Informal and “Informal” Inference

Inferencia informal e inferencia “informal”

Manfred Borovcnik

University of Klagenfurt, Austria

Abstract

“Informal Inference” is an approach to statistical inference based on resampling methods and links Bootstrap as replacement for confidence intervals and re-randomisation tests as alternative to statistical tests. Informal inference, on the other hand, is a conceptualisation of statistical inference by elementarising the full complexity by context that makes the interpretation of the developed concepts meaningful, or by establishing analogies and meta-knowledge that provide insight. Firstly, we illustrate the significance test by an elementary rank test. Secondly, we build informal ideas about inference by an analogy to the medical situation. Thirdly, we highlight the potential of informal ways to statistical inference by examples. Fourthly, we describe the “Informal Inference” approach. Finally, we draw some conclusions about the didactical potential and the drawbacks of “Informal Inference”. Our considerations are signified by the goal of facilitating conceptual understanding.

Keywords: Statistical inference, simulation and resampling, conceptual understanding, statistical thinking, elementarisation

Resumen

“Inferencia Informal” es un enfoque de la inferencia estadística, que se basa en métodos de remuestreo y se vincula a Bootstrap como sustituto de los intervalos de confianza y a las pruebas de aleatorización como alternativa a los contrastes estadísticos. La inferencia informal, por otro lado, es una conceptualización de la inferencia estadística mediante la simplificación de su complejidad mediante contextos que hacen que la interpretación de los conceptos desarrollados sea significativa, o mediante el establecimiento de analogías y metaconocimientos que proporcionen comprensión. Primero, ilustramos la prueba de significación mediante una prueba de rangos elemental. Segundo, construimos ideas informales sobre la inferencia, por analogía con la situación médica. En tercer lugar, destacamos el potencial de las aproximaciones informales a la inferencia estadística mediante ejemplos. En cuarto lugar, describimos el enfoque de “Inferencia Informal”. Finalmente, sacamos algunas conclusiones sobre el potencial didáctico y los inconvenientes de la “Inferencia Informal”. Nuestras consideraciones están significadas por el objetivo de facilitar la comprensión conceptual.

Keywords: Inferencia estadística, simulación y remuestreo, comprensión conceptual, pensamiento estadístico, elementarización

1. Introduction

The paper has two main goals: Firstly, to illustrate ways of elementarising the complex structure in statistical inference; secondly, to compare two different approaches to elementarisation. The difficulties in the concepts and the individual concept acquisition in stochastics in general and in statistical inference are well-known. That has induced the search for new learning forms; for elementarisation, the idea of visualisation and the New Technologies have been integrated into teaching very early. Computer-intensive methods of statistics have also served as incentive for didactic innovations.

Informal inference may be used as a label for endeavours to simplify, visualise, or simulate the *hypothetical* model behind statistical inference. That means, the statistical model in the background is still the target of teaching and forms the background. That

Borovcnik, M. (2019). Informal and “informal” inference. En J. M. Contreras, M. M. Gea, M. M. López-Martín y E. Molina-Portillo (Eds.), *Actas del Tercer Congreso Internacional Virtual de Educación Estadística*. Disponible en www.ugr.es/local/fqm126/civeest.html

implies that the theoretical character of such models is visualised by simpler means. The elementarisation is viewed as a transient stage to statistical inference.

“Informal Inference” – going back to the computer-intensive methods in statistics such as Bootstrap and re-randomisation – is an educational approach that *reduces* statistical *inference* completely *to the observed data* developing the methods solely based on resampling this data. There are only natural null hypotheses of no effect that can be tested for significance, or intervals are calculated from artificially simulated data that mimic confidence intervals.

We illustrate both approaches and present a detailed discussion about the relative merits and show how to build conceptual understanding by meta-knowledge based on simplifications of – the full complexity of – statistical inference.

2. An elementary approach to the significance test

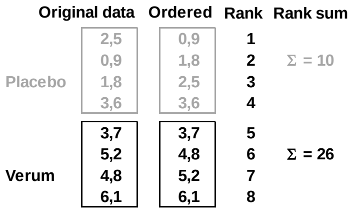
We illustrate the way of thinking in a significance test in a very simple situation, which has also been used by R. A. Fisher in his early justification of the method. The *task* is: The efficacy of an antihypertensive drug should be corroborated by a placebo-controlled, randomised, double-blind clinical study. The *target variable* is: The intra-individual *difference* of blood pressure = systolic blood pressure at baseline minus the value after 4 weeks of medication measured in mm Hg. The *hypotheses* at test: The null hypothesis (H_0) states that Verum (the medication) is equally effective as Placebo (a fake medication that can neither be recognised as such by the patients nor by the medical doctor). The alternative hypothesis (H_1) is that Verum is better than Placebo.

2.1. Re-arrangement and ranks

The basic ideas are illustrated by the Mann-Whitney test for independent samples. We use *ranks* rather than the measurements of the patients and a re-randomisation argument to read a *p* value for H_0 off the data. After ordering and ranking the data (see Figure 1), we find – surprisingly – all the data from the Placebo group in the lowest ranks with a rank sum of 10 while the Verum group attains the maximum rank sum of 26.

The null hypothesis states that there is no difference in the effect of Verum or Placebo so that we should be allowed to perceive any 4 of the 8 persons as the control group (Placebo) and the others as the Verum group. The advantage of the present way to tackle the problem is the following: The null hypothesis has the obvious implication that *any of these ways to recruit a hypothetical control group has the same justification and thus the same probability*. We only have to find all re-attributions of 4 persons from the 8 to a control group. There are $C(8,4)=\binom{8}{4}=\frac{8!}{4!4!}=\frac{8\cdot 7\cdot 6\cdot 5}{4\cdot 3\cdot 2\cdot 1}=70$. In Figure 2 (left), we order these possibilities by the rank sum (only a few to show the principle); in Figure 2 (right), we show the possibilities of the rank sum by a bar chart.

Under the null hypothesis, the distribution in Figure 2 (right) represents the probability distribution of the rank sum. As the probability to get the extreme rank sum of 10 for the Placebo group is only 1/70, we get a *p* value for H_0 of $2/70 = 0.0286 < 0.05$ (if the test is applied two-sided, i.e., if a difference between the two groups could be either way). In the usual perception of significance testing, we can reject H_0 at the 5% level.



