# First PAGE

## Design and statistical analysis of mice experiments

The purpose of this app is to help researchers conducting mice experiments in the Netherlands Cancer Institute with respect to statistical aspects of the studies. One can find here some explanations of basic statistical concepts and tests that can be applied to illustrate data or calculate sample size and power of planned experiments. Moreover, researchers can use this app to calculate sample size and power when an experiment is being designed. The latter is vital and must be performed, because:

More power increases the confidence in the results, whether they are significant or not

Detail information about sample size and power calculations can be found under the Concepts of Power tab .

Sample size and power calculation depend on the type of experiment that is planned. The most common experimental designs used for mice experiments at the NKI compare groups of mice with respect to mean/median values, survival outcomes, proportions and growth curves. Examples of such experiments are listed below and more information can be found under the specific tabs.

**Mean/Median Analysis**

Example: A scientist wants to test the hypothesis that a novel compound had a beneficial effect on reducing high-density lipoprotein (HDL) cholesterol levels in a transgenic C57Bl/6J strain of mice compared with standard treatment. Therefore she randomizes half of the mice to the standard treatment and the other half to the new treatment, in order to compare average HDL cholesterol levels between the two groups",

Go to Mean/Median Analysis

**Survival Analysis**

Example: evaluate whether treatment with paclitaxel improves survival after esophageal adenocarcinoma (EAC), a scientist uses a peritoneal dissemination xenograft mouse model and injects human EAC cell lines intraperitoneally/subcutaneously into SCID mice. Two weeks later, mice are randomly assigned to treatment by vehicle or paclitaxel (20mg/kg, 2 times a week for 2 weeks). Mice are followed until death or the end of the study and the mouse survival times are compared between the 2 groups.

Go to Survival Analysis

**Proportion Analysis(NOT READY YET)**

Example: A scientist wants to test the hypothesis that a novel compound had a beneficial effect on reducing high-density lipoprotein (HDL) cholesterol levels in a transgenic C57Bl/6J strain of mice. Therefore she randomizes half of the mice to control group and the other half to treatment group, in order to compare average HDL cholesterol levels from the two groups

Go to Proportion Analysis

**Growth Curve Analysis(NOT READY YET)**

Example: A scientist wants to test the hypothesis that a novel compound had a beneficial effect on reducing high-density lipoprotein (HDL) cholesterol levels in a transgenic C57Bl/6J strain of mice. Therefore she randomizes half of the mice to control group and the other half to treatment group, in order to compare average HDL cholesterol levels from the two groups

Go to Growth Curve Analysis

# Second page – First Tab

Concepts of Power

Interpretation of power

An experiment is conducted to answer a particular research question, for instance, to investigate whether there is an effect of a new treatment. While analyzing the data, a researcher can make two types of correct decisions and two types of errors, which is shown in the table below.

TABLE 1\_HYPOTHESIS TESTING

The effect either exists or not in nature, while the result of the experiment is either significant or non-significant. Therefore, a researcher either makes a correct inference about the effect or a false one. The type I error, which is denoted by the Greek letter α , is the probability of finding an effect when it does not exist in nature α is also called the significance level of a test and the Type II error, denoted by the Greek letter β , is the probability of not finding an effect when the effect exists in nature. So, the Type I error is a chance of a false positive finding, while the Type II error is a chance of a false negative finding. Complements of the two probabilities, 1- α and 1- β , are probabilities of correctly not finding an effect (true negative finding) and correctly finding an effect (true positive finding), respectively. The latter probability, 1- βis called the statistical power of a test. If there is an true effect of a treatment, researchers would like to detect it with as large probability as possible but a power level of 0.8 or 0.9 is usually considered sufficient. Intuitively, one can think about these probabilities in contexts of replications of one experiment. If we have 100 replications of an experiment where the effect does not exist and α=0.05, then we expect to have a false positive finding in 5 of these experiments. On the other hand, if we have 100 replications of an experiment where the effect truly exists and β=0.2 so power=0.8, then we expect to have a false negative finding in 20 of these experiments

Statistical power gives us a measure of confidence that we will be able to detect a significant effect if it truly exists. The power highly depends on the sample size of an experiment and in the process of designing an experiment, a researcher can assess how many mice should be included in order to achieve a certain chance of making a correct decision. This assessment is of high importance because an experiment can be underpowered if there are too few mice and that can cause missing an effect that truly exists. An experiment can also be overpowered if there are too many mice and that can cause detecting an effect that truly exists but is so small that is of no interest and use in practice. In both situations, resources spent on an experiment, such as money, time or mice life’s, are wasted.

