COMP320 Research Practice

04 - Experimentation and Hypotheses

An experiment can be characterized as a test (or a series of tests) wherein changes are introduced in the state of a system or process, enabling the observation and characterization of effects that can occur as a result of these changes.

Usually performed with an objective in mind:

- Uncovering influential variables in a given system or process;
- Determining desired values for certain parameters
- Characterize behavior of the system or process under study.

Experiments

Data gathering

- Retrospective study;
- Observational study;
- Designed experiment;

Characteristics

- Use of historical data;
- Investigating correlations;

Problems

- Data representativeness;
- Availability of data;

- Retrospective study;
- Observational study;
- Designed experiment;

Characteristics

- Observation of the system with minimal disturbance;
- Investigation of usual behaviors;

Problems

- Low representativeness of extreme cases;
- Low variability can affect observation of interesting effects;

Experiments

Data gathering

- Retrospective study;
- Observational study;
- Designed experiment;

Characteristics

- Introduction of deliberate changes in the system;
- Inference on the *causality* of the effects;

Problems

- Requires rigorous experimental design and data analysis;
- Usually more expensive.

Experimentation strategies

Educated guessing

- Select arbitrary combination of levels for the factors;
- Test and observe behavior;
- Change one or two factors at a time, then re-test;

- Widely used in industry;
- Can achieve good results, but has a lot of limitations;

Experimentation strategies

COST: Change One Separate factor at a Time

- Select a reference point;
- Change each factor individually, keeping all others constant;

- Also widely used;
- Can achieve good results as long as there are no interaction effects;

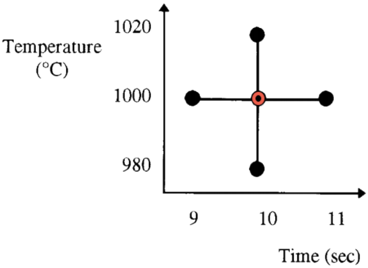


Image: (c) D.C. Montgomery

Experimentation strategies

Factorial designs

- Select **levels** for each factor;
- Vary the factors simultaneously, in a systematic way;

- Estimation of main effects and interactions;
- Greater precision in the effect estimates;
- More efficient use of resources (information/observation);

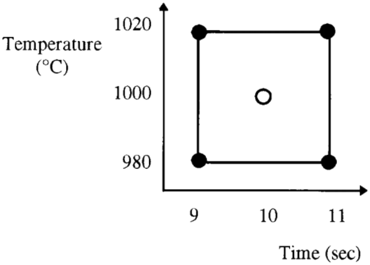


Image: (c) D.C. Montgomery

Process of designing data gathering protocols to enable accurate analyses by statistical tools, capable of supporting sound and objective conclusions.

- Applicable to systems and processes subject to noise, experimental errors, uncertainties, etc.
- Necessary for the conclusions to have a quantifiable meaning;
- Helpful in avoiding errors due to personal biases or other artifacts of experimentation and analysis.

Design of the experiment

- Scientific/technical question of interest;
- Selection of variables and values;
- Definition of the desired confidence level;
- Sample size calculations;
- Determination of protocols for data gathering;

Statistical analyses of the data

- Calculation of a test statistic;
- Validation of the assumptions of the statistical model;
- Calculation of the magnitude of effects;
- Drawing of conclusions and recommendations;

Fundamental principles

Design of experiments (DoE)

- Repetition and replication;
- Randomization;
- Blocking.

- Repeated measurements - estimation of within-group variability;
- Replication - estimative of the experimental error;
- Greater precision in estimating the model parameters;

- Repetition and replication;
- Randomization;
- Blocking;

- Avoids contamination of the data by order-dependent effects such as:
  - Heating effects;
  - Wear and tear effects;
  - External interferences;

Fundamental principles

Design of experiments (DoE)

- Repetition and replication;
- Randomization;
- Blocking;

- Isolation of nuisance variables (those that influence the response, but are not interesting for the analyses) that can be controlled;
- Improvement in the estimation of effects for the factors of interest;
- Reduction or eliminations of inconvenient factor effects;

Fundamental principles

The role of experimental design

Experimental design is useful for avoiding the influence of spurious factors and personal biases on the results, by performing experiments in a impartial and objective way.

