An\_Introduction\_Statistics\_Python\_c05

**Part II Distributions and Hypothesis Tests**

This part of the book moves the focus from *Python* to statistics.

The first chapter serves to define the statistical basics, like the concepts of *popula- tions* and *samples*, and of *probability distributions*. It also includes a short overview of *study design*. The design of statistical studies is seriously underestimated by most beginning researchers: faulty study design produces garbage data, and the best analysis cannot remedy those problems (“Garbage in–garbage out”). However, if the study design is good, but the analysis faulty, the situation can be fixed with a new analysis, which typically takes much less time than an entirely new study.

The next chapter shows how to characterize the position and the variability of a distribution, and then uses the normal distribution to describe the most important *Python* methods common to all distribution functions. After that, the most important discrete and continuous distributions are presented.

The third chapter in this part first describes a typical workflow in the analysis of statistical data. Then the concept of *hypothesis tests* is explained, as well as the different types of errors, and common concepts like *sensitivity* and *specificity.*

The remaining chapters explain the most important hypothesis tests, for con- tinuous variables and for categorical variables. A separate chapter is dedicated to survival analysis (which also encompasses the statistical characterization of material failures and machine breakdowns), as this question requires a somewhat different approach than the other hypothesis tests presented here. Each of these chapters also includes working *Python* sample code (including the required data) for each of the tests presented. This should make it easy to implement the tests for different data sets.

**Chapter 5 Background**

This chapter briefly introduces the main concepts underlying the statistical analysis of data. It defines discrete and continuous probability distributions, and then gives an overview of various types of study designs.

**5.1 Populations and Samples**

In the statistical analysis of data, we typically use data from a few selected *samples* to draw conclusions about the *population* from which these samples were taken. Correct *study design* should ensure that the sample data are representative of the population from which the samples were taken.

The main difference between a *population* and a *sample* has to do with how observations are assigned to the data set (see Fig. 5.1).

**Population** Includes all of the elements from a set of data.

**Sample** Consists of one or more observations from the population.

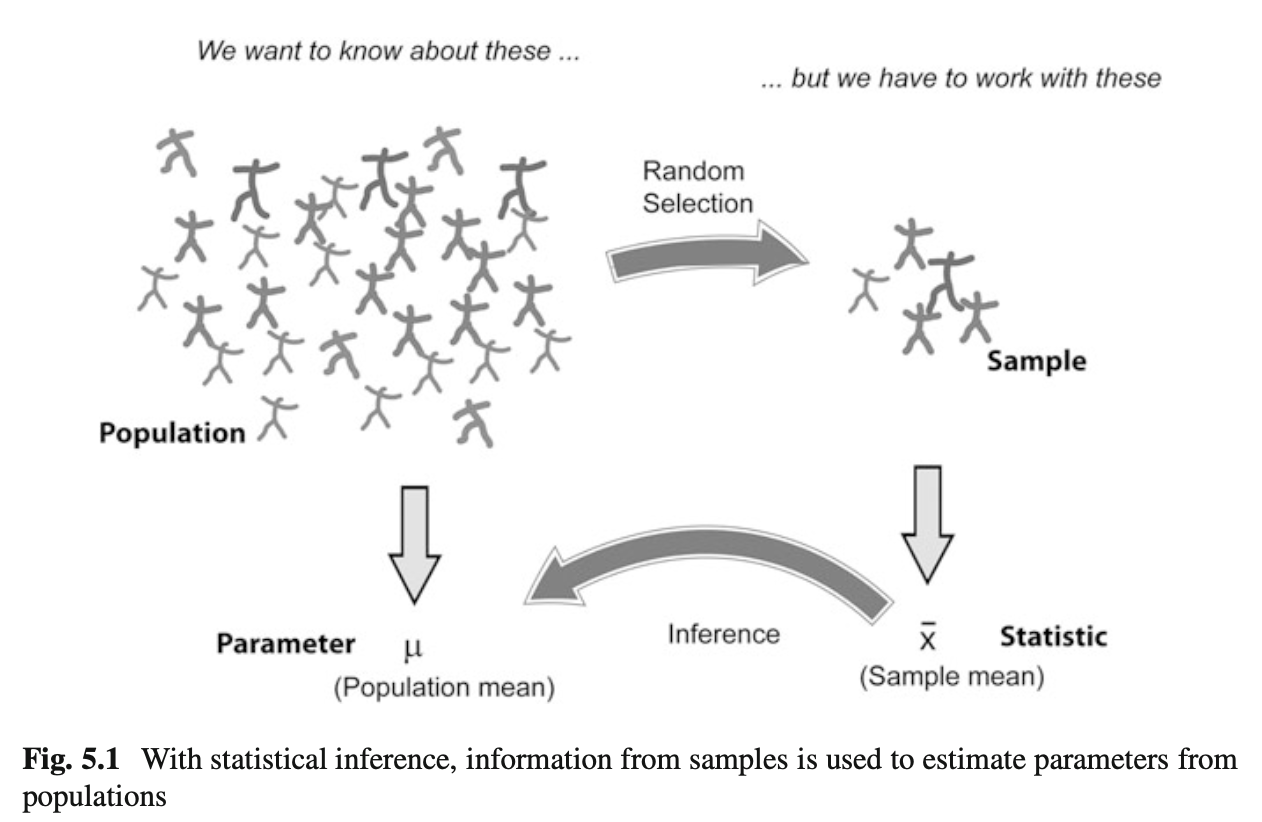
More than one sample can be derived from the same population.

