Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

Basic statistical models

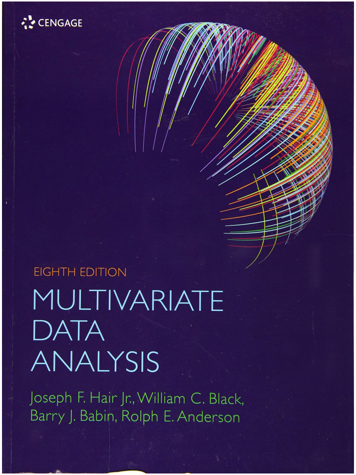
Julius Fenn[1]

[1]University of Freiburg
PhD student at the Institute of Experimental Psychology Freiburg (Cognition, Action, and Sustainability)

27.01.2022

book: multivariate statistic

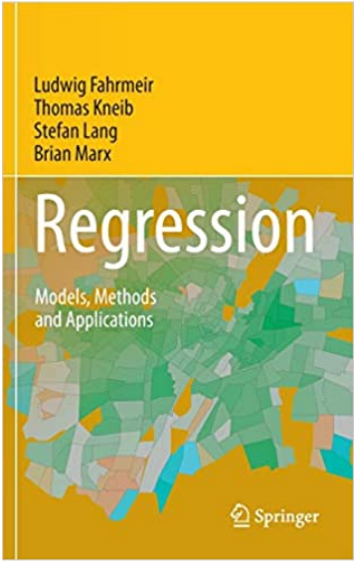
Hair et al. (2019): *Multivariate data analysis*



book: regression models

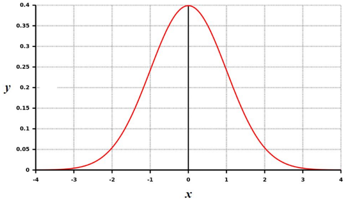
Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

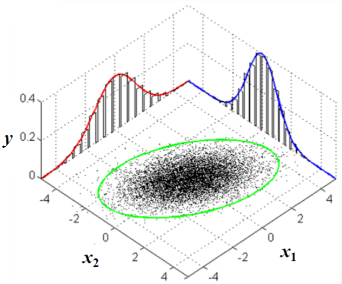
Fahrmeir, Kneib et al. (2013): *Regression Models, Methods and Applications*



univariate vs. multivariate

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

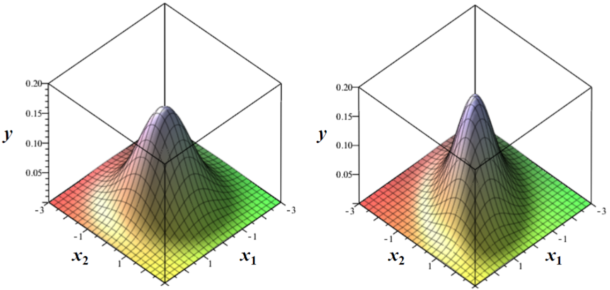




If two variables are bivariate normally distributed, then the two variables are are also univariate normally distributed, which can be seen here in the marginal.

multivariate: dependencies

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \cdot e^{-0.5(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$



only in the right bivariate distribution the variables are correlated

scope of statistics

Statistics may be regarded as

- the study of populations,
- the study of variation,
- the study of methods of the reduction of data.

The third aspect ... is introduced by the practical need to reduce the bulk of any given body of data. - Ronald A. Fisher (1925)

possible: to answer descriptive, predictive, inferential or causal questions, see What is the question?

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

What are multivariate data?

