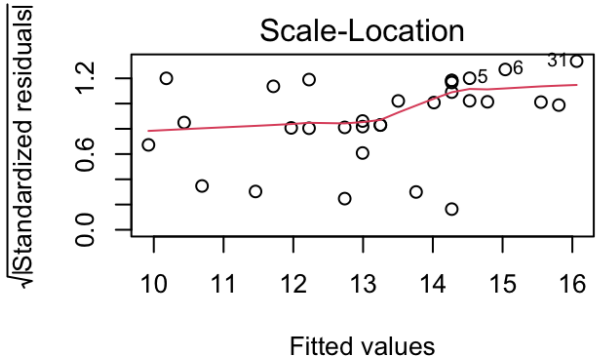
**- Outlier values**

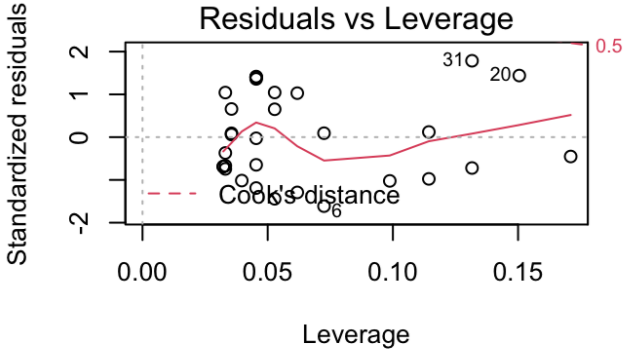
**- Normality of residuals**

# Testing the assumptions of a linear regression



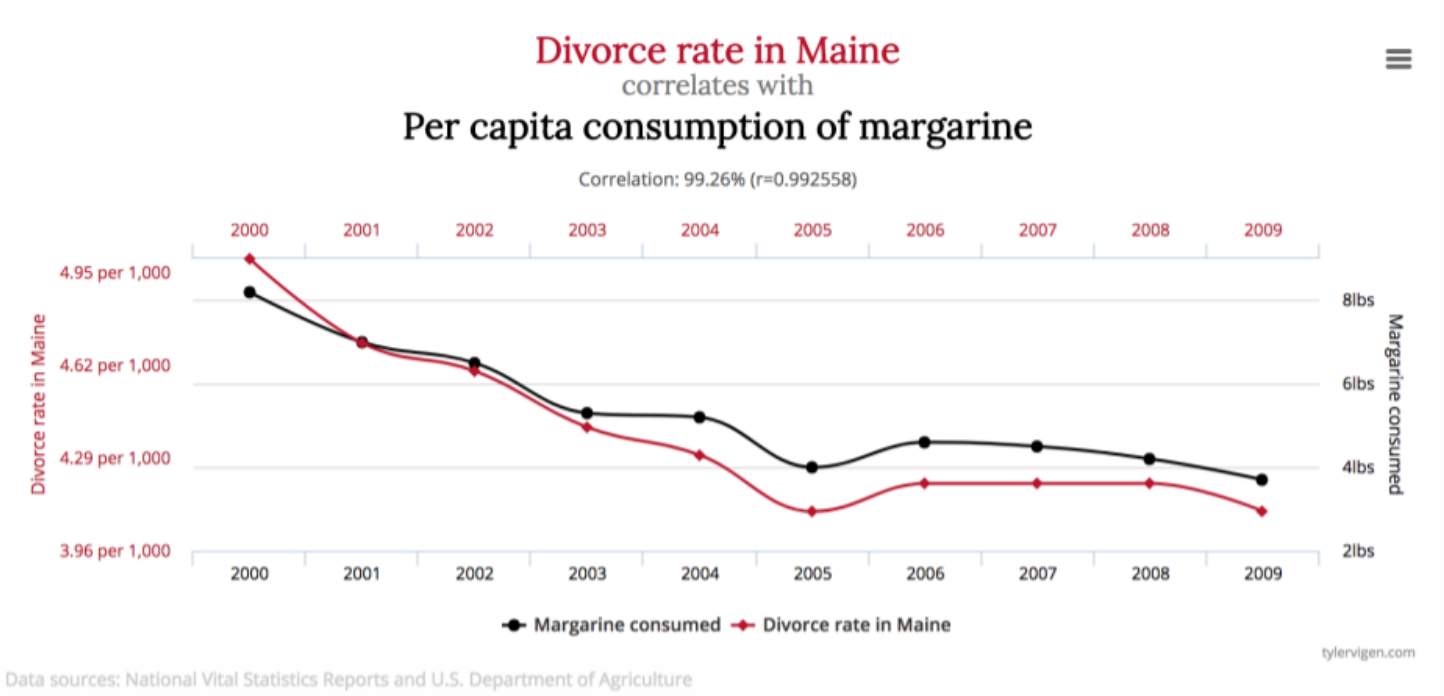

**1) Residuals vs Fitted**
Checks linear relationship assumption of linear regression.

**2) Normal Q-Q**
Checks if the residuals of the model are normally distributed.

**3) Scale-Location**
Checks the homoscedasticity of the model.

**4) Residuals vs Leverage**
Helps to identify outliers (extreme values) that may disproportionately affect the model's results.

# Correlation Is Not Causation!



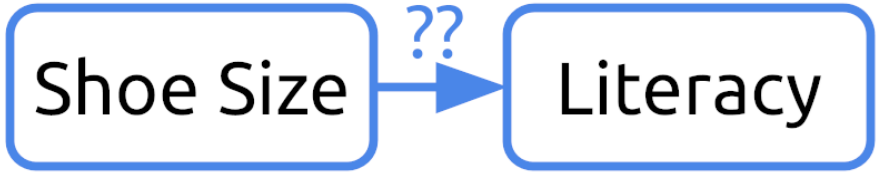

