Universität Zürich[UZH]

Prof. Dr. Reinhard Furrer
Department Mathematical Modeling and Machine Learning (DM^{3}L)
Applied Statistics group
https://dm3l.uzh.ch/research

Dr. Bernadetta Tarigan
Office for Animal Welfare and 3Rs
Biostatistics Consulting
www.tierschutz.uzh.ch

Contact: www.math.uzh.ch/animalwelfare

Planning Experiments and Sample Size Estimations

B. Tarigan

Compiled on May 20, 2025 (v3.1)

This document is a companion to our statistical methods consultation service, mainly –but not limited– for animal research. Thus, the described methods/approaches here can be used and generalized for clinical research and all other fields of research.

Contents

1 Introduction **2**

2 Planning an experiment **3**

3 Sample size determination **4**

4 Continuous outcome **4**
4.1 Confirmatory 4
4.2 Exploratory 7

5 Categorical outcome **9**
5.1 Confirmatory 9
5.2 Exploratory 11

6 Time-to-event outcome **11**

7 Multiple-testing problem **12**

8 Crude bounds for standard deviation **13**

9 Sex as biological variable (SABV) **14**

1 Introduction

In statistics we distinguish two type of study: experimental study versus observational study. An **experimental study** is a study in which the experimenter introduces some *intervention* group(s) and investigates the varying effects of these intervention(s) and typically with comparison to a control group on the response variable (outcome/endpoint). As opposed to experimental study, in **observational study** the researcher *merely observes* and records what she sees, thus there is no intervention.

The purpose of an observational study is to investigate an *association*, while the purpose of an experimental study is to investigate a *causal-effect* relationship between the treatment factor and the outcome. Therefor, the environment of an experiment has to be supervised/manipulated/controlled, keeping all the subjects (aka experimental units) similar except the interventions.

Planning an experiment is an exciting process. It is also extremely time-consuming that creates a temptation to begin collecting data without giving the experimental design sufficient thought (see Dean *et al.*, 2017, Chapter 2). However, this approach will rarely yield data that is collected in the right way and in sufficient quantity to allow a good analysis with the required precision. In experimentation, careful planning and execution of the plan are the most important steps. If anything can go wrong then it will (Murphy's Law), and analysis of data can never compensate for botched experiments. Following van Belle, 2008, Chapter 1, we must address the four basic questions:

1. What is the questions?
2. Can it be measured?
3. Where, when, and how will you get the data?
4. What do you think the data are telling you?

Statistical design of experiments refers to the process of planning the experiment so that appropriate data can be collected and analyzed by statistical methods, resulting in valid and objective conclusion. The Reduction principle in 3R means *appropriately designed and analysed animal experiments that are robust and reproducible, and add to the knowledge base.*

An experimental study can be distinguished further into two general goals: *to learn something new* (which is called **exploratory** experiment) or *to confirm a specific hypothesis* (which is called **confirmatory** experiment). The distinction between the two is important because it affects the decisions made about the design, the subsequent analysis and the way sample size is calculated (see Lazic, 2016, Chapter 2).

The following examples are respectively research question and research hypothesis for exploratory experiment and confirmatory experiment.

- Does stress affect rodent behavior? What kind of stress, for how long, on what behavioral tasks?
- Does fox urine odor (a stressor) affect the amount of food Wistar rats consume during the first 24 hours after exposure?

In the first question we want to learn something new, while the second one is to test a specific hypothesis.

A typical experiment contains an "error". This so-called experimental error is not a mistake, but a variability/uncertainty that is inherent to any stochastic process such as an experiment. There are **two sources of an experimental error**: random error and systematic error (aka bias). Large random error will lead to large imprecision and large systematic error will lead to large inaccuracy. The random error is almost unavoidable, but it can be improved by increasing the number of **replications** of the experimental unit, that is the sample size. The systematic error can be eliminated by **randomization, blinding** and **stratifying** when possible.

Some glossary:

- *Experimental unit (EU)* is the smallest experimental material that is subjected to the treatment groups.
- *Randomization (aka allocation)* is a process by which each EU has the same chance of being assigned to any of the treatment groups.

- *Blinding* is keeping the identity of the treatment assignment from experimenters administering treatments, experimenters assessing outcome, caregiver, animal facilities staff and others who interact with the animals, whenever possible.
- *Stratifying* is the arranging of experimental units in blocks that are similar/homogeneous to one another in order to reduce variability, but in itself it is not of primary interest to the experimenter.

More details can be found, for example, in Lazic, 2016, Chapter 3 and Bate and Clark, 2014, Chapter 3.

2 Planning an experiment

Essentially every study has a main question/objective. Planning an experiment obviously starts with formulating the **objective of the study** in a clear and compact way. The objective can be formulated with an help of PICO framework as a mnemonic. It consist of four elements: *population* of interest which is the experimental unit (P), the *intervention* groups (I), the *control* group (C) and the primary *outcome* or the variable of interest (O).

As an example, consider the following objective of a confirmatory experiment. "To confirm effect of diet-induced obesity by comparing effect of three types of diet (high-fat, low-fat and standard) administered by food in a cage on weight gain (average gr/week) of male B6J mice."

Here the population is the male B6J mice, the intervention groups are the high-fat diet and the low-fat diet, the control group is the standard diet and the outcome is the average weekly weight gain in gram. Moreover, the **experimental unit** is the cage.

