What does my data say? – Stichprobenstrategien

Inhalt

Objectives

This Chapter offers a gentle introduction to the sample size calculation. It covers the importance of sampling from a target population, briefly introduces popular sampling methods in medical research, and addresses core components of the sample size calculation. In the Module Clinical Studies and Biobanking, you will have the opportunity to delve into the sample size calculation using real-life examples from published clinical trials.

Importance of sampling

The Chapter '*Einführung Studiendesign: Woher kommen die Daten?*' emphasised the importance of carefully formulating the research questions of the research project. The research questions dictate the PICOT framework and hence, the target population. Ideally, the researcher would collect all necessary information from the target population and use these data to proceed with the statistical analysis. Then, the results would be fully accurate and precise. In practice, this ideal plan is not feasible, mostly for financial and logistical reasons. Investigating the whole target population would be a resource-intensive task, as it would require a considerable budget, time and workload. Not to mention the challenge of locating and accessing the target population across the whole globe. Therefore, the investigator has to resort to a fraction of the target population, known as the ***sample***. The sample is (ideally) a representative part of the target population that aims to derive the truth for a specific research question of the study. Having access to a fraction of the target population naturally raises the following questions:

- Is the acquired sample **representative** of the target population?
- Is the sample **large enough** to ensure accurate and precise results?

The first question can be addressed by referring to the proper **sampling frame** and selecting a **sampling method** that would introduce as little bias as possible to the sample. The **sampling frame** refers to the source from which the population units will be collected—for example, the medical records of a specific hospital or the hospitals in a specific geographical region. The collected information from the sampling frame should *always* be evaluated for possible errors and accidental omissions that would contribute to *sampling bias*. The **sampling method** refers to the procedures to collect the population units from the sampling frame (see next section).

We will use our simple fictional example of hypothesis testing to exemplify the target population, sampling frame, sampling method and sample. The research question was:

'Which dietary plan does improve fasting insulin substantially in adults with diabetes: eating more meat than veggies or eating more veggies than meat?'

Adults with diabetes are the **target population**. If we plan a retrospective cohort study, the **sampling frame** may be the medical records of the MHH. If we plan a prospective cohort study, the **sampling frame** may be all patients in the Clinic of Gastroenterology in the MHH. The **sampling method** is convenience sampling (see next section), as we considered only the MHH for being employed there (hence, convenience). Suppose we decide to collect information on 150 adults following a veggie-based diet and 150 adults on a meat-based diet. These 300 adults comprise the **sample** from a larger population of adults with diabetes who have not been admitted to the hospital. What do you think about the representativeness of this sample?

Both the sampling frame and the sampling method will reinforce or compromise the representativeness of the sample and hence, the generalisability of the results to the target population. Knowledge of the available sampling frame's advantages and disadvantages is the key to collecting a representative sample. The ultimate goal of the sampling is to ensure the study has *high external validity*, namely, that the results can be applied beyond the collected sample.

The second question can be addressed using objective methods to determine the sample size. This requires (i) knowledge of the available, relevant evidence and (ii) a decision framework on how much error is allowed from collecting a fraction of the target population. Keeping up with the published literature and current practices is necessary to determine the minimum clinically important difference (MCID) of the compared interventions (or exposures). In the context of randomised controlled trials, MCID has been recognised as a core element of the trial design. The interested reader may refer to the ICH-E9 guideline (Statistical Principles for Clinical Trials) of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. Specifying the MCID and important design elements, such as power, and type I error, are required to determine the necessary sample size while maintaining ethical standards. The sample size should be large enough to detect the MCID without exposing an unnecessarily large number of participants.

The close collaboration of clinicians with biostatisticians, already from the protocol development, is crucial for successfully determining and collecting the necessary sample. A poor sample size calculation or a sample size calculation not performed under the assistance of an experienced biostatistician will sink the chances for an ethics approval. Submitting a research proposal for funding without ethics approval may be impossible.

