Model Assisted Statistics and Applications 7 (2012) 143–157
DOI 10.3233/MAS-2011-0222
IOS Press

Teaching Statistics

A tutorial on teaching hypothesis testing

W.J. Post*, M.A.J. van Duijn and A. Boomsma
Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands

Abstract. How can we teach graduate-level students the principles of hypothesis testing in order to improve their skills in application and interpreting hypothesis test results? This was one of the main challenges in our course *Applied Statistics*. Although most students, all potentially future researchers in social and behavioural sciences, were not specifically interested in statistics, it seemed a good idea to teach them the essentials of three approaches to statistical inference introduced by Fisher, Neyman and Pearson, and Bayesian statisticians.

To make the rather subtle differences between the inferential approaches and associated difficult statistical concepts more attractive and accessible to students, a chance game using two dice was used for illustration. We first considered an experiment with simple hypotheses showing the three inferential principles in an easy way. The experiment was then extended to a more realistic setting requiring more complicated calculations (with R-scripts), to satisfy the more advanced students.

We think that our lectures have enabled a deeper understanding of the role of statistics in hypothesis testing, and the apprehension that current inferential practice is a mixture of different approaches to hypothesis testing.

Keywords: Fisher, Neyman-Pearson, Bayesian inference, hypothesis tests, p-values, likelihood ratio

1. Introduction

As statisticians in a faculty of behavioural and social sciences, we work with colleagues and students whose first interest is not statistics but rather psychology, education or sociology. For them, statistical analysis is primarily about how to get substantive results and conclusions. Hypothesis testing procedures are performed rather routinely and results of those tests are summarized simply by interpreting non-significant findings as support for the null hypothesis, which implies that the presumed theory is false. Significant results, on the other hand, are interpreted bluntly as proof of the validity of the theory, or of the existence of 'an effect' of whatever size; see Snijders [28], who discussed several interpretative problems in hypothesis testing. Most of our students do not acknowledge that their significant or non-significant findings may be due to chance, i.e., there are errors of the first and second kind, although they may mention lack of power when hypothesized effects are not found.

We would like to contribute to a more thoughtful application and interpretation of hypothesis testing procedures. In line with the bachelor education endpoints in the social sciences, teaching statistics largely amounts to instructing students how to compute descriptive statistics, and to explain the basic principles of estimation and hypothesis testing. We strongly believe, however, that master students should be trained to critically assess the statistical design and analysis. A critical attitude is even more important for research master students who may pursue a Ph.D. education and become academic researchers. In our faculty, this is a selective group of high-achieving and motivated students. They are heterogeneous, however, with respect to discipline, country of origin (typically about half of them are foreign students), and most importantly, statistical training.

*Correspondence author: Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands. Tel.: +31 503 636 588, Fax: +31 503 636 521, E-mail: w.j.post@rug.nl.

ISSN 1574-1699/12/$27.50

With the aim to refresh and improve statistical knowledge, we designed a course *Applied Statistics*, compulsory for research master students. The purpose of the course is to teach students how to apply the principles of statistical design and analysis to empirical data, and how to interpret and report the results of such efforts, following Wilkinson and the Task Force [33]. The course consists of a theoretical part that offers a review of the main topics in statistics, and a practical part where students work individually on a research project from their own discipline. Basic statistical techniques are practiced with R [23], which software is new to most students. To enhance and extend students' understanding we decided to pay attention to the three main approaches to hypothesis testing. By comparing the testing procedures due to Fisher, Neyman and Pearson, and Bayes, we were able to reiterate the basic concepts of statistical analysis, and to emphasize that what may seem to be a single, elementary way of statistical analysis, is in fact a practice rooted in different philosophical approaches to hypothesis testing.

In the next section, we consider the initial content of our lecture, and then explain why this setup did not work for our students. To make the apparently tough theory more attractive and less abstract without losing its essential information, we were looking for a more appealing illustration. Since a direct application of the theory to the students' own individual research project was not feasible due to the large diversity of research topics, we took an example related to a board game familiar to our students. In Section 3, we explain the game setting, and present four key players, called Ronald, Jerzy-Egon, Thomas and Victor. From this setting, a research question is defined, leading to a – for didactical reasons – relatively simple experiment with a corresponding hypothesis testing problem. In Section 4, a theoretical introduction of each testing approach is given, followed by its application to the experiment. In Section 5, the simple experiment is extended to a more realistic setting. In this way different concepts associated with the three approaches are highlighted and explained several times, although more complex computations are needed. Next, the current statistical practice is discussed critically in Section 6. The paper ends by reflecting on our teaching experiences (Section 7). In the Appendix, an R-script is given for the more complicated calculations and plots in this paper.

2. Initial content of the lecture

An article written by Christensen [6] was our starting point. The paper illustrates the various approaches to hypothesis testing by a simple example. First, Fisher's significance test is explained, emphasizing that only one hypothesis is considered. The p-value of the test statistic is the criterion to evaluate the validity of the hypothesis, and interpreted as a measure for the (strength of) evidence against this hypothesis. Next, Neyman-Pearson (NP) hypothesis tests are discussed, in which an alternative hypothesis is introduced, as well as Type I and Type II errors. It is also explained that a rejection region is determined, using likelihood ratio statistics to accomplish that the Type I error equals a pre-specified level α. Finally, the Bayesian testing approach is explained briefly, starting with Bayes' theorem on conditional probabilities. The concept of the prior and posterior distributions of the model parameters is introduced and it is emphasized that hypotheses are evaluated using their posterior distribution.

During the theoretical lecture we emphasized the different philosophical points of view to demonstrate that a particular statistical inference and substantive conclusion may depend on the testing procedure. We quoted Christensen [6, p. 122] who claimed that "the philosophical basis of a Fisherian test is akin to proof by contradiction". We linked this idea to the falsification principle of Popper [22], known to most students. While Fisherian testing can thus be considered as a type of validation test for the null hypothesis, we stated that Neyman-Pearson's hypothesis tests are more suitable for choosing between two hypotheses. Their criterion is minimization of probabilities of wrong decisions, while defining these probabilities in terms of long-run relative frequencies [30]. We emphasized that the most salient feature of the Bayesian approach is that model parameters are considered and treated as random variables, not as unknown but fixed population quantities. We also discussed with the students that the current practice of statistical test procedures is actually a combination of Fisherian and NP tests, immediately adding that researcher's intuitive reasoning is inclined to be Bayesian; this in turn leads to the wrong interpretation of a p-value as the probability that the null hypothesis is true.

We were surprised to find that neither Christensen's paper nor our own lecture were well understood by the students. The students saw the substance of the lecture as a mere technical issue, not of any practical interest. As lecturers, on the other hand, we were getting more and more convinced that the material was essential for the

students' understanding of what they were doing when performing hypothesis tests and making statistical inferences in general. Moreover, in our opinion, the growing impact and applicability of Bayesian statistics makes it necessary to explain and discuss this approach in far more detail, even though most students were not at all familiar with Bayesian statistics, and also had a hard time remembering properties of conditional probability. Therefore we designed a lecture with a more appealing example of a chance game, explained in the next section.