> "*Never have too much love for your hypotheses.*"

> "*The great tragedy of Science - the slaying of a beautiful hypothesis by an ugly fact.*"
> – Thomas H. Huxley

Image: `http://www.iep.utm.edu/huxley/`

Discussion

Jacques Benveniste and the memory of water

- Nature (1988);
- Investigation committee: Maddox, Stewart, Randi;
- Retracted by Nature due to evidence of misconduct.

Methodological problems

- Experimenter bias (absence of proper blinding);
- Cherrypicking (selective recording of results);
- Unaccounted sampling errors;
- Possible contamination;
- Complete lack of prior physical/ chemical plausibility;
- **Non-reproducibility**.

Image: `http://www.jacques-benveniste.org/inmemoriam/photo1.html`

- Pre-experimental design:
  - Identification and definition of the problem;
  - Selection of experimental and response variables of interest;
  - Choice of experimental protocols;
- Choice of the experimental design;
- Collection of the data;
- Statistical data analyses;
- Conclusions and recommendations;

Pre-experimental design

Before we start

- Is the investigation relevant?
- Would the results be interesting for the research community?
- Practical relevance?
  - Employ exploratory experiments;
- Placement within the literature;
  - Avoid repetition and irrelevance.

- The translation *scientific question* $\rightarrow$ *test hypothesis* requires special attention, and a solid knowledge of the technical area in which the experiment is being performed;

Actual Experiment

Data gathering

- Must be consistent with design, otherwise the validity of the results may be compromised - data collection must always follow the plan:
  - No premature stops;
  - *No-peeking rule* (except when planned, of course);
- Use of pilot experiments:
  - Gathering of preliminary information;
  - Practice with the experimental conditions;

Analysis of the experimental data

Statistical modeling

- General procedure for testing the experimental hypotheses:
  - Definition of a *null-model* (absence of effects) and of a desired level of significance;
  - Determination of $P(\textit{data}|\textit{null-model})$;
  - Decision by rejection (or not) of the null hypothesis;
  - Validation of model assumptions;
  - Estimation of the *magnitude* of differences - **practical significance**;

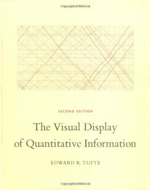
> *Statistical methods do not* ***prove*** *anything, but they allow an objective definition of margins of plausibility for certain statements.*

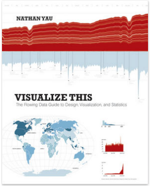
Reporting of results

Presentation

Combine textual, numeric and graphical elements to tell a story with your data. It simplifies the understanding and analysis of the results.

- Strive to achieve graphical excellence;
- Coherence of notation - special attention to figures and tables;
- Display simultaneous confidence intervals and other graphical indicators of effect size.





Other great resources on graphical excellence:

Flowing Data (`http://flowingdata.com/`)

Information is Beautiful (`http://www.informationisbeautiful.net`)

Conclusions

Drawing and reporting conclusions

- Conclusions should be based on solid evidence from the data;
- Be conservative - it is common to exaggerate the generality of the results;
- Report significance levels and the assumptions under which the results are valid;
- *Suggest explanations* to the observed results;
- Careful with *anomaly hunting*;

> *Always let the science drive the statistics. If you get a statistically significant result, go back and describe what it means in the scientific context.*
>
> – Aaron Rendahl

Bibliography

1. Dept. Biochemistry & Cell Biology, Rice University , *Common Errors in Student Research Papers.* - http://www.ruf.rice.edu/~bioslabs/tools/report/reporterror.html
2. T. Brady, *Reviewer's quick guide to common statistical errors in scientific papers.* https://goo.gl/IKvAYc
3. R.M. Szydloa et al., *Sign of the Zodiac as a predictor of survival for recipients of an allogeneic stem cell transplant for chronic myeloid leukaemia (CML): an artificial association.* Transplantation Proceedings 42 (8):3312-3315, 2010.
4. D.C. Montgomery, *Design and Analysis of Experiments*, Chapter 1. 5th ed., Wiley, 2005
5. S.C. Stearns, *Some Modest Advice for Graduate Students* - http://goo.gl/jtT4hA
6. A. Rendahl, *Experimental design and statistical analysis: questions to consider as you write your grant.* - http://goo.gl/yT0TK7
7. J. Maddox et al., *"High-dilution" experiments a delusion.* Nature 334, 287-290, 1988
8. V. Czitron, *One-Factor-at-a-Time Versus Designed Experiments.* The American Statistician, 53(2) 126-131, 1999.
9. B. Dunning, *An Enthusiast's Primer on Study Types.* Skeptoid Podcast, Skeptoid Media, Inc., 2013. - http://skeptoid.com/episodes/4381

A *hypothesis* is a proposed explanation for an observable phenomenon.

