Introduction to Statistics
MedILS School in Bioinformatics

Nikolina Pleić

MSc Mathematics
University of Split, School of Medicine
Department of Biology and Human Genetics

Monday 21st August, 2023

Outline

Variables & data management

Types of variables:

- Numeric
  - Continuos (measures, e.g. height)
  - Discrete (e.g. couting occurences of an event)
- Categorical
  - Dichotomous (sex or Yes/No answers)
  - Nominal (do not have a natural order or ranking e.g. genotype)
  - Ordinal (ranked: school grades, Likert scales)

How to enter data

The incorrect way:

	A	B	C	D	E	F	G	H	I	J	K	L
1	**August 2023.**											
2						**Data on cholesterol levels**						
3						Diabetes study						
4												
5		**ID**	1000	1001	1002	1003	1005	1008	1011	1015	1016	1022
6		**age**	46	29	58	67	64	34	30	37	45	55
7		**gender**	female	female	female	male	male	male	male	male	male	female
8		**cholesterol**	203	165	228	78	249	248	195	227	177	263
9		**location**	Buckingha	Buckingha	Buckingha	Buckingha	Buckingha	Buckingha	Buckingha	Buckingha	Buckingha	Buckingha
10		**height**	62	64	61	67	68	71	69	59	69	63
11		**weight**	121	218	256	119	183	190	191	170	166	202
12												

Key takeaway: R (or any other programme) does not care about color-coding

How to enter data

The correct way:

	A	B	C	D	E	F	G	H	I
1	**ID**	**age**	**gender**	**cholesterol**	**location**	**height**	**weight**		
2	1000	46	female	203	Buckingham	62	121		
3	1001	29	female	165	Buckingham	64	218		
4	1002	58	female	228	Buckingham	61	256		
5	1003	67	male	78	Buckingham	67	119		
6	1005	64	male	249	Buckingham	68	183		
7	1008	34	male	248	Buckingham	71	190		
8	1011	30	male	195	Buckingham	69	191		
9	1015	37	male	227	Buckingham	59	170		
10	1016	45	male	177	Buckingham	69	166		
11	1022	55	female	263	Buckingham	63	202		
12	1024	60	female	242	Louisa	65	156		
13	1029	38	female	215	Louisa	58	195		
14	1030	27	female	238	Louisa	60	170		
15	1031	40	female	183	Louisa	59	165		
16	1035	36	male	191	Louisa	69	183		
17	1036	33	female	213	Louisa	65	157		

Columns represent variables and each row represents one observation (or participant)

Descriptive statistics

Visualizing the data

For discrete numerical and categorical variables, we use the frequency distributions in a form of a:

- table (frequency table)
- graph (bar chart)

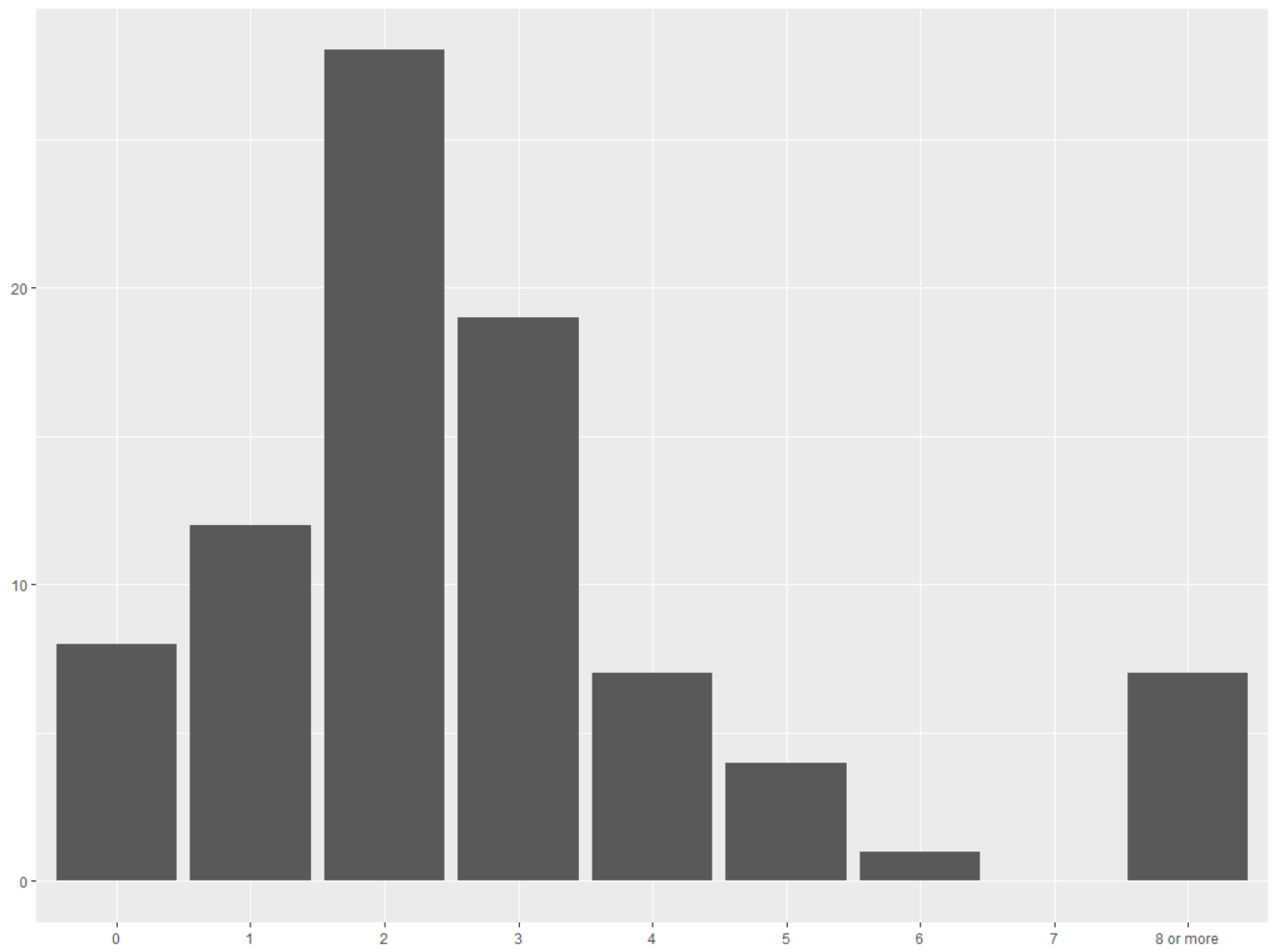
Descriptive statistics

This frequency table represents the frequency distribution of a variable X, which represents the number of children in a family, based on a sample of 80 families:

No. of children	frequency	relative frequency
0	8	0.1
1	12	0.15
2	28	0.35
3	19	0.2375
4	7	0.0875
5	4	0.05
6	1	0.0125
7	0	0
8 or more	7	0.0875
Total	80	1

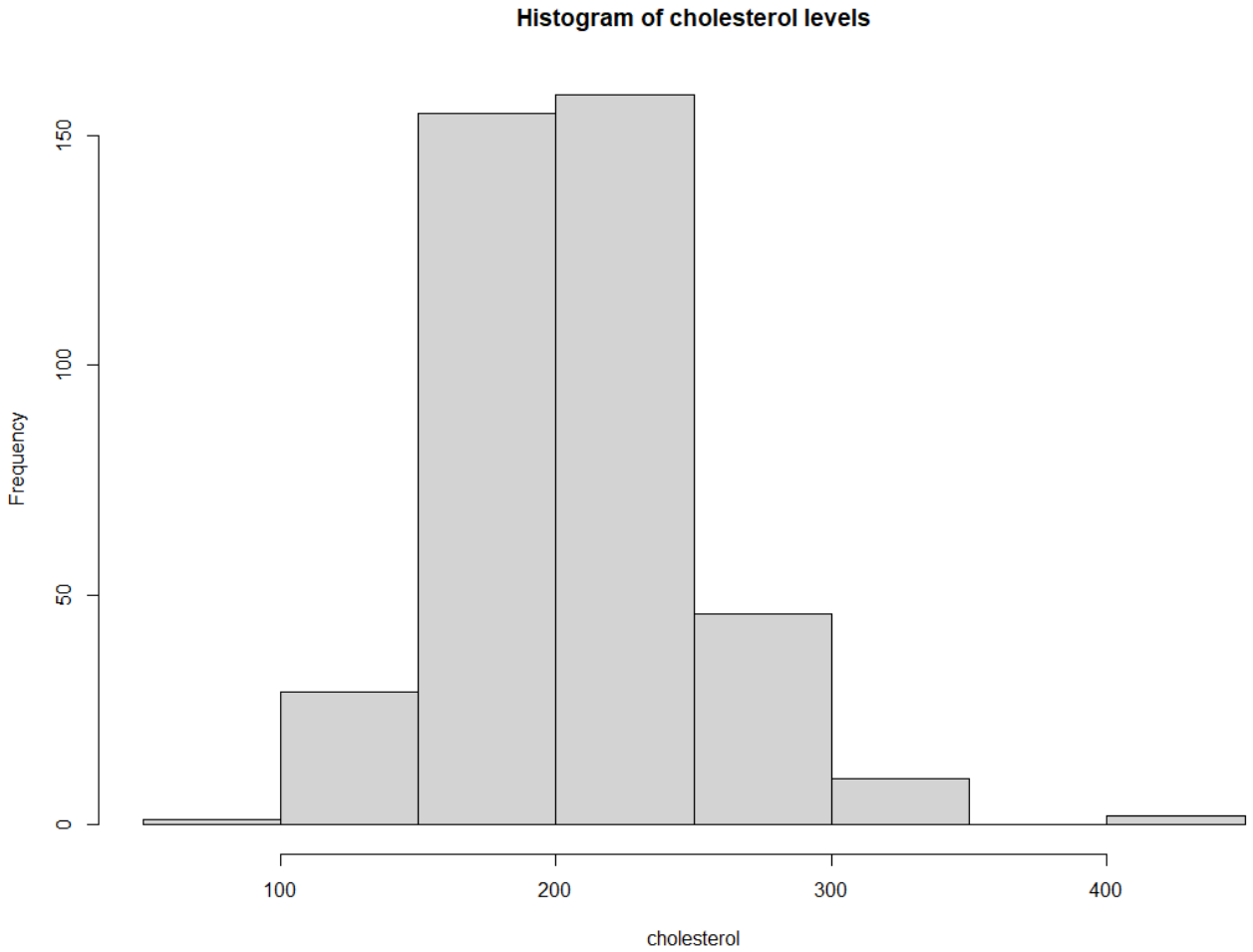
Descriptive statistics

Example of a bar chart corresponding to the frequency table:



Descriptive statistics - continuous variables

For continuous numerical variables, we use histograms. An example of a histogram:



What is the difference between bar charts and histograms?