3. Game setting

Consider a chance game where at the start three numbers between 2 and 12 are randomly assigned to each player. Next, each player throws two dice in turn. A player gains a victory point if the sum of the dice matches any of the assigned numbers. (In case multiple players are assigned that number, these players all gain a victory point.) The first player gaining 10 victory points is the winner of the game. This setup can be regarded as a simplified version of the board game *The Settlers of Catan* [29], used as the main example in our lecture. This game has become very popular worldwide [19] and is played by many of our students. For details of the game the interested reader is referred to [31].

Four students are introduced as players in the game: Ronald, J-E (short for Jerzy-Egon), Thomas, and Victor. The first three students are named after the statisticians whose testing approach we want to illustrate. The fourth student is Victor, who happens to be in a very weak position to earn victory points: the numbers 2, 3 and 4 were assigned to him. The probability for the outcomes is 1/36, 2/36, and 3/36, respectively, implying that in each trial of the game Victor's probability of winning a victory point is only 6/36. (For an overview of all combinations and probabilities of results, see the first columns in Table 1.) Nevertheless, Victor wins the game. The other players become suspicious and wonder whether the dice are fair. (In our lecture we present this as that they wonder whether Victor plays fair, because he was named after Victor Lustig, a famous fraudster who sold the Eiffel tower twice [32].) Ronald, J-E, and Thomas therefore decide to call for a simple and fast experiment to determine whether the dice are fair: just one throw of the pair of dice. The outcome of this experiment turns out to be 3. The question to be answered is: What will they conclude about the fairness of the dice?

Being loyal to their namesakes, the three students strongly disagree about the evaluation of the outcome of the experiment. Which test procedure should be applied? Of course, Ronald proposes Fisherian testing, J-E prefers a Neyman-Pearson test, and Thomas favours a Bayesian procedure. They decide to do all three procedures, and to compare the results and conclusions in terms of fair or unfair dice. Unfortunately, but not accidentally, it turns out that each procedure not only involves different decision rules but also leads to different conclusions about all outcomes, in particular the one regarding outcome 3. The procedures are described and discussed in the next section.

4. Three approaches to hypothesis testing

We start by reviewing Fisherian, Neyman-Pearson and Bayesian testing. We pay attention to the selection of a specific simple alternative, the choice of one-sided versus two-sided alternatives and to the specification of a prior distribution, and we illustrate these issues using the game example.

4.1. Fisherian testing

Fisher's approach to hypothesis testing requires the specification of only one hypothesis, the null or null model (H_0). If the sample data are *not consistent* with H_0, i.e., if improbable outcomes are obtained, H_0 is rejected. In this sense, Fisherian tests can be characterized as a validation procedure for the null model, and this makes only the distribution of the outcome under H_0 relevant. The criterion for rejecting H_0 is based on the p-value: the probability of observing a certain outcome or any more extreme outcome given that H_0 is true. The p-value can be regarded as a measure for the (strength of) evidence against H_0 [9, p. 80]. If the p-value is smaller than or equal to a pre-specified value, the so-called significance level, H_0 will be rejected. This reasoning is consistent with the falsification principle of Popper [22], who stated that one never can prove a theory, only falsify it.

In case of the data being *consistent* with the null model, i.e., a p-value larger than the significance level, not rejecting the null hypothesis does *not* prove that the null model is true: it just cannot be rejected.

In the setting of the game, applying Fisher's approach to conclude whether or not the dice are fair results in the specification of the null hypothesis as: the dice are fair. Suppose player Ronald chooses a significance level of 3/36, i.e., the null hypothesis must be rejected if $p \leqslant 3/36$. (Since there is only one throw of dice, the more common level of 5% would be too small.) For the outcome defined as the sum of two dice, the most unlikely values are 2 and 12, each with probability 1/36. Outcomes 3 and 11 are each obtained from two combinations of the dice, so both are twice as likely with probabilities 2/36. With the significance level set on 3/36, the decision rule is: reject H_0 if the outcome is 2 or 12. Ronald's decision rule is therefore: reject the hypothesis that the dice are fair, when the outcome is 2 or 12.

Since the outcome of the experiment is 3, Ronald's conclusion is: do not reject the hypothesis that the dice are fair. The question now is: Should he conclude that the dice are indeed fair? The answer is definitely NO. Not rejecting the null hypothesis means that the outcome of the experiment is consistent with the null hypothesis but it does not *prove* that the hypothesis is true.

Note that focusing on extreme outcomes of the distribution under the null hypothesis when there is no alternative hypothesis naturally leads to consideration of both large and small values. In this game setting, however, Fisher might have considered only small numbers as extreme, i.e., improbable, outcomes, because only small numbers would benefit Victor. For such a one-sided case, the most extreme small value is 2, with probability 1/36, and the probability of the nearest less extreme outcome 3 is 2/36 (see Table 1). The decision rule that exactly meets the significance level of 3/36 is now: reject H_0 that the dice are fair with an outcome equal to 2 or 3. It follows that choosing between one-sided and two-sided cases clearly affects the decision rule.

4.2. Neyman-Pearson testing

In Neyman and Pearson's approach not only a null hypothesis H_0 is considered but also a second, alternative hypothesis H_1. Two different types of errors could occur: falsely rejecting H_0 (Type I, occurring with probability α) and falsely accepting H_0 (Type II, with probability β). The probability of rejecting H_0 while H_1 is true is called the power of the test, equal to $1 - \beta$. Controlling for Type I error by specifying an upper bound for α, the best NP test is the one with the highest power, i.e., the smallest β. The most powerful test is determined by the likelihood ratio (LR), defined as the ratio between the probability of the outcome under H_1 and the probability of the outcome under H_0, as established by the Neyman-Pearson lemma [18, p. 74]. If the outcome probability is larger under H_1 than under H_0, the outcome is in favour of H_1. The NP decision rule is therefore: reject H_0 if the likelihood ratio exceeds a critical value, and accept H_0 otherwise. The critical value is determined by setting the probability of rejection under H_0 equal to a pre-specified upper bound of α. Note that the null and alternative hypothesis are treated differently: given an upper bound for α, β is minimized.

In their joint paper *On the use and interpretation of certain test criteria for purposes of statistical inference*, Neyman and Pearson [20] used terms as "accept" and "reject" hypotheses without any reference to some falsification principle. This is in contrast with 'do not reject H_0', a phrase more in line with Fisherian testing. Neyman and Pearson viewed the test procedure as a behavioural rule for choosing between two hypotheses, so that in the long run one "shall not be too often wrong" [21, p. 142]. This interpretation reveals another contrast to Fisher's approach as a validation of H_0.

