Introduction to statistics

Bo Markussen
bomar@math.ku.dk

Data Science Laboratory
Department of Mathematical Sciences

November 23, 2022

Welcome to everyone!

- Applied statistics using R via the RSTUDIO interface.
  - 6 course days with lectures and (computer) exercises.
  - Frequentist statistics with univariate endpoints.
  - Statistical models for categorical and continuous data.
- Lectures and exercises given jointly for two courses:
  - Applied Statistics (Master Course, 7.5 ECTS).
  - Statistical methods for the Biosciences (PhD Course, 4.5 ECTS).
- Background:
  - Teaching level and course aims.
  - Data Science Laboratory

Who are we?

- Bo Markussen:
  - Course lecturer.
  - Associate professor.
  - Mathematical education, PhD in statistics from 2002.
  - Experience on applied statistics from Data Science Lab.
  - Office 04.3.16, 3'rd floor E-building at HCØ (Nørre Campus).

- Ulises Bercovich Szulmajster:
  - PhD student in statistics at KU-MATH.
  - Will be present at the exercise class in the afternoon.

Course material

- Computer software:
  - R: www.r-project.org + RStudio: www.rstudio.com
- Main literature:
  - The slides!
  - The help pages in R.
  - Course book: **Martinussen, Skovgaard, Sørensen, “A first guide to statistical computations in R”, Biofolia 2012.**
  - Sterne & Smith (2001), “Sifting the evidence—what’s wrong with significance tests?”, British Medical Journal, 226–231.
  - Gelman & Carlin (2014), “Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors”, Perspectives on Psychological Science, 1–11.
- Also used:
  - Your old book on basic statistics.
  - Grolemund, Wickham (2017), R for Data Science, O’Reilly.
  - Wickham (2016), ggplot2, Springer.

Statistical software

Provides validity, reliability, reproducibility

Programming: R, SAS, Stata, Python, MatLab, ...

Menu based: Excel, Graphpad Prism, SPSS, SAS Enterprise, JMP, Stata, R-commander, ...

Pros and Cons:

	Programming	**Menu based**
$+$	Full control, direct reproducibility	Good overview of models and possibilities
$-$	Syntax, commands, options, etc.	Mouse clicking, limited flexibility, reproducibility difficult

RStudio: An interface to R successfully encountering many of the cons in programming.

R markdown: File format for making dynamic documents that integrate code, output, graphs and text.

Basic concepts of frequentist statistics

Why do statistics?

Well, to answer four important questions

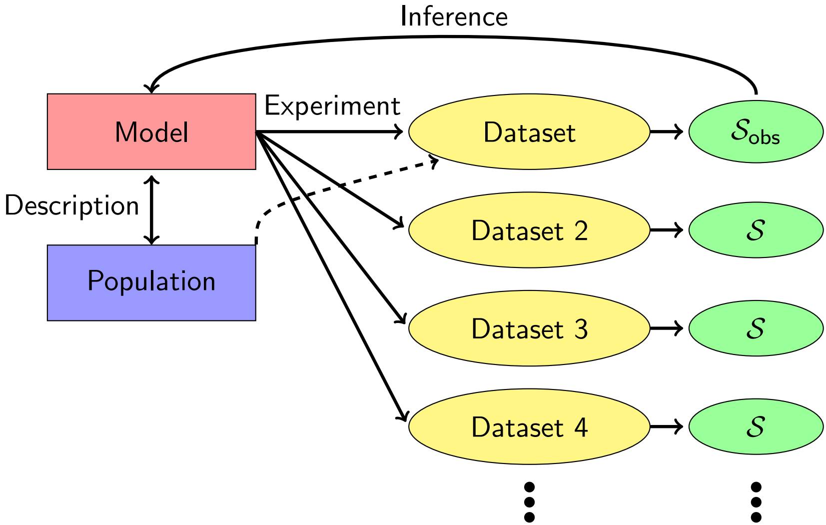
1. Is there an effect?
   - Answered by p-values.
2. Where is the effect?
   - Answered by p-values from post hoc analyses.
3. What is the effect?
   - Answered by confidence and prediction intervals.
4. Can the conclusions be trusted?
   - Answered by model validation.

