Some Guidelines for Teaching Hypothesis Testing to Undergraduate Students

Alex Van Breedam [*]
KPMG Orinoco, University of Antwerp
University of Limburg, Université de Valenciennes

Kristel Van Rompay [†]
University of Antwerp

Abstract

This article contains some advice for teaching hypothesis testing in the introductory college courses in statistics. For these students, hypothesis testing has to be based on the translation of a problem into a hypothesis to be tested, and the choice of the appropriate statistical test, in order to solve a problem at hand. In order to formulate the correct hypotheses, the rationale behind hypothesis testing has to be made perfectly clear, which is often neglected. Secondly, most general textbooks on business statistics split the chapter on hypothesis testing into a separate chapter on parametric hypothesis testing and another one exclusively dedicated to non-parametric hypothesis testing. Consequently, the problem-solver student is almost incapable to draw relationships between parametric and non-parametric hypothesis testing. Some guidelines to overcome both problems are presented.

Key words: Hypotheses formulation, Parametric tests, Non-parametric tests, Statistical test selection.

[*] Alex Van Breedam is partner of KPMG Orinoco, a company specialised in optimisation problems. He is also part-time associate professor in Statistics at the University of Antwerp - RUCA, in Operations Research at the University of Antwerp - UFSIA and the University of Valenciennes and in Logistics at the University of Limburg. Correspondence address: KPMG Orinoco, Bourgetlaan 30, B-1130 Brussels, Belgium, Tel.: + 32 2 7 08 46 47, Fax: + 32 2 7 08 46 66, E-mail: alexvb@ruca.ua.ac.be or avanbreedam@kpmg.com.

[†] Kristel Van Rompay is teaching assistant in Mathematics and Statistics at the Faculty of Applied Economics of the University of Antwerp, (Ruca,Ufsia), E-mail: vanrompk@ruca.ua.ac.be

1 Introduction

Statistical education is a fundamental part of the curriculum of undergraduate students in business economics, psychology, science, engineering... Depending on the university, the program of the undergraduate (and first-year graduate) level usually contains far more than one hundred class hours of statistics. Hypothesis testing, including parametric and non-parametric tests, forms a major part of the statistical training. As applied economists, scientists, engineers... are typically educated as *problem-solvers*, the aim of their statistical education should be to improve their problem-solving skills. They should be trained to make decisions based on (very) limited information. Most of the time, these decisions should be made in such a way that costs are minimized or, alternatively, profits are maximized.

We believe that statistical education can play a major role in improving a student's problem-solving skills. However, this statistical training should be organized in such a way that it helps the decision-maker in formulating the actual problems into a statistical problem. A thorough understanding of hypothesis testing requires that the rationale behind hypothesis testing is perfectly clear to the future decision-maker. If the rationale is perfectly clear, it becomes much easier to formulate the null hypothesis and the alternative hypothesis for a specific problem. In this article, some guidelines are given to help lecturers in statistics to improve the student's ability to formulate an actual problem as one or more hypotheses to be tested.

Having solved the problem of the hypotheses formulation, the decision-maker is confronted with his next major problem: the selection of the appropriate statistical test. Decision support tools such as classification tables, flowcharts, and expert-like systems may help him with his selection process.

This article is organized as follows. Section two considers the way traditional textbooks on statistics for business students are organized. In section 3.1, the rationale behind hypothesis testing is presented in detail. Guidelines for selecting the appropriate statistical test are given in section 3.2. Finally, some conclusions are drawn.

In this article, we choose all examples in the field of Business Economics, and we refer to some textbooks on business statistics, but the reader can easily apply the mentioned techniques to his preferred field of interest.

2 Traditional textbooks on Business Statistics

Lecturers in statistics can choose from many textbooks on statistics. Unfortunately, most of the textbooks treat hypothesis testing inappropriately, as far as the education of decision-makers is concerned. First, textbooks often fail to clarify the concept behind hypothesis testing in a way that is appropriate for decision-makers. Thus making it hard to formulate the correct hypotheses.

Secondly, parametric and non-parametric hypothesis tests are considered in separate chapters, as if there were no relationship between them. Moreover, the parametric tests, as there are the Z-test, the t-test and the F-test, are subjected to constraints with respect to the nature of the data (interval scale, normality, homoscedasticity, large samples...). These constraints are not always satisfied. Consequently, non-parametric tests are at least as important as their parametric counterparts. However, this is not the way things are represented in the business statistics textbooks. The reader is referred to some recent books on Statistics for Business Economics to verify the above statements, e.g. [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], etc...

In what follows, some guidelines are formulated to overcome these problems in courses for business economics students.

3 The rationale behind hypothesis testing

Generally, the hypothesis testing procedure is subdivided into a number of subsequent steps, minimally including the following sequence:

1. The formulation of a practical problem in terms of a null hypothesis versus an alternative hypothesis.

2. The selection of an appropriate test statistic.

3. Determination of the level of significance, or analogously, the critical region.

4. Decision making in favour of one of the hypotheses.

Additionally, it is recommended to finish with an evaluation of the actual significance level, the so-called p-value. This can help to decide whether the data provides considerable or just some evidence against the null hypothesis ([12]).

In this article the main emphasis is put on the first and the second step of the hypothesis testing process. The formulation of hypotheses and the choice of an appropriate statistical test are considered to be the most crucial steps.