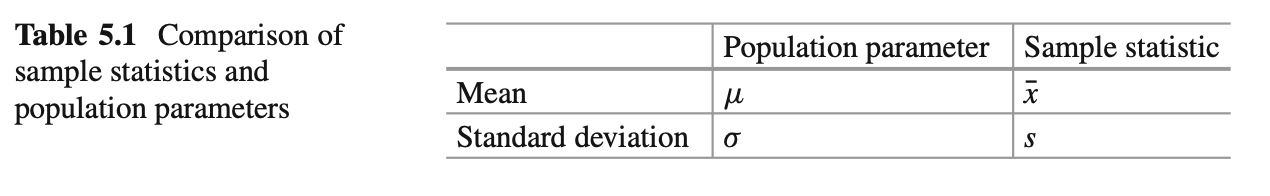
When estimating a *parameter* of a *population*, e.g., the weight of male Euro- peans, we typically cannot measure all subjects. We have to limit ourselves to investigating a (hopefully representative) random *sample* taken from this group. Based on the *sample statistic*, i.e., the corresponding value calculated from the sample data, we use *statistical inference* to find out what we know about the corresponding parameter in the population.

**Parameter** Characteristic of a population, such as a mean or standard deviation. Often notated using Greek letters.

**Statistic** A measurable characteristic of a sample. Examples of statistics are:

• the mean value of the sample data





* the range of the sample data
* deviation of the data from the sample mean

**Sampling distribution** The probability distribution of a given statistic based on a random sample.

**Statistical inference** Enables you to make an educated guess about a popu- lation parameter based on a statistic computed from a sample randomly drawn from that population.

Examples of parameters and statistics are given in Table 5.1. Population param- eters are often indicated using Greek letters, while sample statistics typically use standard letters.

**5.2 Probability Distributions**

The mathematical tools to describe the distribution of numerical data in populations and samples are *probability distributions*.

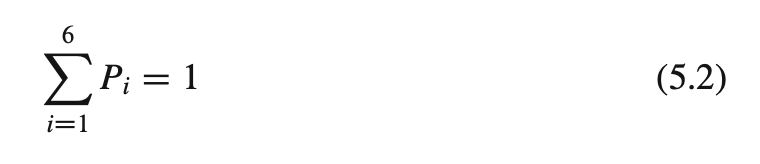
***5.2.1 Discrete Distributions***

A simple example of a discrete probability distribution is the game of throwing dice: for each of the numbers *i* D 1;:::;6, the probability that at the throw of a die the side showing the number *i* faces upward, *Pi*, is

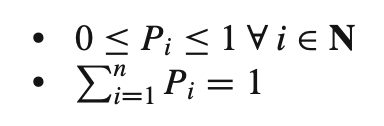
*Pi = ⅙, i=1,...6*  ( 5 . 1 )

The set of all these probabilities f*Pi*g makes up the *probability distribution* for rolling dice.