Sampling methods

The sampling methods are divided into **probability** and **non-probability methods**. The probability methods are based on selecting the sampling units randomly and are prone to less sampling error. Every unit in the population has a probability of being selected in the sample; the researchers can determine this probability. Due to the random selection of the population units, probability methods are used in randomised controlled trials. In contrast, the non-probability methods lack the element of randomness, and thus, they are prone to sampling error; however, they are more straightforward to employ. Non-probability methods are frequently encountered in observational studies. The following table presents probably the most popular sampling methods in medical research.

Probability sampling methods	**Non-probability sampling methods**
• Simple random sampling	• Convenience sampling
• Systematic random sampling	• Judgment sampling
• Stratified random sampling	• Quota sampling
	• Snowball sampling

The next sections briefly delineate the sampling methods mentioned in the table.

Probability sampling methods

Simple random sampling

In simple random sampling, each unit has **the same** probability of being selected. For the units to have the same probability of being selected, it is required that the population (from which the sampling occurs) is relatively homogeneous and small. Homogeneity requires narrower inclusion and exclusion criteria, limiting the generalisability of the results. As the name indicates, simple random sampling is the simplest probability sampling method: we decide how many units to sample from the target population, assign a unique number to each unit, and randomly select the units using a random number or lottery system. For instance, from 150 bioinformaticians with a minimum of 5 years experience (the target population) working in the company Gizmo (the sampling frame), we randomly select the names of 15 employees (the sample) using a lottery system (the sampling method).

Systematic random sampling

In systematic random sampling, the target population is initially ordered by a specific characteristic. Then, the sample is selected using a skip interval that has been pre-determined to achieve the required sample. In the previous example, we sort the 150 bioinformaticians in alphabetical order by last name. Then, we select every 10th name to obtain a sample of 15 employees.

Systematic random sampling is intuitive and can sample evenly from the target population. However, it requires relatively more effort than simple random sampling. Furthermore, systematic random sampling may be prone to hidden periodicity. For instance, most female employees were found in the skipping interval. Hence, the gender was seriously unbalanced in the collected sample. In this case, the sample fails to represent the target population.

Stratified random sampling

In stratified random sampling, the target population is initially divided into subgroups (known as *strata*) based on a characteristic (e.g., age or gender). Then, the units are sampled randomly from each stratum separately. The sample comprises all units sampled from each stratum. We may apply the same or different sampling approaches to different strata.

Stratification with varying sampling probability across strata can allow minority strata to be represented in the sample. This is an advantage of stratified random sampling. However, stratified random sampling may be cumbersome if many stratifying characteristics are considered so that sampling occurs in each combination of these characteristics. Like systematic random sampling, stratified random sampling requires relatively more effort than simple random sampling since the target population needs to be prepared into distinct strata.

Non-probability sampling methods

Convenience sampling

As the term indicates, convenience sampling relies upon convenience and access. It is frequently implemented in pilot studies. The sample is drawn from a readily available and convenient population for the researcher. With the convenience sample, the sample is not representative of the target population. Consequently, the researcher cannot generalise the total population beyond the acquired sample.

For example, if the interviewer were to survey a shopping centre early in the morning on a given day, the interviewees would be limited to those present at that given time. These interviewees would not represent the views of other people living in the area. If the survey had been conducted frequently during the day and several times per week, the sample would have more likely represented the target population.

Judgment sampling

Judgment sampling relies upon the belief that participants fit specific characteristics (based on the researcher's judgment). The researcher chooses the sample based on whom they think would be appropriate for the study. Judgment sampling is preferred when a limited number of people are eligible in the researched area. Due to the undue reliance on the researcher's subjectivity, judgment sampling introduces serious selection bias. Therefore, the collected sample does not represent the target population.

Quota sampling

Quota sampling emphasises the representation of specific characteristics. Initially, the target population is divided into subgroups, similar to stratified sampling. Then the researchers use their judgment to select subjects or units from each segment based on a pre-specified number of units. For example, the researcher may be told to sample 120 and 100 bioinformaticians with a minimum and maximum experience of 5 years, respectively. Then the researchers approach only those candidates that adhere to their subjective criteria, such as acquaintanceship. Hence, not everyone gets a chance to be selected. Like judgment sampling, quota sampling introduces a selection bias into the sample—particularly at the second stage of the method.