Given this theoretical outline, we return to the game. J-E formulates the same null hypothesis as specified in Fisher's case: the dice are fair. Regarding the upper bound α, he chooses Ronald's significance level of 3/36. He might have specified the alternative hypothesis as general as: the dice are not fair. However, in order to facilitate calculating the set of outcome probabilities under the alternative hypothesis, he prefers a specific, simple alternative instead. There are many ways to manipulate the dice. Therefore, many alternative hypotheses can be specified, with associated sets of outcome probabilities. J-E has to choose a sensible and relevant alternative hypothesis for the game setting. He assumes that only one of the dice is manipulated so that the probability of the outcome (sum of dice) smaller than five is higher than in the case of a fair die. This implies that for the manipulated die some of the numbers 1, 2 or 3 must have a higher probability than 1/6. Furthermore, he assumes that the distinct outcomes 2 to 12 are still possible after manipulation: otherwise the unfairness of the die would have become evident during

Table 1
Probabilities for different outcomes for fair dice (H_0) and unfair dice (H_1) dice with its likelihood ratios

Combination	Outcome	Probability		Likelihood Ratio
	X	Fair dice	Unfair dice	
11	2	1/36	$1/6 \times 1/6 = 1/36$	1.00
12 21	3	2/36	$(1/6 \times 2/6) + (1/6 \times 1/6) = 3/36$	1.50
13 22 31	4	3/36	$2 \times (1/6 \times 2/6) + (1/6 \times 1/6) = 5/36$	1.67
14 41 23 32	5	4/36	$0 + (1/6 \times 1/6) + 2 \times (1/6 \times 2/6) = 5/36$	1.25
15 51 24 42 33	6	5/36	$0 + (1/6 \times 1/6) + 0 + 2 \times (1/6 \times 2/6) = 5/36$	1.00
16 61 25 34 52 43	7	6/36	$2 \times (1/6 \times 1/6) + 0 + 0 + 2 \times (1/6 \times 2/6) = 6/36$	1.00
26 62 53 35 44	8	5/36	$(1/6 \times 1/6) + 2 \times (1/6 \times 2/6) + 0 + 0 = 5/36$	1.00
36 63 45 54	9	4/36	$(1/6 \times 1/6) + (1/6 \times 2/6) + 0 + 0 = 3/36$	0.75
46 64 55	10	3/36	$(1/6 \times 1/6) + 0 + 0 = 1/36$	0.33
56 65	11	2/36	$(1/6 \times 1/6) + 0 = 1/36$	0.50
66	12	1/36	$(1/6 \times 1/6) = 1/36$	1.00
Gaining victory points		6/36	9/36	1.33
Gaining no victory points		30/36	27/36	0.90

the many throws in the game. This implies that the outcomes 1 and 6 of the manipulated die still have a positive probability. Hence, he chooses the following probability distribution for the unfair die: probabilities of an outcome of 2 and 3 are doubled, i.e. each 2/6, the probabilities of getting 1 and 6 are still 1/6. It follows that the manipulated die will never give the outcome 4 or 5.

In Table 1 it is shown that Victor's probability of gaining victory points is only 6/36 under the null hypothesis, and 9/36 under the alternative. According to Neyman-Pearson the null hypothesis is rejected for the largest values of the likelihood ratio. In this single-case experiment the highest value equals 1.67 for the outcome 4 with a Type I error probability of 3/36 (the pre-specified upper bound α). This means that the decision rule according to NP is: reject H_0 if the outcome is 4, otherwise accept H_0. Note that this is a different decision rule than Ronald's rejecting H_0 for an outcome equal to either 2 or 12 (two-sided case), as well as Ronald's rejecting H_0 for an outcome equal to either 2 or 3 (one-sided case). It can easily be verified (left as an exercise) that it would have made no difference for the decision rule of J-E, if he had only considered the relevant outcomes 2, 3 and 4, and 'other' (combining outcomes 5 through 12).

Since the outcome of the experiment is 3, the conclusion of J-E is: accept the hypothesis that the dice are fair. The question now is: Should he conclude that the dice are indeed fair? And the answer here is YES. In concluding so, however, J-E willingly takes the risk of making a wrong decision.

Note that the present results of the NP approach are largely determined by the choice of the particular simple alternative (as opposed to the specific form of 'unfairness' in the alternative). For example, if we would have doubled the probabilities of the outcomes 1 and 2 for the manipulated die, i.e., each 2/6, and each other number would have a probability of 1/12, the outcomes 2 and 3 would correspond to the largest likelihood ratio values. The decision rule according to Neyman-Pearson would then have been: reject H_0 if the outcome is 2 or 3; otherwise accept H_0, which is a decision rule similar to the one resulting from the one-sided case of Fisher's approach, except for a different phrasing of the decision rule. In case of choosing as alternative an unfair die with probability 1/3 for all outcomes smaller than 4, and with probability 1/12 for each of the outcomes greater than 3, the largest likelihood ratio value would have occurred for the outcomes 2, 3 and 4. To get the exact pre-specified upper bound α of 3/36 in this case, we would have to use a randomized test, for example by flipping a coin: if it comes up heads, reject H_0 when the outcome is 2 or 3, accept H_0 otherwise; if it comes up tails, reject H_0 for outcome 4, accept H_0 otherwise. As Christensen already concluded, it is rather difficult to convince anyone that such a randomized procedure is reasonable [6, p. 122]. The complete calculations of these two alternative experiments are left as an exercise. The most natural alternative is the composite alternative discussed in Section 5.2 below.

4.3. Bayesian testing

Harold Jeffreys, a contemporary of Fisher, Neyman and Pearson, introduced the principles of Bayesian statistics in his book *Theory of probability* [14]. According to Robert, Chopin and Rousseau [25, p. 141], "this book is rightly considered as the principal reference in modern Bayesian statistics".

Table 2
The probabilities given the hypotheses, the Bayes factor, and posterior probabilities for different outcomes

Outcome				Prior 1: odds 2		Prior 2: odds 1		Prior 3: odds 1/2	
X	P(X|H_0)	P(X|H_1)	LR	P(H_0|X)	P(H_1|X)	P(H_0|X)	P(H_1|X)	P(H_0|X)	P(H_1|X)
2	1/36	1/36	1.0	1/3	2/3	1/2	1/2	2/3	1/3
3	2/36	3/36	1.5	1/4	3/4	2/5	3/5	4/7	3/7
4	3/36	5/36	1.7	3/13	10/13	3/8	5/8	6/11	5/11
>4	30/36	27/36	0.9	15/42	27/42	10/19	9/19	20/29	9/29

Essential in Bayesian statistics is Bayes' theorem on conditional probabilities: Let H_1, H_2, ..., H_k be mutually exclusive and exhaustive events, and let D be some other event. Then, for each j = 1, 2, ..., k, the conditional probability of one of the events H_j given D can be written as

$$P(H_j|D) = \frac{P(D|H_j)P(H_j)}{P(D)} = \frac{P(D|H_j)P(H_j)}{\sum_{i=1}^{k} P(D|H_i)P(H_i)},$$

i.e., as a (weighted) conditional probability of D given H_j.

In Bayesian statistics and reasoning, the theorem is commonly applied to a set of hypotheses H_1, H_2, ..., H_k, and the data D. In this way, the parameters specified in the hypotheses, or the hypotheses themselves, are regarded as random variables with unknown probability distributions instead of fixed but unknown quantities. P(H_j) is called the prior probability of the j-th hypothesis H_j: the probability that H_j is true before observing or considering the sample data. P(D|H_j) is the probability of observing the data given that H_j is the true hypothesis. The posterior probability, P(H_j|D), is a measure how likely H_j is conditional on the observed data. The comparison of posterior probabilities of different hypotheses can now be used as a criterion for deciding which of these hypotheses under consideration is most likely given the data. Most importantly, Bayesian inference involves both prior information and empirical data.

In the Bayesian approach to hypothesis testing just a null and an alternative hypothesis, H_0 and H_1, are considered. The ratio between the probability of the data under each hypothesis, P(D|H_1)/P(D|H_0), is called the Bayes factor. It is related to the odds of the posterior probabilities, as follows:

$$\frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_0)P(H_0)}.$$

In words, the posterior odds are equal to Bayes factor times prior odds. And, in case of equal prior probabilities for the hypotheses this equation simply reduces to *posterior odds = Bayes factor*.