	Original data	Ordered	Rank	Rank sum
Placebo	2,5	0,9	1	
	0,9	1,8	2	Σ = 10
	1,8	2,5	3	
	3,6	3,6	4	
Verum	3,7	3,7	5	
	5,2	4,8	6	Σ = 26
	4,8	5,2	7	
	6,1	6,1	8	

Figure 1. Data and ordered data of the Placebo-Verum experiment

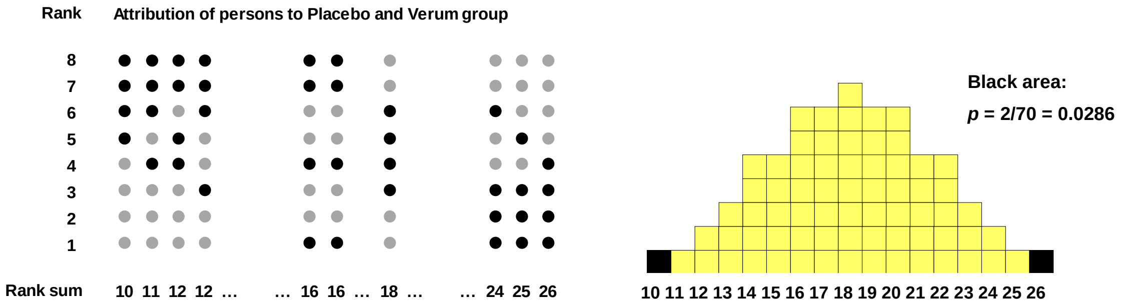


Figure 2. Left: Possible rankings ordered by the rank sum –
Right: Possibilities of each of the rank sum as probability distribution under H_0

2.2. The *p* value: Some initial concerns

We have calculated the probability for an observed result if the null hypothesis H_0 applies and use this so-called *p* value to judge the credibility of H_0. If *p* is smaller than 5%, H_0 is rejected; *p* is the probability for a false positive statement, i.e., the test yields a significant result if the drug is not effective:

$$p = P(\text{Test significant} \mid \text{Drug is not effective}).$$

We have observed something that has less probability than 5% if H_0 applies (drug not effective). *Yet, we are only interested in the following probability*:

$$\text{P (Drug is effective} \mid \text{Test significant).}$$

But this number cannot be calculated from the givens! The conclusion about a clinical study is based on statistical methods. Doctors are no experts in statistics and need not be. Yet, they should know the principles of scientific methods. Neyman and Pearson (1933) clearly limit the scope of statistical tests. They state that “No test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis.” A dialogue between a doctor and a statistician illustrates the conflict:

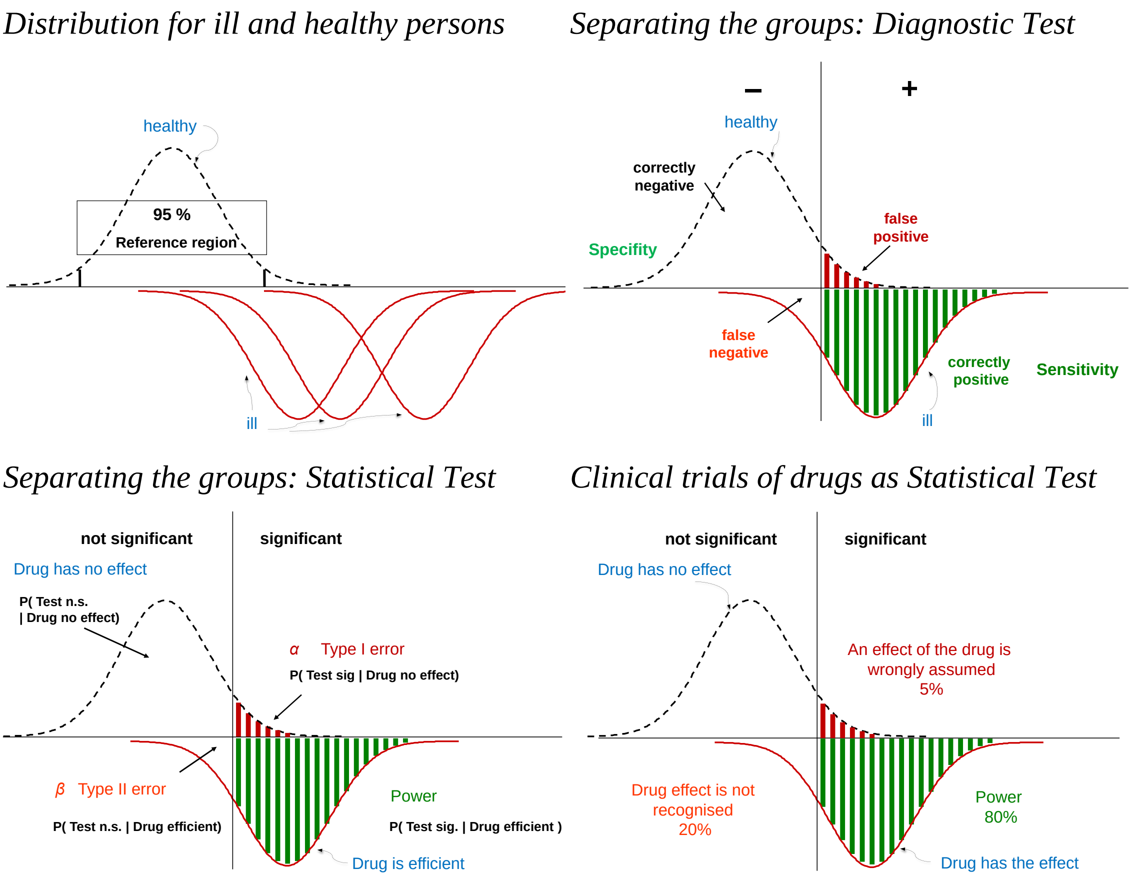
> Do: You tried hard to explain the statistical test to me – but what does it mean if my test yields a significant result? May I claim that the drug is effective?
>
> St: No – You can only calculate how probable such a test result is IF the drug is not effective.
>
> Do: The ethics commission has approved this study to investigate the efficiency of this drug. I did ask you whether you can prove this by a statistical test. As the result now is significant, I thought, that the probability that my drug is effective is 95% – because the *p* value is 5%.
>
> St: You did ask me something to which the *p* value has no answer. The error probability for your statement is higher – yet, I cannot calculate it.
>
> Do: You may be right but I have done it as all do – why should this be wrong? The result of the statistical test is significant and will be published: The drug is effective ($p < 0.05$).

3. Informal inference – An analogy to the medical situation

We explore the situation in medicine where there is always a decision that may lead to various errors whatever the decision is. A diagnostic test may be compared to a statistical test. This serves to understand statistical tests better. It may also serve to understand and investigate the medical decision better.

3.1. Separating the distribution of a variable among healthy and ill people

The other standard task in medicine is the diagnosis of a disease under scrutiny according to the outcome of a medical test, i.e., a biometric variable. We compare the task of separating two groups from each other by the values of a variable in the different settings (Figure 3): the diagnostic test, the clinical trial of a drug, the statistical test.



Distribution for ill and healthy persons

Separating the groups: Diagnostic Test

Separating the groups: Statistical Test

Clinical trials of drugs as Statistical Test

Figure 3. Contexts for separating two groups: Medical diagnosis – Statistical test – Clinical trials: Different terminology for the same concepts

For drug testing, a standard has emerged with magic figures: power (probability to detect an effect of the drug if there is one) should attain 80%, type-I (drug wrongly assumed to have an effect) should not be larger than 5%.

3.2. Medical test as decision

We extend the test in Section 2 to include the alternative hypothesis (shift in the mean of the variable under scrutiny) to allow for power considerations. The clinical trial, the diagnosis problem, or the quality control (Section 4.4), all bear the structure of a decision situation. A decision has to be made about H_0 or H_1 based on data.

Table 1. The clinical trial as decision-making situation with the different errors [1]

			Reality		
			Drug not effective Patient is healthy Quality of lot is good		Drug effective Patient is ill Quality of lot is bad
Test decision			**Null hypothesis H_0**		**Alternative hypothesis H_1**
“Drug not effective” – “Patient is healthy” Lot accepted	**Not reject H_0**	**Correct $1{-}\alpha$**	Ok Specificity Ok	**β error**	Missing that drug is effective False negative diagnosis False acceptance–consumer risk
“Drug effective” + “Patient is ill” Lot rejected	**Reject H_0**	**α error**	False decision for the drug False positive diagnosis False rejection–producer risk	**Correct $1{-}\beta$ power**	Power Sensitivity Ok

[1] The order of H_0 and H_1 as well as the decisions is different from before!