3.1 Null hypothesis - alternative hypotheses

To explain the rationale behind hypothesis testing, we consider the following example. In order to guarantee the high quality of computer chips, a quality engineer who is held responsible for quality control, takes, at regular times, randomly chosen batches of 100 products. The rate of defectives (sample proportion p) for each batch of products is calculated and compared with the quality standard (e.g. not more than 1% defectives). Based on the principles of hypothesis testing, one is able to undertake the appropriate action: if the rate of defectives of the *sample* (the batch of 100 products) indicates that less than 1% of *all products* is defective, then the production process is working properly; while, if the *sample* indicates that the rate of defectives of *all products* is *significant* bigger than 1%, the production process needs remedial attention. (Note the appearence of the word *significant*, which will be important throughout the rest of the text.)

Basically, there are now four alternatives: prove or disprove that the rate of defectives Π of the whole production process satisfies $\Pi < 1$, or $\Pi \leq 1$, or $\Pi > 1$, or finally $\Pi \geq 1$. Note that, in this situation, the two-sided $\Pi \neq 1$ is of no interest, as the knowledge whether $\Pi \neq 1$ does not indicate the direction of the action to be undertaken.

In the sequel, it will be shown that, for logical reasons, only the claims $\Pi < 1$ and $\Pi > 1$ are provable (because the other two contain equalities), and that in this specific situation, the test whether $\Pi > 1$, is to be preferred. Thus, the purpose of these quality inspections will *not* be to find batches with a rate of defectives lower than the quality standard, but to focus at finding batches of products for which the rate of defectives exceeds the quality standard. As we may assume that the quality engineer made a great effort to guarantee high-quality products, his aim at finding batches with too many defectives may seem illogical at first sight. However, this choice is based on logical reasoning solely, as explained in the sequel.

To investigate a statement about a population, e.g. the value of a population parameter (e.g. mean, proportion, variance...), the description of a population distribution, a comparison of two population means... one mathematically constructs an appropriate test statistic (the sample distribution), which explicitely contains this unknown population parameter, distribution function, difference of two population means.... To be able to use this test statistic to perform a statistical test, one has to assume that this unknown component is known, by assigning a specific value to it (normally the value one wants to investigate): e.g. the value of the population mean *is* ..., the population *is* normal distributed, the difference of the two population means *is* .... This assumption is called the null hypothesis H_0, or the *hypothesis of no difference.* To start, one assumes the null hypothesis about the population to be true. The main goal of the test is now to see if a sample taken from that population provides *significant* evidence, to reject the null hypothesis in favour of an alternative hypothesis H_1. The hypothesis H_1 states a *significant* difference in either one or two directions: smaller or larger (two one sided tests) or different from the assumed value (the two sided test). One accepts H_1 when the sample outcome points in the direction of H_1 and when this difference cannot be statistically explained by fluctuations of sample data under the H_0 assumption (thus the meaning of the word *significant*). Thus the policy is to be conservative towards H_0 unless the sample provides *significant* evidence to reject H_0 (under the assumption of H_0, the possibility to reject H_0 is denoted by α, the significance level, with usually $\alpha = 5\%$). It is very important to understand that, for logical reasons, one can only reject or not reject H_0, one can never accept (or proof) H_0: the equality appearing in the null hypothesis can never be shown, one can only perceive a *difference* which statistically cannot be explained by the expected sample fluctuations. The strongest result of a hypothesis test is therefor the rejection of H_0, (that is the acceptance of H_1), which is seen as the positive result. In brief, the goal of a test should be to accept H_1.

From the above paragraph, the hypothesis of a one-sided test with $H_1 : \Pi > 1$ would be

$$H_0 : \Pi = 1 \quad H_1 : \Pi > 1.$$

In some textbooks however, one uses a composite null hypothesis, containing the inequality which is not tested in H_1 (here: $\Pi < 1$), thus writing

$$H_0 : \Pi \leq 1 \quad H_1 : \Pi > 1 \; .$$

To avoid confusions, it should be made clear to students that the underlying assumption is still the equality in H_0 and that one rejects H_0 if the sample provides enough evidence in the direction indicated in H_1 ($\Pi > 1$). Furthermore, a non rejection of H_0 does not mean that $\Pi \leq 1$, but only means that the sample does not contradict that $\Pi \leq 1$ or does not indicate that Π is significant larger than 1 (hence can still be (a little) larger than 1 anyhow).

From the above explanation, three strategies remain:

Test 1: test whether the rate of defectives Π of the whole population is significant smaller than 1:

$H_0 : \Pi = 1 \quad H_1 : \Pi < 1$ (or $H_0 : \Pi \geq 1 \quad H_1 : \Pi < 1$)

Test 2: test whether the rate of defectives Π is significant bigger than 1:

$H_0 : \Pi = 1 \quad H_1 : \Pi > 1$ (or $H_0 : \Pi \leq 1 \quad H_1 : \Pi > 1$)

Test 3: test whether the rate of defectives Π is significant different from 1:

$H_0 : \Pi = 1 \quad H_1 : \Pi \neq 1.$

As the choice between one-sided (case 1, 2) and two-sided testing (case 3) is most often no problem, we further only discuss test 1 and 2. First of all, test 1 and 2 are not equivalent as they are not mutual each others logical opposite. A non rejection of H_0 in test 1 does not necessarily imply the rejection of H_0 in test 2). Secondly, what happens if $\Pi = 1$ and none of H_1 can be accepted? In practice, it could be appropriate to continue the production if $\Pi \leq 1$ and to alter the production if $\Pi > 1$. Thus if one wants to show that either $\Pi > 1$ or $\Pi \leq 1$, then, because of the equality in $\Pi \leq 1$, one can only hope to prove H_1: $\Pi > 1$ and in case of a non rejection of the null hypothesis, one is forced to continu with the present production process, although there is no real positive underlying indication for this.