The second step in planning an experiment is **describing the methods and procedures used to collect and analyze data** on the (primary) outcome specified in the objective of the study.

In the example above, the experimental unit is cage. To minimize the random error, the cages are **completely randomized** into the three treatment groups and each group has the same number of cages (balanced complete randomization). To minimize the systematic/subjective error, the investigators will be **blinded** –they are not aware which group a cage belongs to– from the start of the experiment until the data have been analysed.

At the end of the experiment, we measure the average weight gain in gr/week for each mouse in the cage and take the average over the number of mice in a cage as the primary outcome. This primary outcome is a continuous variable. We then compare the means/averages of the three groups. Thus, the **appropriate statistical model** to analyze the data is oneway ANOVA model. That is, we use F test with **research hypothesis** that at least two groups have different means and **null hypothesis** that the means of the three groups are the same.

In statistics design refers to the randomization technique of the experimental unit. In the example above, the cage was randomized completely –at once– into the treatment groups. Thus, this design is called a Completely Randomized (CR) design with one treatment factor (the diet). The CR is a family of designs, that extend to more than one treatment factors. It is also called *factorial design* family. Apart from CR, there are other families of designs. For example, Block designs and Split-Plot designs. In this note, we focus on CR designs.

The following keywords are a simplification of a very large number of decisions that need to be made for planning an experiment:

1. study objective, source of variability and type of the experiment,
2. randomization,
3. blinding,
4. sample size determination,
5. in- and exclusion criteria,
6. type of the primary response/outcome/endpoint variable,
7. statistical method/model/test and multiple testing.

This list can be seen as a simplified sequence of decisions, or equivalently questions to answer, in planning an experiment, following Section 22 and Section 29 of Form A (Application for licence to perform animal experiments in Switzerland).

See Wilson *et al.* (2023) for a good reading in designing, conducting, and reporting reproducible animal experiments and Lazic, 2016, Chapter 5 for planning for success.

3 Sample size determination

One of the first questions facing an experimenter is "How many occurrences should I take?", or alternatively "Given my limited budget, how can I obtain as much information as possible?". These questions cannot be answered in a couple of sentences. Sample size determinations may be rather complex, and sometimes requiring advanced knowledge and experience.

As indicated in the previous section, sample size determination depends on:

1. the type of the primary outcome variable: either continuous, or categorical, or time-to-event,
2. the type of the study: either confirmatory experiment, or exploratory experiment (including pilot), or observational.

The next three sections describe further the determination based on the type of the primary outcome variable: continuous, categorical and time-to-event. Each section describes further the type of experiment: confirmatory and exploratory (including pilot study), and we give some examples. The power calculation of sample size of the examples can also be carried-out using our shiny R interactive application SampleSizeR (see Astakhov *et al.*, 2025).

For an observational study, typically one would aim as many as as possible for the sample size.

4 Continuous outcome

Continuous outcome is a quantitative measurement. Examples of continuous outcomes are weight gain (gr/week), cholesterol level (mg/dL, milligrams of cholesterol per deciliter of blood), blood flow rate (L/min, liter per minute), etc.

4.1 Confirmatory

Recall that a confirmatory experiment is to test a specific statistical hypothesis. Thus, there is an appropriate statistical test. Sample size determination for confirmatory experiment is then based on **power-to-achieve of the corresponding test** (aka power calculation). Power is the probability of detecting an effect given there is an effect. Thus, we aim for a large power, which is typically not smaller than 0.80. Power depends on:

1. the significant level (type I error) with correction for multiple testing,
2. the alternative hypothesis,
3. the minimum effect size (aka signal) to detect,
4. the estimation of the variability of the population (aka noise) which is measured with standard deviation,
5. the sample size.

The relationship of these elements –see Figure 1 below– is an explicit mathematical equation, based on the appropriate statistical test, that we solve numerically with a statistical software.

More details are in Lazic, 2016, Section 5.2 and Bate and Clark, 2014, Section 3.7.2.

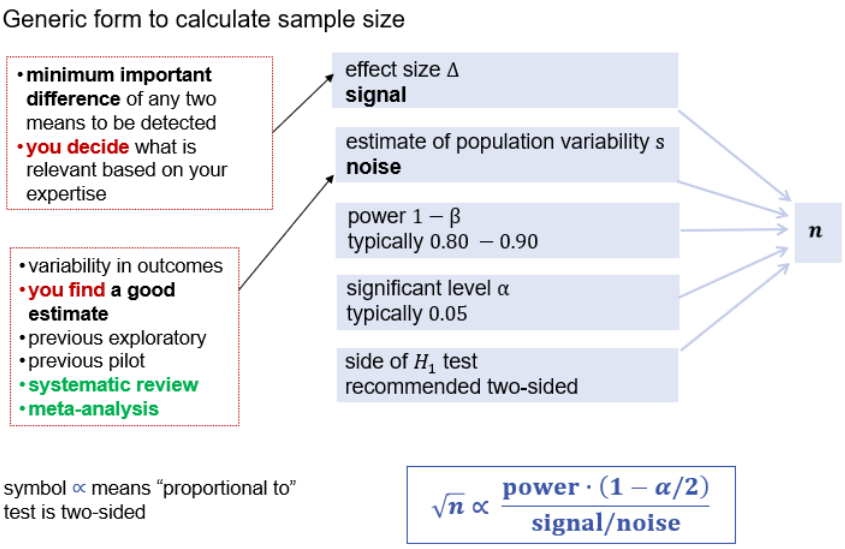

Figure 1: The elements of power

Example 1

We consider the example in Section 2, a CR design with one treatment factor, and recall its objective.