\* More Power increases chances of finding significant result

\* More Power increases chances of replicating prior findings

\* More Power increases confidence about results, either significant or not

Probabilities of correct and false inference about the effect should be corrected with inclusion of the probability of the effect being true, R , if there is an uncertainty whether the effect actually exists in nature:

TABLE 2\_POWER

A scientist who performs 100 experiments with 50% chance of the effect existence in nature, is expected to detect the effect in 50 of these experiments. However, when the statistical power is 80%, then it is expected to capture 40 out of these 50 experiments and when the power is 50%, then only 25 out of the 50 experiments will likely be identified.

For each mice experiment, four important measures are considered:

1. True positive rate = Power\*R / Power\*R + (1-Power)\*R = Power

It is the probability of a significant result given that the effect truly exists in nature

1. True negative rate = (1-α) \* (1-R) / (1-α)\*(1-R) + α\*(1-R) = (1-α)

It is the probability of a non-significant result given that the effect does not exist in nature. It is the complement of type-I error α

1. Positive Predictive Value(PPV) = Power\*R / Power\*R + α\*(1-R)

It is the probability of the effect existence in nature given positive result of the experiment. As can be seen from the formula and the graph below, that probability is higher with higher power level when probabilities α and R are fixed, and higher with higher R when power and α are fixed.

IMAGE 1\_PPV

1. False Positive Report Probability(FPRP) = 1-PPV = α\*(1-R) / Power\*R + α\*(1-R)

It is the probability of the no effect existence in nature given that positive result of the experiment.

As can be seen from the formula and the graph below, that probability is lower with higher power level when probabilities α and R are fixed, and lower with higher R when power and α are fixed.

IMAGE 2\_FPRP

A researcher can also evaluate a false conclusion rate which is a chance of making either the Type I or Type II error. As is illustrated in the graph below, this rate is lower with higher power level when probabilities α and R are fixed, and lower with lower R when power and α are fixed.

Moreover, when the prior probability of the effect existence is maximum, i.e. R=1, then the false conclusion rate depends only on the power level of the test and more precisely, it is actually equal to the Type II error rate β or equivalently to 1-power. On the contrary situation, i.e. when R = 0, the false conclusion rate is equal to the Type I error rate α . As the power and R increase, this rate approaches zero.

For fixed value of α and power, higher probability R is associated with more false experimental results and the lower the power the higher the influence of R on the false conclusion rate.

IMAGE 3\_FC

## SECOND PAGE – SECOND TAB

How to calculate Sample Size and Power

Statistical power highly depends on three factors

To determine the power of an analysis we need firstly to specify the alternative Hypothesis, Ha, or in other words, the effect size that we are interesting in detecting. Further, and for most of analyses, power is proportional to the following

Effect size : the effect of interest which can be measured as a difference in mean/median values, survival outcomes, proportions or growth rates; the bigger the effect size the higher the power

Sample size : the number of mice included in an experiment; the higher the number of mice the higher the power.

Significance level (α) : the type-I error of a test; the higher the α the higher the power

The power level can be assessed when the three factors are known or the required sample size needed for an experiment can be calculated when the power level and the two other factors are fixed. These elements are used differently with different types of study design and statistical tests used for analysis of the outcome of an experiment. What is very important, the same statistical test should be used for power calculations at the planning phase of the experiment and data analysis applied when the experiment is finalized.