Note that the smallest possible value for *Pi* is 0. And since one of the faces of the die has to turn up at every throw of the dice, we have



Generalizing this, we can say that a discrete probability distribution has the following properties



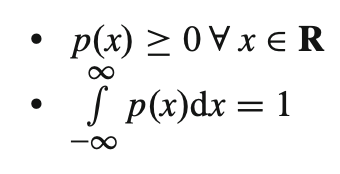
For a given discrete distribution, the *Pi* are called the *probability mass function* (*PMF*) of that distribution.

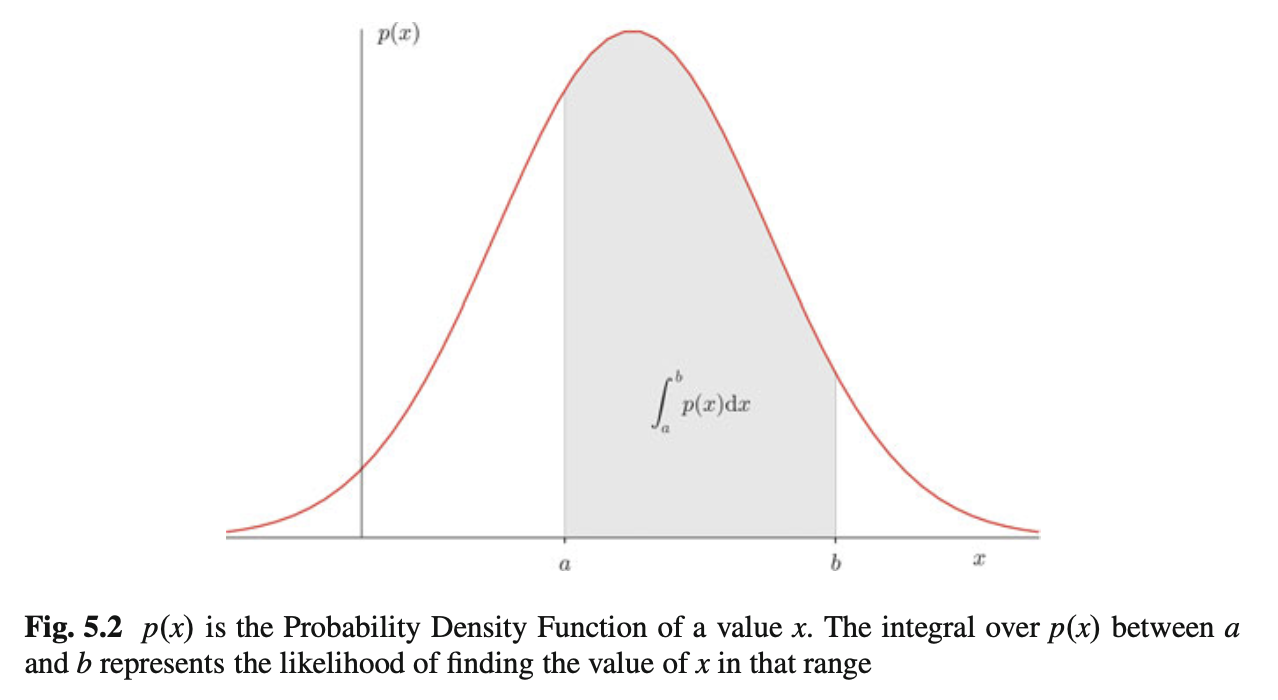
***5.2.2 Continuous Distributions***

Many measurements have an outcome that is not restricted to discrete integer values. For example, the weight of a person can be any positive number. In this case, the curve describing the probability for each value, i.e., the *probability distribution*, is a continuous function, the *probability density function* (*PDF*).

The PDF, or *density* of a continuous random variable, is a function that describes the relative likelihood of a random variable *X* to take on a given value *x*. In the mathematical fields of probability and statistics, a *random variate x* is a particular outcome of a *random variable X*: the random variates which are other outcomes of the same random variable might have different values.