Remarks:

- Often “effect” should be replaced by “association”.
- Statistical models are also used for other purposes: Which ones?

Model, data, statistic

Examples of statistics $\mathcal{S}$: estimator, confidence interval, test statistic, p-value



Understanding hypothesis testing

Exemplified by a permutation test

Does feed concentrate increase milk yield?

An intervention study

- Does feeding concentrate to diary cows have an effect on milk yield?

We can illustrate this is in a *causal diagram*:

$$\begin{array}{ccc} \text{diet} & & \\ & \searrow & \\ & & \text{yield} \\ & \nearrow & \\ \text{animal} & & \end{array}$$

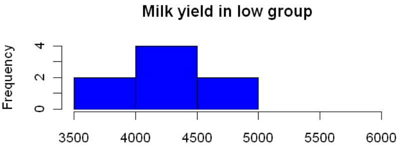
- Objective of the example:
  - Answer the posed question.
  - Learn basic concepts of hypothesis testing doing this.
  - First look at some R code.

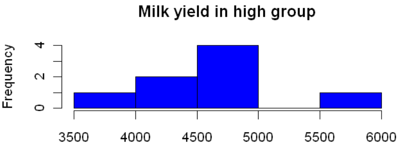
Does feeding concentrate influence milk yield?

Two groups of 8 cows, given low or high amounts of feed concentrate

Concentrate/day	Milk yield (kg) from week 1 to 36							
Low: 4.5 kg	4132	3672	3664	4292	4881	4287	4087	4551
High: 7.5 kg	3860	4130	5531	4259	4908	4695	4920	4727

Reference: V. Østergaard (1978)





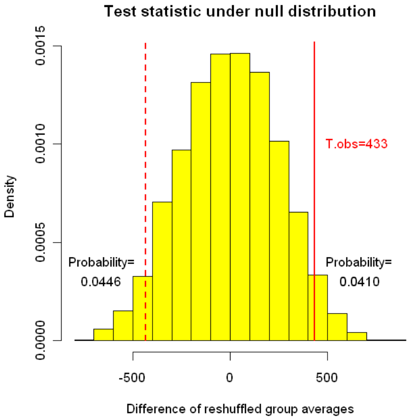
- Apparently there is an effect of feed concentrate.
- Confirmation by falsification of the null hypothesis of no effect.
- A test statistic summarizes the difference between groups, e.g.:

$$\begin{aligned}\mathcal{T}_{\text{obs}} &= \text{mean(high)} - \text{mean(low)} \\ &= 4629 - 4195 \\ &= 433\end{aligned}$$

Is the observed effect ($\mathcal{T}_{\text{obs}} = 433$ kg) significant?

Or might it be due to mere randomness?

- Null hypothesis $\Longrightarrow$ Observed difference of average milk yield due to random allocation of 16 cows into two groups.
- So let's redo a random allocation 10,000 times and inspect the differences of average milk yields...



The p-value is the probability of a more extreme test statistic than $\mathcal{T}_{\text{obs}}$,

$$p = \text{Prob}\big(|\mathcal{T}| \geq |\mathcal{T}_{\text{obs}}|\big)$$
$$= 0.0410 + 0.0446$$
$$= 0.0856$$

R code for the permutation test

For illustration only, in practice use e.g. wilcox.test()

```
# Encode data in two variables
diet  <- c("low","low","low","low","low","low","low","low",
           "high","high","high","high","high","high","high","high")
yield <- c(4132,3672,3664,4292,4881,4287,4087,4551,
           3860,4130,5531,4259,4908,4695,4920,4727)

# Compute test statistic on actual allocation of diets
T_obs <- mean(yield[diet=="high"])-mean(yield[diet=="low"])

# Test statistic after random allocation of diets
T_resample <- replicate(10000,{
  random_diet <- sample(diet)
  mean(yield[random_diet=="high"])-mean(yield[random_diet=="low"])
})

# Compute p-values
(p_value_onesided <- mean(T_resample >= T_obs))
(p_value_twosided <- mean(abs(T_resample) >= abs(T_obs)))
```