Objective
To confirm an effect of diet-induced obesity administered by food in a cage on weight gain (average gr/week) of male B6J mice, comparing three types of diet: high-fat, low-fat and standard.

Statistical test/model
Our research hypothesis is that at least one pair out of the three levels of the treatment factor is different, while the null hypothesis is that there is no difference. Thus, we use F test for one factor (aka oneway ANOVA) where the factor has 3 levels/groups. We are interested in comparing all the three possible pairs for the groups.

Power calculation of sample size
We consider a difference of 4 gr/week of average weight gain among the three groups as meaningful. Previous study shows that the population standard deviation is estimated with 3 gr/week in each group. To count for the three **multiple tests** we use Bonferroni correction, i.e., taking $\alpha = 0.05/3 = 0.0167$ for calculating the sample size (see section Multiple-testing problem). We use the R package *daewr* (see Lawson, 2015, Chapter 2) as below.

```
> library(daewr)
> round(Fpower1(alpha = 0.01667, nlev = 3, nreps = 13:16, Delta = 4, sigma = 3), 3)

     alpha nlev nreps Delta sigma power
[1,] 0.017    3    13     4     3 0.697
[2,] 0.017    3    14     4     3 0.741
[3,] 0.017    3    15     4     3 0.781
[4,] 0.017    3    16     4     3 0.815
```

Thus, 16 cages per group is needed to obtain a power of at least 0.80 with family-wise significant level of 0.05, resulting at $3 \cdot 16 = 48$ cages in total. Since we use male mice, where a cage contains one male mouse, the total number of animals needed in this study is $3 \cdot 16 = 48$ male B6J mice. If we were to use female mice with 3 mice per cage, then the total number of animals we would need is $3 \cdot 3 \cdot 16 = 144$ female B6J mice.

Randomization
The experimental unit (i.e., the cage) is completely randomized (balance randomized) with the use of computer

randomization process. Given the total sample size of 48 for the 3 groups, we can use the R command as below.

```
> set.seed(8057)
> sample(rep(c("HFD", "LFD", "SD"), each = 16), 48)

 [1] "SD"  "LFD" "HFD" "LFD" "SD"  "LFD" "HFD" "SD"  "HFD" "HFD" "LFD" "SD"
[13] "SD"  "HFD" "HFD" "HFD" "SD"  "LFD" "HFD" "SD"  "LFD" "SD"  "SD"  "SD"
[25] "LFD" "HFD" "HFD" "SD"  "LFD" "HFD" "LFD" "LFD" "LFD" "LFD" "HFD" "SD"
[37] "SD"  "HFD" "LFD" "LFD" "HFD" "HFD" "SD"  "SD"  "LFD" "SD"  "LFD" "HFD"
```

Blinding
We blind experimenters administering treatments, experimenters assessing outcome, caregivers, animal facilities staff and others who interact with the animals. The data analyst will also be blinded, she/he knows which observations belong to the same group without knowing which specific diet each group received.

In- and exclusion criteria
All animals will be used and we do not expect any dropout.

Example 2

We extend Example 1 above by adding another treatment factor, that is an exercise factor with two levels (moderate and sedentary). One might think of (i) conducting oneway ANOVA two times, each with 3 groups (one for moderate exercise and one for sedentary exercise for the 3 diets) or (ii) conducting oneway ANOVA with 6 groups. The first approach is typically called one-factor-at-a-time (OFAT). These two approaches are inefficient and unsatisfactory. It is inefficient because it would use a larger sample size than necessary. It is unsatisfactory because one cannot evaluate the effects of the three diets depending on the type of exercise (interaction). The best way is to conduct a 2-factorial design of twoway ANOVA with interaction.

Objective
To confirm an effect of diet with 3 levels (high-fat, low-fat and standard) and an effect of exercise with 2 levels (moderate and sedentary) on weight gain (average gr/week) of male B6J mice. The diet is administered by food in a cage.

Statistical test/model
Our research hypothesis is that at least one pair out of the six combined levels of the two treatment factors is different, while the null hypothesis is that there is no difference. Thus, we use two F tests, each for the two factors (i.e. twoway ANOVA). We are interested in comparing all the three possible pairs of the diet groups for each of the two exercise groups. Thus, in total we have 6 **comparisons**.

Power calculation of sample size
We consider a difference of 4 gr/week of average weight gain among the three groups as meaningful. Previous study shows that the population standard deviation is estimated with 3 gr/week in each combined group. To count for the six multiple tests we use Bonferroni correction, i.e., taking $\alpha = 0.05/6 = 0.0083$ for calculating the sample size (see section Multiple-testing problem). We use R package *daewr* (see Lawson, 2015, Chapter 2) as below.

```
> library(daewr)
> pow <- round(Fpower2(alpha = 0.05/6, nlev = c(3,2), nreps = 6:9, Delta = 4, sigma = 3), 3)
> pow

     alpha a b nreps Delta sigma powera powerb
[1,] 0.008 3 2     6     4     3  0.533  0.870
[2,] 0.008 3 2     7     4     3  0.641  0.929
[3,] 0.008 3 2     8     4     3  0.731  0.963
```

```
[4,] 0.008 3 2     9     4     3  0.804  0.981
```

Thus, 9 cages per combined group is needed to obtain a power of at least 0.80, in each of F test, with family-wise significant level of 0.05, resulting at $6 \cdot 9 = 54$ cages in total. Since we use male mice, where a cage contains one male mouse, the total number of animals needed in this study is $6 \cdot 9 = 54$ male B6J mice.