Descriptive statistics - continuous variables

For a table repesentation of continuous numerical variables, we use *Measures of Central Tendency*.

Measures of Central Tendency provide a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution. There are three main measures of central tendency: the mean, the median and the mode.

Descriptive statistics - continuous variables

- **Mean**: the arithmetic average of the values
- **Median**: the value in the middle of a data set, meaning that 50% of data points have a value smaller or equal to the median and 50% of data points have a value higher or equal to the median
- **Mode**: the value that appears most often in a set of data values

Descriptive statistics - continuous variables

Mean:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + ... + x_n}{n}$$

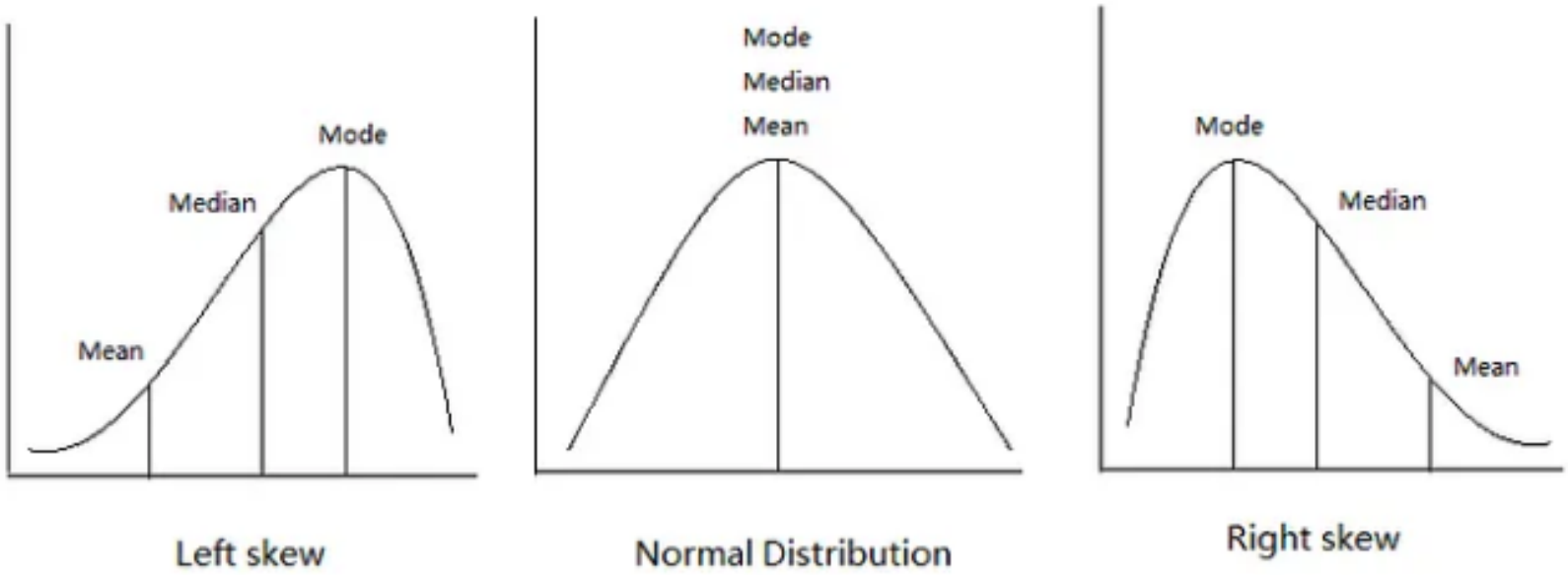
Median: If n is even then:

$$\tilde{x} = \frac{(\frac{n}{2})^{\text{th}} \text{ obs.} + (\frac{n+1}{2})^{\text{th}} \text{ obs.}}{2}$$

If n is odd then

$$\tilde{x} = \frac{n+1}{2}^{\text{th}} \text{ obs.}$$

Descriptive statistics - continuous variables



Descriptive statistics - continuous variables

Measures of central tendency are used in pair with the *measures of variability (or spread)*.

These include:

- variance - the average squared deviation from the mean
- standard deviation - the square root of the average squared deviation from the mean (or the squared root of variance)
- range
- interquartile range (IQR)

Descriptive statistics - continuous variables

Variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

Standard deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

Range:

$$R = x_{\text{Max}} - x_{\text{Min}}$$

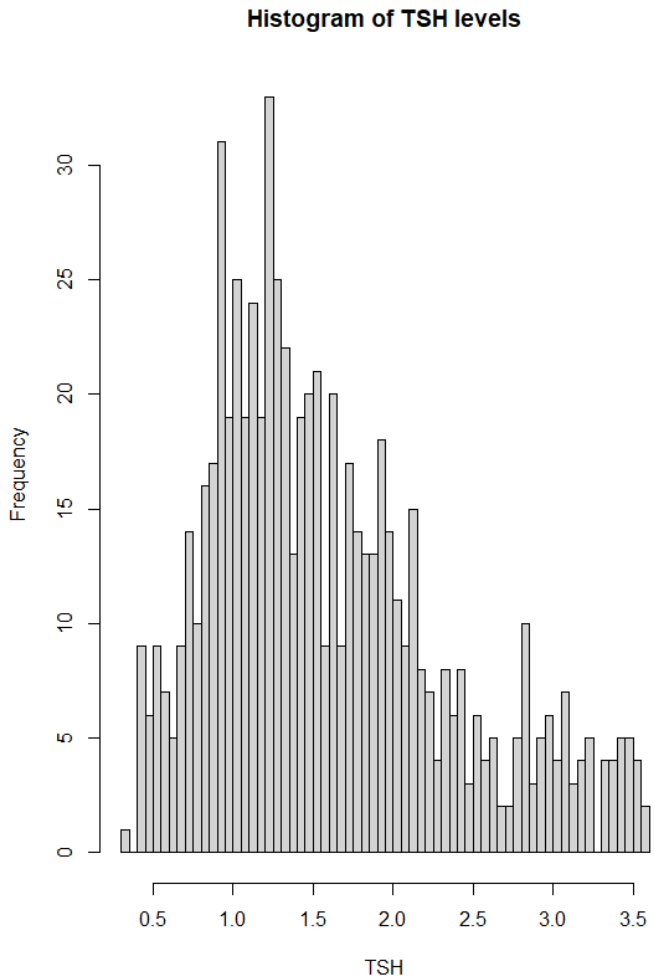
IQR:

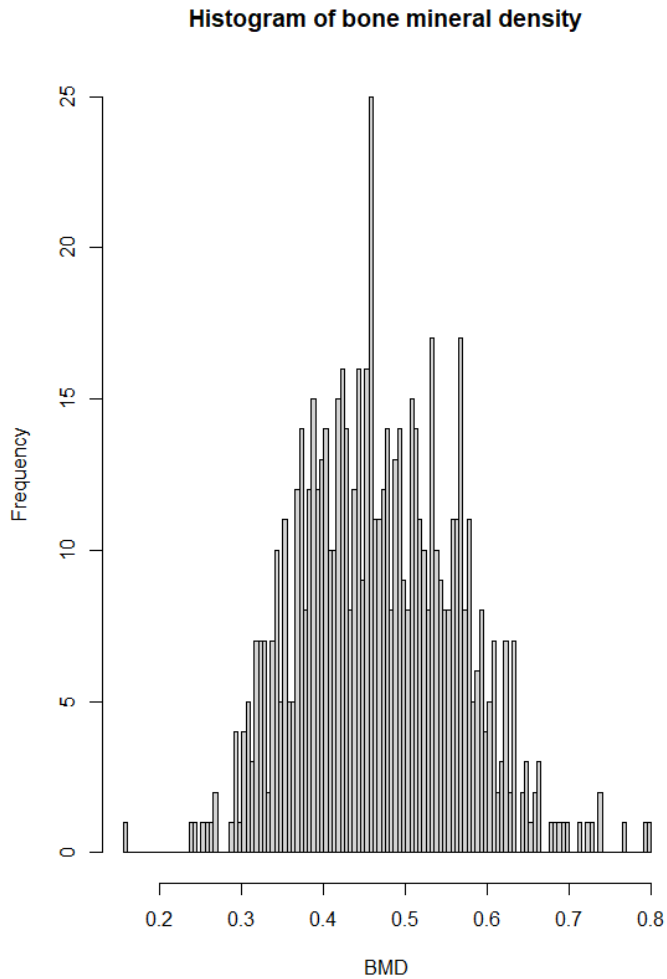
$$\text{IQR} = \text{Q3} - \text{Q1}$$

Note: each of these values is a single number.

Descriptive statistics - continuous variables

Histogram on the left: log-normal distribution (right-skewed)
Histogram on the right: approximately normal distribution





Descriptive statistics - continuous variables

Table 1. Clinical characteristics of study participants.

Variable	Total (*N* = 694)	Reference Interval
Women	396 (57.1%)	-
Age	51.6 (14.8)	-
OC	15.6 (13–18.5)	5–25 ng/mL
CT	5.2 (2.68–8.2)	0–20 ng/mL
TSH	1.43 (1.05–2.01)	0.3–3.6 mIU/L
fT3	4.57 (0.49)	3.39–6.47 pmol/L
fT4	12.7 (11.9–13.9)	10.29–21.88 pmol/L
fT3/fT4	0.36 (0.05)	-
TgAb	7.6 (5–11.2)	5–100 IU/mL
TPOAb	2.7 (1.3–6.3)	1–16 IU/mL
PTH	21.5 (5.7)	12.26–35.5 pg/ml
Total serum Calcium	2.36 (0.1)	2.14–2.53 mmol/L
BMI	27.31 (4.34)	18.5–24.9
Absolute BMD	0.47 (0.1) (g/cm^2)	-

Continuous variables are expressed as means with standard deviations or as medians with lower and upper quartiles, and categorical variables as frequencies (relative frequencies) (Table 1).