Summary of basic concepts

What to be learned from the milk yield example

- Statistical hypothesis tests distinguish real effects from random variation.
- Scientific hypothesis supported by falsifying opposite hypothesis.
- Test statistic measures discrepancy between data and null hypothesis.
- P-value = probability of larger discrepancy than the observed one.
- Small p-value $\Longrightarrow$ significance.
- The observed p-value, 0.0856, is not sufficiently small to claim statistical significance. What to do about that? (Multiple answers!)

Conclusion from a hypothesis test

Remark: Statistical significance is not the same as importance

- p-value measures disagreement with H_0:

$$\text{small p: disagreement=reject}$$
$$\text{large p: agreement=cannot reject}$$

  Conventional labelling ("." in some R outputs):

$$\begin{aligned} p > 0.05 &: \text{NS} && \text{(non significant)} \\ 0.05 < p < 0.10 &: . && \text{(significant at 10\% level)} \\ 0.01 < p < 0.05 &: * && \text{(significant at 5\% level)} \\ 0.001 < p < 0.01 &: ** && \text{(significant at 1\% level)} \\ p < 0.001 &: *** && \text{(significant at 0.1\% level)} \end{aligned}$$

- Small p = strong evidence against H_0. If $p = 0.2\%$, say, then
  - either H_0 is false
  - or H_0 is true and we have been unlucky! (risk $= 2/1000$)
  - or we have tested too many hypothesis (say 1000)
  - or the model is wrong (conclusion cannot be trusted)

Questions?

1. And then a break!

2. After the break we discuss the building blocks of a dataset: Observations, variables, and variable types.
   - Basically, this corresponds to tidy data in the *tidyverse* invented by Hadley Wickham.

Tidy data

Data example 1

Situation: A feature is measured for a collection of molecules *without* and *with* some modification. We have one or several repetitions for each combination of *modification* and *molecule*.

Scientific question: Does the modification have an impact of the individual molecules and for all molecules in general?

Here a data example in a non-tidy organisation:

Without	With
0.0	20.0
48.0	33.0
0.0, 76.2, 82.1, 57.9, 78.4	76.0
0.0	47.0
0.0	69.0
9.0	16.0

Data example 2

Situation: In an experiment multiple features are measured, e.g. *br6* and *br74*, at *multiple repetitions*.

Challenge: There are systematic difference between repetitions due to experimental environments and molecule synthesizing batches.

Data example in a non-tidy organisation. There are *two repetitions* of br6 and br74, and *each row is one molecule*:

br6_rep1	br74_rep1	br6_rep2	br74_rep2
0.6423	0.6129	0.5507	0.5359
0.4004	0.2456	0.3336	0.2749
0.1135	0.0403	0.0424	0.0529

What is a tidy data organization?

A **data matrix** (think of a spreadsheet like Excel) such that

1. each column is a **variable**, i.e. a (physical) quantity that can be measured or chosen by design in the experiment,
2. each row is an **observation**, i.e. the values of the variables for a particular experimental unit,
3. all the relevant information is explicitly represented in the data matrix, e.g. not implicitly given by ordering of rows or columns.

Exercise: Which of these properties are violated by the non-tidy data organizations shown in Data Example 1 and 2?