This 3-by-2 factorial design is much more effective than the two approaches mentioned in the beginning of the example. Keeping all other parameters value the same, the first approach (OFAT) would use total experimental unit of $2 \cdot (3 \cdot 16) = 96$ and the second approach $6 \cdot 23 = 138$, which are much larger than 54 resulting from twoway ANOVA with interaction approach.

A good mathematical reading on factorial designs is Lawson, 2015, Chapter 3.

Randomization
The experimental unit (i.e., the cage) is completely randomized (balance randomized) with the use of computer randomization process. Given the total sample size of 66 for the 6 combined groups, we can use the R command as below, where A and B respectively denote the **diet** and **exercise** factors and the numeric are the levels.

```
> set.seed(80)
> sample(rep(c("A1B1", "A2B1", "A3B1", "A1B2", "A2B2", "A3B2"), each = 9), 54)

 [1] "A2B1" "A2B2" "A1B2" "A3B2" "A1B1" "A2B2" "A1B2" "A3B2" "A1B2" "A2B1"
[11] "A1B1" "A1B1" "A3B1" "A2B1" "A3B2" "A2B1" "A2B2" "A1B1" "A3B1" "A1B1"
[21] "A1B2" "A3B2" "A1B2" "A1B1" "A2B1" "A1B2" "A1B1" "A1B2" "A3B1" "A2B1"
[31] "A3B2" "A2B2" "A3B2" "A3B1" "A3B1" "A2B2" "A2B1" "A2B2" "A1B1" "A1B2"
[41] "A2B2" "A3B2" "A3B1" "A2B1" "A2B2" "A1B2" "A1B1" "A3B2" "A3B1" "A3B1"
[51] "A2B2" "A2B1" "A3B1" "A3B2"
```

4.2 Exploratory

Purposes of exploratory experiments typically are exploration of potential biomarkers, assessing feasibility, assessing dose-finding, and investigating whether or not there is an affect. If there is an effect, then to estimate effect sizes, to generate hypotheses; all of which will have to be tested and confirmed later. Thus, sample sizes should be large enough to give investigators confidence in the direction of effect (worth finding) and small enough to optimize time and money.

Since there is no statistical test in an exploratory experiment, we cannot –and it does not make sense– use power calculation to calculate sample size. Sample size is therefore estimated from available pragmatic information, not from a power calculation. An interesting reading about the abuse of power is Hoening and Heisey (2001). Three approaches to estimate sample size are:

1. **Educated/expert estimation**. This estimation is based on the researcher's previous experience and the literature and/or conventional rule of thumb in the research field. This might work well with an experienced investigator who is thoroughly familiar with his or her material and has already performed several experiments like the one being proposed. This method is discussed in Reynolds (2019).

2. **Backwards of power calculation**. This can be used to quantify how your chosen sample size (based on the educated/expert estimation) can detect some minimum effect size under how much power if an effect exists.

3. **Precision estimation of effect size**. This approach estimates sample size based on the half-width of $(1 - \alpha)\%$ confidence interval (CI) that you deem to be of clinically importance. Thus, it is not about power, but the precision of your biologically meaningful effect size.

The formula of $(1-\alpha)\%$ CI for an effect size –minimum important difference in any two means– out of a means is $\bar{y}_1 - \bar{y}_2 \pm t_{a(n-1);(1-(\alpha/2))}\sqrt{2s^2/n}$, with s the estimated standard deviation and n is sample size per group. Hence, the precision is *the half-width of the CI*:

$$t_{a(n-1);1-\frac{\alpha}{2}}\sqrt{2s^2/n}\ .$$

That is, the $(1-(\alpha/2))$-quantile of t distribution with degree of freedom $a(n-1)$ multiplied with $\sqrt{2s^2/n}$. Thus, we can calculate n given the desired precision, α, a, and s.

We typically don't have an estimate of s, but we can use $\sqrt{2}\cdot(\text{Range}/2)$ as a rudimentary estimate. Range is the difference between the maximum observation and the minimum observation. See section Crude bounds for standard deviation for more details.

The precision above is a one-at-a-time precision, it applies only one particular estimate. In many situations, an experimenter wishes to calculate the precision for several estimates simultaneously. Using Bonferroni method, the formula for simultaneous confidence intervals –precision for several estimates simultaneously– is the same as the formula above by replacing the quantile of t distribution with $(1-(\alpha/2m))$. We refer to Montgomery (2020), Section 3.3.3, for more details.

Be aware that ideally your CI should not contain 0. That means the precision should be strictly smaller than the estimated effect size.

In the last few years we have seen more appeals in animal research to using precision estimate as a way to determine sample size in exploratory studies. We refer Piper *et al.* (2023) as an example.