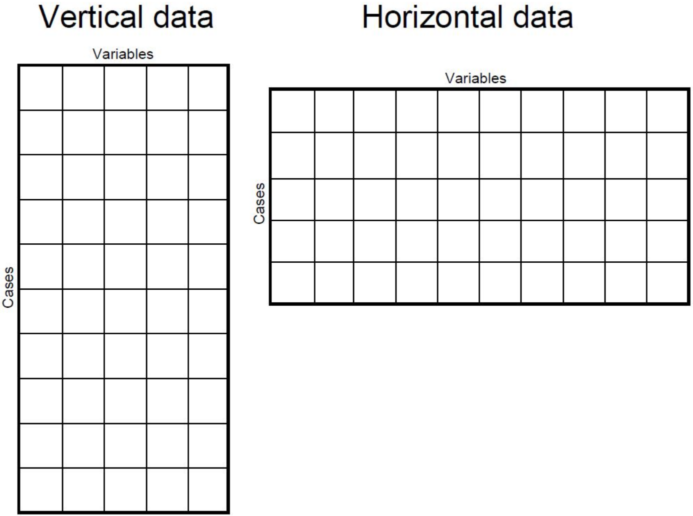
- Multivariate data arise when researchers record the values of several variables on a number of units in which they are interested.
- This leads to a **vector-valued or multidimensional observation** for each unit.
- In some studies, the variables are chosen by design because they are known to be essential descriptors of the system under investigation.
- In other studies, many variables may be measured simply to collect as much information as possible.
- ...

Examples of multivariate data

Multivariate data are ubiquitous as is illustrated by the following three examples:

- Psychologists and other behavioural scientists often record the values of several different cognitive variables on a number of subjects.
- Educational researchers may be interested in the examination marks obtained by students for a variety of different subjects.
- Environmentalists might assess pollution levels of a set of cities along with noting other characteristics of the cities related to climate and human ecology.

Format of multivariate data



The data matrix / data frame

A multivariate data matrix $X \in \mathbb{R}^{n*p}$ will have the form

$$\begin{vmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{11} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{vmatrix}$$

where the element x_{ij} is the value of the jth variable for the ith unit.

- The number of units under investigation is n and the number of measurements taken on each of these n units is p.
- The theoretical entities describing the univariate distributions of each of the p variables and their joint distribution are denoted by random variables X_1, ..., X_p

Hypothetical example

Individual	sex	age (years)	IQ	depression	health	weight (lbs)
1	Male	21	120	Yes	Very good	150
2	Male	43	NA	No	Very good	160
3	Male	22	135	No	Average	135
4	Male	86	150	No	Very poor	140
5	Male	60	92	Yes	Good	110
6	Female	16	130	Yes	Good	110
7	Female	NA	150	Yes	Very good	120
8	Female	43	NA	Yes	Average	120
9	Female	22	84	No	Average	105
10	Female	80	70	No	Good	100

Note: NA = **N**ot **A**vailable

Levels of measurements

- Nominal: unordered categorical variables. Examples include the sex of the respondent and hair colour.
- Ordinal: where there is an ordering but no implication of equal distance between the different points of the scale. Examples include social class (coded from I to V, say) and educational level (no schooling, primary, secondary, or tertiary education).
- Interval: equal differences between successive points on the scale but the position of zero is arbitrary. Example: measurement of temperature in Celsius or Fahrenheit.
- Ratio: one can investigate the relative magnitudes of scores. The position of zero is fixed. Examples include the measure of temperature in Kelvin, weight and length.

Missing values

- Missing values in multivariate data may arise for a number of reasons:
  - Non-response in sample surveys.
  - Dropouts in longitudinal data.
  - Refusal to answer particular questions in a questionnaire.

- Complete-case analysis: omit any case with a missing value on any of the variables.
- Available-case analysis: use all the cases available to estimate quantities of interest.
- Imputation: the practice of „filling in“ missing data with plausible values, see https://stefvanbuuren.name/fimd/

Common statistical tests are linear models

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

Common statistical tests are linear models

Last updated: 28 June, 2019.Also check out the Python version!