Just because you see two things with the same trend does not mean that one caused the other. These are simply **spurious correlations** – things that trend together by chance.

Always keep this in mind when you're doing inferential analysis, and be sure that you never draw causal claims when all you have are associations. Be careful with the language you choose and do not overstate your findings.

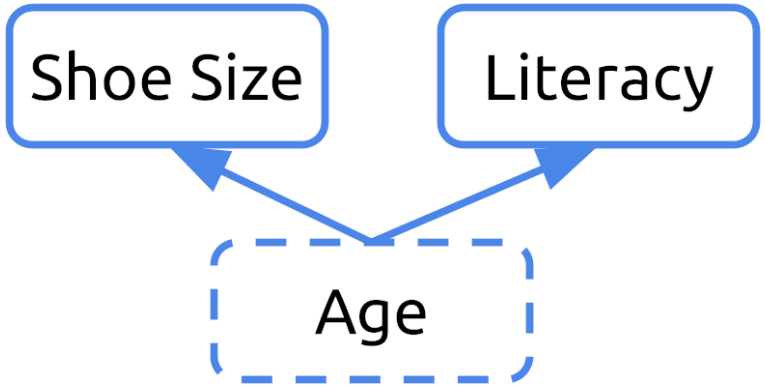
In other cases, you can have two variables that are associated because they are both related to a third hidden variable, which is causing the correlation of the first two ones. In this case, we can have a **confounding variable**.