There is also the so-called **Pilot study**. Pilot experiments are typically designed to allow researchers to gain familiarity with the experimental material, to make sure instructions are understandable and can be followed, to ensure that all steps in the procedure can be performed, to check that the staff is sufficiently trained in the necessary procedures, to assess that the level of intervention is appropriate (e.g., drug dose), to identify adverse effects (pain, suffering, distress, lasting harm) and effectiveness of the actions to reduce them (e.g., analgesia dose & schedule), to verify that the procedures are not too mild or severe, and to define early humane end-points.

In a pilot experiment, typically, the end point is qualitative and the conclusion/decision is not based on any statistical method but biological assessment. Thus, sample size is the minimum needed to achieved the purpose of the experiment.

Example 3

We consider two treatment factors.

Objective
To explore in mice the effect of administration route (3 groups) and dosage (2 groups) of a drug, with a biomarker as primary outcome.

Sample size estimation

1. **Expert estimate.** We estimate a sample size $n = 5$ for each of the 6 combined groups because based on our experience the total 30 data points would allow us to get an insight of the underlying mechanism as the purpose of these experiments. Furthermore, this choice of sample size allows us to conduct the randomization of the 30 animals at once and this is the maximum number we can handle.
2. **Backwards power calculation.** The choice of $n = 5$ gives a power of at least 0.80 to detect a signal-to-noise ratio of 25/15 (the ratio of minimum important difference to standard deviation of the outcome), assuming $\alpha = 0.05$ and two-sided F tests in twoway ANOVA with interaction, and we do not adjust for multiple testings. This calculation is performed with online tool SampleSizeR.

3. **Precision estimation of effect size.** Our focus lies on estimating effect size with 95% CI. Previously we observed s about 15 in this outcome and we are interested to estimate the effect size of 25. With sample size of $n = 5$ in each combined group, either the best dosage or the best route would give an half-width of 95% CI strictly less than 25. That is, the 95% CI is (0.58, 49.42). We are particularly interested to compare the 3 routes for one drug and in this calculation we adjust the 3 comparisons with Bonferroni method. This calculation is performed with online tool SampleSizeR. This would be sufficient to describe the effect for the first time.

Example 4

We consider an hypothetical pilot study. We need to train two experimenters to performance a new surgical technique on the animals. One of the experimenters is an experienced one who needs only 2 trials to obtain the required surgery skill, whereas the other one is a new one who needs 4 trials. Each trial is performed on one animal. Thus, 6 animals in total is necessary and sufficient to achieve the purpose of the study.

5 Categorical outcome

An outcome is called categorical if it has natural, mutually exclusive, non-overlapping groups. It is a qualitative measurement. An example is severity, defined as mild, moderate, or severe. Another example is success or failure of a procedure. Such categories may be ordinal (having an inherent order) or nominal (no inherent order) and can range from two groups or more. We focus on **binomial** outcome, two groups without inherent order.

5.1 Confirmatory

We give two examples. First, a standard example. Secondly, a non-standard example in which either probability is close to either 1 or 0. The test for binomial outcome is χ^2 **test**, which is parallel to a two-sample t test for a continuous outcome with two groups.

Example 5

Objective
To confirm an effect in mice of a new drug compared to a standard drug for a certain tumor disease on death vs alive (a mouse is either alive or dead at the end of the experiment).

Statistical test/model
Our research hypothesis is that the new drug will reduce the proportion of death compare to the standard drug, while the null hypothesis is the two proportions of death are the same. Thus, we use two sample comparison of proportion (relative frequency), which is called χ^2 **test** (using Gaussian as limiting distribution of binomial), with one-sided test.

Power calculation of sample size
The standard drug is known to have 60% proportion of death ($p_1 = 0.60$). A reduction of 50% for the new drug compare to the standard drug is meaningful for us. That is, $p_2 = p_1 - (p_1 * 0.50) = 0.60 - (0.60 \cdot 0.5) = 0.60 - 0.30 = 0.30$. We use R with the function *power.prop.test()* in standard library stats as below.

```
> power.prop.test(p1 = 0.60, p2 = 0.30, sig.level = 0.05, power = 0.80,
+                 alternative = "two.sided")
```

```
     Two-sample comparison of proportions power calculation

              n = 41.97025
             p1 = 0.6
             p2 = 0.3
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

The output is 41.97, rounded up to 42. Hence, 42 mice per group is needed to obtain a power of at least 0.80 with significant level of 0.05. The total number of animals needed in this study is $2 \cdot 42 = 84$ mice.

Randomization
The experimental unit (i.e., the mouse) is completely balance randomized into the two groups of new drug (NEW) and standard drug (STD) with the use of computer randomization process. We use R with the code below, which is a similar fashion as in the previous example, thus we do not show the output.

```
> set.seed(8057)
> sample(rep(c("NEW", "STD"), each = 42), 84)
```

Example 6

When either one of the survival-rates, denoted as p_1 for the control group and p_2 for the intervention group, is close to either zero or one, then one should use Fisher's exact test, instead of χ^2 test. Let n be the sample size calculated based on the χ^2 test. The correction for Fisher exact test is to replace n with

$$n' = \frac{n}{4} \cdot \left(1 + \sqrt{1 + \frac{4}{n|p_1 - p_2|}}\right)^2$$

as shown by Wittes (2002) in equation (7C).

As an example we use the above scenario but now we assume that $p_1 = 0.50$ (instead of 0.60) and that a reduction of 80% (instead of 50%) of the new drug is meaningful. That is, $p_2 = p_1 - (p_1 * 0.50) = 0.50 - (0.80 \cdot 0.5) = 0.50 - 0.40 = 0.10$ which is close to zero. The output of the χ^2 test with power of at least 0.90 and significant level of 0.05 would result in sample size of $n = 26$ per group is shown below.