See worked examples and more details at the accompanying notebook: https://lindeloev.github.io/tests-as-linear

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: lm(y ~ 1 + x)	**y is independent of x** P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y_1, y_2, paired=TRUE) wilcox.test(y_1, y_2, paired=TRUE)	lm(y_2 - y_1 ~ 1) lm(signed_rank(y_2 - y_1) ~ 1)	✓ for N >14	One intercept predicts the pairwise **y_2-y_1** differences. - (Same, but it predicts the *signed rank* of **y_2-y_1**.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y_1, y_2, var.equal=TRUE) t.test(y_1, y_2, var.equal=FALSE) wilcox.test(y_1, y_2)	lm(y ~ 1 + G_2)[A] gls(y ~ 1 + G_2, weights=...[B])[A] lm(signed_rank(y) ~ 1 + G_2)[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
Multiple regression: lm(y ~ 1 + x_1 + x_2 + ...)	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	lm(y ~ 1 + G_2 + G_3 +...+ G_N)[A] lm(rank(y) ~ 1 + G_2 + G_3 +...+ G_N)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G_2 + G_3 +...+ G_N + x)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	aov(y ~ group * sex)	lm(y ~ 1 + G_2 + G_3 +...+ G_N + S_2 + S_3 +...+ S_K + G_2*S_2 + G_3*S_3 + ... + G_N*S_K)	✓	Interaction term: changing **sex** changes the **y ~ group** parameters. *Note: **$G_{2 \text{ to } N}$** is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for **$S_{2 \text{ to } K}$** for sex. The first line (with G_i) is main effect of group, the second (with S_i) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.*	*[Coming]*
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	**Equivalent log-linear model** glm(y ~ 1 + G_2 + G_3 + ... + G_N + S_2 + S_3+ ... + S_K + G_2*S_2 + G_3*S_3 +...+ G_N*S_K, family=...)[A]	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments:* `glm(model, family=poisson())` *As linear-model, the Chi-square test is $\log(y_i) = \log(N) + \log(\alpha_i) + \log(\beta_j) + \log(\alpha_i\beta_j)$ where α_i and β_j are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	chisq.test(y)	glm(y ~ 1 + G_2 + G_3 +...+ G_N, family=...)[A]	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation y ~ 1 + x is R shorthand for $y = 1 \cdot b + a \cdot x$ which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_i and S_i are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when $\Delta x = 1$ between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. lm requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at https://lindeloev.github.io/tests-as-linear.

[A] See the note to the two-way ANOVA for explanation of the notation.
[B] Same model, but with one variance per group: `gls(value ~ 1 + G₂, weights = varIdent(form = ~1|group), method="ML")`.

Jonas Kristoffer Lindeløv
https://lindeloev.net

see https://lindeloev.github.io/tests-as-linear/

The General Linear Model

In a general linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} + \epsilon_i$$

the response y_i, $i = 1, ..., n$ is modelled by a linear function of explanatory variables x_j, $j = 1, ..., p$

General and Linear

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

Here **general** refers to the dependence on potentially more than one explanatory variable, v.s. the simple linear model:

$$y_i = \beta_0 + \beta_1 x_1 + \epsilon_i$$

The model is **linear** in the parameters, e.g.

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon_i$$

Error structure

ϵ_i is the deviation of a measurement y_i from the ideal straight line $\beta_0 + \beta 1 x_i$ (called error or residuals)

We assume that the errors ϵ_i are independent and identically distributed such that

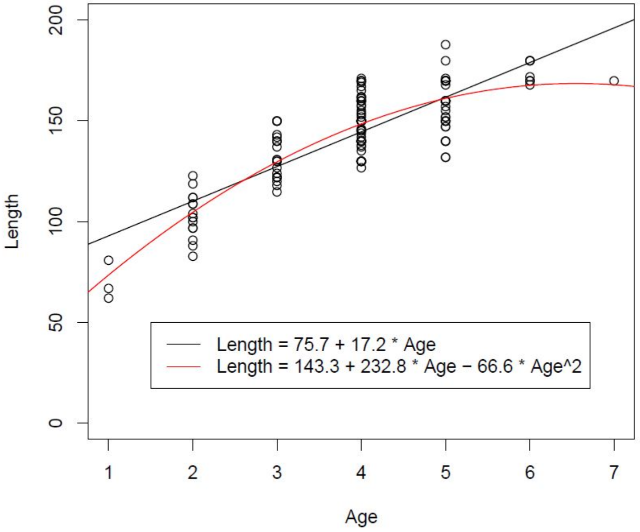
$$E[\epsilon_i] = 0$$
$$Var[\epsilon_i] = \sigma^2$$

Typically we assume

$$\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

as a basis for inference, e.g. t-tests on parameters.