It is helpful to recognise that the structure of the decision problem and the potential errors remains the same in all three contexts. The analogy (Table 1) illustrates the meaning of the same concept in various contexts:

$p = P(\text{Test} + \mid \text{Ill})$ Sensitivity in the diagnosing context,

$p = P(\text{Test significant} \mid \text{Drug effective})$ Power in statistical tests.

Missing is any information on the *Positive or Negative Predictive Value* (PPV or NPV):

$P(\text{Ill} \mid \text{Test} +)$, or $P(\text{Drug effective} \mid \text{Test significant})$ and

$P(\text{Healthy} \mid \text{Test} -)$, or $P(\text{Drug not effective} \mid \text{Test not significant})$.

This probability describes the quality of the decision procedure. Not only that we do not know it, it is also dependent on the prevalence of the disease or the quality of research hypotheses (in drug testing as in statistical tests). We use data on mammography in radiologic clinic and in screening in Table 2, which shows the absolute numbers of the various combinations of disease and diagnosis. If we read the columns, we obtain sensitivity and specificity. The table allows also calculating the proportions in rows, which are the most interesting figures from above, namely the PPV and NPV. Strikingly, the PPV – the probability that a person has a carcinoma after a positive diagnosis – depends on the context and the prevalence of the disease under scrutiny (the same holds for NPV).

Table 2. Expected values of status (carcinoma or no carcinoma) and diagnosis (positive or negative) in radiologic clinic and in screening

	Ca	**No Ca**	**All**
+	80 Sensitivity ↑	4 False pos. ↑	84
–	20 False neg. ↑	96 Specificity ↑	116
All	100	100	200

	Ca	**No Ca**	**All**
+	640 PPV →	3 968	4 608
–	160	95 232 NPV →	95 392
All	800	99 200	100 000

Prevalence	**Clinic**	**50%**	**Screening**	**0.8%**	
Sensitivity ↑	80/100 =	80.0%		80.0%	P(+|Ca)
Specificity ↑	96/100 =	96.0%		96.0%	P(–|No Ca)
PPV →	80/84 =	95.2%		13.9%	P(Ca|+)
NPV →	96/116 =	82.8%		99.8%	P(No Ca|–)

Gigerenzer (2002) has advocated reformulating probabilities (sensitivity and specificity are usually known) by expected values, which he calls natural frequencies. They allow for a quick orientation on relevant probabilities (see Batanero & Borovcnik, 2016).

3.3. Cut points to separate the groups of healthy and ill

We show the difficulty to separate between the groups of tumour patients and tumour-free patients by introducing a cut point. The faecal occult blood test (FOBT) is used to detect colon cancer. We use data on 20 patients in each group (Figure 4). If we diagnose a patient as positive if the FOBT exceeds 75 and negative else, we see that in the tumour group three persons are falsely misclassified as negative, which amounts to a sensitivity of 17/20 = 85%. On the other hand, this cut point leads to two cases of false positive diagnosis in the tumour-free group, which corresponds to a specificity of 18/20 = 90%.

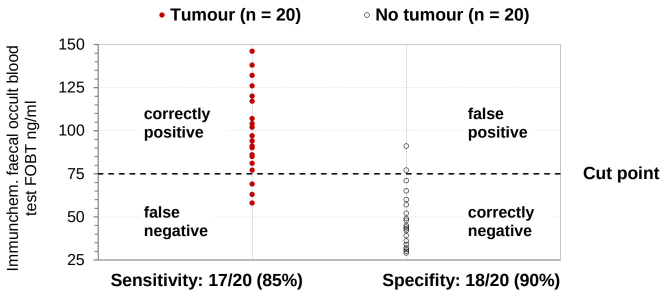
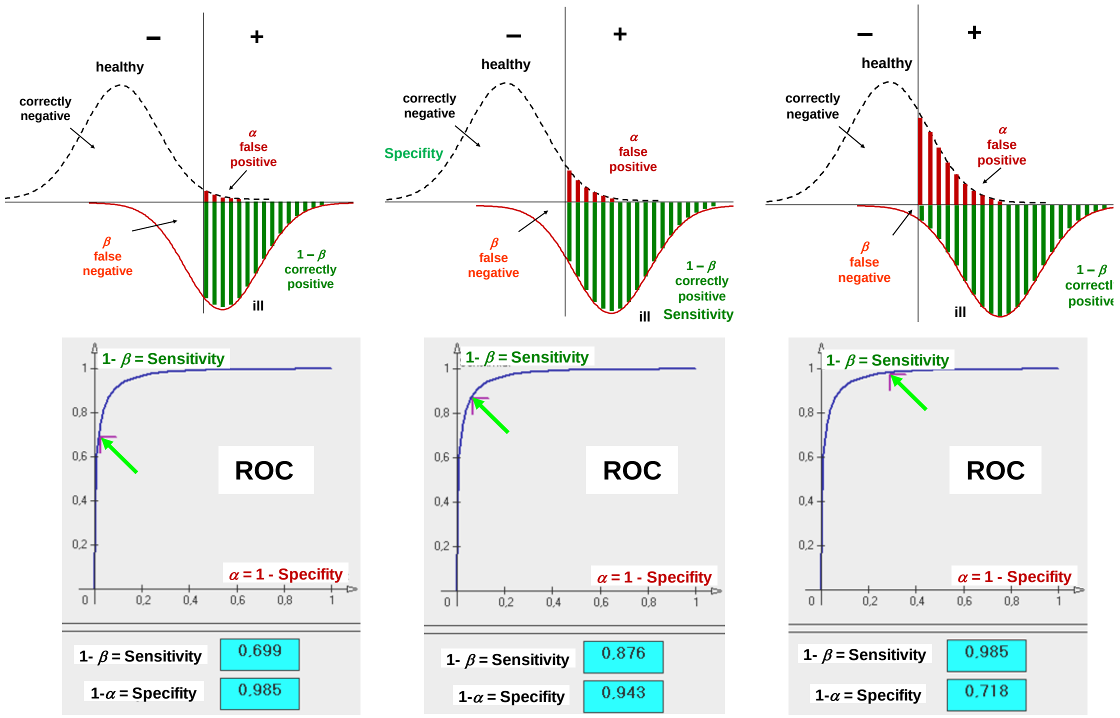


Figure 4. The choice of the cut point leads to a separation of the two groups with distinct quality as measured by the concepts of sensitivity and specificity

Which cut point should be used for diagnosis? If we vary the cut point, we generate several procedures for the diagnosis, all with different properties (Figure 5).



1- β = Sensitivity 0,699
1-α = Specifity 0,985

1- β = Sensitivity 0,876
1-α = Specifity 0,943

1- β = Sensitivity 0,985
1-α = Specifity 0,718

Figure 5. Cut points lead to different methods of diagnosis with a different point on the ROC curve – a point in the upper left corner reflects a good diagnosis

It is customary, to illustrate the quality of the diagnosis by the so-called ROC curve, which shows to each cut point the corresponding point (α, 1 – β), i.e., the type-I error on the first coordinate and the sensitivity on the second. That means a point far left and up is linked to a diagnosis with good properties to separate the two groups.