# Confounding variables and multiple linear regression

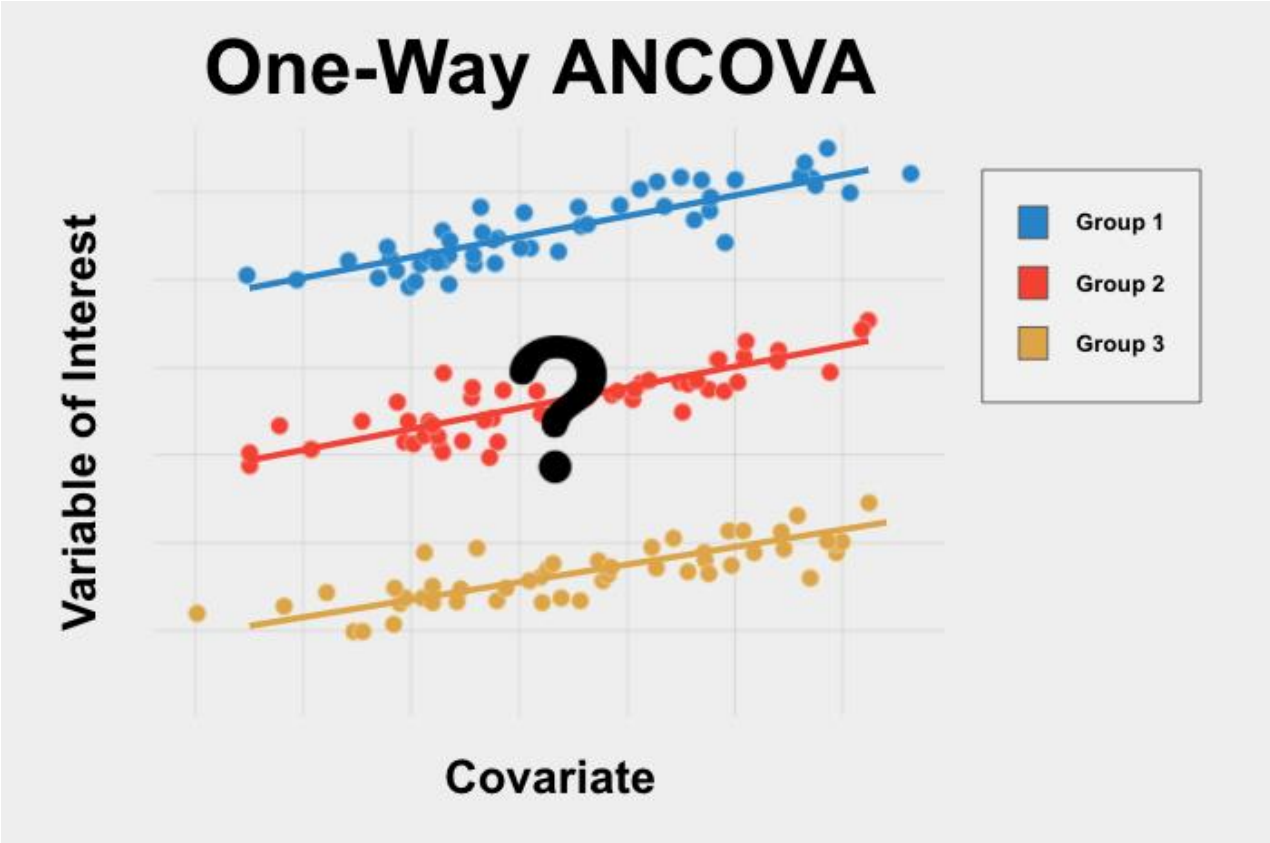

Can we infer literacy rates from shoe size?




Age affects their shoe size and their literacy rates.
In this example, age is a **confounding variable**.



# 2.3. Analysis of Covariance




ANCOVA is a hybrid between ANOVA (categorical factors) and regression (numerical factors). ANCOVA is commonly used to remove the effects of a confounding variable. A generalization of ANCOVA (linear models) can be used in a more general way to test the effects of many factors (of different types) in the final values of a dependent variable.

# 3.1 Linear models in R

ANOVA, ANCOVA and correlation analyses are all cases of a more general framework for analysing relationships between variables: the **linear model:**

```
model <- lm(formula = response ~ factors, data = dat)
```

Using different types of formulas, we can model the effects of any factor on the numerical response variable. We can combine categorical factors and numerical factors. We can also model the interactions between categorical factors (e.g.: different levels of a second factor may have different effects depending on the values of the first factor).