Snowball sampling

Snowball sampling relies upon asking existing interviewees to recommend other candidates with very similar characteristics. Namely, existing study participants recruit more subjects into the sample.

Sample size calculation

Having decided on the sampling frame and sampling method to collect the sample, the researchers need to account also for other design elements that will determine the required size of the sample. In principle, the larger the sample, the more likely it is to represent the target population and make generalisations through statistical analysis. To find out how much sample is large enough, we need to pre-specify a series of necessary parameters:

a) the **heterogeneity** of the target population (the more heterogeneous, the more sample we will need);

b) the **expected MCID** (retrieve the relevant literature or perform a pilot study);

c) the **sampling error** we allow (usually at 5%);

d) the **desired confidence** in estimating the effect size (usually at 80%).

Brainstorming

Suppose two researchers compared an intervention with a placebo on reducing body weight using the same PICOT framework. The significance level was selected at 5%. Researcher 1 found a 5 lb reduction favouring the intervention with a p-value of 0.01. This is a statistically significant result; hence, we *reject* the null hypothesis that the *average* weight may be the same in the compared groups. Researcher 2 found the same effect size favouring the intervention but failed to reject the null hypothesis because the p-value equalled 0.35. What do you think the second researcher may have done differently from the first one?

Important elements of sample size calculation

When we use a fraction of the target population to conduct a study, we inevitably have to deal with some uncertainty when interpreting the analysis results. Uncertainty is at the core of the statistical analysis. The average body weight reduction between the compared interventions will dictate which intervention was better. However, the variance of this reduction, in combination with the minimum error (we allow in our decision), will dictate whether we reject or fail to reject the null hypothesis. Rejecting the null hypothesis may allow us to recommend one of the compared interventions. **Failure to reject** the null hypothesis may require us to **revisit** the study plan and conduct that led to statistical non-significance. Therefore, we need to consider more than the effect size (here, the average body weight reduction in the compared interventions) in calculating the required sample size for our study.

Using our simple fictional example, we can construct the following decision table regarding the null hypothesis (i.e., H_0: both diets yield the same fasting insulin on average),

The decision about the null hypothesis	**Null hypothesis is**	
	true	false
Fail to reject	$1 - \alpha$	β
reject	α	$1 - \beta$

Type I error

Type I error (also known as significance level) is expressed with the Greek letter α. It refers to the probability of (falsely) rejecting the null hypothesis when it is true. In other words, a type I error is the probability of falsely concluding an association (effect measure differs from null) when this association does not exist. Hence, a type I error indicates the probability of a '*false positive*' finding. It is typically specified at 5%.

Note that type I error may be specified at 10% (or even 15%) if we know that the outcome is rare and the target population is small or considerably heterogeneous. In this case, it is more likely to observe a p-value above 5% and to fail to reject the null hypothesis even if there is a clinically significant association. Hence, a large variation in the measurements (due to the heterogeneous population) or a small sample size will result in a large p-value. Increasing the type I error is relatively easier to reject the null hypothesis than considering the default significance level at 5%. Contrariwise, the type I error may be specified at a lower level (e.g., 1%) if we know that the outcome is frequent and the target population is large. In this case, it is more likely to observe a p-value below 5% and reject the null hypothesis even if there is a clinically non-significant association. Hence, a smaller variation in the measurements or a large sample size will result in a small p-value. Decreasing the type I error becomes relatively difficult to reject the null hypothesis than considering the default significance level at 5%.

Type II error

Type II error is expressed with the Greek letter β, which refers to the probability of (falsely) failing to reject the null hypothesis when it is false. In other words, the type II error is the probability of falsely concluding no association (effect measure equals null) when this association does exist. Hence, type II error indicates the probability of a '*false negative*' finding. It is typically specified at 20%.

Power

Power is the probability of (correctly) rejecting the null hypothesis when it is false. In other words, power is the probability of correctly concluding an association (effect measure differs from null) when this association does exist. In short, power is one minus the type II error (1 – β), typically specified at 80%.