We recognise that the higher the cut point is chosen, the further to the left (good for the diagnosis) and the further down (bad for the diagnosis) the corresponding point on the ROC curve will lie. We are faced with antagonistic consequences when shifting the cut point. We have to find a compromise between the two targets to get a small type-I error and a high power. Various diseases have different distributions for the target variable and correspond to different ROC curves. The disease with the dotted (blue) distribution, which is the same as the distribution among healthy people, turns the diagnosis to a coin tossing experiment, i.e., simple guessing. The corresponding ROC is the diagonal, all points of which are far off the upper left corner that hosts the points for diagnosing procedures with good properties.

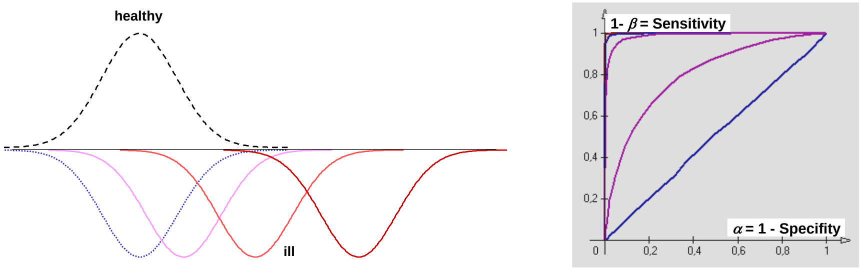


Figure 6. The closer the distributions of ill and healthy the more difficult the diagnosis

3.4. Some conclusions from the analogy to medicine

From the analogy to medicine, we learn that we normally face a decision problem, which may be described – for simplification – by two scenarios. Diagnosing for diseases is a decision problem, which compares distributions under the scenario of healthy and ill people. Whatever we decide, we are prone to committing an error. There are always – at least – two diverging errors in the play:

- Diagnose for the disease when the person is healthy.
- Not recognise the disease despite the fact that the person has it.

Wherever we introduce the cut point between the two groups (scenarios), the errors are influenced by that choice and we should orientate ourselves on their magnitude. Several cut points for separating healthy and ill imply different sizes of these errors. There are diseases that are easy to diagnose. Reducing the complexity of the decision situation, the *p* value is used but it not easy to interpret this number in a practical meaningful way. There is a third error: Whether the decision is a good one, does not only depend on cut points but also on the prevalence of the disease. In summarising, we do not get well interpretable coefficients for the quality of decisions in many cases.

4. Informal ways towards statistical inference

We illustrate various informal ways to explore key statistical concepts. One main problem is to highlight the relevance and the meaning of the sampling distribution of statistics that estimate a parameter of the population. Another idea is to reduce the complexity of statistical tests to a comparison of two distributions that makes sense in the context so that decisions and their implications become a natural issue to discuss just as in the analogy to the medical situation (of diagnosing or testing drugs). The explorations serve to learn about key features, also by establishing meta-knowledge about the method beyond mathematics. The goal is to reduce the complexity of the situation but keep the path to the general situation open.

4.1. Two different methods of estimating the mean

The values of the population are marked by a bar (see Figure 7). Two homogenous strata are visible. If such a case of strata is known, it is advisable to consider that in sampling. We compare two methods: Method 1: Random sample of 6 elements from all neglecting the strata; Method 2: Random sample of 2 from stratum 1 and 4 from stratum 2. For both methods, we can see from the simulation scenario (Figure 7) that the mean of simulated data is roughly equal to the mean of the population (unbiased estimator). We see also that strata sampling (Method 2) delivers much more precise results. The improvement of the estimate by sampling from the strata as compared to sampling neglecting the strata is stable in the repetition of the scenario.

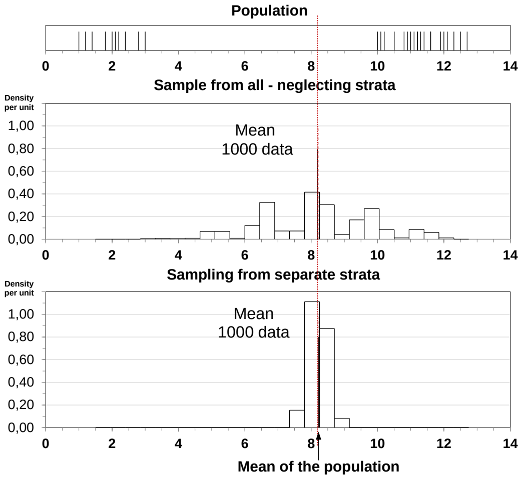


Figure 7. Sampling unrestricted and from strata – sampling distribution of the mean

One can visualise the persons who are sampled and their data and show – like in a video – how these change by renewing the sample to get an impression about the variability of sampling and the changing error in estimating the mean of the population by the mean of the sample. It will get immediately clear that Method 2 (stratified sampling) leads to smaller errors in general. It is important to see how single samples behave before summarising the result of many samples by the sampling distribution of the mean. This is important as in practice we have only one (!) sample. The result of this sampling process (with 1000 samples and their mean each) is shown in Figure 7: What has been an impression is now corroborated by the simulation scenario. By Method 1 (unrestricted sampling), the error is in general very large with means between 2 and 12 while for Method 2 (strata sampling) the error tends to be small with means between 7.5 and 9.5.

4.2. The sampling distribution of the mean is an artificial distribution

In the statistical laboratory, we can simulate samples from any population. The sampling distribution of the estimate of any parameter shows how the parameter estimation varies from one sample to another. Usually, we have only one sample and therefore it seems contra-intuitive to speak of the variation of the estimate. Yet, in a thought experiment, we can repeat the sample very often to illustrate the properties of the estimation. Are we lucky in one sample to have an estimate that is close to the pertaining parameter of the population, or may we rely on the fact that the general risk of getting large deviations from the parameter of the population is small?

We can simulate two completely different populations – one with a uniform and one with a J-shaped distribution on the population to highlight the concept of sampling distribution and to illustrate its key features (see Figure 8). Regardless of the parent population, the sampling distribution of the mean (as of many other parameters) restricts to the point of the mean of the population (the parameter of interest) and resembles more and more in shape to a normal distribution with increasing sample size. See also Batanero and Borovcnik (2016) for a scenario to show these properties of the sampling distribution of the mean; Figure 8 illustrates the development from 5 to 20 samples. Note that for the mean, the width of the distribution (as measured by the standard error) halves if we take a sample 4 times as large as before. It is instructive to see that the shape of the sampling distribution is nearly invariant with the repetition.

