S 03

Data Analytics for Data Scientists

Design of Experiments (DoE)

Suggested solutions for Exercise 03: Introduction to Design of Experiments (DoE)

2025

Prof. Dr. Jürg Schwarz
MSc Adrian Bieri

Suggested solution 01

Quality of an Instrument

A study on sustainable development focuses on measuring climate data at a specific location. Among other things, the daytime temperature on a winter day is recorded over a period of 24 hours. See the figure below.

Note: The figure shows an ideal-typical daily course.

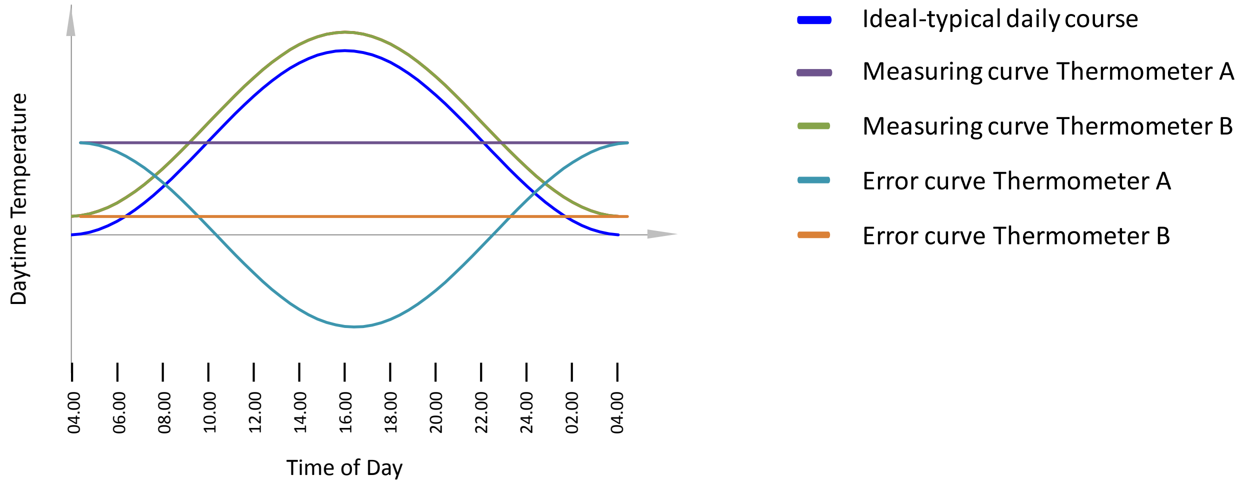


Figure Measurements and deviations are entered

Two independent research teams, A and B, conduct temperature measurements.

…

Suggested answers to the questions[1]

Team	Error Type	Validity	Reliability	Implications
Team A[2]	Random error (stuck at 6 °C)	Low	Low	Data does not reflect real behavior and cannot be corrected.
Team B	Systematic error (always +2° C)	Low	High	Systematic error can be corrected post-hoc, making the data usable.

Conclusion for the use of thermometers

- Team A: Thermometer is fundamentally flawed, as it fails to capture meaningful temperature variations.
- Team B: Thermometer is preferable, as its systematic error can be adjusted in analysis (e.g., subtracting 2° C from all measurements).

[1] See also the figure above, which was completed with the measurements and the deviations.

[2] It could also be argued that the setting of team A cannot be assessed in the same way as for team B because no adequate measurement is available. This means that no statement can be made regarding validity and reliability.

Suggested solution 02

Maximizing / Controlling / Minimizing

Show how variance can be maximized, controlled and minimized in the following descriptions of studies.

First determine the relevant variables:

- Dependent variable (**DV**)
- Central independent variable(s) **(IV)**
- Nuisance variables

Suggested answers to the questions

Description of the study – Summary

Research question: What title for a quarterly newsletter to existing customers (who have made at least one purchase) will increase the open rate?

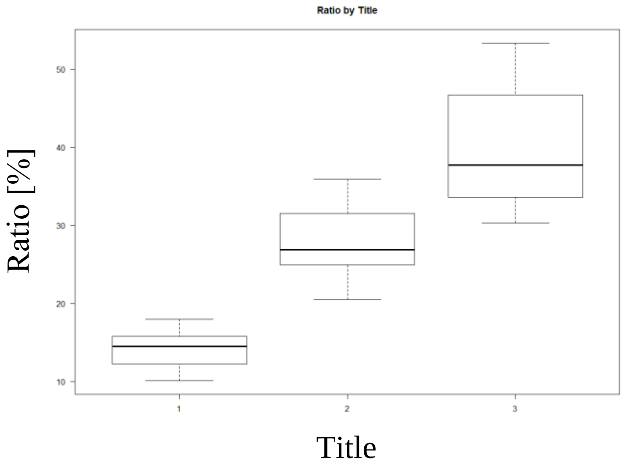
There are three title variations: The current title, a title that announces a competition, and a title that announces a gift. The mailing list has approx. 12,000 customer names.