The Bayes factor can be used as a summary statistic regarding evidence provided by the data in favour of one of the hypotheses [16]. In case of a simple null hypothesis and a simple alternative hypothesis, as in our experiment, the Bayes factor (BF) equals the likelihood ratio (LR), i.e., BF = LR.

Let us return to the game again. Thomas considers the same simple null hypothesis as Ronald and J-E (H_0: the dice are fair), and the same simple alternative of J-E with one single manipulated die (H_1: the dice are unfair). H_0 and H_1 are considered to be two mutually exclusive and exhaustive hypotheses. Thomas states that he thinks that Victor is playing false in order to win. He proposes the prior distribution: P(H_0) = 1/3 and P(H_1) = 2/3, which we call *prior 1* here. So the probability that the dice are unfair is assumed to be twice as large as the probability that the dice are fair (prior odds = 2). Ronald thinks this prior is harsh on Victor, and he proposes to use equal priors instead, with prior odds equal to 1: P(H_0) = 1/2 and P(H_1) = 1/2, called *prior 2*. J-E is even more inclined to give Victor the benefit of the doubt, and proposes *prior 3*: P(H_0) = 2/3 and P(H_1) = 1/3.

The conditional probabilities given both hypotheses and the posterior probabilities for the outcome under the three priors are shown in Table 2. For reasons of clarity, we only consider four relevant outcomes: 2, 3, 4, and > 4 (gaining no victory points). Since a simple null and alternative hypothesis were specified, BF = LR, as indicated. The calculations of the posteriors for the outcomes 5 through 12 under the three priors are left as an exercise.

The comparison of Bayesian posteriors provides answers to the question which hypothesis is most likely given the outcome of the dice. Whatever the outcome, under *prior 1* the posterior probability that the dice are unfair is larger than the posterior probability that the dice are fair. Therefore, Thomas will decide that it is more likely that

the dice are unfair. The opposite is true for *prior 3*: whatever the outcome of the dice, the posterior probability that the dice are unfair is smaller than the posterior probability that the dice are fair. Adopting J-E's prior, Thomas would decide that the dice are fair is more likely than that they are unfair. This clearly demonstrates how the choice of a certain prior may have an impact on substantive conclusions. Under *prior 2*, the odds of the posteriors are equal to the likelihood ratio. For all outcomes with a likelihood ratio larger than 1, the alternative is more likely than the null hypothesis. If the outcome equals 3, the likelihood ratio equals 1.50, leading to the conclusion that it is more likely that the dice are unfair than that the dice are fair. Note that if the outcome would be 2, the Bayesian procedure under *prior 2* would be undecided.

Again, the familiar question arises: Does this mean that Thomas should conclude that the dice are indeed unfair? The answer here is NO. As a Bayesian, Thomas is never absolutely sure which hypothesis is true. Given his personal *prior 1* and the outcome 3, he is *more* convinced that the alternative is true rather than the null hypothesis. Note that this conclusion depends on his prior assumptions regarding Victor's behaviour, because if Thomas would adopt J-E's prior, he would come to different conclusions.

5. Extension to a more realistic setting

The exploration of a larger and more realistic experiment requires more complicated calculations. In the Appendix, the calculations in this section are presented in an R-script; running the R-script will also reproduce the figures illustrating the different testing procedures. The script was written and executed under R version 2.13.1, adapting some examples from Crawley [7, Ch. 5].

5.1. More trials

In empirical research, investigators would seldom base conclusions on an experiment with one single throw, i.e., with a sample size of just one. Increasing the sample size increases the reliability of the results, and in case of Fisher and Neyman-Pearson, one could use the more common significance levels of 1% or 5%. A natural question is: How would the previous results, or rather the substantive decisions change for an experiment with 10 throws and a significance level of 5%?

For a single throw with two dice, we have 11 possible outcomes, 2 through 12, each with a certain probability under both hypotheses (see Table 1). In case of 10 independent throws, the number of possible outcomes is 11^{10}. Under both hypotheses, the outcomes have a multinomial distribution, each with a different set of probabilities. Hence we need to consider all 11^{10} outcomes. For didactical reasons and for simplicity, we transform this question into a binomial problem, defining a success as Victor's gaining victory points and a failure as his not gaining victory points. The outcome variable then is the total number of successes in ten trials, which may thus vary from zero to ten. In case of fair dice (H_0) the success probability π equal to 1/6. Under the alternative hypothesis, the outcome is also binomially distributed, with π equal to 1/4 (see Table 1: 1/36 + 3/36 + 5/36).

The binomial probabilities for the number of successes X = k for $n = 10$ independent trials are defined as

$$\text{P(number of successes} = k|n = 10) = \binom{10}{k} \pi^k (1-\pi)^{(10-k)},$$

where π is the success probability. In Fig. 1 a bar plot is shown with the probabilities for different numbers of successes for both hypotheses.

According to Ronald (supporter of Fisher) only the null hypothesis is relevant. The most extreme observations are the largest numbers of successes. In Fig. 2, the cumulative probabilities P(X $\geqslant$ k) are given for various numbers of successes along with the rejection region. This region consists of the numbers of successes for which P(X $\geqslant$ k) $\leqslant$ 0.05, i.e., smaller or equal to the significance level, and indicated by the shaded area. For X $\geqslant$ 5, the p-value is only 0.015 (i.e., the sum of all probabilities for which X $\geqslant$ 5); for X $\geqslant$ 4 the p-value is 0.070. Using a significance level of 5%, Ronald's decision rule is: reject the hypothesis that the dice are fair if the number of successes is larger than or equal to five, and do not reject this hypothesis if the number of successes is smaller than five.

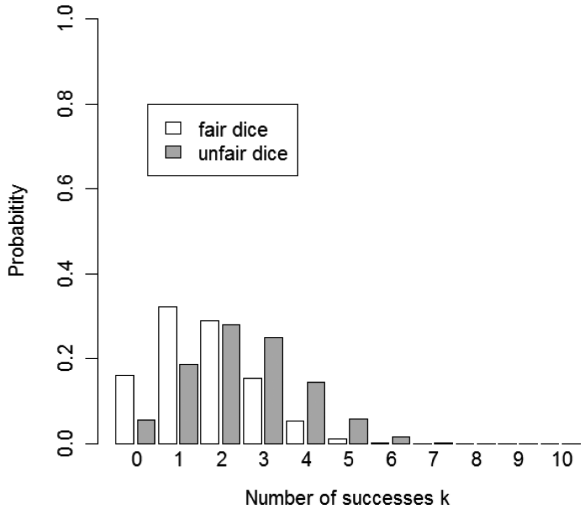

Fig. 1. Bar plots of probability masses under both hypotheses.

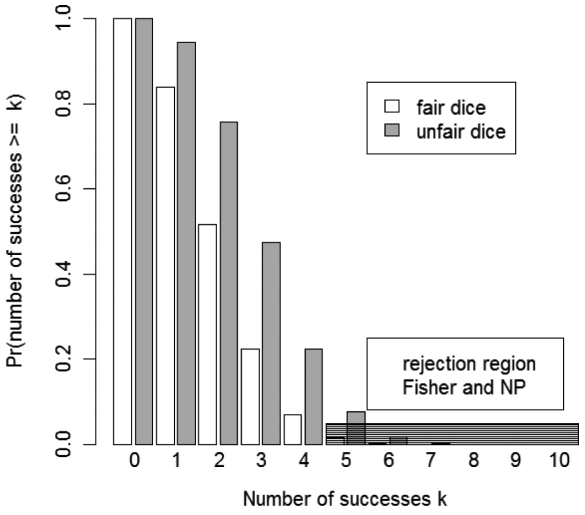

Fig. 2. Bar plots of cumulative probabilities under both hypotheses with rejection regions of Fisher and Neyman-Pearson.

