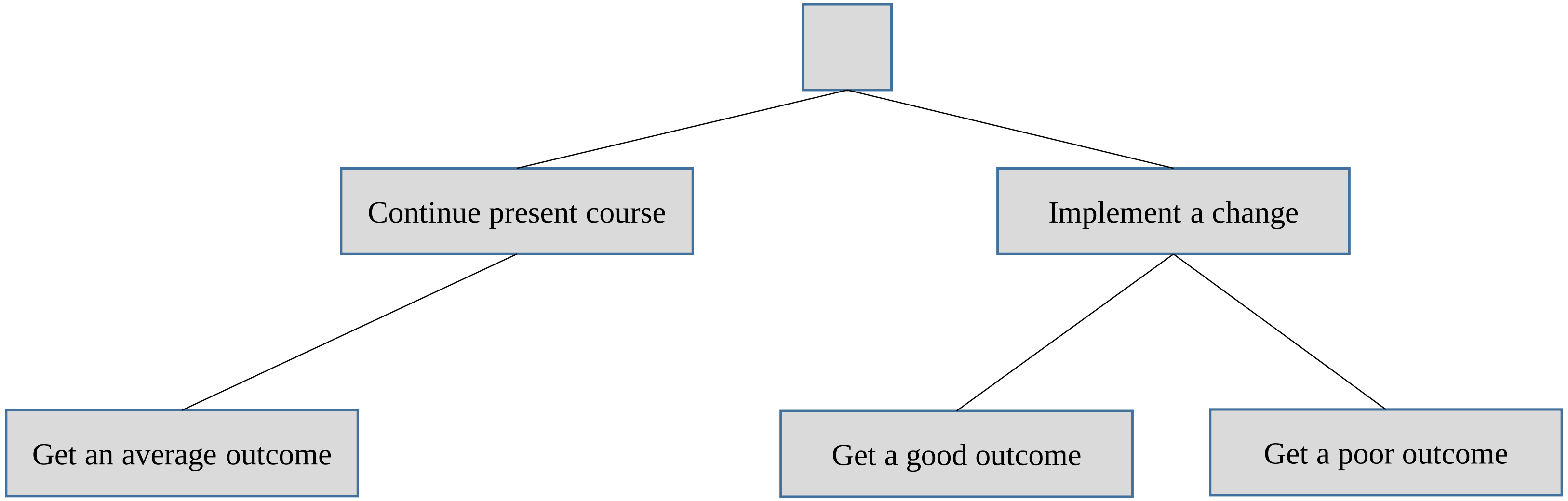
4.3 How to collect and get data

The focus is on identifying which kinds of data collection will have the most favourable impact on the quality of the actions and recommendations supported by the data analysis.

Such an analysis is typically done by a decision tree.

4.3 How to collect and get data

Example of a decision tree:



Continue present course

Implement a change

Get an average outcome

Get a good outcome

Get a poor outcome

4.3 How to collect and get data

Thus, if the chance of getting a good outcome is high enough, it is better to implement a change.

Otherwise, the change will not be implemented.