Example


Length
Age
0
50
100
150
200
1
2
3
4
5
6
7
Length = 75.7 + 17.2 * Age
Length = 143.3 + 232.8 * Age − 66.6 * Age^2

Interpretation of the parameters

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, ..., n$$

- Intercept β_0: expectation of the response if all covariates are set to zero.
- Slope of a continuous covariate x_j: expected difference in the response when comparing two observations with x_j differing by one unit.
- Slope of a binary covariate x_j: expected difference in the response between two observations with $x_j = 1$ and $x_j = 0$.

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

descriptive statistics

see „Repetition of statistics“ slide 20 and the following

$\rightarrow$ modal value, median, mean value, quantiles, variance, standard deviation; graphics (histogram, boxplot); skewness, kurtosis

compute mean score

Example Satisfaction with Life Scale (SWLS), developed by Diener et al. (1985):

		Strongly Disagree	Disagree	Slightly Disagree	Neither Agree nor Disagree	Slightly Agree	Agree	Strongly Agree
1.	In most ways my life is close to my ideal.	1	2	3	4	5	6	7
2.	The conditions of my life are excellent.	1	2	3	4	5	6	7
3.	I am satisfied with my life.	1	2	3	4	5	6	7
4.	So far I have gotten the important things I want in life.	1	2	3	4	5	6	7
5.	If I could live my life over, I would change almost nothing.	1	2	3	4	5	6	7

$\rightarrow$ compute the mean of the scale (precondition: unidimensional construct): $\sum_{i=1}^{m} y_{vi}$
$\rightarrow$ compute the standard deviation of the scale, which indicates the spread of the answers

Moosbrugger und Kelava (2020), 143ff.

compute mean score: example

mean score over variables „stflife, stfeco, stfgov, stfdem“ of the European Social Survey (ESS) in R:

```
# approach 1:
dat %>%
  select(matches("^stf")) %>%
  rowMeans()

# approach 2:
rowMeans(x = dat[, c("stflife", "stfeco",
"stfgov", "stfdem")])
```

! check for inverse coded items

correlation: graphics and measures

see „Repetition of statistics“ slide 91 and the following

$\rightarrow$ contingency table + chi squared test, (pairwise) scatterplot; correlation (Pearson, Spearman, Kendall's τ_α)

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

hypothesis test: basic idea

basic idea behind a hypothesis test:

- State what we think is true.
- Quantify how confident we are about our claim.
- Use sample statistics to make inferences about population parameters.

more technical: process of choosing between two hypothesis (H_0, H_1) about a probability distribution based on observed data from the distribution $\rightarrow$ allows you to make inferences about a population parameter by analyzing differences between the results observed (the sample statistic) and the results that can be expected if some underlying hypothesis is actually true

Example

We want to know, if the average women's weight differs from the average men's weight? Therefore we could do the following:

1. visualizing your data (boxplot)
2. preliminary test to check independent t-test assumptions
3. compute unpaired two-samples t-test

Classical t-test:

$$t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}}$$

useful resources:
Statistical tools for high-throughput data analysis:
http://www.sthda.com/english/wiki/comparing-means-in-r
Methodenberatung [german]: https://www.methodenberatung.uzh.ch/de/datenanalyse_spss.html