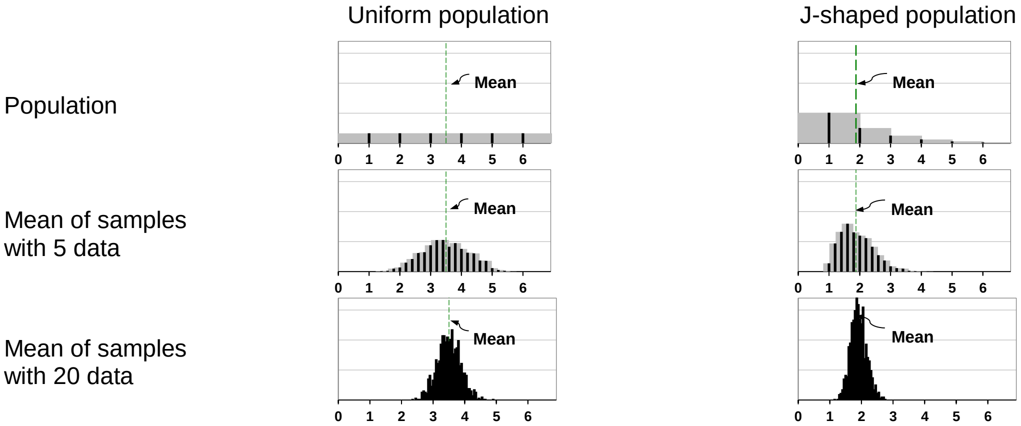


Figure 8. Sampling distribution of the mean of a sample from a uniform (left) and a J-shaped population (right)

4.3. Measuring an unknown probability as path to the Law of Large Numbers

In the following (coin-tossing) experiment, the relative frequencies are investigated along an idea expressed in Batanero and Borovcnik (2016). Rather than showing how the relative frequencies converge (what should that mean and to where they should converge?), a task of estimating the unknown probability is posed. The estimation may be based on samples (blocks) of 5 trials (5 tosses of the coin) or 10 or 20. We show in Figure 9 how the relative frequencies converge with the number of trials and we observe the development until 1000 trials are performed. The current series (in Figure 9) cannot fluctuate much because of the past 1000 values. The curve suggests a great precision of less than 0.5 percentage points of fluctuation. Yet, a new experiment shows – like in a video – another curve with another "limiting point" within +/– 3 % points; a repetition of the series of 1000 trials will also "converge" yet to a different point.

The Law of Large Numbers states that the theoretical (not the empirical) relative frequencies "converge" to the unknown probability. This "convergence" in a real experiment hides that current results are still prone to randomness. What about changing the task and *measuring the unknown probability* by short series and investigate the precision of such a measurement. After each block of 5 (10, or 20), the sample is summarised and used to estimate the unknown probability. The estimates may be 0.0, 0.2, …, 0.8, and 1.0 (according to 0, 1, …, 5 Heads). In Figure 9 (left), we see how these estimates fluctuate; many are beyond the dashed (red) lines with an estimation error larger than 0.2. Of course, the sample is very small, the error should be large.

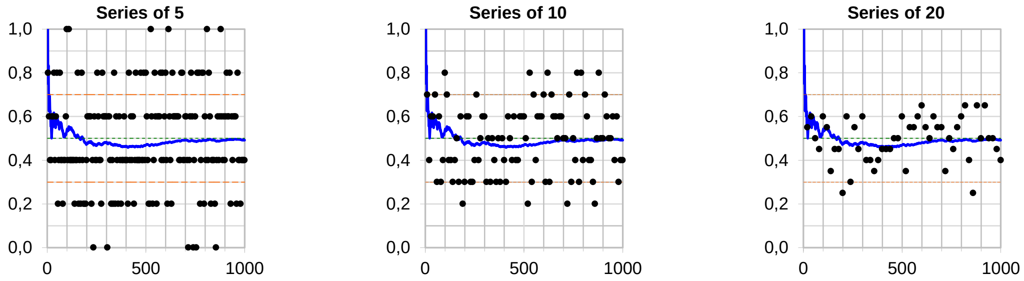


Figure 9. Measuring an unknown probability – Investigation of the precision

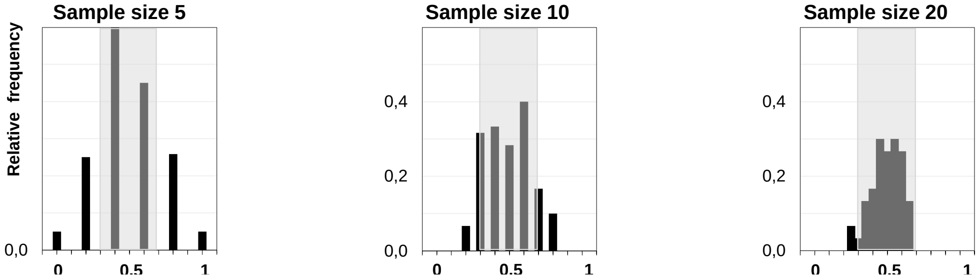


Figure 10. Distribution of the repeated measurements

In the middle diagram of Figure 9, the results of measuring the probability by samples of 10 is shown and the estimate fluctuates much less; even smaller is the variation of the estimate (and the errors are smaller) in the right diagram of Figure 9, which shows the estimates of the probability based on samples of 20. We see only two estimates beyond the dashed (red) lines. By a thought experiment, one may conclude that the precision of the estimate improves the larger the sample and its distribution will restrict to the (unknown) probability. This constriction of the distribution of the estimates, which is represented in Figure 10, may also be seen from the study of the sampling distribution of the mean (in Figure 8). The risk of committing an error larger than 0.2, is decreasing with the size of the series (sample) on which the measurement is based on.

4.4. How to separate good and bad quality – Informal exercises in statistical tests

The following example goes back to Batanero and Borovcnik (2016). The task is to judge whether the current production (or a lot that has come in) has a good or bad quality. Single items can show only this property, which is encoded by 0 (good) and 1 (defective). By inspecting a sample of *n* items, a decision should be made about the quality. The number of defectives in the sample is hyper-geometrically distributed; neglecting that sampling is without replacement, we can use the binomial distribution instead. Two scenarios are compared, which stand for different stakeholders: a lot has come into the consumer’s factory, sent from the producer; good quality is represented by p = 0.04, bad quality by p = 0.10 (p stands for proportion of defectives).

Rather than using the binomial distribution, we have simulated 5000 samples of size n = 100 and determined the relative frequency from the simulation scenario in order to estimate the probabilities. In Figure 11, we show the implication of a rejection number, let us say, reject the lot as bad (reject the null hypothesis of good quality) in favour of the alternative hypothesis. One may change the rejection number (shift the dashed bar in the diagram) and it becomes visible how the two types of error change and recognise that they are antagonistic, i.e., while the one gets smaller, the other gets larger.

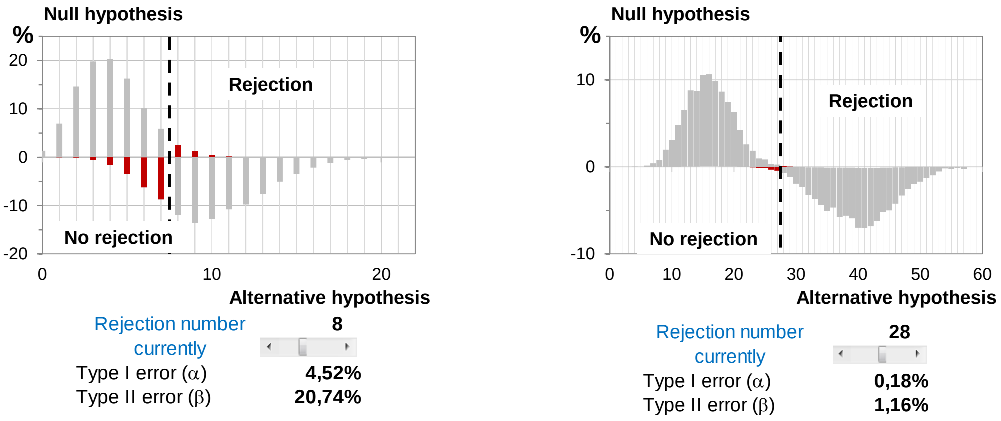


Figure 11. A rejection number (vertical dashed line) is associated with two types of errors: the decision is based on a sample with *n* = 100 (left) and *n* = 400 (right)

On the right side of Figure 11, we see the consequences of a larger sample for the decision. With 400 data, both errors become small. The choice of the rejection number balances the diverging interests of vendor and buyer. The diagram (Figure 11) with the threshold above which the null hypothesis is rejected in favour of the alternative reminds us to the diagrams from Section 3 in medical diagnosis. Basically, it is the same type of decision situation. Being above the rejection number corresponds to being above the cut point; the ensuing decision is that the lot has a bad quality, which corresponds to the positive diagnosis (the patient is classified to have the disease under scrutiny).