The following table summarises the interplay of MCID, variation in the measurements, type I error and power with the sample size:

Expected element		Required sample size
A small effect size (via the MCID)	↓	↑
A large variation in the measurements	↑	
Type I error below 5%.	↓	
A power above 80%	↑	
A large effect size (via the MCID)	↑	↓
A small variation in the measurements	↓	
Type I error above 5%.	↑	
A power below 80%	↓	

Conducting a pilot study or retrieving the relevant published literature is necessary to decide the suitable MCID and measurement variation for our study.

Brainstorming

Would you keep enrolling patients into your study until the p-value is below 5%?

Verdict

No, because such a strategy is unethical and indicates a complete lack of planning. You enrol as many participants as the sample size calculation has indicated. A study with a smaller sample size than required has insufficient power to detect the association between the outcome and the interventions (or exposures). The observed association may be clinically significant; however, it will be statistically non-significant because the p-value is more likely to exceed 5%. On the contrary, a study with a larger sample size than required exposes unnecessarily more participants than it should (this is very problematic when the intervention triggers adverse events). Such a study also has excessive power to detect the association between the outcome and the interventions (or exposures). The observed association may be clinically non-significant; however, it will be statistically significant because the p-value is more likely to be less than 5%.

Further brainstorming …

Would you publish your study if you failed to reject the null hypothesis?

Verdict

Definitely! Regardless of statistical significance, publishing the study is a **good scientific practice** and advances our knowledge in the research field. Unfortunately, this is not the norm. Not publishing your research is another form of **research waste** because resources and time have been consumed without making the 'product' available to the interested end-user. There is extensive empirical evidence on unpublished research due to non-statistically significant results and research that has published only the statistically significant outcomes. Unfortunately, these strategies are prevalent in the research community; they contribute to **publication bias** and **selective reporting outcome bias**, respectively. Both strategies distort the research ecosystem and jeopardise end-users of the published literature. Meta-research can detect and investigate the serious implications of publication bias in any research field. Song et al. (2013) offer a review of publication bias and how we can measure and avoid it.

A simple example: two independent groups t-test

Suppose we want to conduct a prospective cohort study investigating whether a veggie-based or meat-based diet may improve the HbA1c level in adults with diabetes. First, we perform a pilot study using a convenience sample. We measure the change from baseline (CfB) in HbA1c in 200 and 250 participants receiving a veggie-based and meat-based diet, respectively. Namely, for each participant, we measure the difference in HbA1c between the endpoint and baseline. Then we average these differences in each group to obtain the mean CfB. The table presents the mean and standard deviation of the CfB in HbA1c per group (results are fictional and hopefully clinically plausible):

Parameter	Veggie-based diet	Meat-based diet
Mean CfB	-5%	-10%
Standard deviation of CfB	25%	14%

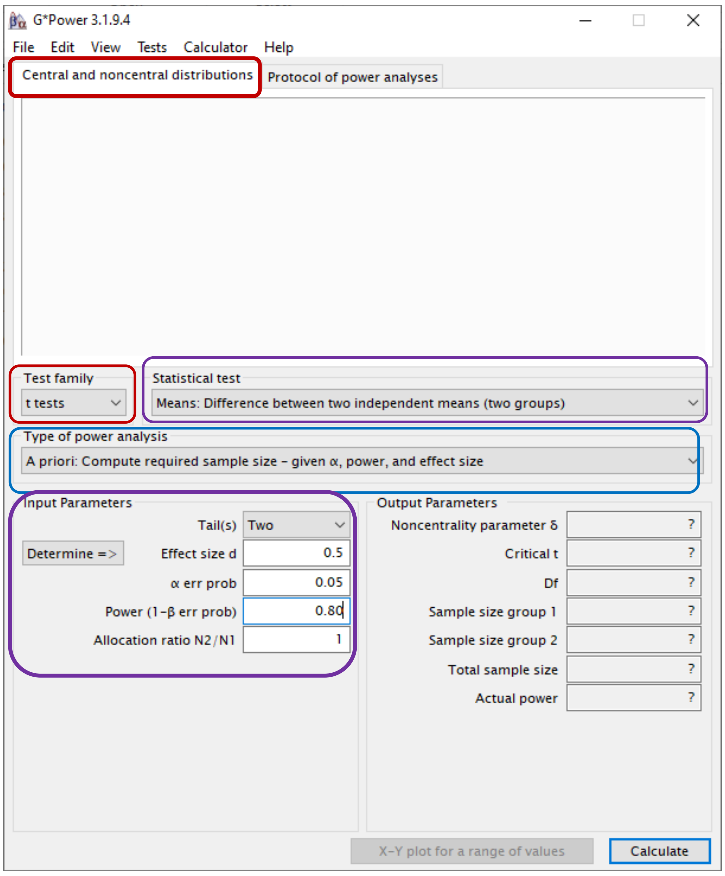
Both diets seem to have improved the HbA1c level by the end of the study because they yielded a negative average change from baseline. However, the meat-based diet achieved a greater reduction than the veggie-based diet. We also observed greater variability in the veggie-based group than in the meat-based diet. We use the results from the table to proceed with the sample size calculation while assuming a type I error at 5% and power at 80%.