Tidy vs. non-tidy data: Data example 1

A tidy data organization:

Molecule	Modification	Feature
A	without	0.0
A	with	20.0
B	without	48.0
B	with	33.0
C	without	0.0
C	without	76.2
C	without	82.1
C	without	57.9
C	without	78.4
C	with	76.0
D	without	0.0
D	with	47.0
E	without	0.0
E	with	69.0
F	without	9.0
F	with	16.0

A non-tidy data organization:

Without	With
0.0	20.0
48.0	33.0
0.0, 76.2, 82.1, 57.9, 78.4	76.0
0.0	47.0
0.0	69.0
9.0	16.0

Tidy vs. non-tidy data: Data example 2

A tidy data organization:

Molecule	br6	br74
A	0.6423	0.6129
A	0.5507	0.5359
B	0.4004	0.2456
B	0.3336	0.2749
C	0.1135	0.0403
C	0.0424	0.0529

A non-tidy data organization:

br6_rep1	br74_rep1	br6_rep2	br74_rep2
0.6423	0.6129	0.5507	0.5359
0.4004	0.2456	0.3336	0.2749
0.1135	0.0403	0.0424	0.0529

Wide data vs. long data: Data example 2

Both organizations below are tidy!

Wide data organization:

Molecule	br6	br74
A	0.6423	0.6129
A	0.5507	0.5359
B	0.4004	0.2456
B	0.3336	0.2749
C	0.1135	0.0403
C	0.0424	0.0529

Long data organization:

Molecule	Experiment	Feature	Y
A	1	br6	0.6423
A	1	br74	0.6129
A	2	br6	0.5507
A	2	br74	0.5359
B	3	br6	0.4004
B	3	br74	0.2456
B	4	br6	0.3336
B	4	br74	0.2749
C	5	br6	0.1135
C	5	br74	0.0403
C	6	br6	0.0424
C	6	br74	0.0529

- When more features are measured, the organization to the left becomes wider, whereas the organization to the right becomes longer.
- Advice: Use physical (and concise) names, and not generic names like Y as done above.

Exercise: Rows and Columns, Observations and Variables

Group I	Group II	Group III
243	206	241
251	210	258
275	226	270
291	249	293
347	255	328
354	273	
380	285	
392	295	
	309	

Table shows red cell folate levels (μg/l).

Group	Level
I	243
I	251
I	275
I	291
I	347
I	354
I	380
I	392
II	206
II	210
II	226
II	249
II	255
II	273
II	285
II	295
II	309
III	241
III	258
III	270
III	293
III	328

Reference: Amess et al. (1978), Megaloblastic haemopoiesis in patients receiving nitrous oxide, Lancet, 339-342.

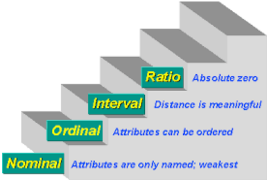
- Are these the same dataset?
- What are the variables? How many observations are there?

Variable types

Nominal, ordinal, interval, ratio

Four categories of variable types with increasing structural information

- Example of a nominal variable:
  - Color (red, green, purple).
- Example of an ordinal variable:
  - Status (healthy, slight symptoms, severe symptoms, dead).
- Example of an interval variable:
  - Temperature measured in degrees of Celsius.
- Examples of ratio variables:
  - Temperature measured in Kelvin.
  - Height (measured in cm).
  - Money on my bank account (measured in Danish kroners).



- Nominal and ordinal variables are subtypes of categorical variables.
- Interval and ratio variables are subtypes of continuous variables.

Table of Variables

Table of Variables

A summary of tidy data in form of another table with one row per variable and 4 columns with meta-information:

1. **Variable:** The name of the variable in the dataset.
2. **Type:** The variable type (nominal, ordinal, interval or ratio).
3. **Range:** State the levels (separated by "," and "<" for nominal and ordinal variables, respectively) and range [min; max] for continuous variables.
4. **Usage:** The role of the variable in the statistical analysis (fixed effect, random effect, response, correlation effect, subject id, not used, . . .)

Today we have only seen fixed effect and response. The response variable is characterized by:

- This is what we are interested in.
- This is what we want to predict knowing the other variables.
- This is where the random variation matters to us.