Since the likelihood of finding any given value cannot be less than zero, and since the variable has to have some value, the PDF *p*.*x*/ has the following properties (Fig. 5.2):

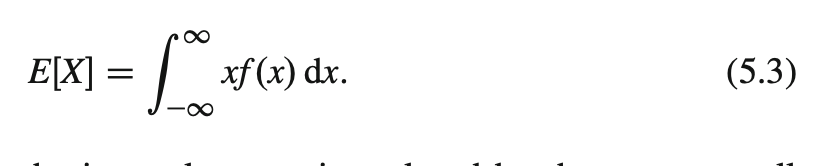




***5.2.3 Expected Value and Variance***

**a) Expected Value**

The PDF also defines the *expected value* E[X] of a continuous distribution of X: Z1



For discrete distributions, the integral over *x* is replaced by the sum over all possible values:

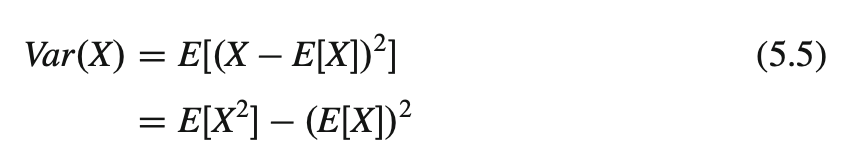


where *xi* represents all possible values that the measured variable can have.

The *expected value* is a function of the probability distribution of the observed value in our *population*. The *sample mean* of our *sample* is the observed mean value of our data. If the experiment has been designed correctly, the sample mean should converge to the expected value as more and more samples are included in the analysis.

**b) Variance**

The *variability* of the data is characterized by the *variance* of the data:



**5.3 Degrees of Freedom**

The concept of degrees of freedom (DOF), which in mechanics appears to be crystal clear, is harder to grasp for statistical applications.

In mechanics, a particle which moves in a plane has “2 DOF”: at each point in time, two parameters (the *x*=*y*-coordinates) define the location of the particle. If the particle moves about in space, it has “3 DOF”: the *x*=*y*=*z*-coordinates.

In statistics, a group of *n* values has *n* DOF. If we look only at the shape of the distribution of the values, we can subtract from each value the sample mean. Then, the remaining data only have *n* 􏰌 1 DOF. (This is clearest for *n* D 2: if we know the mean value and the value of *sample*1, then we can calculate the value of *sample*2 by v*al*2 D2􏰎*mean*􏰌v*al*1.)

The case becomes more complex when we have many groups. For example, in Sect. 8.3.1, there is an example with 22 patients, divided into 3 groups. In the *analysis of variance* (*ANOVA*), the DOFs in this example are divided as follows:

* 1 DOF for the total mean value.
* 2 DOF for the mean value of each of the three groups (remember, if we know the  
     
  mean values of two groups *and* the total mean, we can calculate the mean value  
     
  of the third group).
* 19 DOF (D 22 􏰌 1 􏰌 2) are left for the residual deviations from the group means.