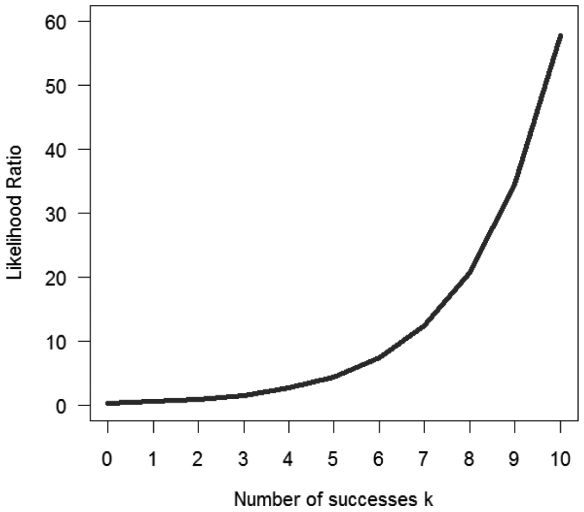

Fig. 3. Likelihood ratio as function of the number of successes.

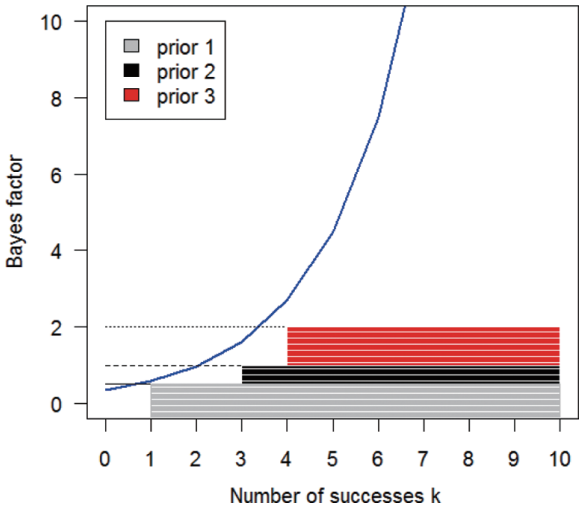

Fig. 4. Rejection regions in Bayesian testing: Simple case.

According to J-E (supporter of Neyman-Pearson), the null hypothesis of fair dice must be rejected in favour of the alternative for large values of the likelihood ratio, controlling for the Type I error, α. In Fig. 3 the likelihood ratio as a function of the number of successes is presented.

Evidently the value of the likelihood ratio increases as the number of successes increases: we are dealing with a monotone likelihood ratio. Using the same upper bound of 5% for α, J-E comes to a similar decision rule as Ronald: reject the hypothesis that the dice are fair if the number of successes is larger than or equal to five, and accept this hypothesis if the number of successes is smaller. The shaded area in Fig. 2 is therefore also the rejection region for J-E.

According to Thomas, the Bayes factor (BF), which equals the likelihood ratio (LR) in this case, plays a crucial role because the posterior odds equals BF times the prior odds. Thomas' conclusion with prior odds equal to two (*prior 1*) is that it is more likely that the dice are unfair if BF $\geqslant 0.5$. For $X = 0$, BF $= 0.349$, while for $X = 1$, BF $= 0.581$. So, when there are any successes, i.e., if X $\geqslant 1$, Thomas will be more convinced that the dice are unfair than that they are fair. For convenience and consistency, we also call this the rejection region in the Bayesian context. In Fig. 4, the value of Bayes factor as function of the number of successes is given, along with the rejection

region. Using the prior odds of one (*prior 2*), Ronald is more convinced that the dice are unfair when BF $\geqslant$ 1. For $X = 2$, BF $= 0.969$, while for $X = 3$, BF $= 1.614$. This means that under *prior 2*, he will be more convinced that the dice are unfair if X $\geqslant$ 3, as shown in Fig. 4. Thomas' conclusion with prior odds equal to 0.5 (*prior 3*) is that is more likely that the dice are unfair if BF $\geqslant$ 2, i.e., if X $\geqslant$ 4; this is also exposed in Fig. 4. The computation for the posteriors with Bayes theorem is left as an exercise, and can be checked by running the corresponding parts of the R-script.

To summarize, the p-value and the upper bound α determine the decision rules of Ronald and J-E. Because of the monotone likelihood ratio property, Ronald en J-E arrive at the same rejection region. They differ only in formulating the decision rule: do not reject the null hypothesis (Fisher) versus accept the null hypothesis (NP). BF is the key statistic for Thomas' decision. Not only his formulation of the decision rule is dissimilar – in terms of how convinced he is about a certain hypothesis – but also the set of values for which the alternative is preferred.

5.2. Composite alternative hypotheses

In our experiment with ten trials, the null hypothesis of fair dice is translated as H_0: $\pi = 1/6$ and the alternative of unfair dice as H_1: $\pi = 1/4$, where π is the probability of Victor's getting resources. A more general formulation would be to use a composite alternative H_1: $\pi > 1/6$. This alternative includes all possibilities of manipulating the pair of dice in favour of gaining victory points. What are the consequences of under the three test procedures under this composite alternative hypothesis?

For Ronald there are no consequences, since he does not consider the alternative hypothesis.

Following Neyman-Pearson's procedure, the null hypothesis is rejected in favour of the alternative for large values of the likelihood ratio. We already observed that the likelihood ratio function is monotone. This implies that the likelihood ratio increases with increasing numbers of successes. This holds for any value of π in the alternative parameter space, e.g., for all $\pi > 1/6$. Since the decision rule is determined by the upper bound of α, large values of the likelihood ratio will occur for the same outcomes, whatever the specified value of π in the alternative hypothesis. This means that for J-E, the evaluation of test results for the composite alternative is also the same as for the simple case.

For Thomas, we need to calculate the posterior probabilities or, more simply, the odds of the posterior probabilities using BF. In the case of a simple null and alternative hypothesis, BF equals LR. In the composite case, however, all possible values of π in the alternative must be considered. This can be accomplished by using a marginal likelihood, averaging the likelihood over all values of π given the prior distribution of π under the alternative. This means that the probability P(X|H_1: $\pi > 1/6$) is determined by integrating (i.e., taking the continuous sum) over the alternative parameter space $\pi > 1/6$ given the prior for π. In the composite case BF is the ratio of *marginal* likelihoods, and is different from LR which is the ratio of maximum likelihoods.

For integrating over the alternative parameter space, we need a prior distribution for π. Assuming that each value of $\pi > 1/6$ is equally likely, a uniform prior distribution on (1/6, 1) is a sensible choice. Under this assumption the marginal likelihood can be written as:

$$P(X|H_1 : \pi > 1/6; n = 10) = 6/5 \int_{1/6}^{1} \binom{10}{k} \pi^k (1-\pi)^{10-k} d\pi$$

This probability can be computed numerically. In Fig. 5, BF is shown as function of the number of successes, together with regions for BF $\geqslant$ 0.5 (*prior 1*), for BF $\geqslant$ 1 (*prior 2*), and for BF $\geqslant$ 2 (*prior 3*).