Data example 2 revisited

Dataset on wide format:

molecule	br6	br74
A	0.6423	0.6129
A	0.5507	0.5359
B	0.4004	0.2456
B	0.3336	0.2749
C	0.1135	0.0403
C	0.0424	0.0529

Table of Variables:

Variable	Type	Range	Usage
molecule	nominal	A, B, C	fixed effect
br6	ratio	[0.0424; 0.6423]	response
br74	ratio	[0.0403; 0.6129]	response

Quiz: Can you imagine a situation, where it e.g. could make sense to use *br6* as a fixed effect?

Dataset on long format:

molecule	experiment	feature	Y
A	1	br6	0.6423
A	1	br74	0.6129
A	2	br6	0.5507
A	2	br74	0.5359
B	3	br6	0.4004
B	3	br74	0.2456
B	4	br6	0.3336
B	4	br74	0.2749
C	5	br6	0.1135
C	5	br74	0.0403
C	6	br6	0.0424
C	6	br74	0.0529

Table of Variables:

Variable	Type	Range	Usage
molecule	nominal	A, B, C	fixed effect
experiment	nominal	6 levels	random effect
feature	nominal	br6, br74	fixed effect
Y	ratio	[0.0403; 0.6423]	response

Questions?

1. And then a break!
2. After the break I give a short introduction to R and RStudio.
3. And we discuss T-tests and data transformations (you might know much of this already).

Introduction to R

The RStudio interface consists of $4 = 2 \times 2$ subwindows

Upper-left The editor, where you write your R programs.

Lower-left The console, where code is executed and results are printed.

Upper-right Overview of objects (variables, vectors, matrices, data frames, lists, functions, "results", etc.) in the (global-) environment.

Lower-right Miscellaneous: overview of working directory, history of plots, administration of packages, and help pages.

- R is a full-scale objected oriented programming language.
- In R your data is typically stored in either vectors, matrices, or most commonly in data frames (*tidyverse* introduce specialization tibbles).
- Results from analyses are stored in associated objects (for well programmed functions):
  - E.g. a call to the `lm()` function results in an lm-object.
  - Such objects may be printed, summarized and/or plotted.

Functions and R packages

R contains many predefined functions for doing statistical computations

- Standard functions: `mean()`, `sd()`, ...
  - These functions may be used without any further ado.
  - Includes so-called generic functions: `print()`, `summary()`, `plot()`, ...

- Functions from pre-installed packages: `MASS::boxcox()`, `cluster::agnes()`, ...
  - The package might be loaded in an R session: e.g. `library(MASS)`.

- Functions from other packages: `nlme::lme()`, `LabApplStat::DD()`, ...
  - The package must be installed once before it can be loaded and used: preferably done using the Install Packages button.
  - Ability to install packages is vital for the functionality of R.
  - Unfortunately, problems installing packages have become more prevalent (possible solutions: "Run as Administrator" + ask for help!)

T-tests

Systematic effects vs. Random variation

- **Systematic effects:** Mean properties of the population. Often the object of interest.
  - For instance the expected life span of men and women, or the difference between the effect of two drugs.

- **Random variation:** The dispersion of the data points around the systematic properties.
  - Natural variation in the population.
  - Measurement errors.
  - Difference between a complex world and a simple model.

- **Hypothesis testing:** Are the systematic effects significant, or can they be explained by the random variation?

Data example 3: Change in glucose level

One sample T-test

- For 8 diabetics the one-hour change in plasma glucose level after some glucose treatment was measured:

```
> change <- c(0.77,5.14,3.38,1.44,5.34,-0.55,-0.72,2.89)
> mean(change)
[1] 2.21125
> sd(change)
[1] 2.36287
```

- Did the treatment change the plasma glucose level?
- Basic statistical concepts: Statistical model, null hypothesis, test statistic, p-value, confidence interval.