The null hypothesis is H_0: $\mu_v = \mu_m$ (or equivalently, $\mu_v - \mu_m = 0$), and the alternative hypothesis is H_a: $\mu_v \neq \mu_m$ (or equivalently, $\mu_v - \mu_m \neq 0$) with μ_v and μ_m being the population mean CfB following a (v)eggie-based and (m)eat-based diet, respectively. We do not give any direction in the alternative hypothesis because we do not have any prior knowledge of the direction of the effect.

G*Power sample size calculator

We use the G*Power sample size calculator (version 3.1.9.4) developed by Faul and colleagues from the Heinrich Heine Universität Düsseldorf. Visit their site to download the software and refer to their manual, which contains all available statistical tests to perform sample size calculations.

For our example, under **Central and noncentral distributions**, we select **t tests** from the **Test family**, **Means: Difference between two independent means (two groups)** from the **Statistical test**, and **A prior: Compute required sample size – given α, power, and effect size** from **Type of power analysis** as pinpointed below. Under **Input parameters**, select **Two** under **Tail(s)** on the main window because we have not specified any direction on the alternative hypothesis. Then, insert the type I error (**α err prob**), power (**Power (1-β err prob)**) and **Allocation ratio N2/N1** equal to 1, which implies an equal sample size in the compared groups. For the moment, ignore the value in **Effect size d**. Click the button **Determine** to open a new smaller window on the right.


G*Power 3.1.9.4
File
Edit
View
Tests
Calculator
Help
Central and noncentral distributions
Protocol of power analyses
Determine =>
X-Y plot for a range of values
Calculate

Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

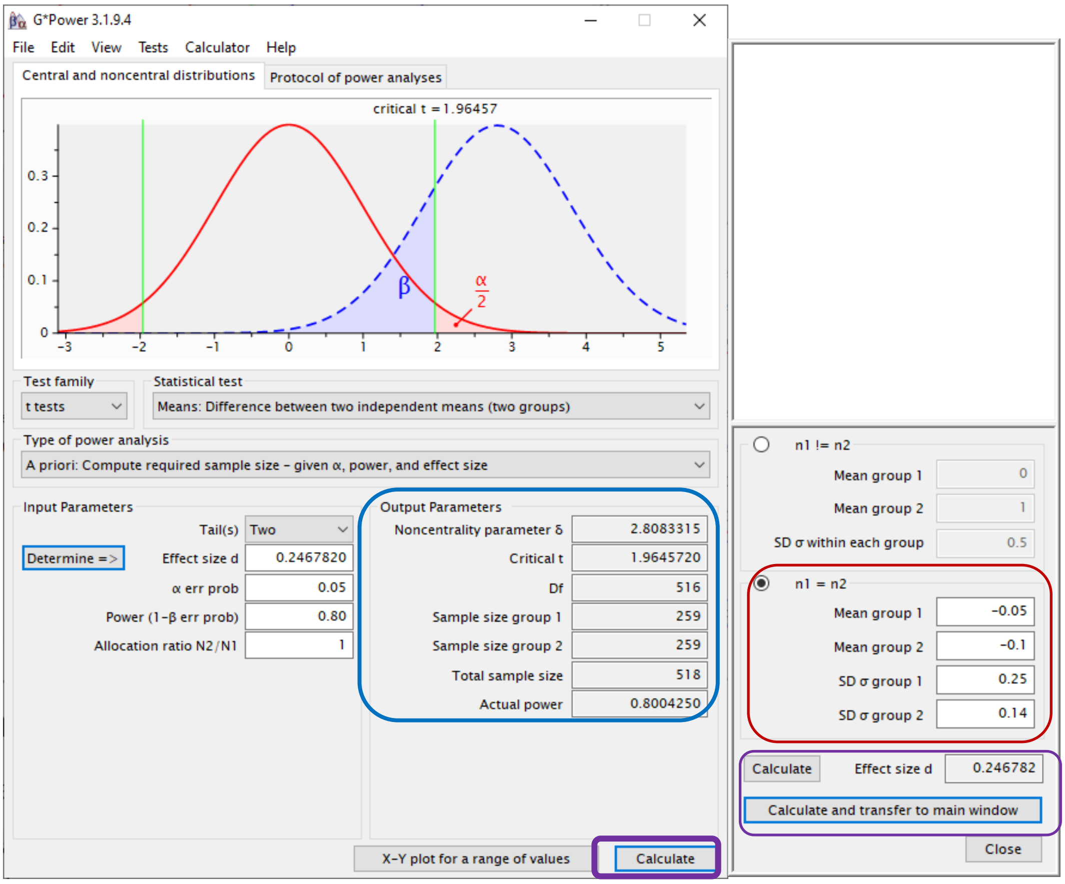
Input Parameters

Tail(s): Two
Effect size d: 0.5
α err prob: 0.05
Power (1-β err prob): 0.80
Allocation ratio N2/N1: 1

Output Parameters

Noncentrality parameter δ: ?
Critical t: ?
Df: ?
Sample size group 1: ?
Sample size group 2: ?
Total sample size: ?
Actual power: ?

In the new smaller window (see picture below), we select the option **n1 = n2** to calculate equal sample sizes in the compared groups. Insert -0.05 and -0.1 for the **Mean group 1** (veggie-based diet) and **Mean group 2** (based diet), respectively, and the corresponding values for **SD σ group 1** and **SD σ group 2** (0.25 and 0.14, respectively). Clicking the button **Calculate** gives an **Effect size d** equal to 0.246782. Then, click the button **Calculate and transfer to main window**. Finally, click the button **Calculate** at the bottom right of the main window to obtain the **Total sample size** under **Output parameters**. For our prospective cohort study, we will need to recruit 518 participants in total or 259 participants per group.


G*Power 3.1.9.4
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical t = 1.96457
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
4
5
β
α/2
Determine =>
X-Y plot for a range of values
Calculate
Calculate and transfer to main window
Close

Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s): Two
Effect size d: 0.2467820
α err prob: 0.05
Power (1-β err prob): 0.80
Allocation ratio N2/N1: 1

Output Parameters

Noncentrality parameter δ: 2.8083315
Critical t: 1.9645720
Df: 516
Sample size group 1: 259
Sample size group 2: 259
Total sample size: 518
Actual power: 0.8004250

○ n1 != n2
Mean group 1: 0
Mean group 2: 1
SD σ within each group: 0.5

● n1 = n2
Mean group 1: -0.05
Mean group 2: -0.1
SD σ group 1: 0.25
SD σ group 2: 0.14

Calculate
Effect size d: 0.246782

R-package' stats'

Initially, install and load the package

```
install.packages("stats")
library("stats")
```

We will use the `power.t.test` function to perform the sample size calculation via the two independent groups t-test. We will use the following arguments of the function:

```
power.t.test(n = NULL,
             delta = 0.05,
             sd = 0.203,
             sig.level = 0.05,
             power = 0.80,
             type = "two.sample",
             alternative = "two.sided")
```

Refer to the function's documentation to access all available arguments (type `help(power.t.test)` in the R console). The argument `n` refers to the sample size. Since we do not know the sample size, we insert`n = NULL`. The argument `delta` refers to the difference in the mean of the compared groups. `delta` can be specified based on published literature or obtained through a pilot study. The argument `sd` is the pooled standard deviation when we assume the same population variance in the compared groups (see the script on Effect Measures, Week 04). `sig.level` and `power` refer to the type I error and the power, respectively. Finally, `type` and `alternative` imply a two independent groups t-test and an alternative hypothesis without direction.