For X $= 2$, BF $= 0.027$, while BF $= 0.636$ for X $= 3$. This means that for *prior 1* Thomas would conclude that it is more likely that the dice are unfair when X $\geqslant$ 3. In the case of *prior 2*, Thomas is more convinced that the dice are unfair when X $\geqslant$ 4, with BF $= 1.961$. And in case of *prior 3*, he is more convinced that the dice are unfair when X $\geqslant$ 5, with BF $= 8.337$. For the composite case we also see that the values for which the alternative of unfair dice is preferred, differs from those in Fisher's and Neyman-Pearson's approach and from the case with one specific alternative hypothesis ($\pi = 1/4$).

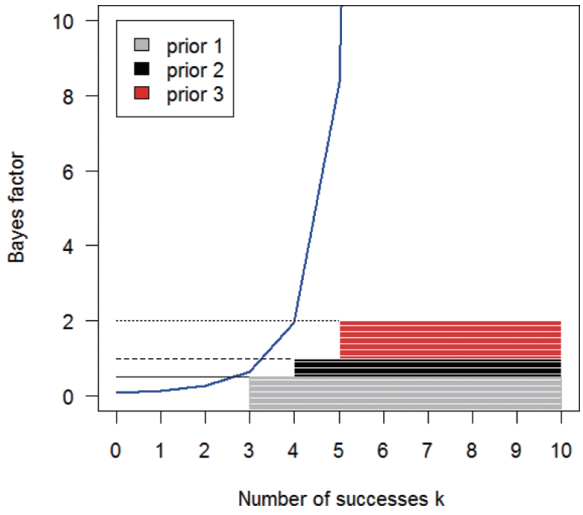

Fig. 5. Rejection regions in Bayesian testing: Composite case.

6. The current practice of hypothesis testing

In the current practice of hypothesis testing, the general procedure of making statistical inferences has elements of both Fisher's significance testing, i.e., p-values, and Neyman-Pearson testing where a null hypothesis is compared to an alternative hypothesis. The alternative hypothesis generally reflects the expected scientific theory, i.e., researchers would like to reject the null hypothesis in favour of the alternative, the model that is expected to hold. The alternative can be formulated as a one- or two-sided hypothesis. It is common to formulate the decision rule as rejecting the null hypothesis *in favour of* the alternative if the p-value is smaller than the specified significance level α. We do not reject the null hypothesis if the p-value is larger than α. Although the choice of the significance level is in principle arbitrary, Fisher himself pointed out that a p-value smaller than 5% is "seldom to be disregarded" [9, p. 80]. In the game example, the current practice mostly resembles the one-sided case of Fisher's approach in our opinion.

Royall [26] discusses several test procedures, including Fisher, Neyman-Pearson and Bayesian. He also explains so-called rejection trials which resemble the current statistical practice of hypothesis testing reviewed here, with the emphasis on rejecting (a sequence of) null hypotheses in favour of the alternative(s).

Many critical comments about Fisher's and Neyman-Pearson's "classical" testing approaches have been put forward. Two types of critique can be distinguished: a) criticism on incorrect inferences due to interpretational difficulties, and b) fundamental criticism questioning whether statistical inference based on Fisher or Neyman-Pearson is adequate in providing the correct and unambiguous answer to the scientific question posed. Note that the former was the main reason why we thought it worthwhile to teach all three approaches. Judging from the many reactions on Fletcher's statistical question posed in 2008: 'Which one statement is true of the p-value?' [10], the interpretation of the p-value still appears to be an issue of debate.

Fundamental criticism comes from Kirk [17], who argued that testing procedures with p-values and significance levels do not tell researchers what they want to know. The probability of observing the data (or even extremer outcomes) under the null hypothesis does not provide a satisfactory answer, because investigators are interested in the probability that the tested theory is true, given the data at hand, i.e., in the posterior distributions. Note, however, that in case of very large sample sizes or when we are dealing with a non-informative prior and unimodal posterior distributions, inferences according to Bayes and to the classical testing approaches are "indistinguishable" [27, p. 431].

Royall stated that the p-value is not a good measure for the strength of evidence against H_0 in favour of H_1, as its dependence on sample size and sample space can lead to different results in situations where the evidence is the same [26, p. 69]. In his view, the correct measure for the strength of evidence is the likelihood ratio because it satisfies the law of likelihood. This needs some further explanation. Define the probability that a discrete random variable X takes the value x by $P_0(x)$ if the H_0 holds, and by $P_1(x)$ if H_1 holds. The law of likelihood now states that

the observation $X = x$ is evidence supporting the null over the alternative hypothesis if and only if $P_0(x) > P_1(x)$. Based on the same law of likelihood, Royall criticises the Neyman-Pearson approach for case where a likelihood ratio $P_0(x) / P_1(x) \leqslant 1$ could still lead to preference for the null hypothesis given the significance level [26, p. 17].

Yet another controversy is caused by the different views on probability by frequentists and Bayesian statisticians. The frequentists define the probability of an event by the limiting value of its relative frequency of occurrence in a number of replications of an experiment [30, p. 14]. In the classical approaches (Fisher and Neyman-Pearson), this view of probability is the *only* probabilistic framework [3, p. 123]. The null and alternative hypotheses are determined by theory, by unknown but fixed model parameters. By contrast, Bayesian statisticians regard probability as a degree of reasonable belief rather than a relative frequency. Moreover, Bayesian inference is based on empirical data *and* prior knowledge instead of on empirical data only. Fisher was a lifelong critic of inverse probability, an ancient name for Bayesian inference [2]. According to Efron [8] the automatic nature of using p-values is one of the reasons why the Fisherian way prevails. One reason to stick to the classical approach is the undisputed fact that there is no universal rule for how to elicit prior knowledge; another is the problem of not knowing how to deal with prior knowledge of different experts.

In the last decade, however, the acceptance and application of Bayesian statistics has become more widespread, facilitated by computational developments (see, e.g. [5,11,13]). Moreover, several statisticians made an effort to bridge the gap between principles of frequentist and Bayesian statistics. Goodman [12] proposed to translate p-values in Bayes factors. Berger [4] tried to reunite the three approaches by computing (posterior or error) probabilities conditional on the strength of evidence in the data. Rice [24] used decision theory to show that Bayesian testing can be connected to Fisher's way of reasoning. Recently, Kass proposed an alternative for the statistical practice of inference, called 'statistical practicism' [15], which may be regarded as an original way of unifying Bayesian and frequentists' views on statistical inference. He suggested to teach this new statistical inference using 'the big picture', distinguishing a 'real world' (data) from a 'theoretical world' (statistical models) rather than the sample from the population. It is clear nowadays that one cannot be ignorant about Bayesian procedures in statistical inference.

7. Discussion

The purpose of this paper was to present the setup and main components of a lecture about the foundations of statistical inference and their principles on hypothesis testing as part of an advanced statistics course. Through a tutorial of the three main approaches we aimed to extend and deepen students' knowledge and understanding of statistical inference. Although the lecture was developed for research master students at our faculty of social sciences, it may also be useful for students in other disciplines, whose primary interest is not in statistics. We would like to encourage other lecturers to try out our classroom example by teaching the material presented in Sections 3 through 6. Our experiences are summarized below.