Scientific hypotheses must satisfy (at least) two conditions:

- Falsifiability;
- Testability;

"*The more we learn about the world, and the deeper our learning, the more conscious, specific, and articulate will be our knowledge of what we do not know, our knowledge of our ignorance.*"

Sir Karl R. Popper
(1902-1994)
Austro-British philosopher

The *hypothetico-deductive model* of construction of scientific knowledge includes:

- Formulation of falsifiable hypotheses;
- Refutation or corroboration of the hypotheses by the data;
- Comparison between alternative hypotheses - principle of parsimony (Ockham's razor);
- Predictive power;

"*Numquam ponenda est pluralitas sine necessitate.*"

William of Ockham

(1287-1347)

English philosopher and theologian

Image: http://www.philosophybasics.com/philosophers_ockham.html

Statistical Hypotheses

Definitions

Statistical hypotheses are defined as objective statements about parameters of one ore more populations;

Attention: the statements in statistical hypotheses are about parameters of the *population* or *model*, **not the *sample***.

On frequentist approaches, the formal test of hypotheses involves the contrast between *null* and *alternative* hypotheses.

Null hypothesis (H_0)

- Absence of effects;
- *Conservative* model.

Example: $H_0 : \mu = 25$

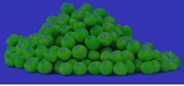
Alternative hypothesis (H_1)

- Presence of some effect;
- Existence of something "new".

Example: $H_1 : \mu \neq 25$

Statistical Hypotheses

Example



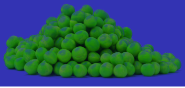
Suppose you own a company that sells green peas to large customers, and that you want to determine whether your 50kg sacks really contain their nominal weight (at least on average).

In this case the null hypothesis could be defined as: *the average net weight of a sack is* $50kg$, and the alternative of interest could be expressed as the complementary inequality.

$$\begin{cases} H_0 : \mu = 50kg \\ H_1 : \mu \neq 50kg \end{cases}$$

Suppose still that $n = 10$ packs are randomly sampled, and their contents are weighted using a calibrated scale;

Image: `http://www.storko.eu/ed_files/image/green-peas.jpg`

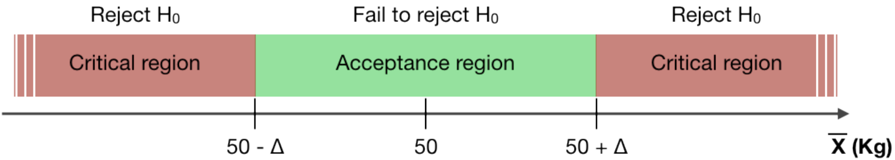


Since the sample mean $\bar{x}$ is a good estimator of the real mean μ, common sense suggests that:

- If $\bar{x} \cong 50$kg - corroboration of H_0;
- If $\bar{x} \ll 50$kg or $\bar{x} \gg 50$kg - refutation of H_0;

That is, we can use $\bar{x}$ as the basis for a statistical test.

But how to define a *critical region* for the rejection of H_0?



Type I error (false positive): rejecting the null hypothesis when it is true.

The probability of occurrence of a false positive in any hypothesis testing procedure is generally known as the *significance level* of the test, represented by Greek letter α:

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true})$$

Another frequently used term is the *confidence level* of the test, given by $(1 - \alpha)$.

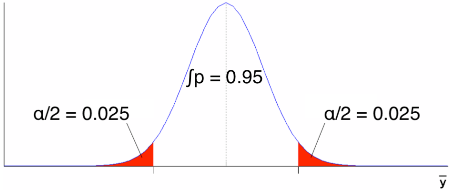
Inferential Errors

Type I error

For a given sample, the selected value of α defines the critical threshold for the rejection of H_0.