Finally, how does one decide between test 1 and test 2? At first sight, one would prefer to prove the H_1 hypothesis that is believed to be true by the quality engineer, who made great efforts to guarantee quality. What happens then if a consumer has reason to believe the opposite? Similarly, one can let the choice depend on the outcome of a sample taken, i.e. for a sample with $p < 1$, perform test 1 (as in this case, the outcome of test 2 is always known: rejection of H_0 in test 2 is impossible); for a sample with $p > 1$, a similar argument leads to the preference of test 2. However, it seems unlogical that 2 different personal opinions or two different samples would leed to different procedures. We now explain that for the above quality control example, test 2 is the most rational choice (although the opposite of H_1 is believed to be true).

Recall that the goal of a statistical hypothesis test should always be to accept H_1, (as this is the stronger result, and the only positive result), so one should prefer the acceptance of the H_1 hypothesis with the most drastic consequences, thus the H_1 hypothesis that you *only* want to accept if it is explicitely supported by the sample data. In the context of the quality control example, the most drastic positive result for both engineer and consumer, is the constatation that the rate of defectives is too high (because then the system needs remedial, which means extra costs/work). Hence, in the above situation, of the two possible rejections of H_0, the second one would be the most drastic and test 2 is to be preferred. Indeed, test 2 is always decisive, no matter what the sample outcome is: if one rejects H_0 then one should definitely change the production line (positive evidence); if one does not reject H_0, it is best (cheapest) to keep the present production as there is no real positive indication to change it, and test 1 can never indicate otherwise. Test 1 is not decisive (hence sometimes useless): if one rejects H_0, then one knows for sure that the production needs no remedial attention, but if one cannot reject H_0, (Π is not sufficiently small, hence Π could still exceed the quality standard), one cannot afford to risk of taking a wrong decision, and one has to perform test 2 anyhow.

Note that the preference for test 2 depends on the actual situation of the problem and the goal of the test. If the quality engineer gets positive response (e.g. a reward) if the rate of defectives is significant smaller than 1, and zero response (no reward, no punishment) if the rate of defectives if bigger than 1, then test 1 is to be preferred.

What happens if none of both possible H_1 hypothesis is to be preferred? Suppose for example that two scientists have opposite believes about a parameter μ. Based on their own experiments, the first scientist wants to show H_1: $\mu < \mu_0$ (because his tests indicate this direction), and scientist two wants to show H_1: $\mu > \mu_0$. If none of both H_1 hypothesis is to be preferred (no clear positive result), then we suggest you do both tests with their own related sample result. Twice a rejection of the H_0 hypotheses is unlikely. A single rejection of H_0 is decisive in favour of one of the two scientists. A non rejection would imply a difference from the value μ_0 that is not significant (two-sided test).

To conclude: as the purpose of a hypothesis test is to find sample evidence to reject the null hypothesis, the H_1 hypothesis of a one-sided test is the hypothesis with the most desired positive result (provided that H_1 does not include a statement about equality). In this case, the null hypotheses contains the hypothesis not believed to be true. For a two-sided test, H_1 always contains the inequality.

An example, similar to the one discussed above, helps to introduce other concepts of hypothesis testing, such as type I error (i.e. false positive) or significance level, type II error (i.e. false negative), critical value or region, one-tailed test, two-tailed test, power of a test, etc...

3.2 Selecting an appropriate statistical test

Once the decision-maker has formulated both the null hypothesis and the alternative hypothesis, an appropriate statistical test should be chosen to solve the problem. The appropriateness of a statistical test for testing a given null hypothesis is primarily affected by:

1. The *nature of the problem*, represented by the particular form of the null hypothesis (e.g. the difference between two group means is assumed to be zero, the correlation between two variables is zero, ... etc.), and the alternative hypothesis (e.g. a one-tailed test versus a two-tailed test).

2. The *nature of the data*, represented by the underlying distribution, measurement level, homoscedasticity (i.e. equality of variances), dependency of measurements,...

As mentioned previously, we noticed that many statistical textbooks reserve two different chapters on hypothesis testing: one including parametric hypothesis tests, and an other one on non-parametric hypothesis tests. The reason why parametric and non-parametric statistical tests are treated separately is of purely technical purposes (e.g. different ways of deriving an appropriate test statistic; differences relating to the nature of the data, differences in applicability, ...). Although this separate treatment of parametric and non-parametric hypothesis tests is beneficial for a statistician, we do not believe that it is very helpful to the decision-maker. The first concern of a decision-maker must be to identify the nature of the problem. After he has found several techniques to solve the problem (e.g. several alternative hypothesis tests), other criteria such as the nature of the data, the robustness against violations of the assumptions, and the power of the hypothesis test may be taken into account to choose the most appropriate test.

For this reason we believe that parametric and non-parametric hypothesis tests should be discussed jointly. Furthermore, the process of analyzing the problem and the selection of an appropriate hypothesis test must be simplified by appropriate tools, such as classification tables, flowcharts, and expert-like systems. Many authors have demonstrated the advantages of such tools in the statistical training of students (e.g. [13], [14]). The main issues that should be dealt with are the following

1. the number of samples: explain to students that if one wants to compair k populations, then one has k samples, each with their own sample size. Especially with paired samples (e.g. the Friedmann test) students often fail to determine the number of samples and the sample sizes. Of course this can depend on the purpose of the test: if a data set contains results of the last 5 years of 10 companies, one can ask to compare the 10 companies (hence 10 samples), or to compare the 5 years (hence 5 samples).