*When a researcher is designing an experiment it is highly advised to follow the following steps.*

1. Formulating the research question, i.e. defining clearly what is the hypothesis of interest

2. Identifying the statistical test that is going to be performed for analyzing the results of the experiment

3. Pre-specifying the effect size, i.e. deciding about the expected effect size based on the researcher’s knowledge of the experiment or information found in the literature, or the smallest effect size that is considered as clinically important

4. Selecting desired α level

5. Selecting desired power level and calculating the required sample size or selecting desired sample size and calculating the power level

## SECOND PAGE – THIRD TAB

Other Statistical issues

Multiple Comparisons

When an experiment involves more than 1 comparison, the overall probability of Type I error in the experiment is higher than the selected α level of one test. This overall probability is also called the familywise error rate or experiment-wise error rate and is the probability that at least one comparison detects a false positive finding. It is calculated with the formula:

1-(1-α)^{n},

where α is the significance level for an individual comparison and n is the total number of comparisons in the experiment. For instance, if one wants to include 4 groups in the experiment and compare each group to each other, then 6 comparisons need to be made. With such an experiment, the probability that at least one comparison detects a false difference is equal to

1-(.95)^{6} = 26%

Many statistical techniques have been developed in order to deal with this issue, i.e. to control the familywise error rate. One can use a lower Type I error for each individual comparison so the overall Type I error is also lower. The most common approach is the Bonferroni correction where the overall desired familywise error rate is divided by the number of comparisons in the experiment to find the individual test α level. So, if a researcher wants to conduct 10 statistical tests with the familywise error rate at 0.05, then the significance level for each individual test should be 0.05/10=0.005, which means that only these comparisons with P < 0.005 are considered significant.

The control of the familywise error rate needs to be taken into account not only in the data analysis phase of an experiment but also when the sample size and power calculations are performed

IMAGE 4\_FWER

One-sided vs Two-sided tests

What is a two-tailed test?

If you are using a significance level of 0.05, a two-tailed test puts half of your \\(\\alpha\\) to testing the statistical significance in one direction and half of your alpha to testing statistical significance in the other direction. This means that .025 is in each tail of the distribution of your test statistic. When using a two-tailed test, regardless of the direction of the relationship you hypothesize, you are testing for the possibility of the relationship in both directions. For example, we may wish to compare the mean of a sample to a given value x using a t-test. Our null hypothesis is that the mean is equal to x. A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x. The mean is considered significantly different from x if the test statistic is in the top 2.5% or bottom 2.5% of its probability distribution, resulting in a p-value less than 0.05."

What is a one-tailed test

If you are using a significance level of 0.05, a one-tailed test puts all of your \\(\\alpha\\) to testing the statistical significance in the one direction of interest. This means that .05 is in one tail of the distribution of your test statistic. When using a one-tailed test, you are testing for the possibility of the relationship in one direction and completely disregarding the possibility of a relationship in the other direction. For example, in comparing the mean of a sample to a given value x using a t-test, our null hypothesis is that the mean is equal to x. A one-tailed test will test either if the mean is significantly greater than x or if the mean is significantly less than x, but not both. Then, depending on the chosen tail, the mean is significantly greater than or less than x if the test statistic is in the top 5% of its probability distribution or bottom 5% of its probability distribution, resulting in a p-value less than 0.05. The one-tailed test provides more power to detect an effect in one direction by not testing the effect in the other direction."

When is a one-tailed test NOT appropriate?

"<br/>", "<br/>", "<br/>",

"Choosing a one-tailed test for the sole purpose of attaining significance is not appropriate. Choosing a one-tailed test after running a two-tailed test

that failed to reject the null hypothesis is not appropriate, no matter how 'close' to significant the two-tailed test was.

Using statistical tests inappropriately can lead to invalid results that are not replicable and highly questionable"

tabPanel("Softwares for Sample size and power calculation",

HTML(paste0("<br/>","<br/>","<br/>",

"<center><strong>G\*Power: Statistical Power Analyses for Windows and Mac</strong></center>")),

HTML(paste0("<br/>","<br/>",

"G\*Power is a tool to compute statistical power analyses for many different t tests, F tests, t tests, z tests and some exact tests.

G\*Power can also be used to compute effect sizes and to display graphically the results of power analyses.",

"<br/>",

"G\*Power is a freely available software and currently it is condidered one of the best for power calculations.