Descriptive statistics - continuous variables

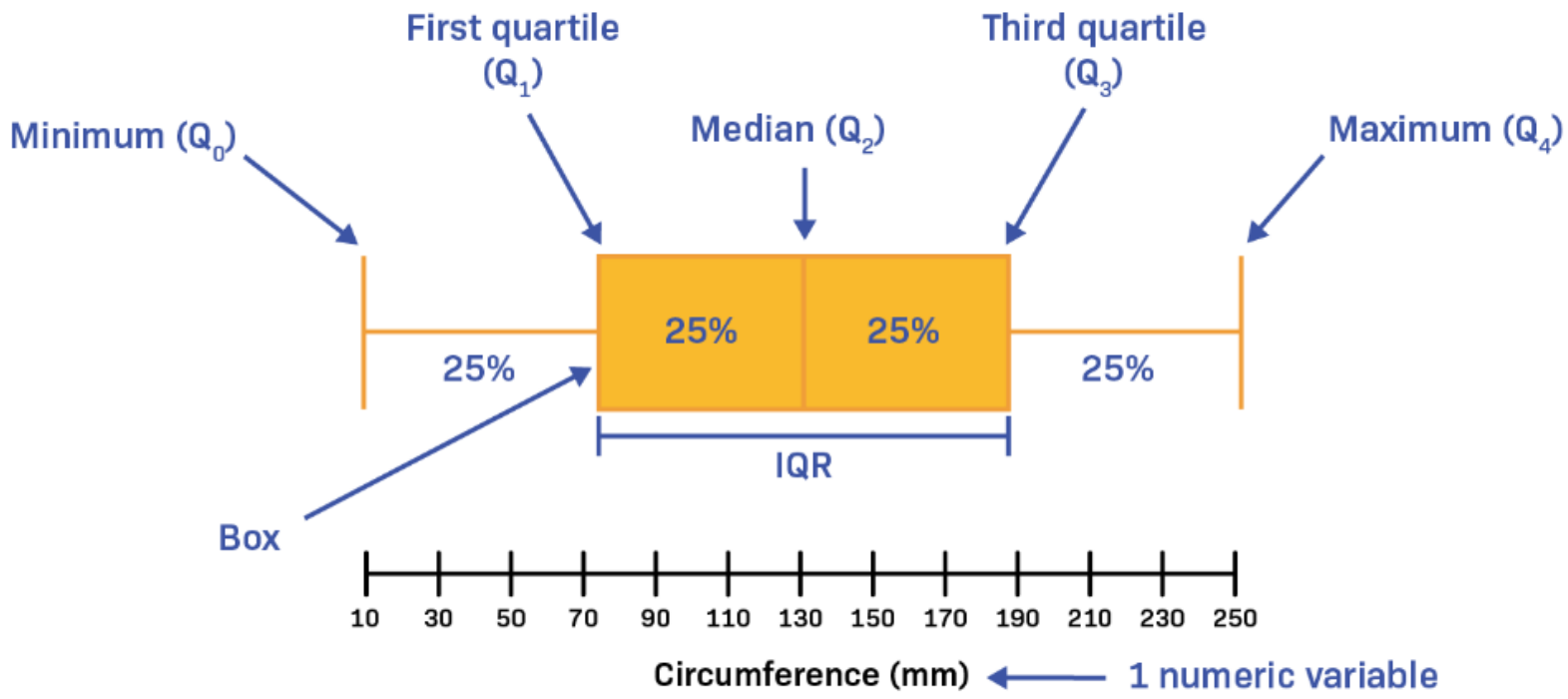
How to deal with reviewers?

3. why for some parameters authors used mean and for other median? Authors were describing mainly biochemical parameters mesured in the blood so the same statistic should be used: decide which one and unify; aspecially as reader cannot clearly see if they are in range of norm (there are no ranges of norms for biochemical parameters – please add, as most of them are diagnostic parameters were ranges are established);

Response: We appreciate the opinion of the reviewer, however, we cannot fully agree. As stated in Lines 104-106, ' The distribution of TSH, fT4 and OC levels was right-skewed, while levels of fT3, fT3/fT4, age, BMI and BMD followed an approximately normal distribution.'. By definition, mean (with SD) is a valid measure of central tendency only in cases when the parameter's distribution is normal or approximately normal. If a parameter's distribution is skewed, the mean is no longer a representative nor a valid value of central tendency because it is over-sensitive to deviations from the normal distribution. In this case, it is necessary to use median along with interquartile range, as this measure gives complete information to the reader. Just like the mean with SD gives us quick numeric information on the percentage of values that lie within an interval estimate in a normal distribution: (68%, 95%, and 99.7% of the values lie within one, two, and three standard deviations of the mean, respectively), the median with the IQR gives us the information on the values that lie in the middle 50% spread of the dana, regardless of distribution.

Descriptive statistics - continuous variables

We can additionally use the box plot:

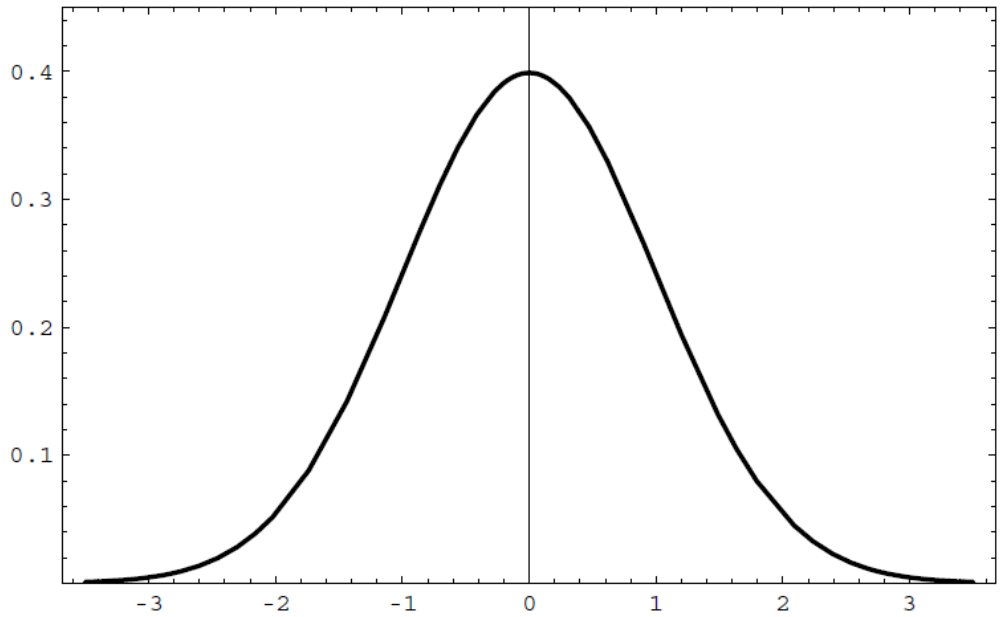


Probability distibutions

Normal Distribution: $X \sim N(\mu, \sigma^2)$
Standardized Normal: $Z \sim N(0, 1)$ where $Z = \frac{X-\mu}{\sigma}$

68-95-99.7 or the 3σ Rule:

- approximately 68% of observations fall within σ of μ
- approximately 95% of observations fall within 2σ of μ
- approximately 99.7% of observations fall within 3σ of μ



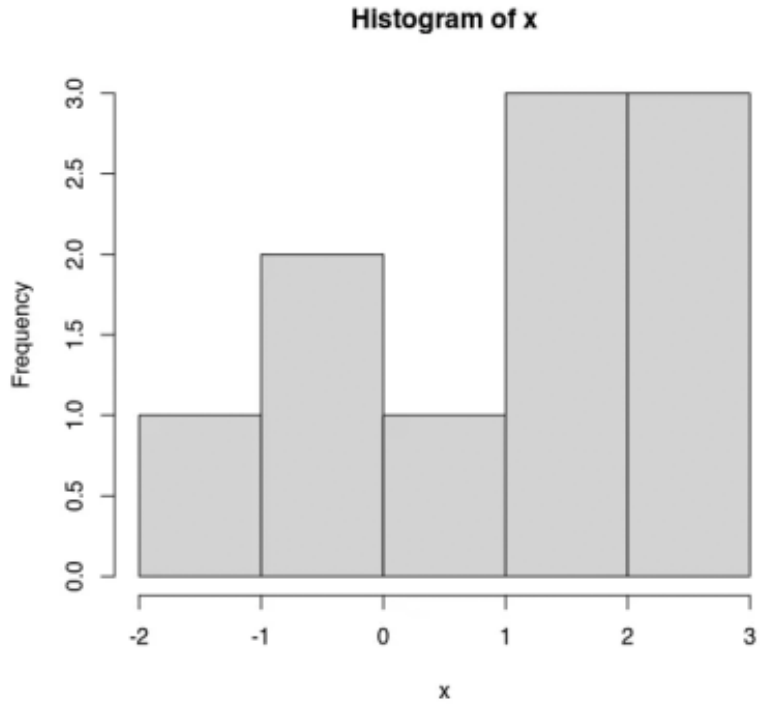
Probability distibutions

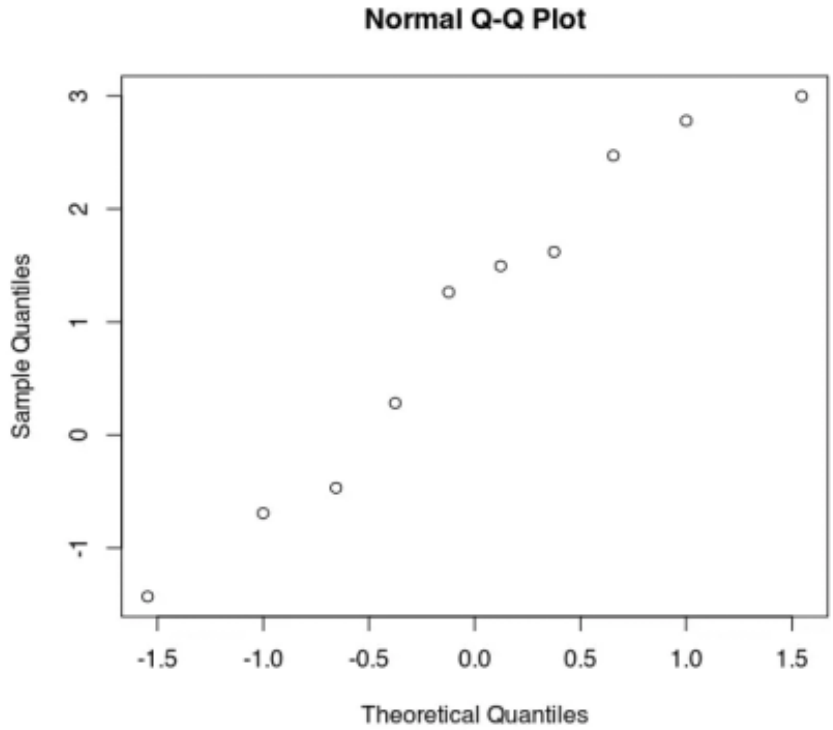
Note: normal distribution is a purely theoretical distribution, however the nice properties can be applied to its approximations.

Important: The assumption of a parametric test is NOT that the examined variable is normally distributed (rather that the fitted residuals are normally distributed, we'll get to that later on).

What's wrong with testing for normality?

Let's consider a small sample (n=10).





From the histogram, we can conclude that it is not normally distributed. In the quantile-quantile (Q-Q) plot, data shows some deviation from normality.

What's wrong with testing for normality?

However, if we perform a Shapiro-Wilk test of normality, we get a p-value of 0.53.