2. matched or unmatched samples: This is one of the hardest problems for students. It is our experience that the dataset itself can mislead the student (data can be presented in a table as matched, although there is no statistical reason to assume dependent variables). We suggest that a logical analysis of the test problem itself always indicates whether the problem requires matched samples. To end, one explains in brief why the non-matched theory is not correct.

3. the object of the hypothesis test: population mean, location, spread, proportion, distribution, difference, dependence...

4. the scale of the data (nominal, ordinal, interval, ratio). It should be made clear that the scale of the data is decisive for the test choice, hence very important (thus avoiding for example the use of a dichotome variable while assuming normality of this variable)

5. the constraints: mention the constraints that have to be verified (and the tests that can be used to check these constraints), and mention alternative tests (using a lower scale assumption).

The above five steps are easily put into a flowchart. Using the information of the flowchart, the reader should be able to find the correct hypothesis test on the more detailed table on pages 9-16.

One random sample: $X_1, \ldots, X_n$							
Test involves	Hypotheses	Min. scale	Assumptions	Sample variable	Test statistic under H_0	Name	p-value
population mean μ	a) $H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$ b) $H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$ c) $H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	interval	i) $X \sim N(\mu, \sigma^2)$, σ^2 unknown ii) $X \sim N(\mu, \sigma^2)$, σ^2 known iii) $n \geq 30$, σ^2 unknown iv) $n \geq 30$, σ^2 known	mean $\overline{X}$ variance (S^2)	i) $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$ ii) $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ iii) $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$ iv) $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$	Z-test or T-test for mean	a) $2P(T \geq \lvert T_{obs}\rvert)$ or $2P(Z > \lvert Z_{obs}\rvert)$ b) $P(T \geq T_{obs})$ or $P(Z > Z_{obs})$ c) $P(T \leq T_{obs})$ or $P(Z < Z_{obs})$
population median m or population location	a) $H_0 : m = m_0$ $H_1 : m \neq m_0$ b) $H_0 : m \leq m_0$ $H_1 : m > m_0$ c) $H_0 : m \geq m_0$ $H_1 : m < m_0$	ordinal		$N = \#$ data $< m_0$	$N \sim b(n, \frac{1}{2})$ for n large: $Z = \frac{2N-n}{\sqrt{n}} \sim N(0,1)$	Binomial test or sign test	a) $2P(N \leq N_{obs})$ if $N_{obs} \leq n/2$, $2P(N \geq N_{obs})$ if $N_{obs} \geq n/2$ b) $P(N \leq N_{obs})$ c) $P(N \geq N_{obs})$
	a) $H_0 : m = m_0$ $H_1 : m \neq m_0$ b) $H_0 : m \leq m_0$ $H_1 : m > m_0$ c) $H_0 : m \geq m_0$ $H_1 : m < m_0$	ordinal		rank deviations from m_0 $T^+ = \sum$ pos. rankings $T^- = \lvert\sum$ neg. rankings$\rvert$ $T = Min(T^+, T^-)$	$T_{obs} = Min(T^+_{obs}, T^-_{obs})$ $\sim$ see tables if $n > 20$: $Z = \frac{T-n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$ $\sim N(0,1)$	Wilcoxon signed rank test for median	a) $2P(T \leq T_{obs})$ b) $P(N \leq T^-_{obs})$ c) $P(T \leq T^+_{obs})$
population proportion Π	a) $H_0 : \Pi = \Pi_0$ $H_1 : \Pi \neq \Pi_0$ b) $H_0 : \Pi \leq \Pi_0$ $H_1 : \Pi > \Pi_0$ c) $H_0 : \Pi \geq \Pi_0$ $H_1 : \Pi < \Pi_0$	nominal (0-1)	$np > 5$ and $n(1-p) > 5$	proportion P	$Z = \frac{P-\Pi_0}{\sqrt{\Pi_0(1-\Pi_0)/n}} \sim N(0,1)$	Z-test for proportion	a) $2P(Z \geq \lvert Z_{obs}\rvert)$ b) $P(Z \geq Z_{obs})$ c) $P(Z \leq Z_{obs})$

One random sample: $X_1, \ldots, X_n$ (continued)