More information about it, such as a manual, as well as guidelines about how to download it can be found in the following link:",

"<br/>",

a("G\*Power link", href="http://www.gpower.hhu.de/", target = "\_blank")

## SECOND PAGE – FOURTH TAB

## 

***Software for Sample size and power calculation***

**G\*Power: Statistical Power Analyses for Windows and Mac**

G\*Power is a tool to compute statistical power analyses for many different t tests, F tests, t tests, z tests and some exact tests. G\*Power can also be used to compute effect sizes and to display graphically the results of power analyses.  
G\*Power is a freely available software and currently it is considered one of the best for power calculations. More information about it, such as a manual, as well as guidelines about how to download it can be found in the following link:  
[G\*Power link](http://www.gpower.hhu.de/)

## THIRD PAGE - FIRST TAB

*(Basic Information)*

A T-test or Wilcoxon-Mann-Whitney test can be used for comparison of two groups, while an ANOVA or Kruskal-Wallis test can be used for comparison of more than two groups. based on mean or median values of the outcome. T-test and ANOVA are parametric tests that rely on certain assumptions and these assumptions need to be met to get reliable test results. Validation of these assumptions becomes impossible when sample is small and a researcher should use non-parametric alternative tests instead, namely Mann-Whitney-Wilcoxon or Kruskal-Wallis test. Since animal experiments are usually conducted with small number of mice, the recommendation is to use Wilcoxon-Mann-Whitney test or Kruskal-Wallis test that are described in more details here. In an experiment with more than two groups of mice, a researcher can first use Kruskal-Wallis test to determine whether there are any statistically significant differences between the groups, i.e. whether at least one group of mice is different from others. This test is called the omnibus test and when significant results are detected, a researcher does not know which group or groups are actually different from each other. Therefore, as the next step, a researcher can perform pairwise comparison tests, also called post hoc pairwise tests, to find the group or groups with significantly different average values of the outcome of interest. In this step, Wilcoxon-Mann-Whitney test can simply be used for comparison of two groups.  
  
  
However, one needs to be aware that conducting multiple pairwise tests increases the probability of a false positive result and a correction of the significance level *α* should be implemented. The most popular correction is the Bonferroni adjustment which divides the *α* by the total number of comparisons that are performed. For instance, in an experiment with three treatment groups (A, B, C), three pairwise comparisons can be performed (A-B, A-C, B-C). To have the overall *α* of 0.05, i.e. 5% chance that at least one of the comparison is false positive, a significance level of *α*/(number of comparisons)=0.05/3=0.0167 for each pairwise test should be used.  
  
  
[Here](https://www.nki.nl/), you can find more about this and some alternative solutions to the multiple comparisons issue.

## THIRD PAGE - SECOND TAB

*(Example)*

***For 2 groups***

##### Example:

A scientist wants to test the hypothesis that a novel compound had a beneficial effect on reducing high-density lipoprotein (HDL) cholesterol levels in a transgenic C57Bl/6J strain of mice. Therefore she wants to conduct a new study by randomizing mice to control treatment groups, in order to compare the average HDL cholesterol levels from the two groups. From a previous experiment, following measurements of HDL are observed:

TABLE\_EXAMPLE DATA

These data can be used to calculate power and required sample size for The new experiment. The following information from the observed previous data is needed:

1. Mean in group A (Control)
2. Mean in group B (Treatment)
3. Standard Deviation in group A
4. Standard Deviation in group B

Further, the power level and the significance level(*α*) of a test, or in other words the desired Type-I error, need to be specified. Usually, power is set to 80% and *α* to 5%.

***For more than 2 groups***

##### Example:

A scientist wants to test the hypothesis that a novel compound had a beneficial effect on reducing high-density lipoprotein (HDL) cholesterol levels in a transgenic C57Bl/6J strain of mice. Therefore she wants to conduct a new study by randomizing mice to control and two treatment groups, in order to compare the average HDL cholesterol levels from the three groups. From a previous experiment, following measurements of HDL are observed:

TABLE\_EXAMPLE DATA 2

In case of more than two groups, for the power calculation we use information about two groups with the smallest difference because a larger sample is required to detect a smaller effect. If an experiment is powered for the smallest difference, it is also powered to detect a larger difference. After finding these 2 groups, the procedure is exactly the same with the case of 2 groups.