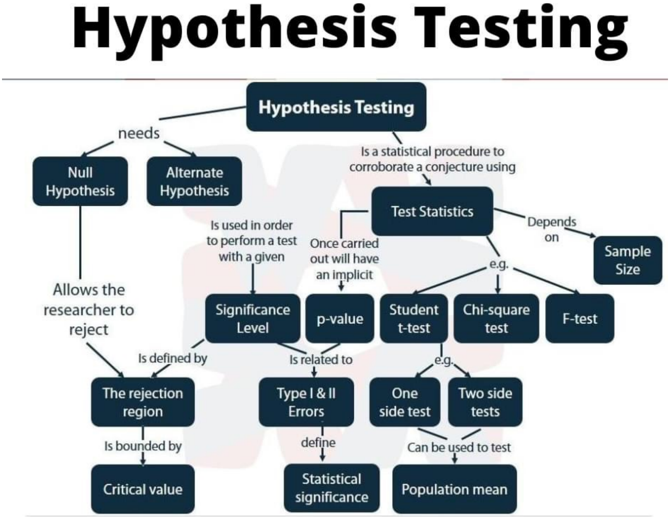
how to: overview I

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

The methodology behind hypothesis testing:

1. State the null hypothesis.
2. Select the distribution to use.
3. Determine the rejection and non-rejection regions.
4. Calculate the value of the test statistic.
5. Make a decision.

how to: overview II



from: https://mobile.twitter.com/DataScienceDojo/status/1469050218360578049/photo/1

1. State the null hypothesis

- null hypothesis (H_0): statement containing a null, or zero, difference; this hypothesis that undergoes the testing procedure; represents the status quo or what is assumed to be true
- alternative hypothesis (H_1): statement must be true if the null hypothesis is false; represent what you wish

Example fair coin:
H_0: statement about the value of a population parameter, such as the population mean (μ) or the population proportion (p)

$$p = .50$$

H_1: the claim to be tested, the opposite of the null hypothesis

$$p \neq .50$$

error types I

Tabularised relations between truth/falseness of the null hypothesis and outcomes of the test:

Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision about null hypothesis (H_0)	**Don't reject**	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)

error types II

- α-error (Type I): Null hypothesis is rejected although it is true
- β-error (Type II): Null hypothesis is not rejected although it is false

$\rightarrow$ logic of hypothesis tests: only reject the null hypothesis if the sample data are not consistent with the null hypothesis AND keep the α error small while accepting the disadvantage of a higher probability of a β error (different logic: *compromise power analyses*)

Zucchini et al. (2009), 244

2. Select the distribution to use

select a sample (Latin letters) or the entire population (Greek letters):
Mean: $\overline{x} \rightarrow \mu$; Standard deviation: $s \rightarrow \sigma$

- if you know the standard deviation σ for a population, then you can calculate a confidence interval (CI) for the mean of that population, sample mean $\overline{x}$ plus or minus a margin of error
- for a population with unknown mean μ and known standard deviation σ, a confidence interval for the population mean, based on a simple random sample of size n, is

$$\overline{x} \pm Z_{\alpha/2} * \underbrace{\frac{\sigma}{\sqrt{n}}}_{\text{standard error}}$$

, where $Z_{\alpha/2}$ is the upper $(1 - alpha)/2$ critical value for the standard normal distribution

see R-code: confidence intervals

Three ways to test hypotheses

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

1) Confidence interval

$$\beta \pm Z_{\alpha/2} * \underbrace{\frac{\sigma}{\sqrt{n}}}_{\text{standard error}}$$

$\rightarrow$ contains the true parameter with a probability of $1 - \alpha$
$\rightarrow$ for linear regression: if 0 contained non-significant predictor / test

2) Test statistics exceed critical value
$\rightarrow$ for linear regression rule of thumb: empirical value $\geq |2|$

3) Exceedance probability / p-values
$\rightarrow$ $p < \alpha$ reject the H_0

Zucchini et al. (2009), 161ff.; Fahrmeir, Heumann et al. (2016), 381ff.