If H_0 is true (i.e., if $\mu = 50$kg), the distribution of values of $\bar{x}$ is aproximately normal (assuming the Central Limit Theorem holds), with average 50kg and standard error $(\sigma/\sqrt{n})$ kg;

For a desired Type-I error probability $\alpha = 0.05$, the critical values of the distribution of $\bar{x}$ are the ones for which the probability content within the acceptance region under the null hypothesis is $1 - \alpha = 0.95$.



Inferential Errors

Type II error

Type II error (false negative): failure to reject the null hypothesis when it is false.

The probability of occurrence of a false negative in any hypothesis testing procedure is generally represented by the Greek letter β:

$$\beta = P\,(\text{type II error}) = P\,(\text{not reject } H_0 | H_0 \text{ is false})$$

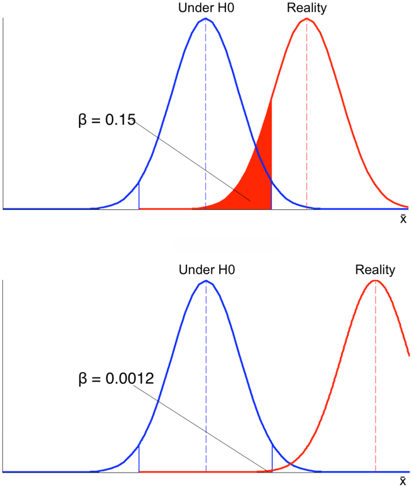
The quantity $(1 - \beta)$ is known as *power* of the test, and quantifies its sensitivity to effects that violate the null hypothesis.

Inferential Errors

Type II error

Unlike the Type-I error, the definition of the Type-II error rate requires further specification of the value of the parameter being investigated under the alternative hypothesis;

The probability of failing to reject a false H_0 is strongly dependent on the magnitude of the difference between the value under H_0 and the real value of the parameter.



Inferential Errors

Type II error

The power of a test is governed by several factors:

- Controllable: significance level, sample size, directionality of H_1;
- Uncontrollable: real value of the parameter, variance;

If H_0 is false, the smaller the magnitude of the difference between the real value of the parameter and the one under the null hypothesis, the greater the probability of a type II error - ***but the practical importance of the effect gets smaller***.

Type I error (α) depends only on the distribution of the null hypothesis - easier to control;

Type II error (β) depends on the real value of the parameter - more difficult to specify and control;

These characteristics lead to the following classification of the conclusions obtained from the test of hypotheses:

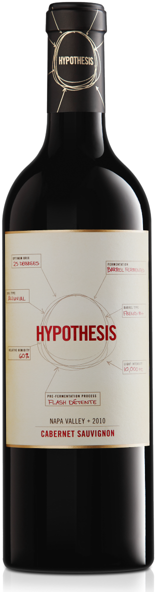
- Rejection of H_0 - *strong* conclusion;
- Failure to reject H_0 - *weak* conclusion (but we can fortify it);

It is important to remember that failing to reject H_0 does not mean that there is evidence in favor of H_0 - it only suggests that it is a better model than the alternative.

Hypothesis Testing

General procedure

- Identify the parameter of interest;
- Define H_0 and H_1 (one- or two-sided);
- Determine desired α, β;
- Define minimally interesting effect δ^*;
- Calculate sample size;
- Determine the test statistic and critical region;
- Compute the statistic;
- Decide whether or not to reject H_0;



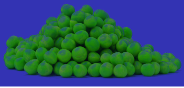
Back to the green peas example, we want to determine if there is any significant deviation on the mean weight of the sacks. Assume (for now) that the variance of the process is known. The test hypotheses are defined as:

$$\begin{cases} H_0 : \mu = 50kg \\ H_1 : \mu \neq 50kg \end{cases}$$

Let the desired significance level be $\alpha = 0.05$;

Hypothesis Testing

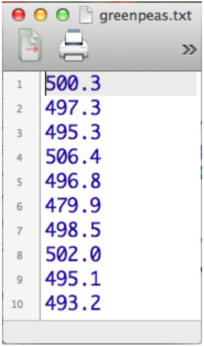
Mean of a normal distribution, variance unknown



```
> sample <- as.numeric(scan("../data files/greenpeas.txt"))

> t.test(sample,
+        alternative = "less",
+        mu = 50,
+        conf.level = 0.99)

One Sample t-test
data:  sample
t = -1.5969, df = 9, p-value =0.07237
alternative hypothesis: true mean is less than 50
99 percent confidence interval:
-Inf 50.2699
sample estimates:
mean of x
49.648
```


greenpeas.txt

```
500.3
497.3
495.3
506.4
496.8
479.9
498.5
502.0
495.1
493.2
```

Description of the results:

(In)Sufficient evidence for rejecting H_0
at the significance level α.