## THIRD PAGE - THIRD TAB

*(Example – Power calculation)*

*(Here is the power calculation example with interaction with the app)*

Here it is an example of how this app can be used in order to perform power calculations. In the left side, there is a panel where you can provide the input that is required for it, and after clicking on **Go!**, the results will appear in the right side(it might take a couple of seconds). Along with that, a plot will also be provided with a range of possible sample sizes and their associated power values.   
  
  
For illustration, we will use the example that we saw Here for both cases (2 groups & >2 groups)

***For 2 groups***

For this particular example, we have:

1. Mean in group A (Treatment) = 267.39
2. Mean in group B (Control) = 283.46
3. Standard Deviation in group A = 14.38
4. Standard Deviation in group B = 11.83

Finally, we specify *α* at 5% and the desired power to be 80%.

***For more than 2 groups***

For this particular example, we have:

1. Mean in group A (Treatment A) = 267.39
2. Mean in group B (Treatment B) = 256.48
3. Mean in group C (Control) = 283.46
4. Standard Deviation in group A = 14.83
5. Standard Deviation in group B = 9.75
6. Standard Deviation in group c = 11.83

As we explained, in this case we should find the two groups with the smaller difference. We observe here that these two groups are Treatment A and Treatment B, and therefore these will be used for the power calculation. Since three pairwise tests will be performed, the *α* is adjusted by 0.05/3 = 0.0166. Finally, we specify the desired power to be 80%.   
  
  
If we now provide the input to the panel at the left, we will see below the results.

***(Below in the text, the values are changing according to the input in the panel)***

**In order to achieve** 0.8 **power to reject the null hypothesis of equal means when the population mean difference is M1-M2 =** -16.07 **with a standard deviation for group A=** 14.38 **and group B=** 11.83 **and with a significance level (a) of** 0.05

#### The required sample size per group is:

*( The result shows up here )*

## THIRD PAGE - FOURTH TAB

*(Power calculation)*

*(Here is the actual power calculation)*

***(Below in the text, the values are changing according to the input in the panel)***

**In order to achieve** 0.8 **power to reject the null hypothesis of equal means when the population mean difference is M1-M2 =** -16.07 **with a standard deviation for group A=** 14.38 **and group B=** 11.83 **and with a significance level (a) of** 0.05

#### The required sample size per group is:

*( The result shows up here )*

## THIRD PAGE - FIFTH TAB

*(Wilcoxon ranks-sum test description)*

(TO BE WRITTEN…)

## FOURTH PAGE - FIRST TAB

*(Basic Information)*

Survival analysis can be used for comparison of groups based on information of an event where main interests of the study is 'whether' and/or 'when' the particular event occurs. In mice experiments, the event of interest is usually death. For each mouse in an experiment, the time until death is measured from the day when an experiment starts, for example when a mouse is randomized to a treatment group, until a mouse dies or is scarified due to reaching a human endpoint. When a mouse is still alive at the end of an experiment, the observation for that mouse is censored. Censoring in survival analysis means that the information about the survival time is incomplete, like for a mouse that is still alive at the end of an experiment, the only available information on survival time is that time of death has not been observed during the duration of the experiment, so death happened later on but an investigator does not known the exact time of death.  
  
  
  
  
Survival times of two or more treatment groups can be compared with the log-rank test that is also called Mantel-Cox test. It is a non-parametric test that does not make any distributional assumptions about observations. It looks at the order in which events happened and calculates number of observed and expected events in each group at each observed event time. Thus, the test compares survival across the whole spectrum of time, not just at one or two time points. If there are no censored observations in an experiment, a researcher can use a Wilcoxon-Mann-Whitney test instead to compare average survival times between groups.