p: probability of getting a good outcome, if a change is made

U: value (utility) of making a change with good outcome

L: value (utility) of making a change with bad outcome

u: value (utility) of continuing with the present course

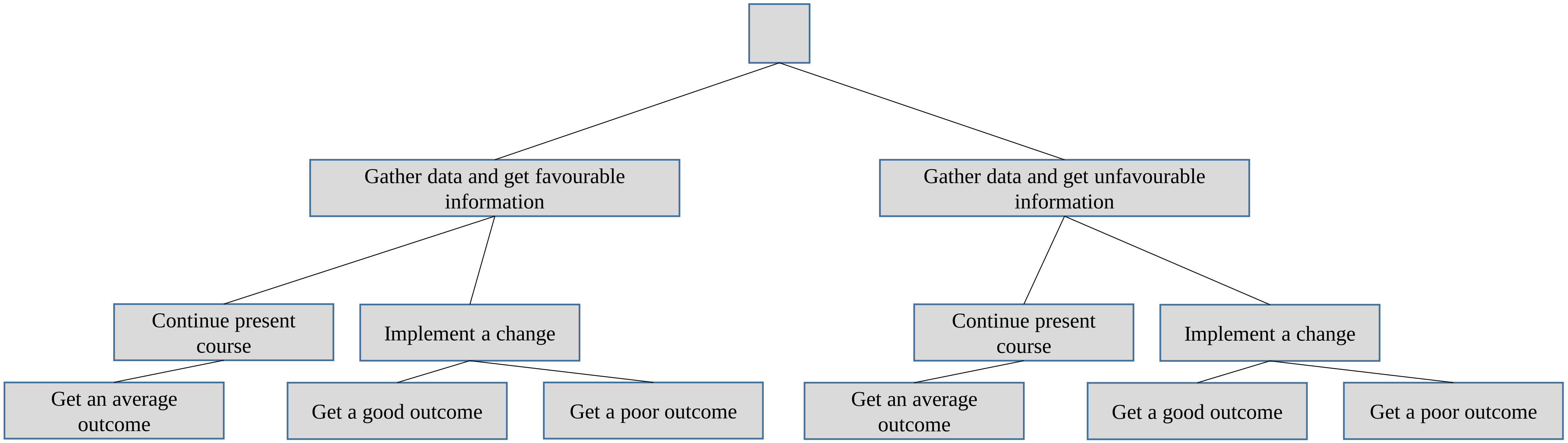
Change will be made if: $pU + (1-p)L = p(U-L) + L > u \quad (1)$

4.3 How to collect and get data

But we can gather data and make the decision based on the results of the data gathering exercise.

4.3 How to collect and get data

Example of a decision tree with gathering data



4.3 How to collect and get data

When would information be valuable?

Favourable information is leading to a substantial change in the probability of getting a good outcome.

If the change is substantial, we will make a change.

u*: value (utility) of implementing a change, given getting favourable information

u: value (utility) of continuing the present course, given getting unfavourable information

q: probability getting favourable information

Utility if we decide to gather data: qu* + (1-q)u (2)

4.3 How to collect and get data

Comparing (1) (no data gathering) with (2) (data gathering), shows that we have improved u to qu* + (1-q)u.

Since u* > u, the utility is improved if we are collecting data.

Thus:

The value of information is non-negative.

4.3 How to collect and get data

But we have seen, that in reality we have costs.

Suppose that these costs are reducing our value (utility) by a factor d.

Utility, if we are collecting information:

$$d(qu^* + (1-q)u)$$

4.3 How to collect and get data

p: probability of getting a good outcome, if a change is made

p*: probability of getting a good outcome, if getting favourable information

p**: probability of getting a good outcome, if getting unfavourable information

q: probability getting favourable information

$$p = qp^* + (1-q)p^{**}$$

Utility of making a change, see (1): $pU + (1-p)L$

Utility of making a change, given a favourable outcome:

$$u^* = p^*U + (1-p^*)L = p^* (U-L) + L \qquad (3)$$

4.3 How to collect and get data

Thus, comparing (1) and (3) it shows, that u* depends on how much p* differs from p.

Thus, the degree of change of p with the new information depends on

- the confidence we have in the original value p and,
- the impact of the data

If new information is telling us something unexpected, i.e. the favourable outcome, how much will our initial belief change?

4.3 How to collect and get data

Why we had a look at the theory of utility functions?

Because the decision to collect data and which data

- is very subjective
- is based on our beliefs
- depends how confident we are in the original probability p

And you must be aware of that in such a situation of decision-making about data gathering.

4.3 How to collect and get data

After the identification of the variables one should collect, the next step is the actual data collection.

The data collection proceeds in five steps:

1. The sample design: determination how to identify the subjects / items
2. The sample plan: determination how many subjects / items to identify
3. Determination the questions to be asked
4. Granularity of the experiment: Determination of the possible answers to the questions
5. Determination of the control group

4.3 How to collect and get data

Sample design:

Choosing the sample out of a population which is of interest.