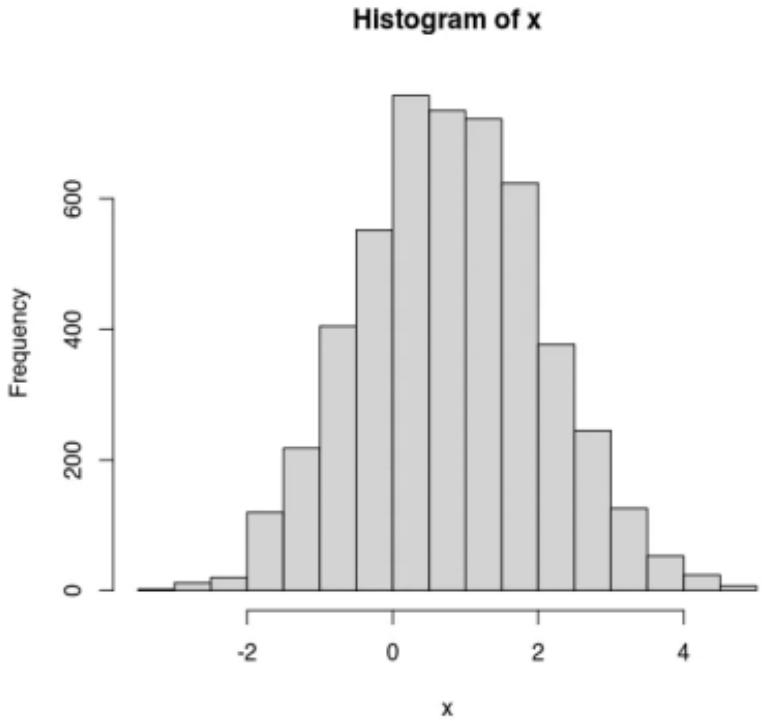
Therefore, we have no evidence to reject the null hypothesis and suggest that x is not normally distributed.

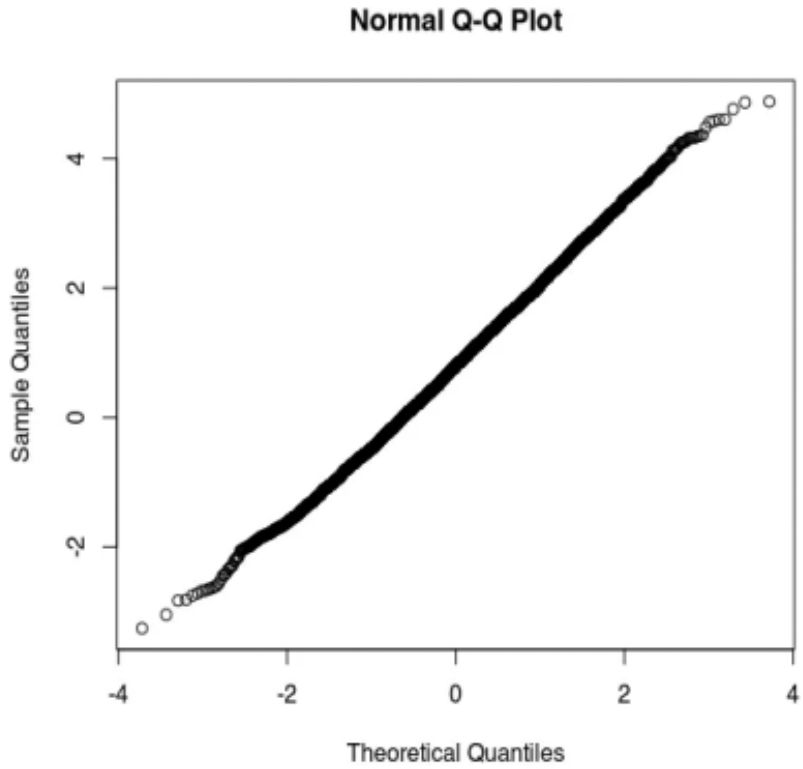
Furthermore, not being able to conclude that x is not normally distributed does not mean that x is normally distributed.

The problem here is that in small samples the 'normality tests' (e.g. Kolmogorov-Smirnov or Shapiro-Wilk) are underpowered to detect deviations from normality.

What's wrong with testing for normality?

Now let's consider a large sample (n=5000).





From the histogram and the Q-Q plot, we can conclude that it is normally distributed.

What's wrong with testing for normality?

However, if we perform a Shapiro-Wilk test of normality, we get a p-value of 0.001. There's very strong evidence that we can reject the null hypothesis and that x is not normally distributed.

The Shapiro-Wilk test (and other normality tests) are designed to test for theoretical normality (i.e. the perfect Gaussian curve).

In small samples these tests are underpowered to detect quite major deviations from normality which can be easily detected through graphical methods. In larger samples these tests will detect even extremely minor deviations from theoretical normality and always reject the null hypothesis.

How to inspect normality?

If you are unsure about your variable's distribution, inspect it visually using **histograms** and **Q-Q plots**, this will give you a much clearer picture about the normality of your data.

Remember that statistical analysis is a research within a research (not a cookbook recipe) and that a lot of it depends on your decisions.

Testing hypotheses

Hypothesis $\neq$ Statistical hypothesis

H_o: Null hypothesis is a tentative assumption about a population parameter.
H_a: Alternative hypothesis is what the test is attempting to establish.

- $H_o : \mu \geq \mu_o$ vs $H_a : \mu < \mu_o$ (one-tail test, lower-tail)
- $H_o : \mu \leq \mu_o$ vs $H_a : \mu > \mu_o$ (one-tail test, upper-tail)
- $H_o : \mu = \mu_o$ vs $H_a : \mu \neq \mu_o$ (two-tail test)

Testing hypotheses

Type I and Type II errors:

Type I error: rejecting H_o when H_o is true
Type II error: not rejecting H_o with H_o is false

$P(\text{Type I error}) = \alpha$
$P(\text{Type II error}) = \beta$

Power is the probability of rejecting H_o, when H_o is false.

$$\text{Power} = 1 - \beta$$

Note: we say nothing about 'accepting' the alternative hypothesis H_a. This is because the p-value tells us nothing about the H_a.

Statistical tests: $\chi^2 - test$

When an analyst attempts to fit a statistical model to observed data, he or she may wonder how well the model actually reflects the data. How "close" are the observed values to those which would be expected under the fitted model? One statistical test that addresses this issue is the **chi-square goodness of fit test**. The test statistic:

$$H = \sum \frac{(observed - expected)^2}{expected}$$

If the computed test statistic is large, then the observed and expected values are not close and the model is a poor fit to the data (we reject the null hypothesis H_o).

Statistical tests: $\chi^2 - test$

- Used for categorical or discrete numerical variables
- We are comparing the observed outcome frequencies to the expected frequencies
- Used for a single variable - goodness of fit

Example: If we want to test if a dice is fair, the appropriate model would be the uniform distribution: $P(X = i) = \frac{1}{6}, for i = 1, 2, 3, 4, 5, 6.$

Statistical tests: ANOVA

Analysis of variance (ANOVA) is a test that is appropriate to compare means of a continuous variable in two or more independent comparison groups (or treatments).
The null hypothesis in ANOVA is always that there is no difference in means.
Additionally, the analysis of variance for k treatments is equivalent to a regression model in which the outcome variable Y depends on k-1 independent binary variables (being either 0 or 1).

ANOVA

Assumptions for One-Way ANOVA Test
There are three primary assumptions in ANOVA:

1. The data are independent.
2. The model residuals are independent and normally distributed.
3. The model residuals are homoskedastic.

Violations to the last two that are not extreme can be considered not serious. A simple data transformation can usually fix both.
The sampling distribution of the test statistic is fairly robust, especially as sample size increases and more so if the sample sizes for all factor levels are equal. If you conduct an ANOVA test, you should always try to keep the same sample sizes for each factor level.

Statistical tests: t-test

F-test in the analysis of variance for the comparison of k = 2 treatments is equivalent to a t-test and the relationship between the test statistics is: $T^2 = F$.
The null hypothesis in t-test is always that there is no difference in means.

Note: Neither the t-test nor the ANOVA require the examined variable to be normally distributed.
Key takeaway: If you have more than 30 samples per group and your data are independent, the assumptions of parametric tests should be satisfied. If you have less than 30 samples per group, then you can perform both parametric tests and their non-parametric versions, just keep in mind that they don't exactly test the same thing (e.g. Mann-Whitney tests the stochastic dominance).

Correlation

Let's say we want to determine the association between our variables. We'll limit ourselves to:

- the bivariate case (two variables)
- linear association (the conditional expectation of the dependent variable is a linear function of the independent variable)

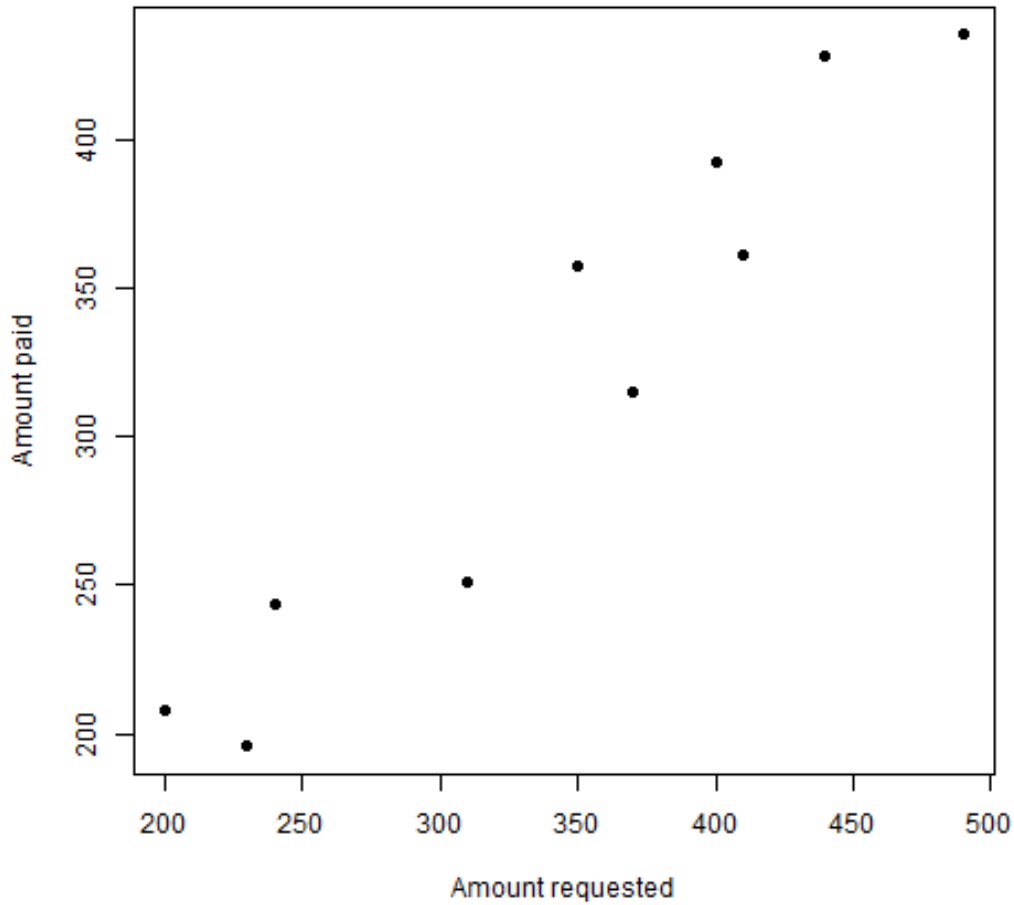
We do this using the correlation analysis.