Test involves	Hypotheses	Min. scale	Assumptions	Sample variable	Test statistic under H_0	Name	p-value
population variance σ^2	a) $H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$ b) $H_0 : \sigma^2 \leq \sigma_0^2$ $H_1 : \sigma^2 > \sigma_0^2$ c) $H_0 : \sigma^2 \geq \sigma_0^2$ $H_1 : \sigma^2 < \sigma_0^2$	interval	$X \sim N(\mu, \sigma^2)$	variance S^2	$\chi = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$	χ^2-test for variance	a) $2P(\chi \geq \chi_{obs})$ if $\chi_{obs} \geq n-1$, $2P(\chi \leq \chi_{obs})$ if $\chi_{obs} < n-1$ b) $P(\chi \geq \chi_{obs})$ c) $P(\chi \leq \chi_{obs})$
population standard deviation σ	a) $H_0 : \sigma = \sigma_0$ $H_1 : \sigma \neq \sigma_0$ b) $H_0 : \sigma \leq \sigma_0$ $H_1 : \sigma > \sigma_0$ c) $H_0 : \sigma \geq \sigma_0$ $H_1 : \sigma < \sigma_0$	interval	$X \sim N(\mu, \sigma^2)$ $n \geq 30$	stand. deviation S	$Z = \frac{S-\sigma_0}{\sigma_0}\sqrt{2(n-1)}$ $\sim N(0,1)$	Z-test for standard deviation	a) $2P(Z \geq \lvert Z_{obs} \rvert)$ b) $P(Z \geq Z_{obs})$ c) $P(Z \leq Z_{obs})$
goodness of fit, population distribution	H_0 : pop. distr. is ... (class freq. Np_i=...) H_1 : pop. distr. is not ...	nominal	$N \geq 50$, at most $5\%N$ cells with $Np_i < 5$	cell freq. f_i	$\chi = \frac{\sum (f_i - Np_i)^2}{Np_i}$ $\sim \chi^2_{(n-r-1)}$ $r = \#$ estim. param.	χ^2-test	$P(\chi \geq \chi_{obs})$
	H_0 : pop. distr. is ... (cum. distr. $F_X = ...$) H_1 : pop. distr. is not ...	ordinal		cum. distr. S_X	$D = \max_x \lvert F_X(x) - S_X(x) \rvert$	Kolmogorof Smirnov I (KS I)	$P(D \geq D_{obs})$
randomness	H_0 : randomness H_1 : non randomness	ordinal (nominal)		# runs R	$R = \#$ runs $\sim$ see tables for n large : $Z = \frac{R-\mu_R}{\sigma_R} \sim N(0,1)$ $\mu_R = \frac{2n_1n_2}{n_1+n_2} + 1$ $\sigma_R = \sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$	Wald Wolfowitz runs test	$2P(R \geq R_{obs})$ and $2P(R \leq R_{obs})$, or $2P(Z \geq Z_{obs})$
simultaneous proportions in 1 sample	$H_0 : p_1 = (p_1)_0, \ldots, p_n = (p_n)_0$ H_1 : not H_0	nominal	$N \geq 50$, at most $5\%N$ cells with $Np_i < 5$	cell freq. f_i	$\chi = \frac{\sum (f_i - Np_i)^2}{Np_i}$ $\sim \chi^2_{(n-1)}$	χ^2-test	$P(\chi \geq \chi_{obs})$

Two random samples: $(X_1)_1, \ldots, (X_1)_{n_1}$ and $(X_2)_1, \ldots, (X_2)_{n_2}$

Test involves	Hypotheses	Min. scale	Assumptions	Sample variable	Test statistic under H_0	Name	p-value
two population means	a) $H_0 : \mu_1 - \mu_2 = \mu_0$ $H_1 : \mu_1 - \mu_2 \neq \mu_0$ b) $H_0 : \mu_1 - \mu_2 \leq \mu_0$ $H_1 : \mu_1 - \mu_2 > \mu_0$ c) $H_0 : \mu_1 - \mu_2 \geq \mu_0$ $H_1 : \mu_1 - \mu_2 < \mu_0$	interval	i) $X_1, X_2 \sim N(\mu, \sigma^2)$, σ_1^2, σ_2^2 known ii) $X_1, X_2 \sim N(m, \sigma^2)$, $\sigma_1^2 = \sigma_2^2$ and unknown iii) $X_1, X_2 \sim N(m, \sigma^2)$, $\sigma_1^2 \neq \sigma_2^2$ and unknown iv) $n_1 \geq 30, n_2 \geq 30$, σ_1^2, σ_2^2 known v) $n_1 \geq 30, n_2 \geq 30$, σ_1^2, σ_2^2 unknown	means $\overline{X}_1, \overline{X}_2$ variances S_1^2, S_2^2	i) $Z = \frac{\overline{X}_1 - \overline{X}_2 - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ ii) $T = \frac{\overline{X}_1 - \overline{X}_2 - \mu_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$ $s_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ iii) $T' = \frac{\overline{X}_1 - \overline{X}_2 - \mu_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ $t'_{1-\alpha} = \frac{\frac{S_1^2}{n_1} t_{(1-\alpha, n_1-1)} + \frac{S_2^2}{n_2} t_{(1-\alpha, n_2-1)}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ iv) $Z = \frac{\overline{X}_1 - \overline{X}_2 - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ v) $Z = \frac{\overline{X}_1 - \overline{X}_2 - \mu_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	Z-test or T-test	a) $2P(Z > \lvert Z_{obs} \rvert)$ or $2P(T > \lvert T_{obs} \rvert)$ or $2P(T' > \lvert T'_{obs} \rvert)$ b) $P(Z > Z_{obs})$ or $P(T > T_{obs})$ or $P(T' > T'_{obs})$ c) $P(Z < Z_{obs})$ or $P(T < T_{obs})$ or $P(T' < T'_{obs})$
two population medians	a) $H_0 : P(X_1 < X_2) = 0.5$ $H_1 : P(X_1 > X_2) \neq 0.5$ b) $H_0 : P(X_1 < X_2) \leq 0.5$ $H_1 : P(X_1 > X_2) > 0.5$ c) $H_0 : P(X_1 < X_2) \geq 0.5$ $H_1 : P(X_1 > X_2) < 0.5$	ordinal	$n_1 < n_2$	rankings within $(n_1 + n_2)$ sample $W_x = \sum \text{rank}(X_1)_i$	W_x $W'_x = n_1(n_1 + n_2 - 1) - W_x$ for n, m large: $Z = \frac{W \pm 0.5 - n_1(n_1+n_2+1)/2}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}}$ $\sim N(0, 1)$ (+0.5 for leftsided test)	Mann-Whitney	a) $2P(W < W_{obs})$ and $2P(W' < W'_{obs})$ b) $P(W < W_{obs})$ and $P(W' < W'_{obs})$ c) $P(W < W_{obs})$ and $P(W' < W'_{obs})$