Typically, there are two kind of sampling:

- (Simple) Random sampling
- Stratified random sampling

4.3 How to collect and get data

(Simple) Random sampling:

To each subject / item the same chance is allocated to be selected for the sample.

Advantage: it is unbiased

Disadvantage: if events / subjects / items of interest are very unlikely, then it is also unlikely to have them in a (simple) random sample.

$\Rightarrow$ It may be advantageous to bias the sampling towards sampling those subjects which are for interest.

Nevertheless, the analysis has to take into account this bias and formulas have to be corrected for this bias.

4.3 How to collect and get data

Stratified sampling:

1) The population is divided into homogenous subpopulations (strata) [stratification]
2) The strata have to be mutually exclusive
3) To each stratum sampling is applied
4) Correction for bias have to be implemented

4.3 How to collect and get data

Example: Stratified sampling

In hospitals for each stationary treatment a code which is representing the treatment and many subcodes for subtreatments have to be allocated, the so-called SwissDRG codes. For each treatment and code a cost figure is allocated. Each hospital has the duty that an external expert has to test the correctness of these SwissDRG cost code allocations, performing certain statistics and submitting a report.

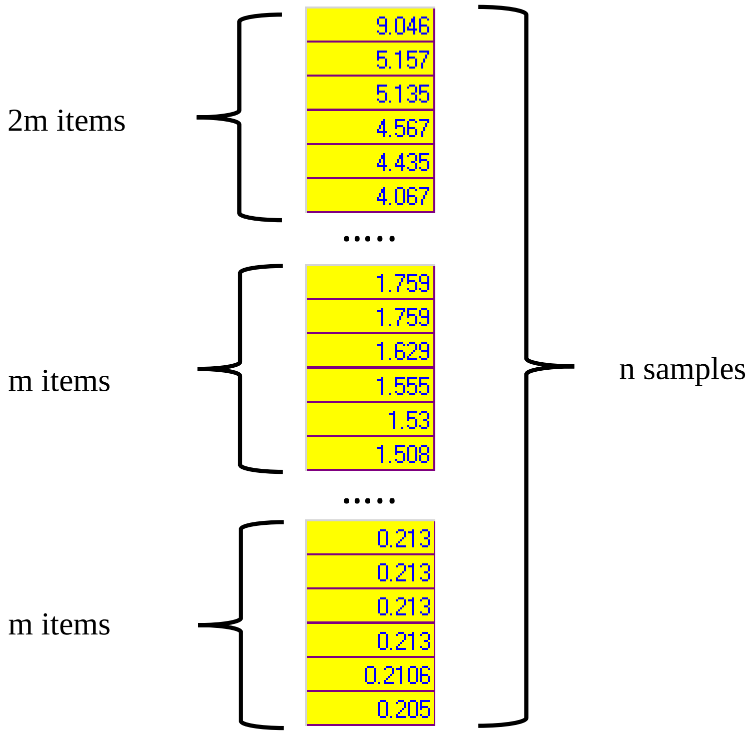
The sampling is done by (a variation of) stratified sampling.

How does this work?

4.3 How to collect and get data

Example: Stratified sampling (cont’d)

1) The cost figures per each SwissDRG code are ordered according their size
2) The population is divided in n classes
3) From each class a sample of m items are taken
4) Except in the class with the highest amounts 2m items are sampled



2m items

9.046
5.157
5.135
4.567
4.435
4.067

.....

m items

1.759
1.759
1.629
1.555
1.53
1.508

.....

m items

0.213
0.213
0.213
0.213
0.2106
0.205

n samples

4.3 How to collect and get data

Sample design:

Each subject / item has different characteristics so-called covariates.

To determine how such a characteristic is affecting the result one could change each and every such stand-alone factor and determining the impact on the result.

But this is inefficient.

It is more efficient to change several of these factors simultaneously.

4.3 How to collect and get data

Sample design:

Thus, typically *full factorial design* is used:

It gives the possibility to identify the impact of each factor as well as of possible interactions between the factors.

Examples:

- Regression analysis
- ANOVA