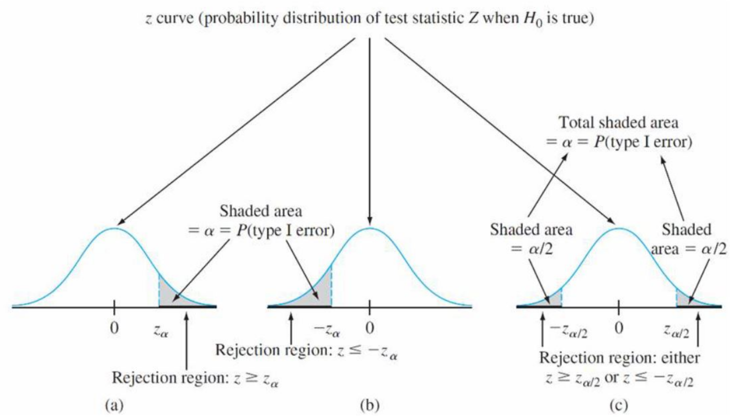
3. Determine the rejection and non-rejection regions

significance level (also denoted as α) is the probability of rejecting the null hypothesis when it is true

Example:
significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference (α-error)

- p-value is the area under the curve to the left and / or right of the test statistic, compared to the level of significance (α)
- critical value is the value that defines the rejection zone (the test statistic values that would lead to rejection of the null hypothesis), defined by the level of significance
- level of significance (α) is the probability that the test statistic will fall into the critical region when the null hypothesis is true, set by the researcher

rejection and non-rejection regions



Rejection regions for z tests: (a) upper-tailed test; (b) lower-tailed test; (c) two-tailed test

4. Calculate the value of the test statistic

values of the test statistic separate the rejection and non-rejection regions:

- Rejection region: the set of values for the test statistic that leads to rejection of H_0
- Non-rejection region: the set of values not in the rejection region that leads to non-rejection of H_0, acceptance of H_1

p-value, also called the probability of chance: the greater likelihood of obtaining the same result; definition: „the probability of obtaining a test statistic equal to or more extreme value than the observed value of H_0"
compare the p-value with α:

- if p-value $< \alpha$, reject H_0
- if p-value $>= \alpha$, do not reject H_0

5. Make a decision

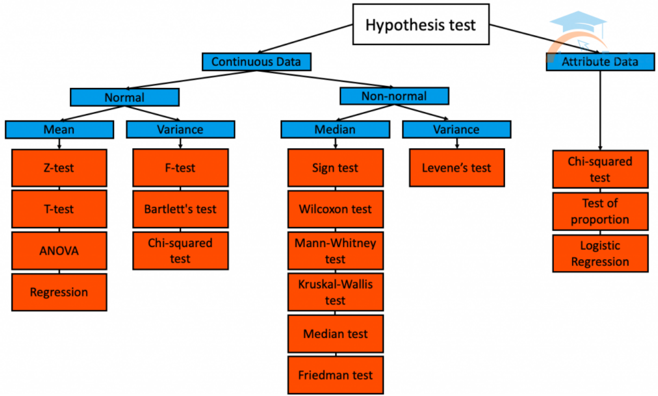
based on the result, you can determine if your test accepts or rejects the null hypothesis using the procedures on slide 32

Example t-Test:
The body weight between male and female twins does not differ significantly:

```
Welch Two Sample t-test

data:  daten$KG by daten$GES
t = 0.59441, df = 13.388, p-value = 0.5622
alternative hypothesis: true difference in means is not
equal to 0 95 percent confidence interval:
 -14.22265  25.06385
sample estimates:
mean in group 1 mean in group 2
        80.5140         75.0934
```

types of hypothesis tests



from: https://leanmanufacturing.online/introduction-to-hypothesis-testing/

! test assumptions

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

Preleminary test to check independent t-test assumptions:

- Assumption 1: Are the two samples independents?
- Assumtion 2: Are the data from each of the 2 groups follow a normal distribution?
- Assumption 3. Do the two populations have the same variances?
    - if not use the classic t-test which not assume equality of the two variances (Welch's t-test)