- Dependent variable (**DV**) → name of variable open rate
  Rate at which readers open the newsletter = number opened / number sent
- Central independent variable(s) **(IV)** → name of variable title
  Selection of titles with three options: Current = 1, competition = 2, gift = 3
- Nuisance variables
  Many (e.g., hours a customer works, because busy people are less likely to open a newsletter email; existence of a spam filter; etc.)

Maximize primary variance

Note: The research design prescribes three title options that cannot be changed.

→ **The three options of title as the central independent variable "automatically" maximize the primary variance as much as possible because they are prescribed.**



Control the secondary variance

Sources of secondary variance include all the characteristics of the recipients of the newsletters. The characteristics are mostly unknown. For this reason, randomization remains one of the few options.

→ **The elements of the three groups are drawn randomly from the mailing list.**

Minimize the error variance

The measurement has no error variance in the strict sense (reading errors or fluctuations).

However, there are a large number of disturbations in the decision whether an email is read.

For this reason, maximize the sample size remains one of the few options.

→ **It should be possible to use as large a subset as possible from the mailing list with around 12,000 names.**
But, reduced interest due to possible "fatigue" (= too many newsletters) is to be expected.

Description of the study – Summary

Research question: Can tuberculosis be cured by treating patients with the antibiotic streptomycin?

Two groups of patients are planned: Treated (Streptomycin) vs. Control (No treatment).
The total sample size includes 100 patients

The example comes from a study[3] that Hill published in 1948.
→ See also "Exercise 01: Introduction"

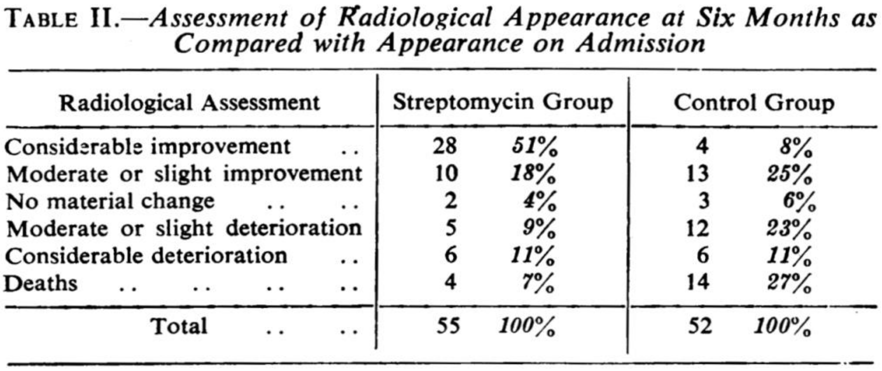


TABLE II.—*Assessment of Radiological Appearance at Six Months as Compared with Appearance on Admission*

Radiological Assessment	Streptomycin Group		Control Group	
Considerable improvement . .	28	*51%*	4	*8%*
Moderate or slight improvement	10	*18%*	13	*25%*
No material change	2	*4%*	3	*6%*
Moderate or slight deterioration	5	*9%*	12	*23%*
Considerable deterioration . .	6	*11%*	6	*11%*
Deaths	4	*7%*	14	*27%*
Total	55	*100%*	52	*100%*

- Dependent variable (**DV**)
  Radiological assessment with 6 options
  ("Considerable improvement", ...)
- Central independent variable(s) (**IV**)
  Vaccination with streptomycin, with 2 options: Treatment (Yes, No)
- Nuisance variables
  Many (e.g., existence of another disease which could influence the effectiveness of the drug; age of patient; other medication of the patient, etc.)

Maximize primary variance

→ **The two manifestations of the central independent variable maximize "automatically" the primary variance as much as possible, since they are predetermined.**

Control secondary variance and minimize error variance

Sources of secondary variance are all possible characteristics of patients. The characteristics are mostly unknown. Therefore, patients are **randomly** assigned to the *treated (Treatment*) and non-treated *(Control*) groups. A major source of secondary variance would be the placebo effect, which is reduced by **blinding** the patients and medical staff.

There are many sources of error variance, but the maximization of sample size in a clinical study is limited. However, the RCT design reduces part of the error variance.

→ **RCT with blinding is the "gold standard" for controlling secondary variance.**

[3] Medical Research Council (1948): Streptomycin Treatment of Pulmonary Tuberculosis. In: BMJ 2 (4582), p. 769. DOI: 10.1136/bmj.2.4582.769.