Basic statistical concepts

Statistical model, null hypothesis, test statistic

- Models are described by parameters. In data example 1 these are the mean μ and the standard deviation σ. And we have the model:

$$Y_1, \ldots, Y_n \quad \text{i.id. } \mathcal{N}(\mu, \sigma^2)$$

- A statistical hypothesis is a simplifying statement about the model. Often formulated in terms of the parameters, e.g.

$$\text{Null hypothesis } H_0\text{: } \mu = 0, \qquad \text{Alternative } H_A\text{: } \mu \neq 0$$

- Test statistic T is a function of the data. Actual value denoted t_{obs}.
  - If T measures disagreement with H_0 and if t_{obs} is too extreme, then we reject H_0.
  - If the observed data is conceived as being random, then T becomes a random variable with a probability distribution.
  - Extremeness quantified by the p-value, calculated assuming H_0 is true,

$$p = P(T \text{ more extreme than } t_{\text{obs}})$$

One sample T-test

Testing in a normal sample: $Y_1, \ldots, Y_n$ i.id. $\mathcal{N}(\mu, \sigma^2)$

- Given prefixed value μ_0, often 0, we pose hypothesis H_0: $\mu = \mu_0$.

$$\text{Estimates:} \quad \widehat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \widehat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- Test statistic and p-value for one-sided test, H_A: $\mu > \mu_0$,

$$T = \frac{\overline{Y} - \mu_0}{s/\sqrt{n}} \sim T_{\mathrm{df}=n-1}, \qquad p = P(T_{\mathrm{df}=n-1} > t_{\mathrm{obs}})$$

- Test statistic and p-value for two-sided test, H_A: $\mu \neq \mu_0$,

$$T = \frac{\overline{Y} - \mu_0}{s/\sqrt{n}} \sim T_{\mathrm{df}=n-1}, \qquad p = P(|T_{\mathrm{df}=n-1}| > |t_{\mathrm{obs}}|)$$

Data example 3: Change in glucose level

One sample T-test

- For 8 diabetics the one-hour change in plasma glucose level after some glucose treatment was measured:

```
> change <- c(0.77,5.14,3.38,1.44,5.34,-0.55,-0.72,2.89)
> mean(change)
[1] 2.21125
> sd(change)
[1] 2.36287
```

- Did treatment change plasma glucose level in data example 3?

$$t_{\text{obs}} = 2.21 \cdot \sqrt{8}/2.36 = 2.65, \quad p = 2 \cdot P(T_{\text{df}=7} > 2.65) = 0.03$$

Quiz: Is this actually a paired T-test?

Data example 4: Phosphor concentration in lakes

Two independent samples, not necessarily of the same length

```
> lakes
# A tibble: 627 x 2
   location     phosphor
   <chr>           <dbl>
 1 East-Denmark    255
 2 East-Denmark    102.
 3 East-Denmark    166.
 4 East-Denmark     42.5
 5 East-Denmark    102.
 6 East-Denmark     60.6
 7 East-Denmark     89.8
 8 East-Denmark    182.
 9 East-Denmark    243.
10 East-Denmark     30.9
# ... with 617 more rows
```

- Is there a difference between East-Denmark (235 observations) and West-Denmark (392 observations)?
- Let's write up the statistical model and do the analysis using R.

Statistical analysis of two independent normal samples

Statistical model:

$$\text{First population} \sim \mathcal{N}(\mu_1, \sigma_1^2), \qquad \text{Second population} \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

Sequence of hypothesis [usually we skip (I) and simply use (IIb)]:

(I) H_0: $\sigma_1 = \sigma_2$

(II) H_0: $\mu_1 = \mu_2$

- (IIa) Assuming equal standard deviations $\sigma_1 = \sigma_2$.
- (IIb) Not assuming equal standard deviations.

Available statistical tests:

(I) `var.test()`, `bartlett.test()`, `lawstat::levene.test()`, `fligner.test()`, and many more.

- Don't do too many tests. Preferably only one test. Why?

(II) T-test, slightly different form in (IIa) and (IIb).

Assumptions and Checking for Normality

All T-tests, and other "normal" models

- **Assumptions**
  - The response variable (more precisely, the error terms) are normally distributed.
  - Possibly *homogeneity of variance* (homoscedasticity), meaning that the variance of the response variable is constant over the observed range of some other variable. This is (I) on slide 40.