The aim of correlation analysis is to determine the **magnitude** of the *linear correlation* between the two variables.

The aim of regression analysis is to determine the **nature** of the association between the dependent and the independent variable.

Correlation

Amount requested (x)	200	230	240	310	350	370	400	410	440	490
Amount paid (y)	208	196	244	251	357	315	392	361	428	435



Correlation

In order to analyse the linear association of two variables, we calculate the following statistics:

$$S_{XX} := \sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n \cdot \bar{X}^2$$

$$S_{XY} := \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - n\overline{XY}$$

$$S_{YY} := \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n \cdot \bar{Y}^2.$$

These produce the *Pearson correlation coefficient*: $r := \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$ which measures the magnitude of the linear association between the two variables. Pearson correlation coefficient for the example data is r=0.958 which indicates a strong, positive linear association.

Correlation

More about the Pearson correlation coefficient:

- Always falls between -1 and +1
- A positive r value indicates a positive association
- A negative r value indicates a negative association
- r value close to +1 or -1 indicates a strong linear association
- r value close to 0 indicates a weak association

Linear regression

The aim of regression analysis is to fit the appropriate model (in our case, the linear model) to the observed data in order to **predict** the values of the outcome variable Y based on the values of the input variable X.

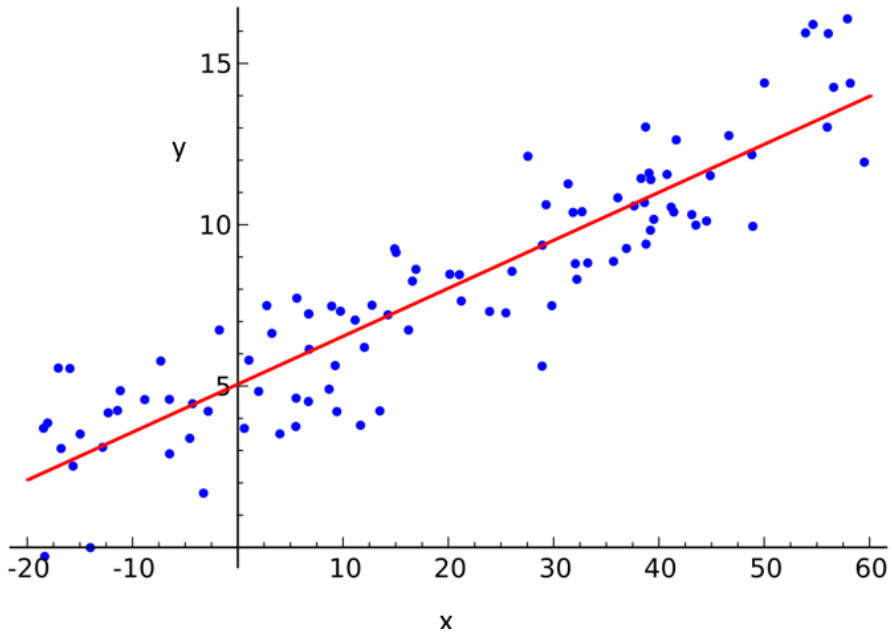
Simple linear regression model: $y = \beta_o + \beta_1 x + \varepsilon$

β_o and β_1 are model parameters (β_o is the intercept and β_1 is the slope coefficient), y and ε are random variables and ε is the error term or the noise.

Linear regression

Simple linear regression model: $y = \beta_o + \beta_1 x + \varepsilon$

Regression Line:



Coefficient of Determination: r^2

The proportion of observed variation in y that can be explained by the simple linear regression model.

Linear regression

Observed data:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

the model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

residuals: $\varepsilon_i = y_i - \hat{y}_i$

Gauss-Markov assumptions for the model **residuals**:
(A1) centered (mean zero): $\mathbb{E}[\varepsilon_i] = 0$ for all i;
(A2) Homoskedastic: $\mathrm{Var}[\varepsilon_i] = \sigma^2$ for all i;
(A3) uncorrelated: $\mathrm{cov}[\varepsilon_i, \varepsilon_j] = 0$ for all $i \neq j$.

In addition, the residuals ε_i should be:
(A4) independent and normally distributed and
(A5) there should be a linear relationship between the two variables X and Y.

Linear regression

If the assumptions are met $\rightarrow$ the model can be used with confidence.

If the assumptions are violated $\rightarrow$ the model should probably be discarded because you cannot confidently assume that the relationships seen in the model are mirrored in the population.

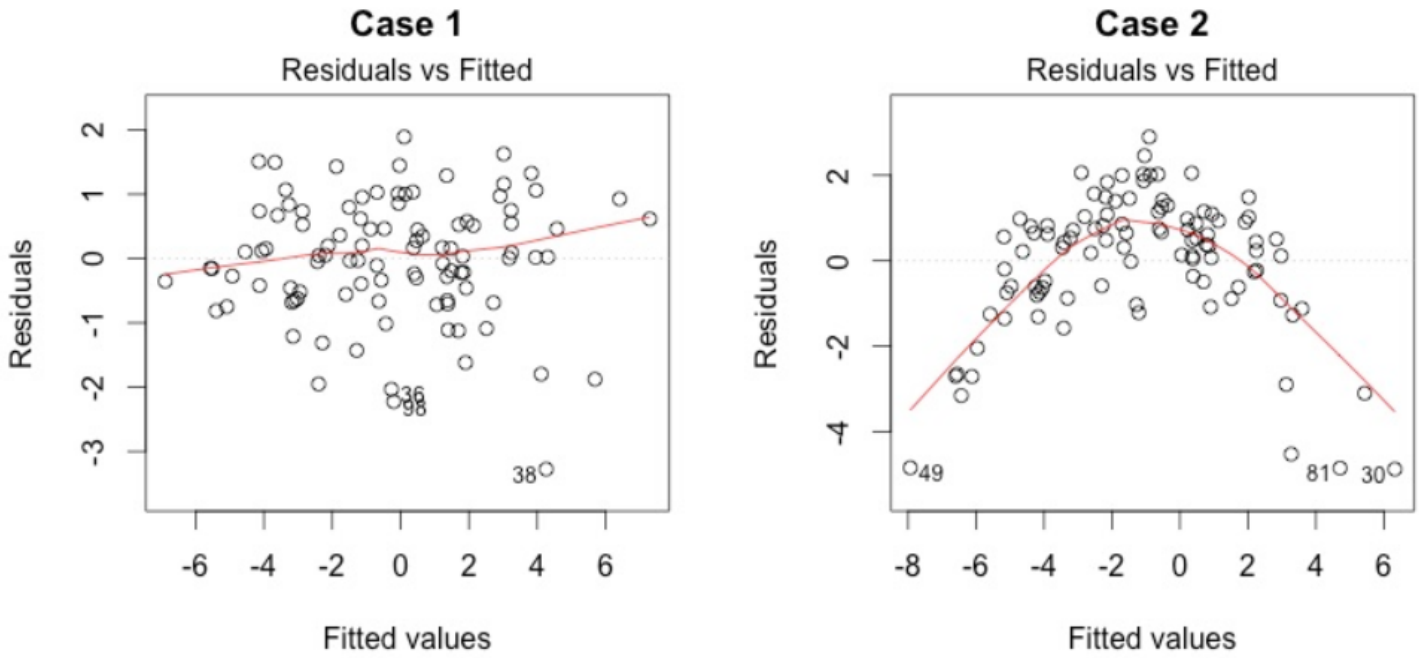
Diagnostic plots allow us to check the assumptions:

1. A scatter plot to inspect the nature of the relationship (A5)
2. Diagnostic plots to check for residuals assumptions (A1-A4) and (A5)

Linear regression

1. Residuals vs Fitted

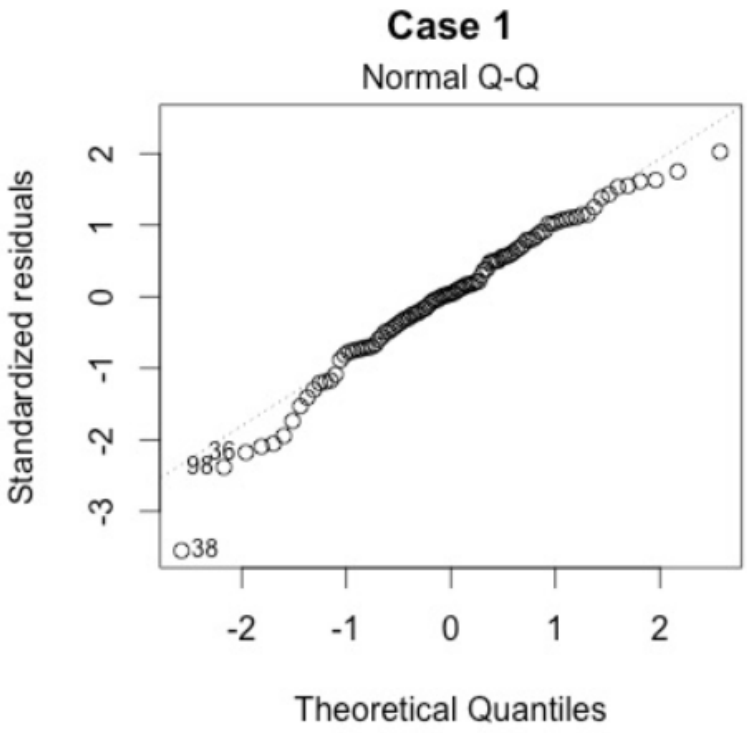
Used to check the linear relationship assumption (A5). There could be a non-linear relationship between predictor variables and the outcome variable, and the pattern could show up in this plot even if the model doesn't capture the non-linear relationship. If you find **equally spread residuals around a horizontal line without distinct patterns**, that is a good indication you don't have non-linear relationships.