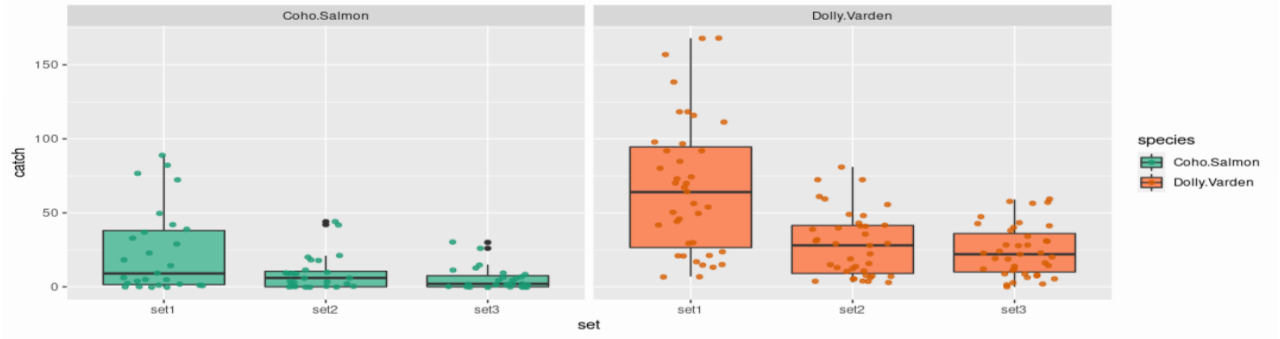
Some examples are:

```
height ~ sex + age + geographical_origin

lobster_density ~ temperature + pH + protected_status + locality

lice_mortality ~ treatment * year
```

# Example of an ANOVA using a linear model




```
anova_test <- lm(catch ~ 0 + set + species, data=dat_green_tiny)
```

This is an ANOVA with 2 categorical factors
(aka 2-way ANOVA)

**summary(anova_test)**

```
Residuals:
    Min      1Q  Median      3Q     Max
-53.481 -18.632  -0.494  10.471 107.519

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
factor(set)set1            33.3429     4.0285   8.277 2.01e-14 ***
factor(set)set2             4.5701     4.0285   1.134    0.258
factor(set)set3             0.4944     4.0285   0.123    0.902
factor(species)Dolly.Varden 27.1377     3.8865   6.983 4.50e-11 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.89 on 194 degrees of freedom
Multiple R-squared:  0.6333,     Adjusted R-squared:  0.6257
F-statistic: 83.76 on 4 and 194 DF,  p-value: < 2.2e-16
```

**anova(anova_test)**

```
Analysis of Variance Table

Response: catch
                 Df Sum Sq Mean Sq F value    Pr(>F)
factor(set)       3 206984   68995  95.431 < 2.2e-16 ***
factor(species)   1  35249   35249  48.756 4.498e-11 ***
Residuals       194 140258     723
```

# The same example including the interaction between the factors

```
anova_test <- lm(catch ~ 0 + set * species, data=dat_green_tiny)
```

Use multiplication sign instead of addition sign when you want to include the interaction between categorical factors in the model

```
summary(anova_test)
Residuals:
    Min      1Q  Median      3Q     Max
-60.128 -16.455  -3.556  11.299 100.872

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
factor(set)set1                                23.741      5.083   4.671 5.63e-06 ***
factor(set)set2                                 9.111      5.083   1.793  0.07463 .
factor(set)set3                                 5.556      5.083   1.093  0.27577
factor(species)Dolly.Varden                    43.387      6.612   6.562 4.82e-10 ***
factor(set)set2:factor(species)Dolly.Varden   -23.934      9.351  -2.560  0.01125 *
factor(set)set3:factor(species)Dolly.Varden   -24.815      9.351  -2.654  0.00863 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.41 on 192 degrees of freedom
Multiple R-squared:  0.6498,     Adjusted R-squared:  0.6389
F-statistic: 59.39 on 6 and 192 DF,  p-value: < 2.2e-16
```