In our example, the effect measure is the mean difference (MD) and is calculated as follows:

$$MD = \bar{x}_v - \bar{x}_m = -0.05 - (-0.10) = 0.05$$

with $\bar{x}_v$ and $\bar{x}_m$ being the calculated mean CfB in the veggie-based and meat-based group. The effect size (`delta`) of 0.05 favours the meat-based diet.

Assuming the same population variance in the compared groups, we calculate the pooled standard deviation (`sd`), which is the square root of the weighted mean of the variances of the compared groups weighted by their sample size minus one:

$$SE = \sqrt{\frac{(n_v - 1) \times V_v + (n_m - 1) \times V_m}{n_v + n_m - 2}}$$

$$= \sqrt{\frac{(200 - 1) \times 0.25^2 + (250 - 1) \times 0.14^2}{200 + 250 - 2}} = 0.197$$

The t-test is relatively robust to violations of this assumption if we assumed equal sample sizes (Chapter 20 in the manual of G*Power). In this case, the formula above coincides with the formula of Cohen (Chapter 20 in the manual of G*Power), and it is the formula used in G*Power:

$$SE = \sqrt{\frac{V_v + V_m}{2}} = \sqrt{\frac{0.25^2 + 0.14^2}{2}} = 0.203$$

The `power.t.test` function yields the same sample size per group with the G*Power software.

```
     Two-sample t test power calculation

              n = 259.7195
          delta = 0.05
             sd = 0.203
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

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

Using `sd = 0.197` results in `n = 244.6498` for each group.

Impact of power and type I error

We will investigate how different values for power and type I error affects the total sample size. We create a sequence of values for type I error in the range [0.001, 0.15] and a sequence of values for power in the range [0.80, 0.95],

```
type.one.err <- seq(0.001, 0.15, 0.005) # increment of 0.005
power.range <- seq(0.8, 0.95, 0.05)     # increment of 0.05
```

We use the `power.t.test` function to obtain the total sample size for all possible combinations of `type.one.err` and `power.range` values. Results return as a matrix (`size.err`) with number of rows and columns equal to the length of `type.one.err` and `power.range`, respectively,

```
## A matrix with the total sample size
size.err <- matrix(NA, nrow = length(type.one.err),
                       ncol = length(power.range))
rownames(size.err) <- type.one.err
colnames(size.err) <- power.range

## Calculate the total sample size for all possible combinations of
## type.one.err and power.range values
for(i in 1:length(type.one.err)) {
  for(j in 1:length(power.range)) {
    size.err[i, j] <- 2*power.t.test(n = NULL, delta = 0.05,
                                     sd = 0.203,
                                     sig.level = type.one.err[i],
                                     power = power.range[j],
                                     type = "two.sample",
                                     alternative = "two.sided")$n
  }
}

head(size.err) # View the first six rows of the matrix

> head(size.err)
            0.8      0.85       0.9      0.95
0.001 1131.2247 1239.8805 1383.7044 1611.4471
0.006  853.2676  947.9802 1074.2602 1276.0005
0.011  758.4269  847.8708  967.5218 1159.4368
0.016  699.5703  785.5774  900.9040 1086.4096
0.021  656.7406  740.1546  852.2187 1032.8889
0.026  623.0331  704.3464  813.7673  990.5198
```

Next, install and load the packages' reshape2' and 'ggplot2'