The elaboration of the board game example appeared to be fruitful in rehearsing basic elements of hypothesis testing, starting with the translation of a practical problem into a statistical hypothesis testing problem. In our example, the outcome measure had a multinomial distribution with known probabilities under the null hypothesis. In calculating the probabilities for all combinations of the pair of dice, some basic elements of probability theory were recaptured. The calculation of the p-value for each outcome was pretty straightforward and became less abstract. Therefore, given a sensible significance level, Fisher's procedure could be easily explained.

In discussing Neyman-Pearson's approach, the formulation of the alternative was an important issue. In our example, the alternative hypothesis (the dice are not fair) could be formulated in many ways, and it was not directly obvious which to prefer. Although in most research the alternative hypothesis may be more straightforward, the issue of relevant alternatives did not appear to be trivial. Our students were forced to think explicitly about the alternative. This turned out to be a good preparation for an adjacent lecture about practical significance (substantively relevant effect sizes) as opposed to statistical significance, interpretations of confidence intervals and power calculations. For the determination of sample sizes, knowledge about the Neyman-Pearson approach is crucial. Some students were reminded of the existence of the likelihood ratio statistic, while for other students this was a completely new topic.

By explaining Bayes' theorem, basic rules of conditional probability calculations were revisited in a different context. Since the majority of our students did not know anything about Bayesian statistics, our initial example

with discrete outcomes and simple null and alternative hypotheses provided a suitable introduction to the subject, in agreement with Albert [1, p. 1].

We would like to emphasize that it was not our intention to direct students to a "best" approach, as we do not have a clear preference ourselves. Our aim was to force students to think about making statistical inferences, emphasizing the differences between the various approaches. In this way, we tried to teach them to interpret statistical test results in a more accurate, meaningful and confident way. In our opinion, formulating reasonable priors, and showing differences in results reflected by the posteriors in Bayesian inference would make them aware of more subtle ways of reasoning. Students should realize that researchers may consider different criteria for the evaluation of sample results, and they should be forced to think about their own preferences.

Regarding the evaluation of our teaching, it was evident that the use of the board game application made the theory more attractive and that students really enjoyed the story. They were able and eager to pose critical questions afterwards. These critical questions did lead to the extension of the simple experiment of one trial to the experiment with more trials and to composite alternative hypotheses, as elaborated in Section 5. A disadvantage of our extension, however, is that students need to have more advanced knowledge of mathematical detail. For some of them such material appeared to reach beyond understanding. By adding the R-script and letting them play around with it, we hope that students will grasp the theory more easily and will continue to experiment with applications.

We are convinced that the majority of our students have learned the difference between p-values and the probability that the null hypothesis is true, and that they understood statistical inferences can be made in various ways. We hope that a deeper theoretical and practical understanding is achieved through the tutorial. We even hope that the students have improved their reporting on conclusions and interpretations of hypothesis testing results in scientific publications. We are, however, not so sure that we have achieved our ambitious goals completely, let alone permanently for all students. A major reason for such uncertainty is the heterogeneity of the group in ability, background and training. Learning statistics takes a lot of time for everyone involved, students and teachers alike. As lecturers we are convinced that teaching statistics deepens our understanding every single occasion. We hope this also holds for our students.

Appendix R-script

```
#### Figures in Section 5

# Preliminaries
options(digits=3, scipen=10)           # setting notation options
n <- 10                                # number of independent experiments n
p0 <- 1/6                              # success probability p0 under H_0
p1 <- 1/4                              # success probability p1 under H_1
k <- c(0,1,2,3,4,5,6,7,8,9,10)         # vector of number of successes k
n; p0; p1; k

#########################################################################
# Fig. 1
(pft0 <- dbinom(k, n, p0))             # P(X=k|H_0)
(pft1 <- dbinom(k, n, p1))             # P(X=k|H_1)
(LR <- pft1/pft0)               # Likelihood ratio LR
n2 <- 2*n+2
both <- numeric(n2)
both[1:n2 %% 2 != 0] <- pft0           # pft0 at odd places
both[1:n2 %% 2 == 0] <- pft1           # pft1 at even places
both
par(mar=c(7,5,3,2) + 0.1)              # setting plot margins
xscale <- barplot(both, col=rep(c("white","grey"),11), ylab="Probabitity",
```

```
          xlab="Number of successes k \n", ylim=c(0,1))
as.vector(xscale)
mtext(as.character(0:n), side=1, at=seq(1.4, 26.3, 2.4))
legend(2, 0.8, c("fair dice","unfair dice"), fill=c("white","grey"))
title(sub="Fig. 1. Bar plots of probability masses", col.sub="red",
font.sub=2, line=4)

#############################################################################
# Fig. 2
k1 <- c(1,1,2,3,4,5,6,7,8,9,10)       # vector of number of successes k,
# where k1[1]==1
cft0 <- 1 - pbinom(k1-1, n, p0)       # cumulative probabilities under H_0
cft1 <- 1 - pbinom(k1-1, n, p1)       # cumulative probabilities under H_1
both <- numeric(n2)
both[1:n2 %% 2 != 0] <- cft0          # cft0 at odd places
both[1:n2 %% 2 == 0] <- cft1          # cft1 at even places
both[1] <- both[2] <- 1
both
par(mar=c(7,5,3,2) + 0.1)             # setting plot margins
xscale <- barplot(both, col=rep(c("white","grey"),11),
          xlab="Number of successes k \n",
          ylab="P(number of successes >= k)", ylim=c(0,1))
as.vector(xscale)
mtext(as.character(0:n), side=1, at=seq(1.4, 26.3, 2.4))
rect(12.2, 0, 26.5, 0.05, density=100, angle=0,
     col="firebrick2", border="black")
legend(14.5, 0.85, c("fair dice", "unfair dice"), fill=c("white","grey"))
legend(14.5, 0.25, c("rejection region ", "Fisher and NP "))
title(sub="Fig. 2. Bar plot of cumulative probabilities", col.sub="red",
      font.sub=2, line=4)

#############################################################################
# Fig. 3
(k <- c(0,1,2,3,4,5,6,7,8,9,10))      # number of successes k
(pft0 <- dbinom(k, n, p0))            # P(X=k|H_0)
(pft1 <- dbinom(k, n, p1))            # P(X=k|H_1)
(LR <- pft1/pft0)                     # Likelihood ratio LR
par(mar=c(7,5,3,2)+ 0.1)              # setting plot margins
plot(k, LR, type="l", ylim=c(0,60), ylab="Likelihood ratio",
     xlab="\n Number of successes k", lty=1, las=1,
     lab=c(10,5,7), lwd=2, col="blue")
title(sub="Fig. 3. Likelihood ratio function", col.sub="red", font.sub=2,
      line=5)

#############################################################################
# Fig. 4
(k <- c(0,1,2,3,4,5,6,7,8,9,10))     # number of successes k
(pft0 <- dbinom(k, n, p0))           # P(X=k|H_0)
(pft1 <- dbinom(k, n, p1))           # P(X=k|H_1)
(LR <- pft1/pft0)                    # Likelihood ratio LR
par(mar=c(7,5,3,2)+ 0.1)             # setting plot margins
```

```
plot(k, LR, type="l", ylim=c(0,10), ylab="Bayes factor", lty=1, las=1,
     xlab="\n Number of successes k", lab= c(10,5,7), lwd=2,
col="blue")
n1 <- n+1                              # number of independent experiments n+1
bsnorm1 <- rep(0.5,n1)
bsnorm2 <- rep(1,n1)
bsnorm3 <- rep(2,n1)
lines(k, bsnorm1, ylim=c(0,n), lty=1)
lines(k, bsnorm2, ylim=c(0,n), lty=2)
lines(k, bsnorm3, ylim=c(0,n), lty=3)
rect(4, 0, 10, 2, density=100, angle=0, col="red", border="white")
rect(3, 0, 10, 1, density=100, angle=0, col="black", border="white")
rect(1, -0.35, 10, 0.5, density=100, angle=0, col="grey", border="grey")
legend(0, 10, c("prior 1","prior 2","prior 3"),
fill=c("grey","black","red"))
title(sub="Fig. 4. Rejection regions in Bayesian testing: Simple case",
      col.sub="red", font.sub=2, line=5)

##############################################################################
# Fig. 5
(k <- c(0,1,2,3,4,5,6,7,8,9,10))       # number of successes k
(pft0 <- dbinom(k, n, p0))             # P(X=k|H_0)
pft1 <- numeric(n+1)                   # vector of marginal probabilities
                                       # P(X=k|H_1: p0 > 1/6)
for (i in 0:n) {
     pft1[i+1] <- (6/5)*(integrate(function(x)
                  {choose(10,i)*(x^i)*(1-x)^(10-i), lower=p0, upper=1)
                  $val)
}
pft1
(BF <- pft1/pft0)                      # Bayes factor BF
par(mar=c(7,5,3,2)+ 0.1)               # setting plot margins
plot(k, BF, type="l", ylim=c(0,10), ylab="Bayes factor", lty=1, las=1,
     xlab="\n Number of successes k", lab= c(10,5,7), lwd=2,
col="blue")
#mtext("\n Number of successes k", side=1, line=2)
n1 <- n+1                              # number of independent experiments n+1
bsnorm1 <- rep(0.5,n1)
bsnorm2 <- rep(1,n1)
bsnorm3 <- rep(2,n1)
lines(k, bsnorm1, ylim=c(0,n), lty=1)
lines(k, bsnorm2, ylim=c(0,n), lty=2)
lines(k, bsnorm3, ylim=c(0,n), lty=3)
rect(5,0,10,2, density=100, angle=0, col="firebrick2", border="white")
rect(4,0,10,1, density=100, angle=0, col="black", border="white")
rect(3,-0.35,10,0.5, density=100, angle=0, col="grey", border="grey")
legend(0, 10, c("prior 1","prior 2","prior 3"),
fill=c("grey","black","red"))
title(sub="Fig. 5. Rejection regions in Bayesian testing: Composite case",
      col.sub="red", font.sub=2, line=5)

##############################################################################
```