two population proportions	a) $H_0 : \Pi_1 - \Pi_2 = \Pi_0$
	$H_1 : \Pi_1 - \Pi_2 \neq \Pi_0$
	b) $H_0 : \Pi_1 - \Pi_2 \leq \Pi_0$
	$H_1 : \Pi_1 - \Pi_2 > \Pi_0$
	c) $H_0 : \Pi_1 - \Pi_2 \geq \Pi_0$
	$H_1 : \Pi_1 - \Pi_2 < \Pi_0$
		nominal (0-1)	$nP_1 > 5$ or $n(1 - P_1) > 5$ and $nP_2 > 5$ or $n(1 - P_2) > 5$	prop. P_1, P_2	if $\Pi_0 = 0$:
					$Z = \frac{P_1 - P_2 - \Pi_0}{\sqrt{P'(1-P')(\frac{1}{n_1} + \frac{1}{n_2})}}$
					with $P' = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$
					if $\Pi_0 \neq 0$:
					$Z = \frac{P_1 - P_2 - \Pi_0}{\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}}$
						Z-test	a) $2P(Z \geq \lvert Z_{obs} \rvert)$
							b) $P(Z \geq Z_{obs})$
							c) $P(Z \leq Z_{obs})$

Two random samples: $(X_1)_1, \ldots, (X_1)_{n_1}$ and $(X_2)_1, \ldots, (X_2)_{n_2}$ (continued)

Test involves	Hypotheses	Min. scale	Assumptions	Sample variable	Test statistic under H_0	Name	p-value
two population variances	a) $H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$ b) $H_0 : \sigma_1^2 \leq \sigma_2^2$ $H_1 : \sigma_1^2 > \sigma_2^2$ c) $H_0 : \sigma_1^2 \geq \sigma_2^2$ $H_1 : \sigma_1^2 < \sigma_2^2$	interval	$X_1, X_2 \sim N(\mu, \sigma^2)$ $S_1^2 > S_2^2$	variances S_1^2, S_2^2	$F = \frac{S_1^2}{S_2^2} \sim F_{(n_1-1, n_2-1)}$	F-test	a) $2P(F \geq F_{obs})$ if $s_1^2 > s_2^2$ b) $P(F \geq F_{obs})$ c) $P(F \leq F_{obs})$
comparison of two population distr.	H_0 : 2 populations are equally distributed H_1 : not H_0	nominal	$N \geq 50$, at most $5\%N$ cells with $F_{ij} < 5$	cell freq. f_{ij}	$\chi = \frac{\sum(f_{ij}-F_{ij})^2}{F_{ij}}$ $\sim \chi^2_{((r-1)(k-1))}$ $r = \#$ rows, $k = \#$ columns $F_{ij} = n_i n_j / N$	χ^2-test	$P(\chi \geq \chi_{obs})$

Paired observations or two matched samples $(X_1, Y_1), \ldots, (X_n, Y_n)$

Test involves	Hypotheses	Min. scale	Assumptions	Sample variable	Test statistic under H_0	Name	p-value
2 population means	a) $H_0 : \mu_D = \mu_0$ $H_1 : \mu_D \neq \mu_0$ b) $H_0 : \mu_D \leq \mu_0$ $H_1 : \mu_D > \mu_0$ c) $H_0 : \mu_D \geq \mu_0$ $H_1 : \mu_D < \mu_0$	interval	$X, Y \sim N(m, \sigma^2)$	$D_i = X_i - Y_i$, $\overline{D}$, s_D	$T = \frac{\overline{D}-\mu_0}{\sigma_D/\sqrt{n}} \sim t_{n-1}$	T-test	a) $2P(T \geq \|T_{obs}\|)$ b) $P(T \geq T_{obs})$ c) $P(T \leq T_{obs})$
2 population medians or population locations	a) $H_0 : P(+) = P(-)$ $H_1 : P(+) \neq P(-)$ b) $H_0 : P(+) \leq P(-)$ $H_1 : P(+) > P(-)$ c) $H_0 : P(+) \geq P(-)$ $H_1 : P(+) < P(-)$	ordinal		$\#(+), \#(-)$	$a)$ $N = \mathrm{Min}(\#+, \#-) \sim b(n, \frac{1}{2})$ $b)$ $N = \#(-) \sim b(n, \frac{1}{2})$ $c)$ $N = \#(+) \sim b(n, \frac{1}{2})$ for n large: $Z = \frac{(K \pm 0.5) - 0.5\,n}{0.5\sqrt{n}}$ with $K + 0.5$ if $K < 0.5n$ and $K - 0.5$ if $K > 0.5n$	Binomial test or sign test	a) $2P(N \leq N_{obs})$ if $N_{obs} \leq n/2$ $2P(N \geq N_{obs})$ if $N_{obs} \geq n/2$ b) $P(N \leq N_{obs})$ c) $P(N \geq N_{obs})$
	a) $H_0 : \mu_1 = \mu_2$ $H_1 : \mu 1 \neq \mu_2$ b) $H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$ c) $H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$D_i = X_i - Y_i$ interval		rank $(X_i - Y_i)$ $T^+ = \sum$ pos. rankings $T^- = \|\sum$ neg. rankings$\|$ $T = Min(T^+, T^-)$	$T = \mathrm{Min}(T^+, T^-) \sim$ see tables if $n > 20$: $Z = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \sim N(0,1)$	Wilcoxon signed rank for median	a) $2P(T \leq T_{obs})$ b) $P(T \leq T^-_{obs})$ c) $P(T \leq T^+_{obs})$
comparison of 2 dichotome distributions	$H_0 : P(X_i = 0, Y_i = 1) =$ $P(X_i = 1, Y_i = 0) \quad \forall i$ $H_1 : P(X_i = 0, Y_i = 1) \neq$ $P(X_i = 1, Y_i = 0) \quad \forall i$	nominal (0-1)		$A = \#\{i \| X_i = 0, Y_i = 1\}$ $B = \#\{i \| X_i = 0, Y_i = 0\}$ $C = \#\{i \| X_i = 1, Y_i = 1\}$ $D = \#\{i \| X_i = 1, Y_i = 0\}$	$\chi^2 = \frac{(\|A-D\|-1)^2}{A+D} \sim \chi^2_1$	McNemar	$P(\chi^2 > \chi^2_{obs})$