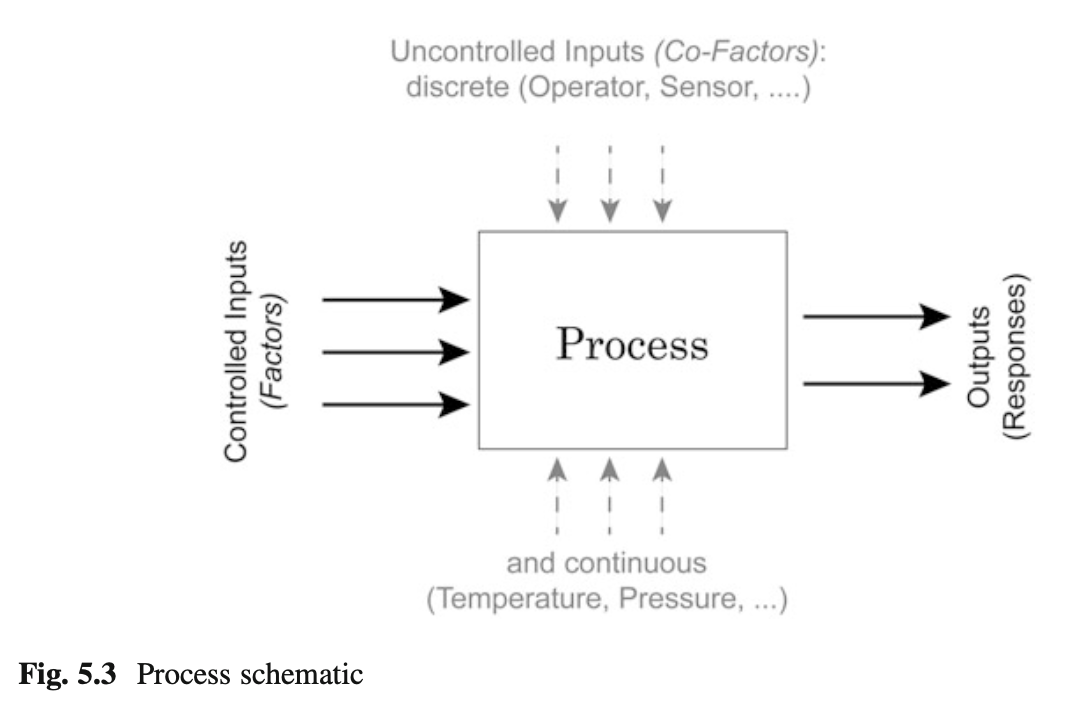
**5.4 Study Design**

The importance of a good study design has been demonstrated recently by an investigation showing the effect of the introduction of the *clinicaltrials.gov* registry (Kaplan and Irvin 2015): A 1997 US law mandated the creation of the registry, requiring researchers from 2000 onwards to record their trial methods and outcome measures *before* collecting data. Kaplan et al looked at studies evaluating drugs or dietary supplements for the treatment or prevention of cardiovascular disease. They found that before the introduction of *clinicaltrials.gov*, 57 % of the studies showed a positive outcome, while after the introduction, this number was reduced dramatically to only 8 %. In other words, without rigorous study design, there is a significant bias towards getting the result that you hope for.

***5.4.1 Terminology***

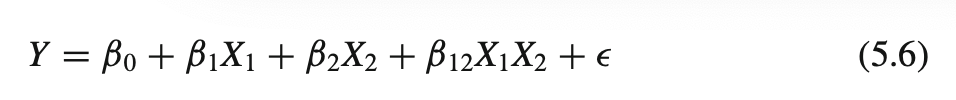
In the context of study design, various terminology can be found (Fig. 5.3):

* The controlled inputs are often called *factors* or *treatments*.
* The uncontrolled inputs are called *cofactors*, *nuisance factors*, or *confoundings*.



The term *covariate* refers to a variable that is possibly predictive of the outcome being studied, and can be a factor or a cofactor.

When we try to model a process with two inputs and one output, we can formulate a mathematical model for example as



The terms with the single *X* (ˇ1;ˇ2) are called *main effects*, and the terms with multiple *X* (ˇ12) *interaction terms*. And since the ˇ parameters enter the equation only linearly, this is referred to as a *general linear model*. The 􏰃 are called *residuals*, and are expected to be distributed approximately normally around zero if the model describes the data correctly.

***5.4.2 Overview***

The first step in the design of a study is the explicit clarification of the goal of the study. Do we want to

1. Compare two or more groups, or one group to a fixed value?
2. Screen the observed responses to identify factors/effects that are important?
3. Maximize or minimize a response (variability, distance to target, robustness)?
4. Develop a regression model to quantify the dependence of a response variable on  
      
   the process input?

The first question leads to a *hypothesis test*. The second one is a *screening investigation*, where one must be watchful of artifacts if the factors in the model are not completely independent. The third task is an *optimization problem*. And the last one brings us into the realm of *statistical modeling*.

Once we have determined *what* we want to do, we have to decide *how* we want to do this. Either *controlled experiments* or *observations* can be used to obtain the necessary data. In a controlled experiment we typically try to vary only a single parameter, and investigate the effects of that parameter on the output.