```
install.packages(c("reshape2", "ggplot2"))
library("reshape2")
library("ggplot2")
```

We use the function `melt` from the 'reshape2' package to properly reshape `size.err` into a data-frame with as many rows as the number of possible combinations of`type.one.err` and `power.range` values,

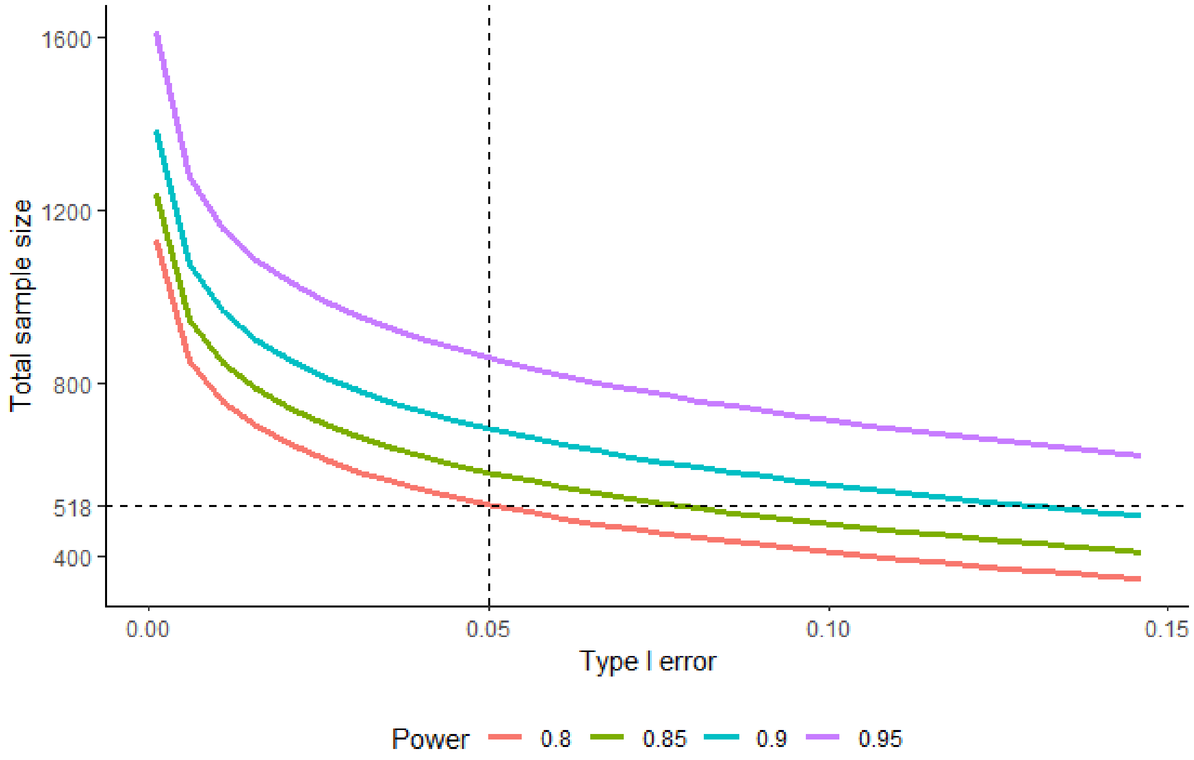
```
data.plot <- melt(size.err)
colnames(data.plot) <- c("type.one.error", "power", "total.n")
```

```
head(data.plot)  # View the first six rows of the data-frame

> head(data.plot)
  type.one.error power   total.n
1          0.001   0.8 1131.2247
2          0.006   0.8  853.2676
3          0.011   0.8  758.4269
4          0.016   0.8  699.5703
5          0.021   0.8  656.7406
6          0.026   0.8  623.0331
```

To visualise the results, we use the 'ggplot2' package (type `help(ggplot2)` in the R console):

```
ggplot(data.plot, aes(x = type.one.error, y = total.n)) +
  geom_line(aes(color = as.factor(power)), size = 1.2) +
  geom_hline(yintercept = 518, linetype = 2) +
  geom_vline(xintercept = 0.05, linetype = 2) +
  scale_y_continuous(breaks = c(400, 518, 800, 1200, 1600)) +
  labs(x = "Type I error", y = "Total sample size", color = "Power") +
  theme_classic() +
  theme(legend.position = "bottom")
```



The dotted vertical and horizontal lines refer to the total sample size of our fictional example for power and type I error equal to 0.80 and 0.05, respectively. As the power increases, the total sample size increases as well. On the contrary, as the type I error increases, the sample size decreases exponentially.