    $\rightarrow$ usually, the results of the classical t-test and the Welch t-test are very similar unless both the group sizes and the standard deviations are very different

simple linear model

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

Classical linear regression analysis aims to identify relationships between a dependent metric variable and one (or more) independent variables to predict values of the dependent variables. It assumes:

- the dependent variable and the independent variable(s) change only in constant relations (linearity)
- the residuals between the statistical units are independent of each other and are normally distributed: $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

simple linear model: interpretation

simple linear model:

$$y_i = \beta_0 + \beta_1 x_1 + \epsilon_i, \quad i = 1, ..., n$$

- Intercept β_0: expectation of the response if all covariates are set to zero.
- Slope of a continuous covariate x_j: expected difference in the response when comparing two observations with x_j differing by one unit.
- Slope of a binary covariate x_j: expected difference in the response between two observations with $x_j = 1$ and $x_j = 0$.
- ϵ_i: the deviation of a measurement y_i from the ideal straight line $\beta_0 + \beta_1 x_1$; the outcome is generally stochastic, so the prediction can typically never be exact

simple linear model: categorical variables I

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

dummy coding: reference group central

- make the categorical variable into a series of dichotomous variables (variables that can have a value of zero or one only)
- for all but one (reference group) of the levels of the categorical variable, a new variable will be created that has a value of one for each observation at that level and zero for all others

Example:

Level of race	New variable 1 (x1)	New variable 2 (x2)	New variable 3 (x3)
1 (Hispanic)	1	0	0
2 (Asian)	0	1	0
3 (African American)	0	0	1
4 (white)	0	0	0

simple linear model: categorical variables II

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

effect coding: no reference group, instead 0 is the average value of all observations

- all of the values in any new variable must sum to zero
- which level is assigned a positive or negative value is not very important

effect coding of C:

$$X_k^e(C) = \begin{cases} 1 & C = k \\ 0 & C \neq k, C \neq K, \\ -1 & C = K \end{cases}$$

various types of contrasts possible, see:

https://stats.oarc.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis/

simple linear model: terms

simple linear model:

$$y_i = \beta_0 + \beta_1 x_1 + \epsilon_i, \quad i = 1, ..., n$$

Thereby

- y_i is the target variable (variable to be explained, response),
- x_i is the (non-stochastic) influencing variable (covariate, explanatory variable, regressor)
- ϵ_i are the errors (residuals),
- $\beta_0 + \beta_1 x_1$ is the regression line,
- β_0 is the intercept of the regression line,
- β_1 is the slope of the regression line (slope)
- n is the sample size

simple linear model: residual assumptions

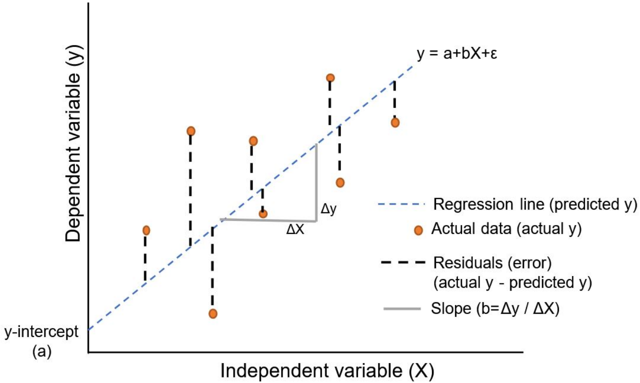
Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

Assumptions: The following conditions are typically placed on the error ϵ_i:

- $E(\epsilon_i) = 0$, the average deviation from the regression line is 0.
- $Var(\epsilon_i) = \sigma^2$, the dispersion around the regression line is the same everywhere (homoscedasticity).
- $\epsilon_1, ..., \epsilon_n$ are stochastically independent, i.e. the individual observations do not influence each other.
- $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, the errors are normally distributed

simple linear model: central aim

Least squares estimation: $KQ(\beta_0, \beta_1) = \sum(y_i - \hat{y}_i)^2 \rightarrow min$