***5.4.3 Types of Studies***

**a) Observational or Experimental**

In an *observational study*, the researcher only collects information, but does not interact with the study population. In contrast, in an *experimental study* the researcher deliberately influences events (e.g., treats the patient with a new type of medication) and investigates the effects of these interventions.

**b) Prospective or Retrospective**

In a *prospective study*, the data are collected from the beginning of the study. In contrast, a *retrospective study* takes data acquired from previous events, e.g., routine tests done at a hospital.

**c) Longitudinal or Cross-Sectional**

In *longitudinal* investigations, the researcher collects information over a period of time, maybe multiple times from each patient. In contrast, in *cross-sectional* studies individuals are observed only once. For example, most surveys are cross-sectional, but experiments are usually longitudinal.

**d) Case–Control and Cohort studies**

In a *case–control study*, first the patients are treated, and then they are selected for inclusion in the study, based on certain criteria (e.g., whether they responded to a certain medication). In contrast, in a *cohort study*, subjects of interest are selected first, and then these subjects are studied over time, e.g., for their response to a treatment.

**e) Randomized Controlled Trial**

The gold standard for experimental scientific clinical trials, and the basis for the approval of new medications, is the *randomized controlled trial*. Here bias is avoided by splitting the subjects to be tested into an *intervention group* and a *control group*. Group allocation is *random*.

In a designed experiment, there may be several conditions, called *factors*, that are controlled by the experimenter. By having the groups differ in only one aspect, the factor *treatment*, one should be able to detect the effect of the treatment on the patients.

Through randomization, confoundings should be balanced across the groups.

**f) Crossover Studies**

An alternative to randomization is the *crossover* design of studies. A crossover study is a longitudinal study in which subjects receive a sequence of different treatments. Every subject receives every treatment. (The subject “crosses over” from one treatment to the next.) To avoid causal effects, the sequence of the treatment allocation should be randomized.

For example, in an investigation that tests the effect of standing and sitting on the concentration of subjects, each subject performs both the execution of tasks while standing and the execution of tasks while sitting. The sequence of standing/sitting is randomized, to cancel out any sequence effects.

***5.4.4 Design of Experiments***

Block whatever you can; and randomize the rest!

I have mentioned above that we have *factors* (which we can control) and *nuisance factors*, which influence the results, but which we cannot control and/or manipulate. Assume, for example, that we have an experiment where the results depend on the person who performs the experiment (e.g., the nurse who tests the subject), and on the time of the day. In that case we can block the factor *nurse*, by having all tests performed by the same nurse. But it won’t be possible to test all subjects at the same time. So we try to average out time effects, by *randomly* mixing the timing of the subjects. If, in contrast, we measure our patients in the morning and our healthy subjects in the afternoon, we will invariably bring some *bias* into our data.

**a) Sample Selection**

When selecting the subjects, one should pay attention to the following:

1. The samples should be representative of the group to be studied.

2. In comparative studies, groups must be similar with respect to known sources of variation (e.g., age, : : :).

3. **Important:** Make sure that your selection of samples (or subjects) sufficiently covers all of the parameters that you need! For example, if age is a nuisance factor, make sure you have enough young, middle aged, and elderly subjects.

Ad (1) For example, randomly selected subjects from patients at a hospital automatically bias the sample towards subjects with health problems.

Ad (3) For example, tests of the efficacy of a new rehabilitation therapy for stroke patients should *not* only include patients who have had a stroke: make sure that there are equal numbers of patients with mild, medium, and severe symptoms. Otherwise, one may end up with data which primarily include patients with few or no after- effects of the stroke. (This is one of the easiest mistakes to make, and cost me many months of work!)

Many surveys and studies fall short on these criteria (see section on *Bias* above). The field of “matching by propensity scores” (Rosenbaum and Rubin 1983) attempts to correct these problems.

**b) Sample Size**

Many studies also fail because the sample size is too small to observe an effect of the desired magnitude. In determining the sample size, one has to know

* What is the variance of the parameter under investigation?
* What is the magnitude of the expected effect, relative to the standard deviation  
   of the parameter?

This is known as *power analysis*. It is especially important in behavioral research, where research plans are not approved without careful sample size calculations. (This topic will be discussed in more detail in Sect. 7.2.5.)  
   