```
> power.prop.test(p1 = 0.50, p2 = 0.10, sig.level = 0.05, power = 0.90,
+                 alternative = "two.sided")


     Two-sample comparison of proportions power calculation

              n = 25.43862
             p1 = 0.5
             p2 = 0.1
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

Based on the χ^2 test we would need $n = 26$ experimental units in each group. However, we should use the correction above since χ^2 test is not appropriate but Fisher's exact test is. Plug-in the correction formula

```
> n <- 26; p1 <- 0.50; p2 <- 0.10
> nprime <- (n/4)*(1+sqrt(1+(4/(n*abs(p1-p2)))))^2; nprime

[1] 30.79706
```

we calculate $n' = 31$. Thus, we should require the total number of animals of $2 \cdot 31 = 62$ mice.

5.2 Exploratory

The approaches for the sample size determination for binomial outcome is similar to the exploratory experiment with a continuous outcome: (1) expert estimate, (2) backwards of power calculation, and (3) precision estimate of the the difference of two proportions.

We focus on approach (3) in which we are interested in the precision estimate of the difference between two independent proportions. Many methods have been devised for computing confidence intervals for the difference: approximation, exact, or based on score. Recently, a new optimization algorithm is developed –see Peer and Azriel (2024)– but it is not available in open source software. After (almost) exhausted literature check, we use the Newcombe method Newcombe (1998), sometimes called method 10 or hybrid score method. It is made available in R package *presize* in Haynes *et al.* (2021). Newcombe's method is generally better than other method. It is typically more accurate; it handles better small to moderate sample sizes, it also handles better proportions near extreme (0 or 1), and provides better confidence interval coverage.

Example 7

Suppose a study is planned in which the researcher wishes to access two methods and to construct a two-sided 95% CI for the difference in two independent proportions of 0.60 and 0.30, in a balanced design. Using SampleSizeR, we need a minimum sample size of $n = 21$ in each group to estimate a 95% CI that does not contain zero. The 95% CI is $(0.001, 0.534)$ with the precision width of 0.533.

6 Time-to-event outcome

For the time-to-event outcome, we refer to Tarigan (2025) where more details on the topic of modeling survival data are given. We consider only the confirmatory case.

Example 8

Objective
To confirm an effect in mice of a new drug on the time-to-recurrence of a tumor compared to a standard drug.

Statistical test/model
Our research hypothesis is that the new drug would decrease the hazard rate compared to the conventional drug, while the null hypothesis is that the hazard ratio of both groups is one. Thus, we use **logrank test** (aka Cox test).

Power calculation of sample size
First we note that two-sided sample size formula of two groups is given by

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{p_0\ p_1\ (\ln(\text{HR}))^2} \times \frac{1}{p_0 d_0 + p_1 d_1}$$

with

- p_0 and p_1 are proportions of the subjects randomized to respectively control group and intervention group,
- d_0 and d_1 are probabilities of the subjects to experience the event in respectively control group and intervention group (which are also called as mortality rates),
- α is the level of the test (probability of type I error),
- $1 - \beta$ is the power of the test.

The details of how this formula is constructed can be seen in Schoenfeld, 1981, Section 3.2.

A decrease of 35% hazard rate of the new drug relative to the conventional drug is meaningful for us. Let λ_1 and λ_0 be the hazard rates of the intervention drug and the control drug, respectively. Thus, the hazard ratio is

$$\text{HR} = \frac{\lambda_1}{\lambda_0} = \frac{\lambda_0 - 0.35\lambda_0}{\lambda_0} = 1 - 0.35 = 0.65\ ,$$

for any value of the hazard rate of control group λ_0.

Under the assumption of logrank test (that the underlying survival curve is assumed to be exponential), the mortality rates are $d_0 = 1 - \exp(-\lambda_0)$ and $d_1 = 1 - \exp(-\lambda_1)$ respectively for the conventional drug and the new drug. From literature, an estimate of the mortality rate of the conventional drug is $d_0 = 0.40$. Therefor, $\lambda_0 = -\ln(1 - d_0) = -\ln(1 - 0.40) \approx 0.51$. Furthermore, the mortality rate of the intervention group is

$$d_1 = 1 - \exp(-\lambda_1) = 1 - \exp(-\lambda_0 \cdot \text{HR}) = 1 - \exp(-0.51 \cdot 0.65) = 1 - \exp(-0.33) \approx 0.28\ .$$

We further assume an equal sample size for each group. Thus, $p_0 = p_1 = 1/2$. Plugging in these values to the sample size formula

```
> my_nsize <- function(HR, d0, d1, P0, P1, alpha, beta){
+   (qnorm(1-alpha/2) + qnorm(1-beta))^2 / (P0*P1*(P0*d0+P1*d1)*(log(HR))^2)}
> n_size <- my_nsize(0.65, 0.40, 0.28, 1/2, 1/2, 0.05, 0.20); round(n_size, 2)

[1] 497.59
```

gives us a total sample size of 498 for both groups (balance size of 249 per group), with power of 0.80 and significant level of 0.05. Further detail on the topic of modeling survival data can be found in Tarigan (2025).