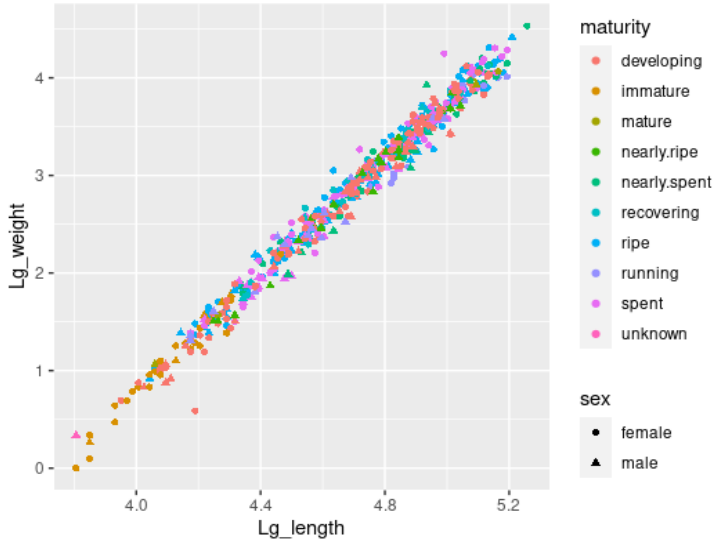
Interaction is significant!
That means that differences for set levels should be studied for each species separately

```
anova(anova_test)
Analysis of Variance Table

Response: catch
                             Df Sum Sq Mean Sq F value    Pr(>F)
factor(set)                   2 206984   68995  98.908 < 2.2e-16 ***
factor(species)               1  35249   35249  50.532 2.245e-11 ***
factor(set):factor(species)   2   6325    3163   4.534   0.01191 *
Residuals                   192 133933     698
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Example of a correlation using a linear model

```
model_ruff <- lm(Lg_weight ~ Lg_length, data=data_ruffe)
```




```
summary(model_ruff)

Call:
lm(formula = Lg_weight ~ Lg_length, data = data_ruffe)

Residuals:
     Min       1Q   Median       3Q      Max
-0.76687 -0.06228  0.00595  0.07397  0.54665