## FOURTH PAGE - SECOND TAB

*(Example)*

A scientist aims to establish a peritoneal dissemination xenograft mouse model of Esophageal adenocarcinoma (EAC). Human EAC cell lines OE19 are injected intraperitoneally/subcutaneously into SCID mice and two weeks after the injection, mice are randomly allocated to vehicle or paclitaxel treatment (20 mg/kg, 2 times a week for 2 weeks). The main interest of that experiment is to compare time of death between the two groups of mice. Data from a previous similar experiment is available and can be used as a basis of a power calculation for the following experiment. Data on 20 mice is accessible with information on number of days between start of the experiment and time of death or censoring, event status (status = 1 when a mouse died, status = 0, when a mouse did not die) and treatment groups (vehicle or treatment group).

TABLE\_EXAMPLE DATA

For the power calculation based on a logrank test information on the proportions of surviving until a particular time point is required. Moreover, information on accrual time and total experiment time is needed. The accrual time is the duration of time of mice enrolment into the study, which is equal to one when all mice are included in the experiment at the same time. The total experiment time is the planned duration of the experiment. Further, the power level and the significance level (*α*) of the test, or in other words the desired Type-I error, need to be specified. Usually, power is set to 80% and *α* to 5%.Therefore, we need:

1. Median Survival in group A (Control)
2. Median Survival in group B (Treatment)
3. Accrual time
4. Follow-up time
5. Power level
6. Significance level

## FOURTH PAGE - THIRD TAB

*(Example – Power calculation)*

Here it is an example of how this app can be used in order to perform power calculations. In the left side, there is a panel where you can provide the input that is required for it, and the results will appear in the right side.   
  
  
For illustration, we will use the example that we saw Here   
  
  
For this particular example, we have:

1. Median Survival in group A (Treatment) = 25.5
2. Median Survival in group B (Control) = 35
3. Accrual time= 1
4. Follow-up time = 60

Finally, we specify *α* at 5% and the desired power to be 80%.

***(Below in the text, the values are changing according to the input in the panel)***

**In order to achieve** 0.8 **%power at a** 0.05 **significance level to detect a hazard ration of** 1.372549 **when the control group median survival is** 25.5 **for a study that lasts for** 60 **time periods of which subjects accrual(entry) occurs in the first time period,**

#### The required sample size per group is:

*( The result shows up here )*

## FOURTH PAGE - FOURTH TAB

*(Power calculation)*

***(Below in the text, the values are changing according to the input in the panel)***

**In order to achieve** 0.8 **%power at a** 0.05 **significance level to detect a hazard ration of** 0.5 **when the control group median survival is** 60 **for a study that lasts for** 100 **time periods of which subjects accrual(entry) occurs in the first time period,**

#### The required sample size per group is:

*( The result shows up here )*

## FOURTH PAGE - FIFTH TAB

*(Log-rank test description)*

In the data that from the example here we observe that all the mice in the control group have experienced the event of interest during the experiment, while in the treatment group we have 3 mice that are censored. Two mice have died early but from reasons unrelated to the experiment and one mouse was still alive at the end of the experiment. Because of the presence of censoring, the log-rank test is the only test that we should use. I order to prove that, we will analyze this data with both the log-rank test and a simple Wilcoxon-Mann-Whitney.

TABLE\_RESULTS

So, we see that the log-rank test is significant at the 5% significance level while the Wilcoxon test is not, based on their p-values. Intuitively this can be explained from the fact that we have 2 early deaths in the treatment group which are not related to our experiment and thus they are considered censored. This can be seen from the log-rank test which can handle the censoring but not from the t-test which just compares the times from the 2 groups and does not take into account the censoring. Therefore, we see that when there is censoring in the data, log-rank test (or any other survival analysis test) is the only choice.   
  
  
  
  
  
Let's now consider another example.

TABLE\_EXAMPLE DATA 2

Here we have 39 mice, 21 in control and 18 in treatment respectively, and all them have experienced the event except one mice in the treatment group which has died from other causes early in the study. We analyze these data again with both log-rank test and Wilcoxon.

TABLE\_RESULTS 2

Here both tests are highly significant in the 5% significance level. Here we have only 1 out of the 39 mice censored and this does not affect the conclusion that we make with both tests. Of course, this would cause a problem if the sample size here was smaller, because each mouse would put more weight on the tests.