Paired observations or two matched samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ (continued)

Test involves	Hypotheses	Min. scale	Assumptions	Sample variable	Test statistic under H_0	Name	p-value

independence, correlation	a) $H_0 : \rho = \rho_0$
	$H_1 : \rho \neq \rho_0$
	b) $H_0 : \rho \leq \rho_0$
	$H_1 : \rho > \rho_0$
	a) $H_0 : \rho \geq \rho_0$
	$H_1 : \rho < \rho_0$
		interval	(X, Y) biv. normal	$R = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2}}$	if $\rho_0 = 0$:
					$T = R\sqrt{\frac{n-2}{1-R^2}} \sim t_{n-2}$
					if $\rho_0 \neq 0$:
					$Z = \frac{Z_R - Z_{\rho_0}}{S_{Z_R}} \sim N(0,1)$
					$Z_R = \frac{1}{2}\ln\frac{1+R}{1-R}$
					$Z_\rho = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$
					$S_{Z_R} = 1/\sqrt{n-3}$
						Pearson correlation	a) $2P(T > \lvert T_{obs}\rvert)$ or $2P(Z > \lvert Z_{obs}\rvert)$
							b) $P(T > T_{obs})$ or $P(Z > Z_{obs})$
							c) $P(T > T_{obs})$ or $P(Z > Z_{obs})$

independence, correlation	a) $H_0 : \rho = 0$
	$H_1 : \rho \neq 0$
	b) $H_0 : \rho \leq 0$
	$H_1 : \rho > 0$
	a) $H_0 : \rho \geq 0$
	$H_1 : \rho < 0$
		ordinal		D_i= rank X_i - rank Y_i
				$R_s = 1 - \frac{6\sum_{i=1}^n D_i^2}{n^3-n}$
					if $n \leq 30$:
					$R_s \sim$ tables
					if $n > 30$:
					$Z = R_s\sqrt{n-1}$
					$\sim N(0,1)$
						Spearman rank correlation	a) $2P(R_s > R_{s_{obs}})$ or $2P(Z > \lvert Z_{obs}\rvert)$
							b) $P(R_s > R_{s_{obs}})$ or $P(Z > Z_{obs})$
							c) $P(R_s < R_{s_{obs}})$ or $P(Z < Z_{obs})$

	a) $H_0 : \tau = 0$
	$H_1 : \tau \neq 0$
	b) $H_0 : \tau \leq 0$
	$H_1 : \tau > 0$
	a) $H_0 : \tau \geq 0$
	$H_1 : \tau < 0$
		ordinal		$S = N_c - N_d$
				$N_c = \#$ concordant pairs
				$N_d = \#$ discordant pairs
					if $n \leq 10$:
					$\tau = \frac{2S}{n(n-1)}$
					$\sim$ tables
					if $n > 10$:
					$Z = \frac{\tau - 0}{\sqrt{2(2n+5)9n(n-1)}}$
					$\sim N(0,1)$
						Kendall's tau correlation	a) $2P(\tau > \tau_{obs})$ or $2P(Z > \lvert Z_{obs}\rvert)$
							b) $P(\tau > \tau_{obs})$ or $P(Z > Z_{obs})$
							c) $P(\tau > \tau_{obs})$ or $P(Z > Z_{obs})$

independence	H_0 : 2 variables are independent H_1 : 2 variables are dependent	nominal (0-1)	$N \geq 50$, at most $5\%N$ cells with $F_{ij} < 5$	cell freq. f_{ij}	$\chi = \frac{\sum(f_{ij}-F_{ij})^2}{F_{ij}}$ $\sim \chi^2_{((r-1)(k-1))}$ $r = \#$ rows $k = \#$ columns $F_{ij} = n_i n_j / n$	χ^2-test	$P(\chi \geq \chi_{obs})$

k random samples ($k \geq 3$)

Test involves	Hypotheses	Min. scale	Assumptions	Sample variable	Test statistic under H_0	Name	p-value
comparison of k population means	$H_0 : \mu_1 = \mu_2 = \ldots \mu_k$ H_1 : not H_0	interval	$X_i \sim N(\mu_i, \sigma^2)$ $\sigma_i^2 = \sigma^2 \quad \forall i$	$S_e^2 = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_j)^2}{n-k}$ $S_k^2 = \frac{\sum_{j=1}^{k} n_j (\overline{X}_j - \overline{\overline{X}})^2}{k-1}$	$F = \frac{S_k^2}{S_e^2} \sim F_{(k-1,n-k)}$	One way anova	$P(F > F_{obs})$
comparison of k populations	H_0 : k populations equally distributed H_1: not H_0	ordinal		$R_j = \sum$ (rankings of sample j within full sample)	$KW = \frac{12}{n(n+1)}(\sum_{j=1}^{k} \frac{R_j^2}{n_j}) - 3(n+1)$ $\sim \chi_{k-1}^2$	Kruskal Wallis	$P(\chi_{k-1}^2 > KW_{obs})$
comparison of k population distributions	$H_0 : k$ pop. distributions are equal H_1 : not H_0	nominal	$N \geq 50$, at most 5%N cells with $F_{ij} < 5$	cell freq. f_{ij}	$\chi = \frac{\sum (f_{ij} - F_{ij})^2}{F_{ij}}$ $\sim \chi^2_{((r-1)(k-1))}$ r = # rows, k = # columns $F_{ij} = n_i n_j / N$	χ^2-test	$P(\chi \geq \chi_{obs})$