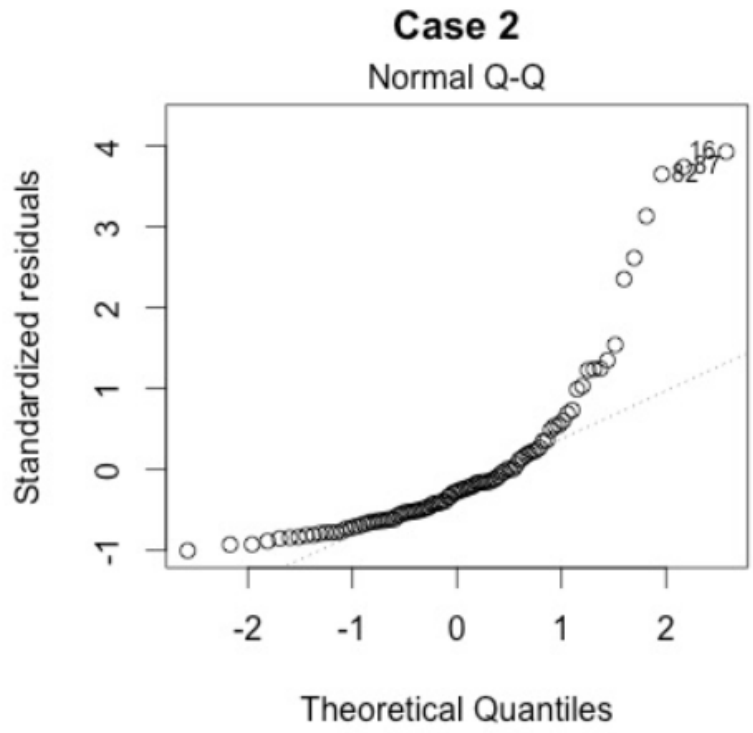


Linear regression

2. Normal Q-Q plot

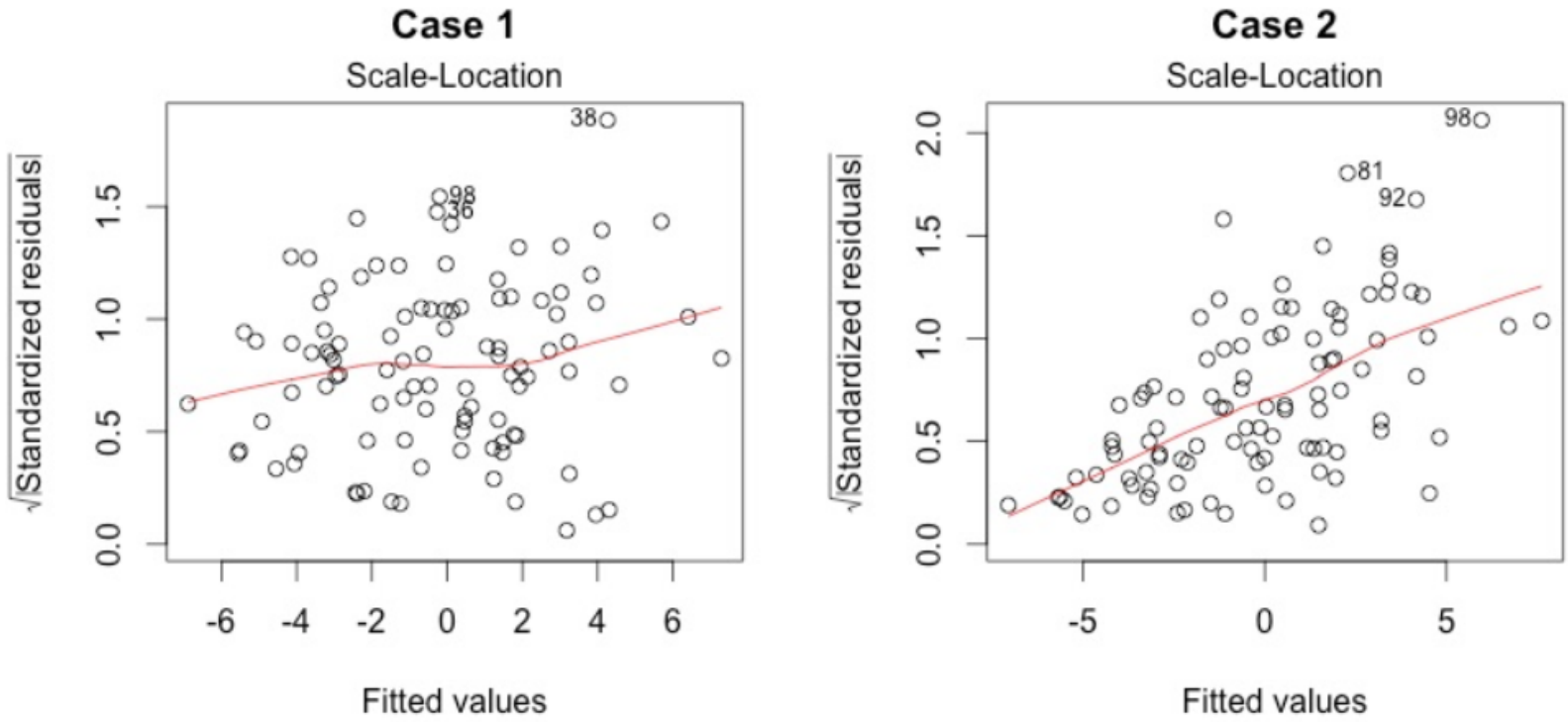
This plot shows if residuals are normally distributed. It's good if residuals follow the straight dashed line.





Linear regression

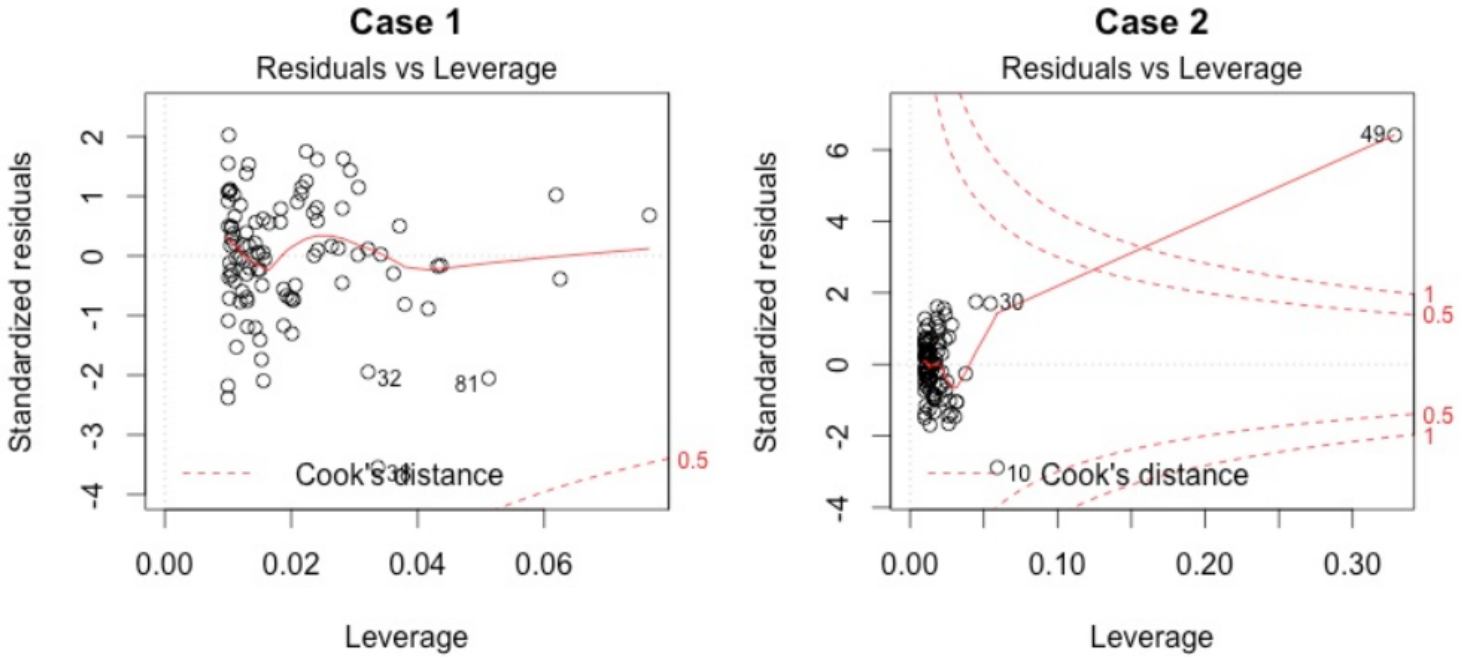
3. Scale-Location plot This plot shows if residuals are spread equally along the ranges of predictors. Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity.



Linear regression

4. Residuals vs Leverage

As well as checking our assumptions, we should also investigate any outliers or influential cases. We search for outlying values at the upper right corner or at the lower right corner. When cases are outside of the dashed lines (meaning they have high "Cook's distance" scores), the cases are influential to the regression results and the regression results will be altered if we exclude them.



Logistic regression

Bninary logistic regression models the probabilities for classification problems with two possible outcomes (a binary outcome).

Instead of fitting a straight line (or a hyperplane), the logistic regression model uses the logistic function to restrict the output of a linear equation between 0 and 1. The logistic function is defined as:

$$\text{logistic}(\eta) = \frac{1}{1+\exp(-\eta)}$$

Logistic regression

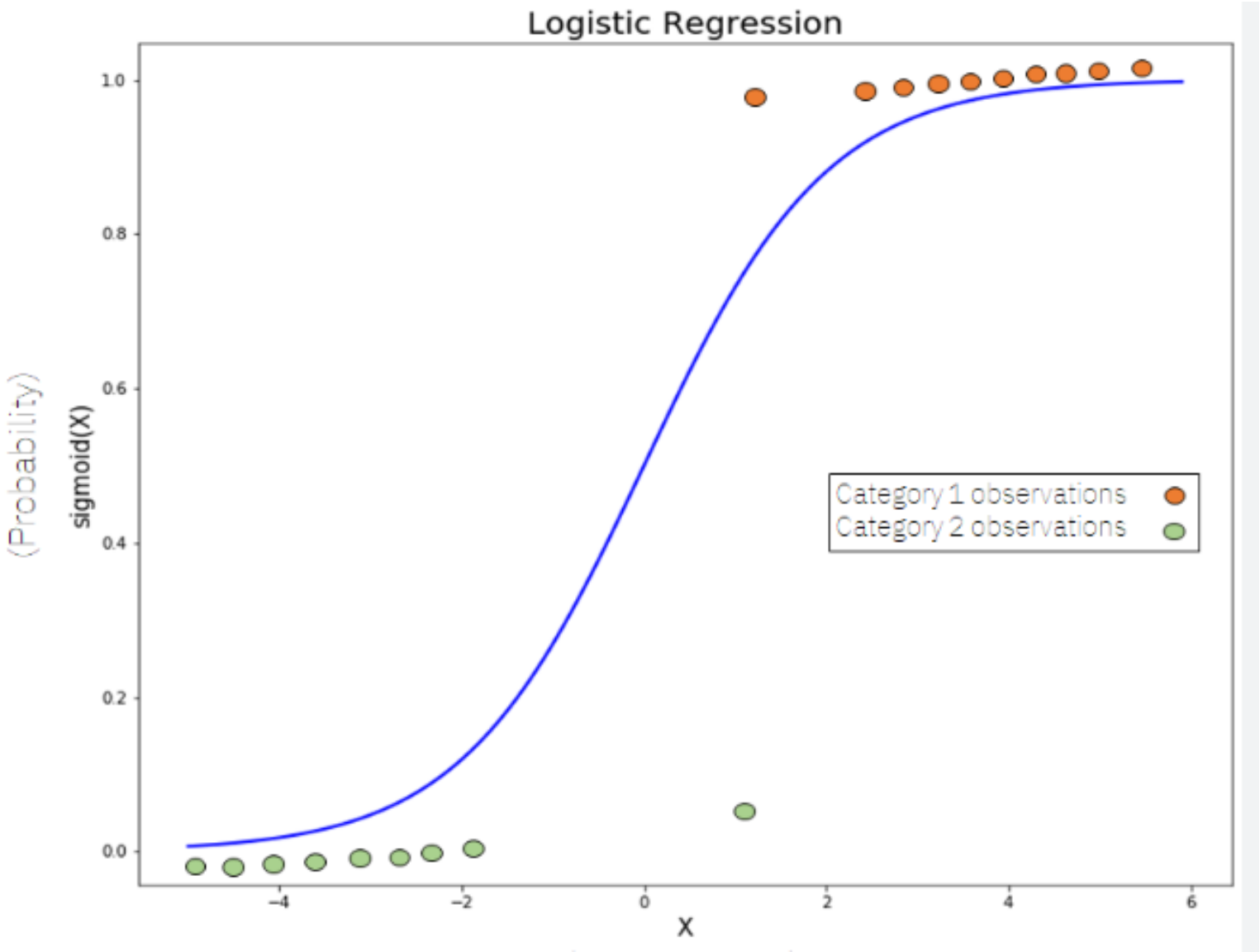
For example, if we want to assess the association between obesity and incident cardiovascular disease, we could fit the following logistic regression model:

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = -2.367 + 0.658(\mathit{Obesity})$$

where is $\hat{p}$ the expected probability that the outcome (CVD) is present.

Obesity is an independent variable in the model, coded as follows: 1=obese and 0=not obese.
The log odds of incident CVD is 0.658 times higher in persons who are obese as compared to not obese.
If we exponentiate the regression coefficient, $\exp(0.658) = 1.93$, we get the unadjusted Odds Ratio. Then, then odds of developing CVD are 1.93 times higher among obese persons as compared to non obese persons.

ROC analysis



ROC analysis

Evaluating the model's performance is a key step in validating it for use in real-world decision-making and prediction. A common evaluative tool is the ROC curve.

ROC curves are graphs that plot a model's false-positive rate against its true-positive rate across a range of classification thresholds; that is, across various cutoffs used to split real-valued model outputs (such as probabilities) into binary predictions of "Yes"/1/"Success"/etc. and "No"/0/"Failure"/etc.

ROC stands for receiver operating characteristic; They came about in World War II as a way of assessing the accuracy of radio operators' determinations of whether radar blips were genuine signals—e.g. fighter planes or noise.

ROC analysis

Let's say that we estimate a logistic regression for a data set containing a binary outcome variable with values of Yes and No, and a set of predictor variables.

We can use that model to estimate the probability that each observation in the original data set—or, even better, in an independent data set will be a Yes case. Let's call these probabilities $P_1, ..., P_n$.

We can convert the probability estimated for each observation into a binary prediction —Yes or No — based on some classification threshold, for example, we might by setting T=0.5.

ROC analysis

$$\text{Binary prediction for the } i^{\text{th}} \text{ observation} = \left\{ \begin{array}{ll} Yes, & \text{if } P_i > T \\ No, & \text{if } P_i \leq T \end{array} \right\}$$

The binary predictions can be compared to the actual values of Y to determine the counts of true positives, false positives, true negatives, and false negatives among the model's predictions at a particular classification threshold. These counts comprise a confusion matrix:

	Actual outcome = Yes	Actual outcome = No
Predicted outcome = Yes	# true positives	# false positives
Predicted outcome = No	# false negatives	# true negatives

From there, true-positive and false-positive rates-the constituent values of a ROC curve-are easily derived:

ROC analysis

$$\text{True-positive rate } (\mathrm{TPR}) = \frac{\text{True positives } (\mathrm{TP})}{\text{True positives } (\mathrm{TP}) + \text{ False negatives } (\mathrm{FN})}$$

$$\text{False-positive rate } (\mathrm{FPR}) = \frac{\text{False positives } (\mathrm{FP})}{\text{False positives } (\mathrm{FP}) + \text{ True negatives } (\mathrm{TN})}$$

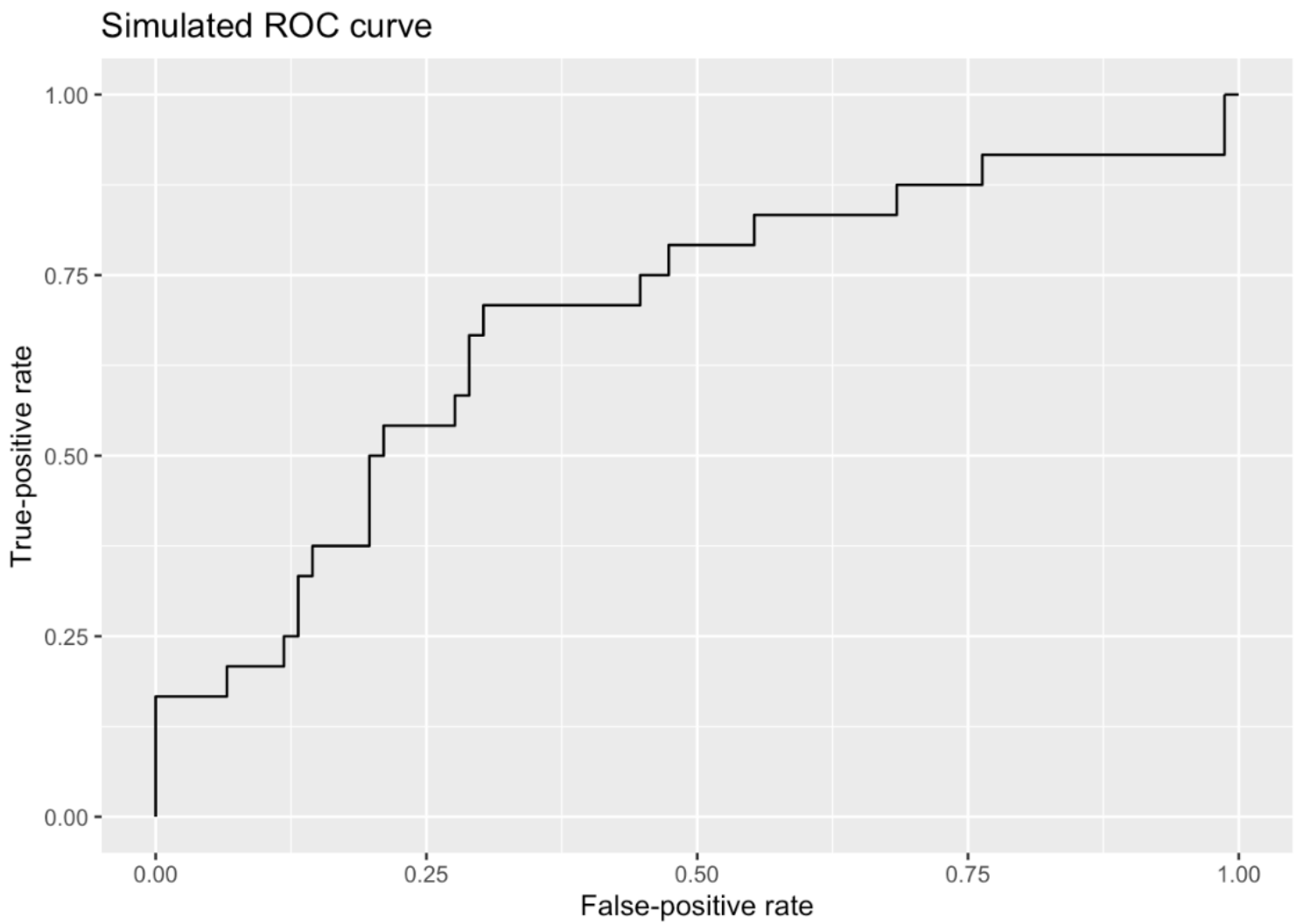
For a given model, we can calculate these rates at a range of classification thresholds.

ROC analysis

- These calculations don't need to be performed manually; software packages like pROC and ROCR in R quickly generate ROC curves by calculating TPR/FPR values for various classification thresholds, using programmatic rules and speedy algorithms to determine thresholds and corresponding TPRs/FPRs.

- Once TPRs and FPRs have been calculated for a range of classification thresholds, generating the corresponding ROC curve is simply a matter of plotting those points, with the classification threshold decreasing—"relaxing"—from left to right on the graph.