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.92577    0.07309  -149.5   <2e-16 ***
Lg_length     2.93113    0.01554   188.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1175 on 605 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9833, Adjusted R-squared:  0.9833
F-statistic: 3.558e+04 on 1 and 605 DF,  p-value: < 2.2e-16
```

This p-value means that the intercept is != 0

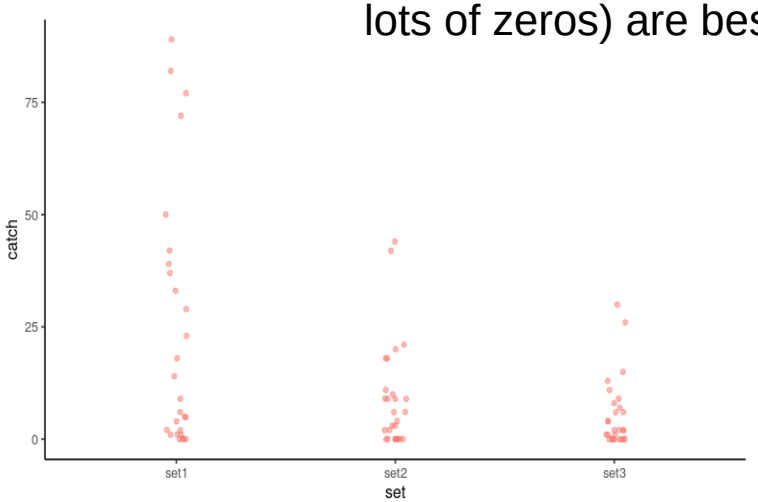
This p-value means that the slope is != 0

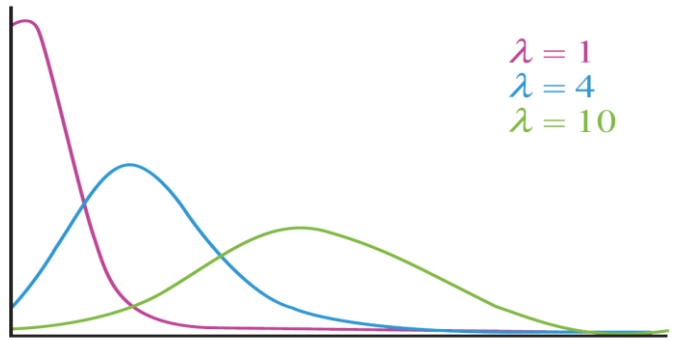
This p-value means that the correlation is signific

# 3.2 Generalized linear models

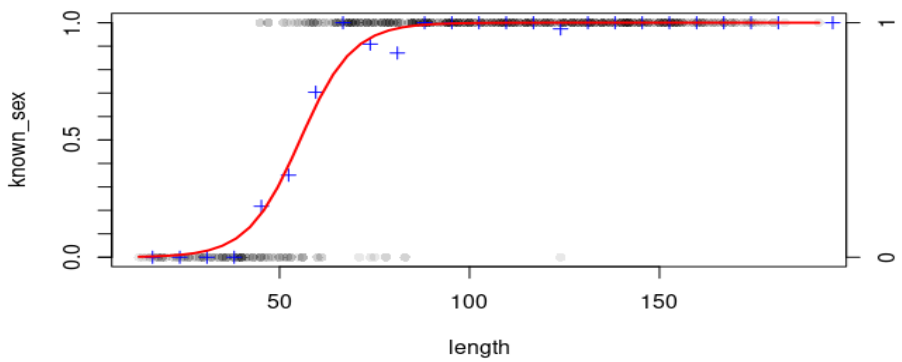
Linear models assume that the residuals will follow a normal distribution. However, this is not usually the case when working with some types of data:

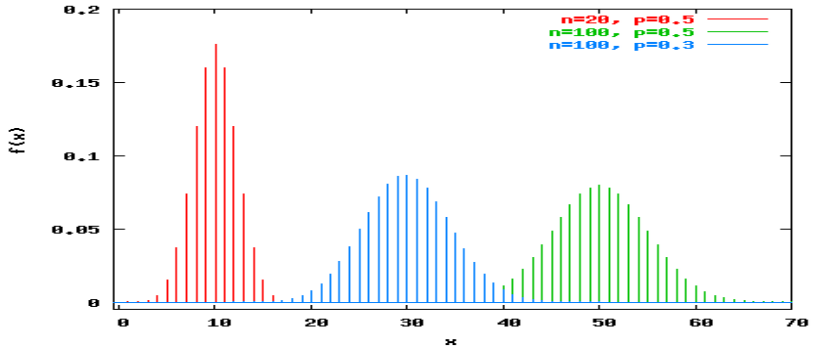
**Count data** (positive integer occurrences, commonly with lots of zeros) are best fitted using a **Poisson distribution**




**Binary data** (Boolean False/True variables) are best fitted using a **Binomial distribution**



# Generalized linear models

These special types of distributions can be modelled in R using **generalized linear models:**

```
model <- glm(formula = response ~ groups, data = dat,  family = poisson)
```

Function glm() will not show p-values! So, deciding whether a generalized linear model is correct is sometimes difficult. It is easier to compare between different models.

We can always compare the proposed model with the *null* model (no differences among groups).

```
model <- glm(formula = response ~ 1, data = dat, family = poisson)
```

Using summary() of a glm object we will get the AIC values (Akaike information criterion). The model with the least AIC value will fit the data best.

# Hands-on Session 4

## Basic hypothesis testing and linear models