We have many other tasks, in which we do not have to separate two or more groups from each other but to study the consequences of scenarios for different groups. For example, the case of single-choice exam papers. While the assumptions for a binomial distribution for the number of correctly solved items by merely guessing are undisputed, the assumptions are not really appropriate to describe someone who has learned. Yet, the scenarios may be played through with valuable insights into the way how such test papers have to be designed in order to keep the risk small that someone passes the exam who is merely guessing and – at the same time – to ensure that the risk gets small to fail in the exam for someone who has learned and has a solving capacity of 0.55 or even 0.80. More can be seen from Borovcnik (n.d.). Another prototype of tasks originates from statistical process control where the hourly inspection for the quality of the measurements should orientate about the present calibration of the production machine with a built-in alarm in case that the calibration has shifted. For details, see Borovcnik (n.d.).

5. “Informal inference”

Informal inference centres all considerations about generalising the information contained in a given data set solely on this data. Early steps of development of the informal approach are the following: Resampling as a didactical technique (Borovcnik, 1996), as a transient stage to statistical inference (Borovcnik, 2006a, b); as method to replace statistical inference (Cobb, 2007); re-randomisation tests as replacement for the significance test (Rossman, 2008); Bootstrap intervals to replace confidence intervals (Engel, 2010). Stohl Lee, Angotti, & Tarr (2010) present a panoramic approach by examples. The methods are explained below; more about the methodology may be consulted from Lunneborg (2000); for extensive critique of the approach, see Howell (n.d.).

5.1. Introduction to the “informal inference” approach

Estimation Bootstrap is used to estimate the standard error. Instead of sampling from a true distribution function F, the estimate of F from the initial sample is sampled (with replacement). Bootstrap yields approximate intervals for the unknown parameter.

Hypothesis testing This is reduced to randomisation tests. Re-randomisation of assignment to groups to be compared provides artificial data that are used for the test. Either all permutations of the data are investigated, or sampling is from the data with no replacement, which is equivalent to sampling all permutations. This approach provides exact nonparametric tests in specific cases.

The case of the natural null hypothesis The intention of “Informal Inference” is to embed the complex situation in statistical inference in a natural, material setting (i.e., the data) leaving out any consideration about hypotheses except the natural null hypothesis (see Section 5.3) of pure random effects on the statistical units.

Inference about one “group” If one data set is to be judged, e.g., for a measure of location, a Bootstrap interval is obtained by repeatedly sampling from the given data (always calculating this statistical measure). This resampling provides an empirical basis (an empirical distribution) for the statistical measure. If a (hypothesised) parameter value falls outside the Bootstrap interval, then it is “rejected”.

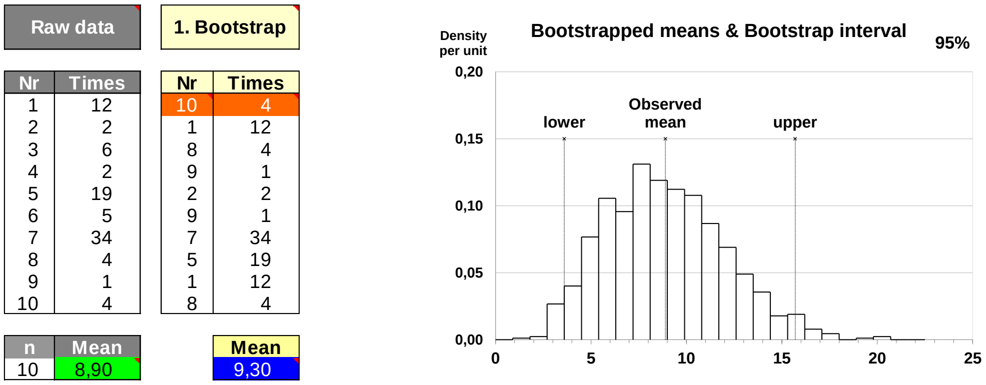
Inference about two groups If two given data sets are to be compared for a measure of location (or any other parameter), then there are two options: First, we can resample from the given data on each group to derive the Bootstrap interval for this parameter. Second (and much more intuitive), we can re-randomise the attribution of single data to one of the groups by a random decision. If the null hypothesis of no difference between the two groups applies, then the data can be pooled and from this pool, the data for group 1 (and 2) can be randomly selected so that again an empirical basis of the statistic of interest can be generated solely by the given data. The initial random attribution is randomly redone on the existing data, which reflects the natural null effect hypothesis.

5.2. Bootstrap interval and classical confidence interval for the mean

Given: A sample of size *n* with mean and SD for a specific variable (data below in Figure 12). How precise is the mean of the sample as a measurement for the population? The variable *Time* = time worked for a seminar.

Rather than sampling again from the population, which is impossible, we sample from the first sample (with replacement). The first Bootstrap yields a new measurement of the mean of the population, which differs not too much from the mean of the original sample. We repeat the Bootstrap and get 1000 (or more) artificial measurements.

The artificial data obtained by this method reflect the *variability of repeated measurements* of the unknown mean of the population. From the Bootstrap distribution for the mean, we can easily cut off the lowest and highest 2.5% of the Bootstrapped means in order to get the 95% Bootstrap interval, which yields (3.60, 15.70) in our simulation scenario. This may be compared to the classical confidence interval of (2.46, 15.34). We see a good match of both methods. Yet, the interpretation differs. The Bootstrap interval reflects the precision of repeated measurements of the mean of the population while the confidence interval contains the mean of the population in 95% of “repeated” samples.



Raw data

Nr	Times
1	12
2	2
3	6
4	2
5	19
6	5
7	34
8	4
9	1
10	4

n	Mean
10	8,90

1. Bootstrap

Nr	Times
10	4
1	12
8	4
9	1
2	2
9	1
7	34
5	19
1	12
8	4

Mean
9,30

Figure 12. Left: Original sample of the time worked on a seminar and first Bootstrap sample from this data – Right: Histogram of 1000 Bootstrap samples

The Bootstrap may be applied to estimate other parameters of the distribution. The procedure is analogue to that with the mean. Estimate the relevant parameter from the original sample, Bootstrap the first sample (by sampling with replacement) and calculate the parameter of the resampled sample in order to get artificial data about the parameter of interest and analyse the Bootstrap distribution of this parameter. Cut off the upper and lower end of this distribution to get a Bootstrap interval. One may also determine a Bootstrap interval for the correlation of two samples in this way.