Even thought it is correct, this description is relatively poor:

- It does not provide information on the intensity of the evidence for rejection/non-rejection;
- It imposes a predetermined significance level to the consumer of the information;
- Does not provide information the magnitude of the effect found or the sensitivity of the test.

p-value: *the lowest significance level that would lead to the rejection of H_0 for the available data.*

Can be interpreted as the probability under H_0 of the test statistic assuming a value at least as extreme as the one obtained;

For the previous example, R gives the p-value as 0.07237. Interpretation:

$$p = P(\bar{x} \leq 49.648 | H_0 = \textbf{TRUE}) = 0.07237$$

A priori definition of the significance level is still important!

Hypothesis Testing

Beware *p*-value fishing

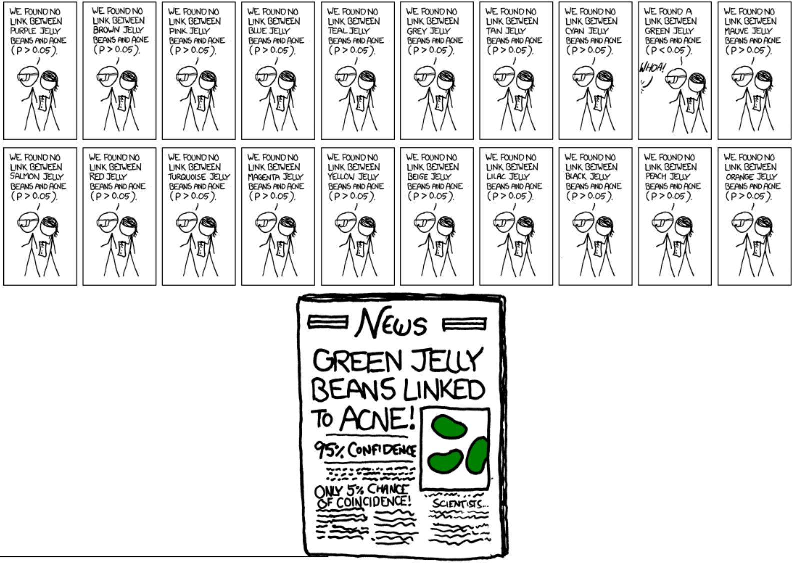


Image adapted from `https://xkcd.com/882/`

Statistical $\times$ practical significance: p-values can be made arbitrarily small, if n is big enough;

As an example, suppose a test of $H_0 : \mu = 500$ against a two-sided alternative, with $n = 5000, \bar{x} = 499,\ s = 5$. In this case we would have:

- $t_0 = -14.142$;
- $p = 1.02 \times 10^{-23}$;

Is it really *that* significant?

To “tell the whole story” of the experiment, it is necessary to use **effect size estimators** alongside the tests of statistical significance;

While there are whole books on the subject[c], the main idea is quite simple - to quantify the magnitude of the observed deviation from the null hypothesis.

Examples of effect size estimators include the simple point estimator for the difference $\bar{x} - \mu_0$, or the dimensionless d estimator:

$$d = \frac{\bar{x} - \mu_0}{s}$$

which quantifies the difference in terms of sample standard deviations.

[c] See, for instance, Paul D. Ellis' *The Essential Guide to Effect Sizes*, Cambridge University Press, 2010.

Hypothesis Testing

p-values, effects sizes and confidence intervals

Point estimators + confidence intervals quantify the magnitude and accuracy of effects, and must be reported alongside the results of significance testing whenever possible.