7 Multiple-testing problem

When an F test for null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$ is rejected, it tells us that there are significant differences between at least two of the means. If there are several levels of the treatment factor ($a \geq 2$), the rejection of the H_0 does not necessarily imply that all means are significantly different from each other. Thus, further investigation should be conducted to find out exactly which means differ. If the comparisons are not planned in advance (unplanned comparison), the investigator might be tempted to chose the comparisons she/he would like to make based on the means of the data. This implicitly means that all possible comparisons have been made. Thus, we must address multiple comparisons before starting the experiment.

When testing all possible comparisons, each at the significance level $\alpha = 0.05$, the overall significance level can be much larger than 0.05, can even be greater than 0.50, depending on the number of comparisons. This implies that even there is no difference in the means $\mu_1, \mu_2, \ldots, \mu_a$ there could be a high chance of finding one or more comparisons significant when each comparison is tested individually.

This overall significance level is called family-wise error rate (FWER). The term "family" refers to the collection of null hypotheses. For the F test with a levels/groups, there are $m = \binom{a}{2}$ null hypotheses (comparing every combination of 2 out of a means) that are jointly tested. Thus, the probability of committing at least one type I error out of m tests is

$$\begin{aligned}\text{FWER} &= P(\text{commiting at least one type I error in } m \text{ tests}) \\ &= 1 - P(\text{commiting no type I error}) \\ &= 1 - (1-\alpha)^m,\end{aligned}$$

where $\alpha = P(\text{commiting type I error})$ for each individual test.

Consider a case where the treatment factor has five groups ($a = 5$) and we are interested in comparing all possible pairs $m = \binom{5}{2} = 10$. In this case with individual $\alpha = 0.05$ the FWER is $1 - (1 - 0.05)^{10} = 0.40$, which is away much larger than the individual error rate. Even with $a = 3$, the overall significant level is $1 - (1 - 0.05)^3 = 0.14$, triple the individual significant level.

Thus, in the case of multiple comparisons, we should correct the individual significant level in determining sample size. **In the planning phase, the Bonferroni method is most conservative but simplest**. For the overall error rate of α, the corrected significant level is α/m with m the number of comparisons that are jointly tested. For example, if we have 3 comparisons, then to achieve an overall significant level of 0.05 the test should be with significant level of $\alpha = 0.05/3 = 0.017$.

The FWER is typically used for small scale of multiple tests. For large scale, like in "omics" study where one often tests hundreds or thousands of hypotheses at once, one should use the so-called false discovery rate (FDR). While FWER is appropriate when you want to guard ANY false positive (type I error), in the "omics" study we typically can live with a certain number of false positives. Thus, the more relevant quantity to control is the FDR because FDR controls the proportion of false positives among the set of rejected hypotheses. For sample size determination in the planning phase using FDR we refer to Ni *et al.* (2024) for more information.

Once the data are collected, one can use a (much) less conservative method for the multiple-testing correction such as Holm-Bonferroni, Tukey's HSD, Sidak, Hochberg, etc. We refer to Menyhart *et al.* (2021) for more details and its correction Menyhart *et al.* (2022).

8 Crude bounds for standard deviation

In an exploratory experiment, it is very often the case that one does not have an estimate for standard deviation s because there is no data yet. Estimating standard deviation without data is challenging because standard deviation is inherently a data-dependent measure. However, when there is some information about the range of the outcome, we can derive a very crude bounds. Range is the difference between the maximum observation and the minimum observation.

For example, if we at least can estimate the smallest possible outcome and the largest possible outcome, then the largest possible value for the standard deviation with these two observations is $\sqrt{2} \cdot (\text{Range}/2)$.

A rule of thumb for crude bounds of s provided in van Belle, 2008, Section 2.4 is

$$\frac{\text{Range}}{\sqrt{2(n-1)}} \leq s \leq \sqrt{\frac{n}{n-1}} \cdot \frac{\text{Range}}{2} \ .$$

For large n the ratio $n/(n-1)$ is approximately 1 and can be ignored, thus s is less than Range/2.

9 Sex as biological variable (SABV)

Currently most preclinical research uses only one sex, typically males. This means many female animals are bred but are not used for research, and many males are used in studies that cannot be generalized to the whole population. This has also created a sex bias in the scientific knowledge base where information about female biology is lacking. Clearly sex bias in preclinical research greatly limits translation.

Since more than five years ago regulators, finders and publishers are pushing researchers to include female and male animals in their study and to analyze the potential of sex influence on the outcomes. It seems that the high importance of considering sex as a biological variable (SABV) in animal research is now well understood, despite persistent misconceptions. We now also understand that considering SABV is not necessarily the same as looking for sex difference. However, sex bias (female animals are still under represented) is still happening in animal studies until today. Where lays the problem? One of main reasons is because of a lack of understanding of how to collect and analyze the data so that the presence or absence of potential sex influence can be assessed properly. This leads to a persisting false idea that including both sexes would double the sample size.