**c) Bias**    
To explain the effects of selection bias on a statistical analysis, consider the 1936 presidential elections in the USA. The Republican Landon challenged the incumbent president, F.D. Roosevelt. *Literary Digest*, at the time one of the most respected magazines, asked ten million Americans who they would vote for. 2.4 million responded, and Literary Digest predicted Landon would win 57 % of the vote compared with 41 % for Roosevelt. However, the actual election results were 62 % for Roosevelt and 38 % for Landon. In other words, despite the huge sample size, the predictions were a whopping 19 % off!  
   
What went wrong?  
   
First, the sample was poorly chosen, and not representative of the American voter: the mailing lists for the survey were taken from telephone directories, club membership lists, and lists of magazine subscribers. Thus, they were strongly biased towards the American middle- and upper-class. And second, only about one-fourth of the people asked responded. And people who respond to surveys are different from people who don’t, the so-called *non-response bias*. This example shows that a large sample size alone does not guarantee a representative response. One has to watch out for selection bias and non-response bias.

In general, when selecting the subjects one tries to make them representative of the group to be studied; and one tries to conduct the experiments in a way representative of investigations by other researchers. However, it is very easy to get biased data.

Bias can have a number of sources:

* The selection of subjects.
* The structure of the experiment.
* The measurement device.
* The analysis of the data.  
     
  Care should be taken to avoid bias in the data as much as possible.

**d) Randomization**

This may be one of the most important aspects of experimental planning. Random- ization is used to avoid bias as much as possible, and there are different ways to randomize an experiment. For randomization, *random number generators*, which are available with most computer languages, can be used. To minimize the chance of bias, the randomly allocated numbers should be presented to the experimenter as late as possible.

Depending on the experiment, there are various ways to randomize group assignment:

Simple Randomization

This procedure is robust against selection and accidental bias. The disadvantage is that the resulting group size can differ significantly.

For many types of data analysis it is important to have the same sample number in each group. To achieve this, other options are possible:

**Block Randomization**

This is used to keep the number of subjects in the different groups closely balanced at all times. For example, with two types of treatment, A and B, and a block-size of four, one can allocate the two treatments to the blocks of four subjects in the following sequences:

1. AABB

2. ABAB

1. ABBA
2. BBAA
3. BABA
4. BAAB

Based on this, one can use a random number generator to generate random integers between 1 and 6, and use the corresponding blocks to allocate the respective treatments. This will keep the number of subjects in each group always almost equal.

**Minimization**

A closely related, but not completely random way to allocate a treatment is *minimization*. Here one takes whichever treatment has the smallest number of subjects, and allocates this treatment with a probability greater than 0.5 to the next patient.

Assume, for example, that you are conducting a randomized controlled trial of a new medication, with a “placebo-group” and a “real medication group.” Halfway through the trials you realize that your placebo-group already contains 60 subjects, while your medication-group only has 40. You can now solve this imbalance, by giving each remaining subject with 60 % probability (instead of the previously used 50 %) the medication instead of the placebo.

**Stratified Randomization**

Sometimes one may want to include a wider variety of subjects, with different characteristics. For example, one may choose to have younger as well as older subjects. In this case, one should try to keep the number of subjects within each *stratum* balanced. In order to do this, separate lists of random numbers should be kept for each group of subjects.

**e) Blinding**

Consciously or not, the experimenter can significantly influence the outcome of an experiment. For example, a young researcher with a new “brilliant” idea for a new treatment will be biased in the execution of the experiment, as well in the analysis of the data, to see the hypothesis confirmed. To avoid such subjective influence, ideally the experimenter as well as the subject should be blinded to the therapy. This is referred to as *double blinding*. When also the person who does the analysis does not know which group the subject has been allocated to, we speak about *triple blinding*.

**f) Factorial Design**

When each combination of factors is tested, we speak of *full factorial design* of the experiment.

In planning the analysis, one must distinguish between *within subject* compar- isons, and *between subjects* comparisons. The former, *within subject comparisons*, allows to detect smaller differences with the same number of subjects than *between subject comparisons*.