Suppose we are testing $H_0 : \mu = 50$ against the two-sided alternative hypothesis, with $n = 10$ and $\alpha = 0.01$. Assume that the population is known to be normal, with unknown variance. We'll use the same data as before:

```
> t.test(sample, mu = 50, conf.level = 0.99)
(...)
t = -1.5969, df = 9, p-value = 0.1447
alternative hypothesis: true mean is not equal to 50
99 percent confidence interval:
 48.93166 50.36434
sample estimates:
mean of x
   49.648
```

Sample size and Type-II error

Some considerations

The probability of Type-II error can be easily (and often wrongly) evaluated *a posteriori*, but its definition *a priori* requires some care;

Given a desired test, its power is essentially a function of 4 elements:

- Actual size of the difference;
- Variability of the observations;
- Significance level;
- Sample size.

The experimenter generally have very little control over the first two.

Sample size and Type-II error

Some considerations

A strategy for estimating an effective lower bound for the power of a test includes a definition of an *minimally interesting effect* δ^*.

This value must be derived from technical and scientific knowledge about the phenomenon or system under experimentation.

> It is essential to have a good understanding of the field in which the experiment will be conducted.

Once δ^* is defined, the experimenter can obtain an estimate of the variability of observations (e.g., by a pilot study), which can then be used to obtain an approximate power value for the experiment;

Sample size and Type-II error

Some considerations

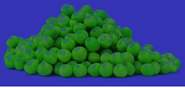
Having obtained this estimation of the Type-II error probability, one can run his/her experiment with a better understanding of its ability to detect effects of interest.

The test will have lower power for differences smaller than δ^*, but these differences are below the minimally interesting effect; any effect greater than δ^* will result in a higher power for the test;

This technique can also be used as a way to compute the maximum necessary sample size for the experiment.

Sample size and Type-II error

Example



Suppose that on the green peas example we are really interested in detecting negative deviations from the nominal value greater than 1%, i.e., $\delta^* = 0.01 \times 50 = 0.5$kg. The researcher defines that, for this minimally interesting effect, a test power of 0.85 is desired. The desired significance is $\alpha = 0.01$.

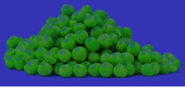
The same sample of $n = 10$ sacks is used. Assume that you estimated a reasonable standard deviation for this sample as $s = 1$kg. From this data, we can compute the power of this test as:

```
> s <- sd(sample)
> power.t.test(n = 10,
+         delta = 0.5,
+         sd = 1,
+         sig.level = 0.01,
+         type = "one.sample",
+         alternative = "one.sided")
```

```
One-sample t test power calculation
n = 10
delta = 0.5
sd = 1
sig.level = 0.01
power = 0.1654013
alternative = one.sided
```

Sample size and Type-II error

Example



What is the smallest sample size needed to obtain the desired power of 0.85?

```
> power.t.test(power = 0.85, delta = 0.5, sd = 1, sig.level = 0.01,
                type = "one.sample", alternative = "one.sided")

One-sample t test power calculation
n = 47.98044 ⟵ (round this value up)
delta = 0.5
sd = 1
sig.level = 0.01
power = 0.85
alternative = one.sided
```

We need at least 48 observations to detect a $-5g$ (1%) or larger deviation on the mean weight of the green peas packages with a power level of 0.85.

Bibliography

1. D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, Ch. 9. 5th ed., Wiley, 2010.
2. J.T. Mordkoff (2011), *The assumption(s) of normality* - `http://goo.gl/Z3w8ku`
3. A. Reinhart, *Statistics Done Wrong: the woefully complete guide* - `http://www.statisticsdonewrong.com`
4. W. Thalheimer and S. Cook, *How to calculate effect sizes from published research articles: A simplified methodology* - `http://goo.gl/c0g1oK`
5. E. Ernst and S. Singh, *Trick or Treatment*, Norton & Company, 2009.

About this material

Conditions of use and referencing

Please reference this work as:

Felipe Campelo (2015), *Lecture Notes on Design and Analysis of Experiments*. Online: https://github.com/fcampelo/Design-and-Analysis-of-Experiments Version 2.11, Chapter 5; Creative Commons BY-NC-SA 4.0.

```
@Misc{Campelo2015-01,
  title={Lecture Notes on Design and Analysis of Experiments},
  author={Felipe Campelo},
  howPublished={\url{https://github.com/fcampelo/Design-and-Analysis-of-Experiments}},
  year={2015},
  note={Version 2.11, Chapter 5; Creative Commons BY-NC-SA 4.0.},
}
```