- **Checking for Normality**
  - Visual inspection : `QQ-plot`.
  - Goodness-of-fit tests: `Shapiro-Wilks test`, `Kolmogorov-Smirnov test`, `Cramer-von Mises test`, `Anderson-Darling test`.

Shapiro-Wilks and Kolmogorov-Smirnov tests are available in `base R`, the others in the package `nortest`.

Transforming data

Often a solution when normality assumptions fails

- **Standard transformations (for $x > 0$):**
  - `log` transform: $x \mapsto \log(x) = y$.
  - Square root transformation: $x \mapsto \sqrt{x} = y$.
  - The inverse transformation: $x \mapsto \frac{1}{x} = y$.
    This transformation changes the order of the observations.
  - Box-Cox transformation with index λ:

$$x \mapsto y_\lambda = \begin{cases} \frac{x^\lambda - 1}{\lambda} & , \quad \lambda \neq 0 \\ \log(x) & , \quad \lambda = 0. \end{cases}$$

    Note the order of the observations is changed when $\lambda < 0$.
    Some particular cases:

$\lambda = -1$	$\lambda = 0$	$\lambda = 0.33$	$\lambda = 0.5$	$\lambda = 1$
Inverse	`log`	*cubic root*	*square root*	*no transformation*

- **Arcus sinus transformation (for $x \in [0, 1]$):**
  - $x \mapsto \arcsin(\sqrt{x})$.
  - May be appropriate when x measures the *proportion of successes out of a number of trials*.

Data example 4: Using R

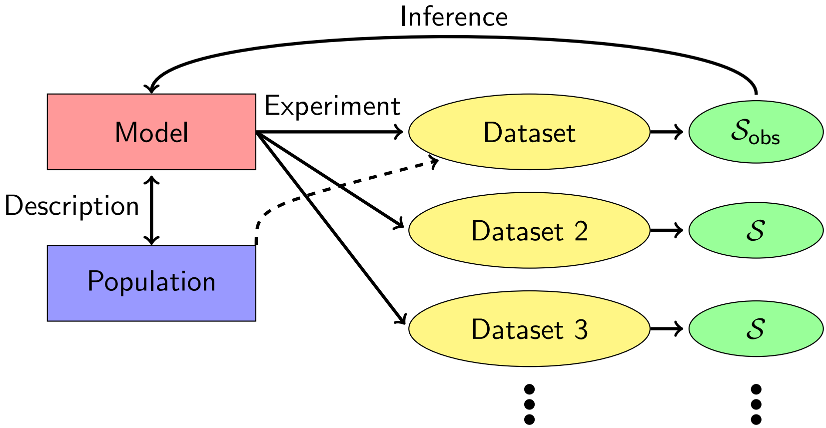
- Reading an Excel sheet.
- Validation of normality:
  - Graphical: `qqnorm(); abline(mean(),sd())`
  - Shapiro-Wilks test: `shapiro.test()`
  - Kolmogorov-Smirnov test: `ks.test(,"pnorm",mean(),sd())`

  Statisticians often prefer the graphical method.
- Data transformation.
- The actual two sample T-test.
- Keyboard shortcuts (Windows): `Ctrl-Enter`, `Ctrl-Shift-Enter`, `Ctrl-1`, `Ctrl-2`

Summary + Exercises

Today's summary: Model, data, statistic

Examples of statistics $\mathcal{S}$: estimator, confidence interval, test statistic, p-value



- What is the distribution of the p-value?
- Where do standard deviation and standard error reside?
- What is the interpretation of a confidence interval?

Homework

- Exercise class November 23 from 13.00 to 15.45.
  - Exercise sheets: `ex_day1.pdf`, `ggplot2_exercise.pdf`
  - Exercises not completed at class should be completed at home.

- Before the lectures on November 30 you should read the papers:
  - Sterne & Smith (2001), "Sifting the evidence—what's wrong with significance tests?", British Medical Journal, 226–231.
  - Gelman & Carlin (2014), "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors", Perspectives on Psychological Science, 1–11.