***5.4.5 Personal Advice***

1. Be realistic about your task.
2. Plan in sufficient control/calibration experiments.
3. Take notes.
4. Store your data in a well-structured way.

**1) Preliminary Investigations and Murphy’s Law**

Most investigations require more than one round of experiments and analyses. Theoretically, you state your hypothesis first, then do the experiments, and finally accept or reject the hypothesis. Done.

Most of my real investigations have been less straightforward, and often took two rounds of experiments. Typically, I start out with an idea. After making sure that nobody else has found the solution yet, I sit down, do the first rounds of measurements, and write the analysis programs required to analyze the data. Through this I find most of the things that can go wrong (they typically do, as stated by *Murphy’s Law*: “Anything that can go wrong will go wrong.”), and what I should have done differently in the first place. If the experiments are successful, that first round of investigation provides me with a “proof of principle” that my question is tractable; in addition, I also obtain data on the variability of typical responses. This allows me to obtain a reasonable estimate of the number of subjects/samples needed in order to accept or reject my hypothesis. By this time I also know whether my experimental setup is sufficient or whether a different or better setup is required. The second round of investigations is in most cases the real thing, and (if I am lucky) provides me with enough data to publish my findings.

**2) Calibration Runs**

Measurements of data can be influenced by numerous artifacts. To control these artifacts as much as possible, one should always start and end experimental recordings with something known. For example, during movement recordings, I try to start out by recording a stationary point, and then move it 10 cm forward, left, and up. Having a recording with exact knowledge of what is happening not only helps to detect drift in the sensors and problems in the experimental setup. These recordings also help to verify the accuracy of the analysis programs.

**3) Documentation**

Make sure that you document all the factors that may influence your results, and everything that happens during the experiment:

* The date and time of the experiment.
* Information about the experimenters and the subjects.
* The exact paradigm that you have decided on.
* Anything noteworthy that happens during the experiment.

Be as brief as possible, but take down everything noteworthy that happens during the experiment. Be especially clear about the names of the recorded data-files, as they will be the first thing you need when you analyze the data later. Often you won’t need all the details from your notes. But when you have outliers or unusual data points, these notes can be invaluable for the data analysis.

**4) Data Storage**

Try to have clear, intuitive, and practical naming conventions. For example, when you perform experiments with patients and with normals on different days, you could name these recordings “[p/n][yyyy/mm/dd]\_[x].dat,” e.g., *n20150329\_a*. With this convention you have a natural grouping of your data, and the data are automatically sorted by their date.

Always store the raw data immediately, preferably in a separate directory. I prefer to make this directory read-only, so that I don’t inadvertently delete valuable raw- data. You can in most cases easily redo an analysis run. But often, you will not be able to repeat an experiment.

***5.4.6 Clinical Investigation Plan***

To design a medical study properly, a clinical investigation plan is not only advisable, it is even required by ISO 14155-1:2003, for *Clinical investigations of medical devices for human subjects*. This norm specifies many aspects of clinical studies. It enforces the preparation of a *clinical investigation plan* (*CIP*), specifying

1. Typeofstudy(e.g.,double-blind,withorwithoutcontrolgroup,etc.).

2. Discussionofthecontrolgroupandtheallocationprocedure.

3. Descriptionoftheparadigm.

4. Descriptionandjustificationofprimaryendpointofstudy.

5. Descriptionandjustificationofchosenmeasurementvariable. 6. Measurementdevicesandtheircalibration.

7. Inclusioncriteriaforsubjects.

8. Exclusioncriteriaforsubjects.

9. Pointofinclusion(“Whenisasubjectpartofthestudy?”)

10. Descriptionofthemeasurementprocedure.

11. Criteriaandproceduresforsubjectswhodropout.

12. Chosensamplenumberandlevelofsignificance,andtheirjustification.

13. Procedurefordocumentationofnegativeeffectsorside-effects.

14. Listoffactorsthatcaninfluencethemeasurementresultsortheirinterpretation. 15. Procedurefordocumentation,alsoformissingdata.

16. Statistical analysis procedure.

17. The designation of a *monitor* for the investigation.

18. Thedesignationofa*clinicalinvestigator*.

19. Specificationsfordatahandling.