k matched samples $(k \geq 3)$							
Test involves	Hypotheses	Min. scale	Assumptions	Sample variable	Test statistic under H_0	Name	p-value
comparison of 2 treatments factors	H_0 : no main effects from row factors (row means equal) H_1 : not H_0 and H_0 : no main effects from column factors (colum means equal) H_1 : not H_0	interval	$X_i \sim N(\mu_i, \sigma^2)$ $\sigma_i^2 = \sigma^2 \ \forall i$	$S_e^2 = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij}-\overline{X}_{.j}-\overline{X}_{i.}+\overline{\overline{X}})^2}{(n-1)(k-1)}$ $S_k^2 = \frac{n\sum_{j=1}^{k}(\overline{X}_{.j}-\overline{\overline{X}})^2}{k-1}$ $S_n^2 = \frac{k\sum_{i=1}^{n}(\overline{X}_{i.}-\overline{\overline{X}})^2}{n-1}$	a) $F = \frac{S_k^2}{S_e^2} \sim F_{(k-1,(n-1)(k-1))}$ b) $F = \frac{S_n^2}{S_e^2} \sim F_{(n-1,(n-1)(k-1))}$	Two way anova	$P(F > F_{obs})$
comparison of k related populations	H_0 : k populations equally distributed H_1: not H_0	ordinal		$R_j = \sum$ (rankings within sample j)	$\chi^2 = \frac{12}{nk(k+1)}\sum_{j=1}^{k} R_j^2 - 3n(k+1)$ $\sim$ tables for k or n large: $\chi^2 \sim \chi^2_{k-1}$	Friedmann	$P(\chi^2_{k-1} > \chi^2_{obs})$
comparison of k related populations	H_0 : k populations equally distributed H_1: not H_0	nominal (0-1)		C_j= column total L_i= row total	$Q = \frac{(k-1)(k\sum_{j=1}^{k}C_j^2-(\sum_{j=1}^{k}C_j)^2)}{k\sum_{i=1}^{n}L_i-\sum_{i=1}^{n}L_i^2}$ $\sim \chi^2_{k-1}$	Cochran Q	$P(Q > Q_{obs})$

4 Experiments

Expermiments can be done (and are done by the authors) to demonstrate the positive influence of the above table and flowchart. Indeed, this showed that students using the above aid perform better on exams than students having a classical table (with parametric and non-parametric tests separated). However, this does not provide sufficient proof that the students have also a better understanding of hypothesis testing. Therefor we prefer not to pursue this matter.

5 Conclusions

The aim of this article is to demonstrate two important interrelated problems in statistical education. Both problems are linked with the hypothesis testing chapters in traditional business statistics textbooks.

The first problem includes the formulation of the null hypothesis as the hypothesis of no difference, and the alternative hypothesis which is represented as the hypothesis with the strongest positive result.

The second problem has to do with the separate treatment of parametric and non-parametric hypothesis testing. The use of some aids, as there are classification tables, flow-charts and expert-like systems can partly overcome this problem.

References

[1] Abranovic, W. A., *Statistical Thinking and Data Analysis: Methods for Managers*, Reading, Massachusetts, Addison-Wesley, 1977.

[2] Anderson D.R, Sweeney D.J. and Williams T.A, Statistics for Business and Economics, (Fifth ed.) West Publishing Company, Minneapolis, 1993.

[3] Barrow, M., Statistics for Economics Accounting and Business Studies, London, Longman, 1991.

[4] Keller G.B, Warrack H. and Bartel H., Statistics for Management and Economics: A Systematic Approach, (Second ed.), Wadsworth, 1990.

[5] Kohler H., Statistics for Business and Economics,(Third ed.), Harper Collins Publisher, 1994.

[6] Mason R.D and Lind D.A., Statistical Techniques in Business and Economics, (Eight ed.), Duxbury Press, Boston Massachusetts, 1993.

[7] Mendenhall W. and Reinmuth J.E., Statistics for Management and Economics, (Fourth edition), Duxbury Press, Boston Massachusetts 1992.

[8] Newbold P., Statistics for Business and Economics, (Fourth ed.) Prentice-Hall, Englewood Cliffs, New Jersey, 1995.

[9] Triola M.F and Franklin L.A., Business Statistics: Understanding Populations and Processes, Addison-Wesley, Reading, Massachussets 1994.

[10] Wonnacot T.H. and Wonnacot R.J., Introductory Statistics for Business and Economics, (Fourth ed.), J. Wiley, New York 1990.

[11] Henkel R.A., Tests of Significance, Quantitative Applications in the Social Sciences, Sage, London 1990.

[12] Kanji G.K, 100 Statistical Tests, Sage, London 1994.

[13] Hand D.J., Expert Systems in Statistics, The Knowledge Engineering Review, 1, p2-10, 1986.

[14] Andrews F.M., Klem L., Davidson T.N., O'Malley P.M. and Rodgers W.L., A Guide for Selecting Statistical Techniques for Analyzing Social Science Data, The Institute for Social Research, The University of Michigan, Ann Arbor, Michigan 1981.