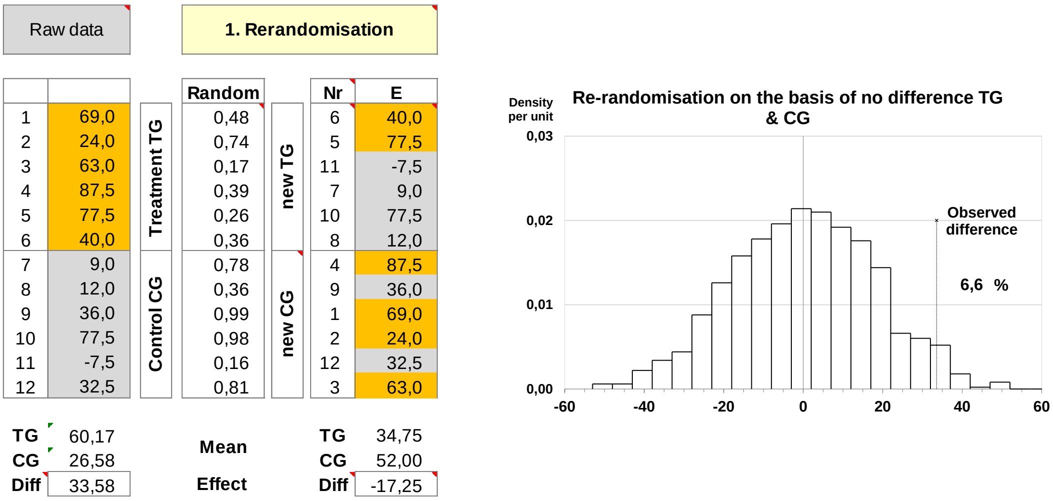
5.3. Re-randomisation test for the difference in means

Is treatment effective (with respect to a target variable)? Treatment group (TG) gets Verum – control group (CG) gets Placebo. Re-randomisation offers an alternative to the two-sample t test. The procedure is similar to the significance test in Section 2. Rather than dealing with ranks of observations, we analyse the values of the data here. Rather than determining all possibilities of various rank sums, we simulate from the distribution of all possibilities.

Under the null hypothesis of *NO DIFF*, it is intuitive that *any re-arrangement of persons to treatments should have NO EFFECT*. We therefore permute the persons and the next treatment group consists of 6, 5, 11, 7, 10, and 8. The first re-attribution yields a new measurement of the difference of the means (as measuring the effect of treatment); the difference between treatment and control group in the original sample is 33.58 while the first re-attribution yields a difference of –17.25 (see Figure 13).

The distribution for the repeated re-randomisation is shown in Figure 13 (right); it yields the artificial results based on the hypothesis of *NO DIFFERENCE*, i.e., the null hypothesis. We can fit the result of the first sample into this distribution and see that the *p* value of it is 6.6% (two-sided). The whole simulation scenario may be repeated to show that the result is stable. Again, we can compare this re-attribution result to the classical two-sample t test, which yields 2.16 with a *p* value of 5.6% or 2.16 with 5.9% (depending on the additional assumption of equal or unequal variances).

Again, the similarity of the classical results to the re-attribution test is striking. The procedure can be applied to other comparisons as well; for example, the correlation task may be rephrased as a test of the hypothesis that the correlation in the population is zero. See Borovcnik (n.d.) for details.



Raw data

1. Rerandomisation

			Random		Nr	E
1	69,0	Treatment TG	0,48	new TG	6	40,0
2	24,0		0,74		5	77,5
3	63,0		0,17		11	-7,5
4	87,5		0,39		7	9,0
5	77,5		0,26		10	77,5
6	40,0		0,36		8	12,0
7	9,0	Control CG	0,78	new CG	4	87,5
8	12,0		0,36		9	36,0
9	36,0		0,99		1	69,0
10	77,5		0,98		2	24,0
11	-7,5		0,16		12	32,5
12	32,5		0,81		3	63,0

TG	60,17	Mean	TG	34,75
CG	26,58		CG	52,00
Diff	33,58	Effect	Diff	-17,25

Figure 13. Left: Original sample of the target variable in treatment & control group and first re-attribution of persons to treatment – Right: Histogram of 1000 re-attributions

5.4. “Informal inference” contrasted to key statistical notions

“Informal inference” is not inference in a mathematical sense. It is NOT an informal approach to what the discipline of statistics calls inference. It presents a rather restricted approach to making inferences with no obvious links how to proceed from there to formal inference (unless one omits traditional statistical inference).

Within “Informal Inference” it is impossible to address key concepts. In Table 3 (Borovcnik, 2017), Bootstrap and Re-randomisation are compared to key statistical ideas to highlight the deficits if taken as the sole approach towards statistical inference.

Table 3. “Informal” inference and key statistical ideas

Concepts	*Re-randomisation*	*Bootstrap*
Hypotheses – scenarios	Only NULL effect hypothesis	Not possible to conceptualise
Type I (or α) error	Yes	No
Type II (or β) error	No	No
Alternative hypotheses	Not possible to conceptualise	Not possible to conceptualise
Methods	Only significance test of NULL effect	No link to significance tests

Statistical inference implies a hypothetical approach. There have been several endeavours to compare the various schools of inference starting from Barnett (1982). They differ according to which hypotheses are included and how they are treated.

- Alternative hypotheses. There is no way to introduce alternative hypotheses except by probabilistic assumptions and by simulation (or probability calculations). Any alternative cannot be resampled as is has not been sampled. The data are not suitable for investigating alternative hypotheses by resampling.
- Hypothetical comparison of models. Modelling involves comparing scenarios (described by probability distributions) and not only judging a single one. Hypothesis tests are comparisons of models (while the model family is restricted).

Resampling marks a change from probability models to data; it causes a shift in connotation from hypotheses to facts (data as facts); models are absorbed in (resampled) data. Yet, how judgement of hypotheses is done, lies at the core of statistical inference. Whether it is done by classical or Bayesian methods, there is no link from resampling.

Resampling reduces probability to the frequentist conception. As statistical inference is reduced to resampling the data, there is no obvious link for other connotations of probability though they are relevant for the various schools of inference (see Barnett, 1982). This narrows also the connection to decision theory. Many problems of statistical inference are enhanced if seen from a decision-theoretic perspective.

The case of small probabilities. In Bootstrap, a new error is introduced. If it is about the tails of a distribution (small probabilities), one would not get data about them so that the tails will not be resampled. If the first sample is not big enough, more regions of the distribution cannot be sampled well. If the first sample is big, then anyway the Central Limit Theorem delivers better results. If one resamples, then the additional variation is big unless one generates more than 10,000 re-samples. That makes it intractable for teaching. Simulation is misplaced for the problem of small probabilities; a problem, which is underestimated in statistics education (see Batanero & Borovcnik, 2016).

6. Educational concerns about resampling and conclusions

“Informal Inference” has been suggested as a way to revolutionise teaching statistical inference (Cobb, 2007; delMas, 2017; Ben-Zvi, Makar, & Garfield, 2018). Yet, there are issues to re-consider not only from an educational point of view.