References

[1] J. Albert, Discrete Bayes with R, *Technology Innovations in Statistics Education* **3**(2) (2009), 1–18.
[2] J. Aldrich, R.A. Fisher on Bayes and Bayes' theorem, *Bayesian Analysis* **3**(1) (2008), 161–170.
[3] V. Barnett, *Comparative Statistical Inference*, Chichester, UK: Wiley, 1999.
[4] J.O. Berger, Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science* **18**(1) (2003), 1–32.
[5] B.P. Carlin and T.A. Louis, *Bayesian Methods for Data Analysis*, London, Chapman & Hall, 2009.
[6] R. Christensen, Testing Fisher, Neyman, Pearson, and Bayes, *The American Statistician* **59**(2) (2005), 121–126.
[7] M.J. Crawley, *The R Book*, Chichester, UK: Wiley, 2007.
[8] B. Efron, Why isn't everyone a Bayesian? *The American Statistician* **40**(1) (1986), 1–5.
[9] R.A. Fisher, *Statistical Methods for Research Workers* (*14th ed.*), Edinburgh, UK: Oliver and Boyd, 1970.
[10] J. Fletcher, Statistical question: P-values, *British Medical Journal* **337** (2008), a201.
[11] A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin, *Bayesian Data Analysis*, London: Chapman & Hall, 2004.
[12] S.N. Goodman, Of p-values and Bayes: A modest proposal, *Epidemiology* **12**(3) (2001), 295–297.
[13] H.J.A. Hoijtink, Bayesian data analysis, in: *The SAGE Handbook of Quantitative Methods in Psychology*, R.E. Millsap and A. Maydeu-Olivares, eds, London: Sage, 2009, pp. 423–443.
[14] H. Jeffreys, *Theory of Probabillity,* (*1st ed.*), Oxford, UK: The Clarendon Press, 1939.
[15] R.E. Kass, Statistical inference: The big picture, *Statistical Science* **26**(1) (2011), 1–9.
[16] R.E. Kass and A.E. Raftery, Bayes factors, *Journal of the American Statistical Association* **90**(430) (1995), 773–795.
[17] R.E. Kirk, Practical significance: A concept whose time has come, *Educational and Psychological Measurement* **56**(5) (1996), 746–759.
[18] E.L. Lehmann, *Testing Statistical Hypotheses*, New York: Springer, 1997.
[19] L. Levy, Special K-Klaus Teuber, *The Games Journal, A Magazine About Boardgames* (2001), http://www.thegamesjournal.com/articles/SpecialK3.shtml, accessed at 1 May, 2007.
[20] J. Neyman and E.S. Pearson, On the use and interpretation of certain test criteria for purposes of statistical inference. Part I, *Biometrika* **20** (1928), A, 175–240.
[21] J. Neyman and E.S. Pearson, On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society A* **231** (1933), 289–337.
[22] K.R. Popper, *The Logic of Scientific Discovery*, London: Hutchinson, 1968.
[23] R Development Core Team, R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria; http://www.Rproject.org, accessed at 16 September 2011.
[24] K. Rice, A decision-theoretic formulation of Fisher's approach to testing, *The American Statistician* **64**(4) (2010), 345–349.
[25] C.P. Robert, N. Chopin and J. Rousseau, Harold Jeffreys's theory of probability revisited, *Statistical Science* **24** (2009), 141–172.
[26] R.M. Royall, *Statistical Evidence: A Likelihood Paradigm*, London: Chapman & Hall, 1997.
[27] A.A. Rupp, D.K. Dey and B.D. Zumbo, To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling, *Structural Equation Modeling* **11**(3) (2004), 424–451.
[28] T.A.B. Snijders, Hypothesis testing: Methodology and limitations, in: *International Encyclopedia of the Social & Behavioral Sciences*, N.J. Smelser and P.B. Baltes, eds, Amsterdam: Elsevier, **10**, 2001, pp. 7121–7127.
[29] K. Teuber, *The Settlers of Catan*, Stuttgart: Franckh-Kosmos Verlag, 1995.
[30] R. von Mises, *Probability, Statistics, and Truth*, New York: Dover, 1939.
[31] Wikipedia, the free encyclopedia, The Settlers of Catan, http://en.wikipedia.org/wiki/The_Settlers_of_Catan accessed at 4 February 2011.
[32] Wikipedia, the free encyclopedia, Victor Lustig, http://en.wikipedia.org/wiki/Victor_Lustig accessed at 4 February 2011.
[33] L. Wilkinson and the Task Force on Statistical Inference, Statistical methods in psychology journals: Guidelines and explanations, *The American Psychologist* **54**(8) (1999), 594–604.