ROC analysis



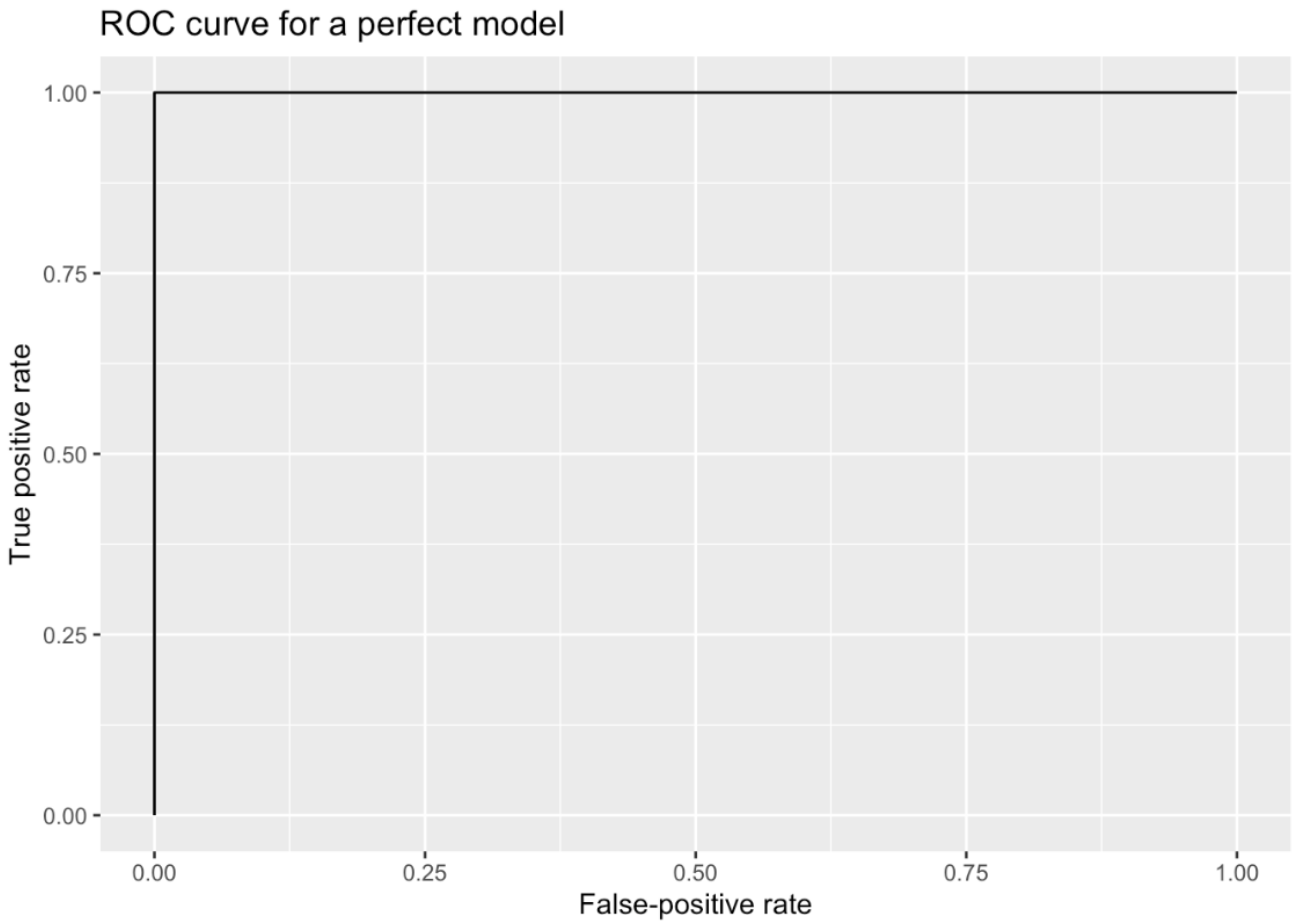
ROC analysis

Fundamental fact about models used for binary classification: *The dual interests of maximizing true-positive rates and minimizing false-positive rates are in tension.*

1. if we set the classification threshold for a prediction of Yes at a probability of 1, the threshold is so strict that we're going to miss all of the true Yes's, but in exchange, we're not going to mistakenly predict that any true No's are Yes's (This is reflected on the far left of the ROC curve).
2. Conversely, if we set the classification threshold at 0, we're going to predict that every observation is a Yes. We're therefore going to achieve a true-positive rate of 100%, but that will be in exchange for suffering from a false-positive rate of 100% as well (This is reflected on the far right side of the ROC curve).

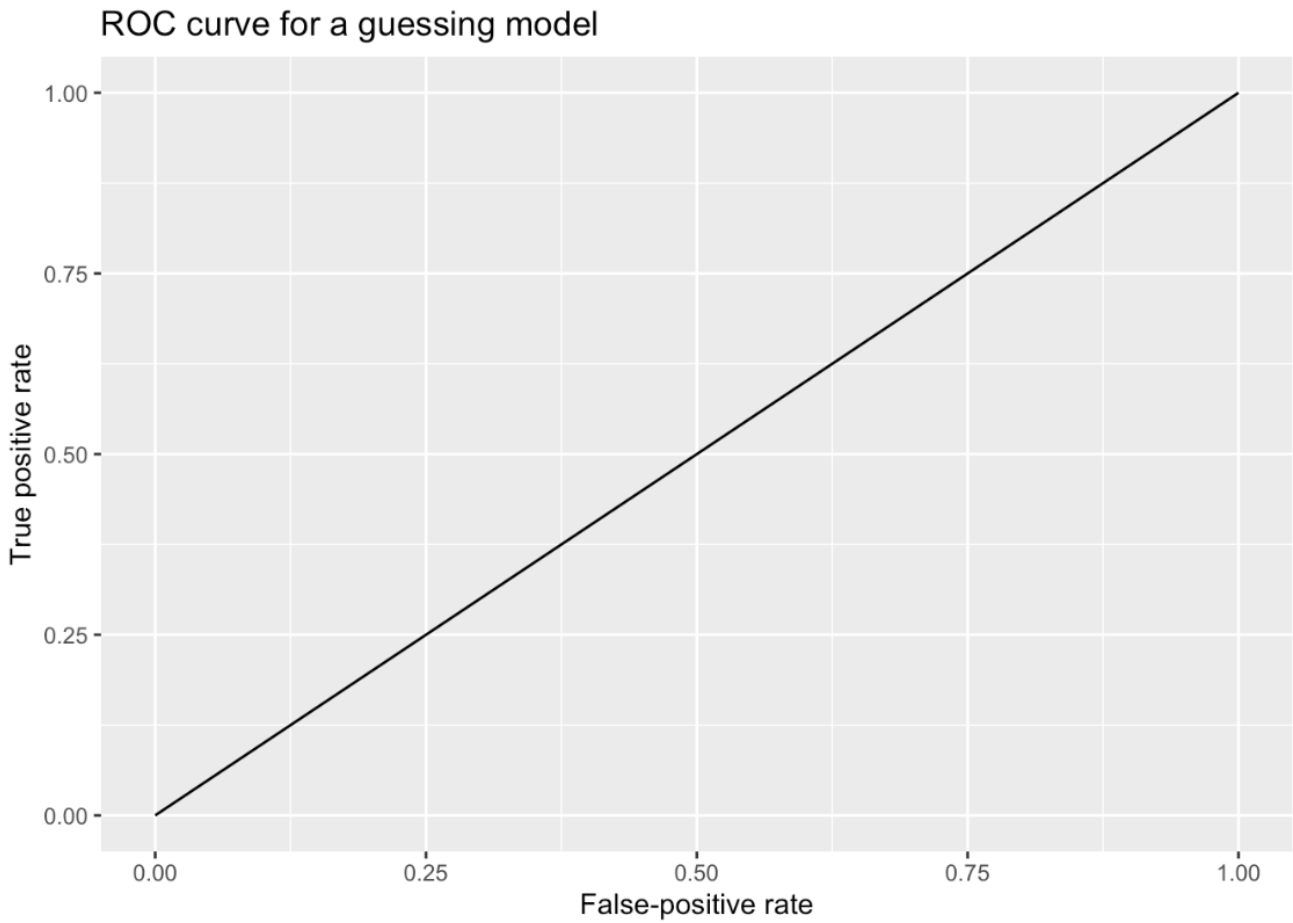
ROC analysis

A perfectly predictive model, for example, a model that assigned a probability of 0 to every true No case and a probability of 1 every true Yes case — would generate the following ROC curve:



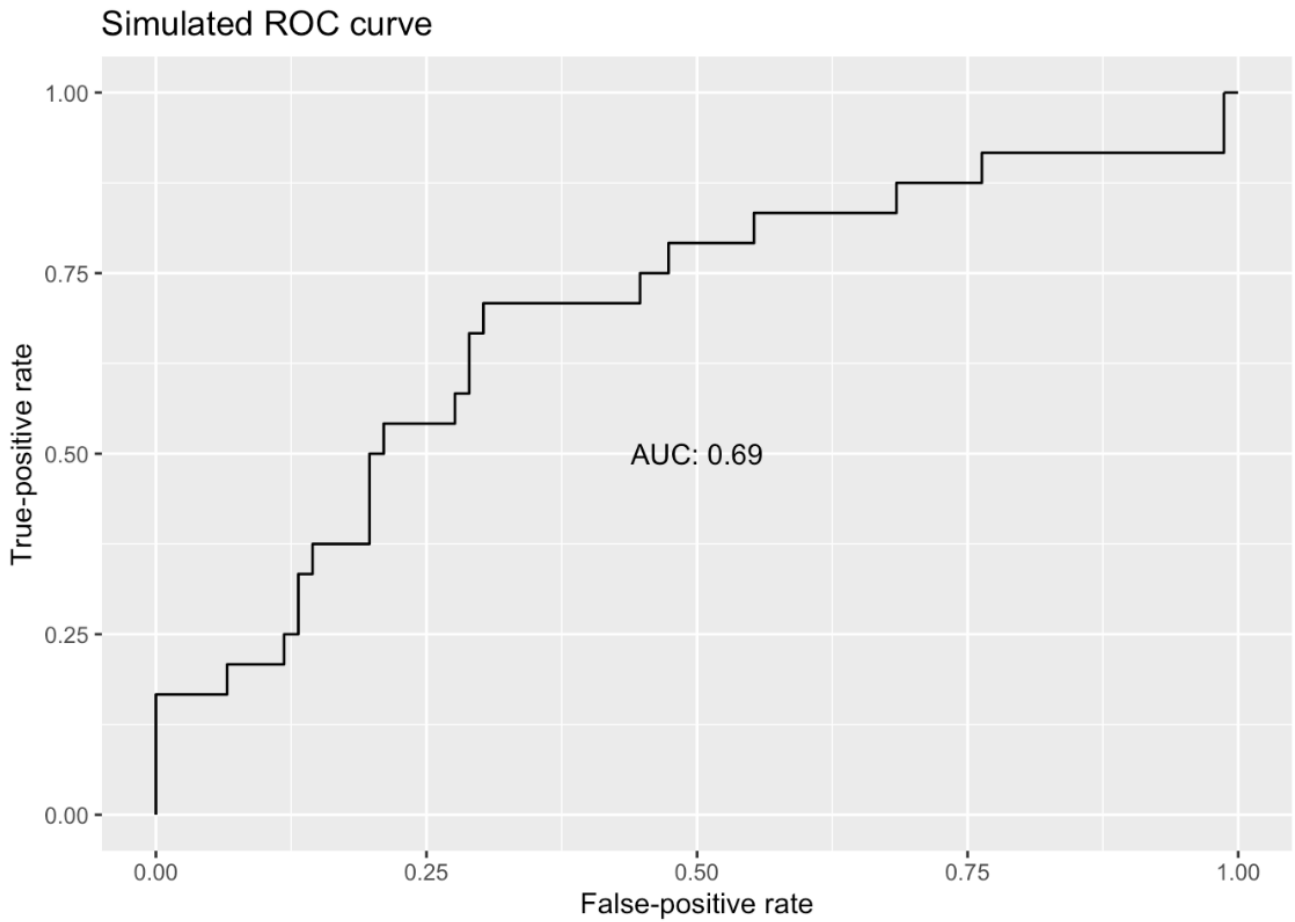
ROC analysis

A useless, guessing model - a model that simply assigned an identical probability of Yes to every observation would generate a diagonal ROC curve. The model has no discriminant ability, so its FPR and TPR are equivalent.



ROC analysis

The area under the ROC curve (AUC) or the amount of space beneath it, scales with overall classification performance.



ROC analysis

The AUC is the probability that the real-valued model output (e.g., the probability) for a randomly selected Yes case will be higher than the real-valued model output for a randomly selected No case.

We should see, then, that if we repeatedly sample one true Yes case and one true No case at random from the simulated data, the long-run proportion of times that the Yes case's predicted probability of being a Yes is greater than the No case's predicted probability of being a Yes will converge to 0.69.

The AUC can assist in comparing the overall performance of models used for binary classification.

The End!