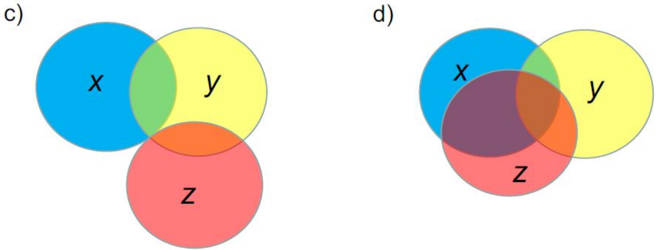
simple linear model: all assumptions

1. Number of predictors is smaller than number of cases (identifiability)
2. Assumption of linearity
3. The expected value of the errors is 0: $E(\epsilon_i) = 0$
4. The variances of the errors are equal (homoscedasticity): $Var(\epsilon_i) = \sigma^2$
5. No correlation between the errors: $Cov(\epsilon_i, \epsilon_j) = 0, foralli \neq j$
6. No exact multicollinearity of predictors
7. Normal distribution of errors: $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$
8. Reliability: The predictors are measured without error ($Rel(x_j) = 1$).

If assumptions 2-6 apply, the model is **BLUE** (smallest variance, linear, not biased).

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

Problems of multicollinearity



- Fig. c) y correlates with x and z, but the predictors x and z have no common variance: There is no multicollinearity.
- Fig. d) y correlates with x and z, but the predictors x and z have a very large proportion of shared variance: Multicollinearity is present.

simple linear model: significance I

F-Test: (Omnibus test) $\rightarrow$ Sum-of-squares decomposition

- In multiple regression, significance is usually tested with an F-test
- The F-test is based on a decomposition of the variance of the criterion variable y into an explained and an unexplained portion:

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

$$SS_{\text{total}} = SS_{\text{reg}} + SS_{\text{res}}$$

- SS = sum of squares
- reg = due to regression
- res = residuals (residuals)

simple linear model: significance II

Overview
Literature
multivariate statistics
framework
Models
descriptive
Correlation
hypothesis tests
linear model
Literature

Multiple coefficient of determination R^2:

$$R^2 = \frac{SS_{reg}}{SS_{tot}}$$

The sums of squares are non-standardised measures of variability: Dividing by N-1 results in standardised measures: the variances.

$$\frac{SS_{reg}}{N-1} \Big/ \frac{SS_{tot}}{N-1} = \frac{Var(\hat{y})}{Var(y)}$$

The coefficient of determination is then:

$$R^2 = \frac{Var(\hat{y})}{Var(y)} = \left(\frac{Cov(\hat{y}, y)}{SD(\hat{y}) \cdot SD(y)} \right)^2 = r^2_{y,\hat{y}}$$

simple linear model: significance III

Significance testing of **individual** β **coefficients:**

$$H_0\colon \beta_j = 0 \quad H_1\colon \beta_j \neq 0$$

$$t = \frac{\beta_j}{SE(\beta_j)}$$

with
$df = N - p - 1$ and SE is the estimated standard error

The empirical t-value is compared with the critical t-value:
$t_{emp} >= t_{crit} => H_1$

- The standard error estimation assumes a normal distribution of the predictors
- if the assumption is violated: SE is not estimated correctly

Literature

Fahrmeir, Ludwig, Christian Heumann et al. (2016). *Statistik: Der Weg zur Datenanalyse*. Springer-Verlag.

Fahrmeir, Ludwig, Thomas Kneib et al. (2013). *Regression Models, Methods and Applications*. Springer.

Hair, Joseph F et al. (2019). *Multivariate data analysis*. Annabel Ainscow.

Moosbrugger, Helfried und Augustin Kelava (2020). *Testtheorie und Fragebogenkonstruktion*. Springer.

Zucchini, Walter et al. (2009). *Statistik für Bachelor-und Masterstudenten: eine Einführung für Wirtschafts-und Sozialwissenschaftler*. Springer-Verlag.