Consider a 1-factorial CR design (one treatment factor) with 3 groups. Assume that a sample size n has been determined, in total $3n$ animals, based on power calculation of F test of the given effect size, estimated standard deviation, power and type I error. The proper way to incorporate SABV is to include it as a second fixed factor (biological variable) into the design. Thus, a 2-factorial design in which $3n/2$ females and $3n/2$ males are completely randomized in a balanced way to the 3 treatment groups, each treatment group with $n/2$ for each sex. The analysis would be –instead of oneway ANOVA– using twoway ANOVA with interaction. Keep everything the same, the total sample size $3n$ in the 2-factorial design gives the same power as in the 1-factorial design.

Sex inclusive research is now more of a norm than of an exception. More than often, researchers want to do the best science and generate meaningful data but many think of sex inclusive research as an insurmountable challenge. For more information on sex inclusive research we refer to Karp *et al.* (2025) and the reference therein. The authors of the paper created a framework –a set of resources– to check whether sex has been considered when designing and assessing in vivo research proposals.

References

Astakhov, G., Cherneva, K., Furrer, R., and Tarigan, B. (2025). Samplesizer: calculate sample sizes within completely randomized design. http://shiny.math.uzh.ch/git/reinhard.furrer/SampleSizeR/.

Bate, S. and Clark, R. (2014). *The Design and Statistical Analysis of Animal Experiments*. Cambridge.

Dean, A., Voss, D., and Draguljic, D. (2017). *Design and Analysis of Experiments*. Springer.

Haynes, A. G., Lenz, A., Stalder, O., and Limacher, A. (2021). presize‘: An r-package for precision-based sample size calculation in clinical research. *Journal of Open Source Software*, **6**, 3118. https://doi.org/10.21105/joss.03118.

Hoening, J. M. and Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, **55**, 19–24. https://doi.org/10.1198/000313001300339897.

Karp, N. A., Berdoy, M., and Gray, K. e. a. (2025). The sex inclusive research framework to address sex bias in preclinical research proposals. *Nature Communications*, **16**, . https://doi.org/10.1038/s41467-025-58560-5.

Lawson, J. (2015). *Design and Analysis of Experiments with R*. CRC Press.

Lazic, S. (2016). *Experimental Design for Laboratorist Biologists*. Cambridge University Press.

Menyhart, O., Weltz, B., and Győrffy, B. (2021). Multipletesting.com: A tool for life science researchers for multiple hypothesis testing correction. *PLoS ONE*, **16**, e0245824. https://doi.org/10.1371/journal.pone.0245824.

Menyhart, O., Weltz, B., and Győrffy, B. (2022). Correction: Multipletesting.com: A tool for life science researchers for multiple hypothesis testing correction. *PLoS ONE*, **17**, e0274662. https://doi.org/10.1371/journal.pone.0274662.

Montgomery, D. C. (2020). *Design and Analysis of Experiments*. John Wiley & Sons, 9th EMEA Edition.

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, **17**, 873–890. https://doi.org/10.1002/(sici)1097-0258(19980430)17:8%3C873::aid-sim779%3E3.0.co;2-i.

Ni, Y., Seffernick, A. E., Onar-Thomas, A., and Pounds, S. B. (2024). Computing power and sample size for the false discovery rate in multiple applications. *Genes*, **15**, 344. https://doi.org/10.3390/genes15030344.

Peer, A. and Azriel, D. (2024). Optimal confidence interval for the difference between proportions. *Statistics and Computing*, **34**, . https://doi.org/10.1007/s11222-024-10485-7.

Piper, S. K., Zocholl, D., Toelch, U., Roehle, R., Stroux, A., Hoessler, J., Zinke, A., and Konietschke, F. (2023). Statistical review of animal trials—a guideline. *Biometrical Journal*, **65**, 2200061. https://doi.org/10.1002/bimj.202200061.

Reynolds, P. (2019). When power calculations won't do: Fermi's approximation of animal numbers. *Lab Animal, Nature*, **48**, 249–253.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, **68**, 316–319. https://doi.org/10.1093/biomet/68.1.316.

Tarigan, B. (2025). Comparing survival curves. https://www.math.uzh.ch/consulting/fileadmin/user_upload/Comparing_Survival_Curves.pdf.

van Belle, G. (2008). *Statistical Rules of Thumb*. John Wiley & Sons. https://doi.org/10.1002/9780470377963.

Wilson, E., Ramage, Wever, Sena, Macleod, and Currie (2023). Designing, conducting, and reporting reproducible animal experiments. *Journal of Endocrinology*, **258**, e220330. https://doi.org/10.1530/JOE-22-0330.

Wittes, J. (2002). Sample size calculations for randomized controlled trials. *Epidemiologic Reviews*, **24**, 39–53. https://doi.org/10.1093/epirev/24.1.39.

Version Information

v1.1	04.02.2021	
v1.2	19.01.2022	add examples, add references & change style
v2.1	25.04.2024	update & reorganize sections, add more info & examples for exploratory, add section bounds for std, change D-Math to DM3L
v3.1	May 20, 2025	reorganize & add sections (categorical, bounds for s, SABV), adjusted to the updated SampleSizeR, layout, uzh-logo

```
print( sessionInfo(), locale=FALSE)

## R version 4.3.2 (2023-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.4 LTS
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.so;  LAPACK version 3.10.0
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] daewr_1.2-11 knitr_1.45
##
## loaded via a namespace (and not attached):
## [1] compiler_4.3.2 tools_4.3.2    highr_0.10     xfun_0.42      evaluate_0.23
```