Re-randomisation is null hypothesis significance testing. It deals with testing a null effect hypothesis. The difference in the mean, e.g., between two populations is compared by a random re-arrangement of the units to one of the groups. A distribution for the difference of means is generated, which corresponds to the null hypothesis of no difference between groups. A re-arrangement of the given data cannot be performed to reflect specific differences between the two groups so that the method fails to incorporate issues of errors of type II. We are stuck with a pure significance test with all the confusion that arises from the *p* value (see, e.g., Hubbard & Bayarri, 2003).

Confusion of Bootstrap and confidence intervals. Bootstrap intervals are not easy to implement when corrections are applied to adapt their coverage probability (for the sophisticated corrections, see Efron & Tibshirani, 1993). Bootstrap intervals are not easy to connect to traditional topics further on. Often, the specification “Bootstrap” is omitted. The same terms for widely different concepts may confuse learners.

Simplifying the approach to a discipline for didactical purpose. Brousseau (1984) warns of the implications of elementarisation in didactical considerations. In simplifying, one finally might end up with teaching a new object that does not even exist in mathematics (“glissement methadidactique”). Biehler (2014) states that: “[…] formal inferential reasoning as such is controversial itself […] This raises questions with regard to which view of formal […] inference we design […] informal inference activities for.” “Informal Inference” goes beyond informally exploring probabilistic models; it aims to replace traditional statistical inference. The advantages of an intuitive approach towards inference get lost if it is not seen as a transient stage in teaching and learning.

Issues to re-consider for an” Informal Inference” approach:

- “Informal Inference” is very convincing but leads to a restricted methodology that is a strict subset of statistical inference.
- Bootstrap intervals differ from classical confidence intervals; to adapt them, requires sophisticated methods so that their intuitive advantage is lost (see Efron &

Tibshirani, 1993, Lunneborg, 2000, or Howell, n.d.). Re-randomisation does not permit errors of type II so that testing is reduced to the disputed significance test.

- Bootstrap fails with small (tail) probabilities, which are not covered in the data unless the sample is very large so that one cannot resample them. It is a matter of modelling small probabilities and risks rather than only dealing smartly with data.
- The “Informal Inference” approach unfeasibly reduces the conception of probability to the frequency aspect as inference is reduced to resampling and relative frequencies of artificial data. Here, we face the dilemma formulated by Carranza and Kuzniak (2008) with a pure frequentist approach to stochastics where the applied problems require a decision-theoretic approach and a qualitative (subjectivist) connotation of probability.
- How to adapt the probability curriculum? Should we leave the normal distribution behind? Probabilistic modelling uses many other distributions (e.g., for risk analysis). How to deal with other approaches and interpretations (e.g., Bayes).
- How to continue the curriculum within such a setting? There is no path from resampling to decision theory, which is much closer to many problems of everyday concern but also to many applications such as in medicine or economy. There is no connection from resampling to Bayes methods (though Bayesians use simulation a lot), which form a relevant approach for problems from real world.
- Conceptual understanding differs from easier access and solving of tasks. Furthermore, modelling is absorbed in simulation. This may result in data as facts while models represent a hypothetical way of thinking. It might be better to teach classical and Bayesian inference in parallel to highlight the differences in the concepts and thus allow for a sustainable concept acquisition (see Vancsó, 2009).

“Informal Inference” narrows the view on probabilistic modelling later. General educational questions that arise with the approach are: The statisticians use ever more sophisticated models but we have not even managed to teach the simplest. How will anyone be able to challenge the experts if educated only on this side-track? Should statistics for secondary level be a field that has nearly nothing in common with statistics at university and the abundant applications that intrude every sector of public and private life? Are we going to distract people from critically appraising and challenging those applications of statistics? We suggest using resampling (Bootstrap and re-randomisation) as *a transient stage* to statistical inference and focus on ways of elementarising the full complexity of statistical inference.

References

Barnett, V. (1982). *Comparative statistical inference (2nd ed.)*. New York: Wiley.

Batanero, C. & Borovcnik, M. (2016). *Statistics and probability in high school*. Rotterdam: Sense Publishers.

Ben-Zvi, D., Makar, K., & Garfield, J. (2018). *International handbook of research in statistics education*. Cham, Switzerland: Springer International.

Biehler, R. (2014). On the delicate relation between informal statistical inference and formal statistical inference. In K. Makar (Ed.), *Proceedings of the Ninth International Conference on Teaching Statistics*. The Hague: ISI.

Borovcnik, M. (1996). Trends und Perspektiven in der Stochastik-Didaktik [Trends and perspectives in the didactics of stochastics]. In: G. Kadunz, H. Kautschitsch, G. Ossimitz, & E. Schneider (Eds.): Trends und Perspektiven (pp. 39-60). Wien: HPT.

Borovcnik, M. (2006a). Daten – Zufall – Resampling [Data–randomness–resampling]. In J. Meyer (Ed.), *Anregungen zum Stochastikunterricht Band 3. Tagungsband 2004/5 des Arbeitskreises "Stochastik in der Schule"* (p. 143-158). Berlin: Franzbecker.

Borovcnik, M. (2006b). On outliers, statistical risks, and a resampling approach towards statistical inference. *Paper presented at* CERME V). Larnaka.

Borovcnik, M. (2017). Informal inference – Some thoughts to reconsider. In *Proceedings of the 61st World Statistics Congress*. The Hague: ISI.

Borovcnik, M. (n.d.). *Spreadsheets in Statistics Education*. Online: wwwg.uni-klu.ac.at/stochastik.schule/Boro/index_inhalt.

Brousseau, G. (1984). *Le rôle central du contrat didactique dans l'analyse et la construction des situations*. Non-published paper.

Carranza, P. & Kuzniak, A. (2008). Duality of probability and statistics teaching in French education. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education.* Monterrey: ICMI and IASE.

Cobb, G.W. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education* 1(1).

delMas, R. (2017). A 21st century approach towards statistical inference – Evaluating the effects of teaching randomization methods on students' conceptual understanding. In *Proceedings of the 61st World Statistics Congress*. The Hague: ISI.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York – London: Chapman & Hall.

Engel, J. (2010). On teaching bootstrap confidence intervals. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society.* Voorburg: International Statistical Institute.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster.

Howell, D. (n.d.). Resampling statistics: Randomization & Bootstrap. *Statistical page Howell*. Online: www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html.

Hubbard, R. & Bayarri, M. J. (2003). Confusion over measures of evidence (p) versus errors (α) in classical statistical testing. *The American Statistician 57*(3), 171-182.

Lunneborg, C. E. (2000). *Data analysis by resampling: concepts and applications*. Pacific Grove, CA: Duxbury Press.

Neyman J. & Pearson E. (1933). On the problem of most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A, 231*, 289-337.

Noether, G. (1967). *Elements of noparametric statistics*. New York: Wiley.

Rossman, A. J. (2008). Reasoning about informal statistical inference: one statistician's view. *Statistics Education Research Journal 7*(2), 5-19.

Stohl Lee, H., Angotti, R. L., & Tarr, J. E. (2010). Making comparisons between observed data and expected outcomes: students' informal hypothesis testing with probability simulation tools. *Statistics Education Research Journal 9*(1), 68-96.

Vancsó, Ö. (2009). Parallel discussion of classical and Bayesian ways as an introduction to statistical inference. *International Electronic Journal of Mathematics Education 4*(3), 291-322.