DESIGN AND EXPLORATORY EVALUATION OF A LEARNING TRAJECTORY LEADING TO DO RANDOMIZATION TESTS FACILITATED BY TINKERPLOTS

Daniel Frischemeier & Rolf Biehler

University of Paderborn

We investigate the reasoning of preservice teachers about uncertainty in the context of randomization tests. Last winter term (2011/2012), we developed a seminar about data analysis with TinkerPlots, where group comparisons have played a fundamental role. A typical task in this context was: "Is there a difference between two groups or could that difference have occurred at random due to the selection of our sample?" In the first part of this article we describe a possible method to answer such a question with the help of a randomization test (see Rossman, 2008) facilitated with TinkerPlots. In the second part we point out how the participants of our course conducted a randomization test in a statistical project work and which problems and (mis-)conceptions were observed.

INTRODUCTION

The preservice education of teachers of mathematics at the University of Paderborn consists of three domains: mathematics, didactics of mathematics and pedagogy. An obligatory course called "Elementary Statistics and Probability Theory" is part of the programme. In addition the student teachers can participate in a seminar which deepens the course "Elementary Statistics and Probability Theory". The authors of this article have designed a seminar for preservice teachers in mathematics called "Developing statistical reasoning with using the software TinkerPlots" (Frischemeier & Biehler, 2012). In this course, the participants go through the whole PPDAC-cycle (Wild & Pfannkuch, 1999) which includes analysing data with the software TinkerPlots 2.0 (Konold & Miller, 2011) and writing down findings in a statistical report. Alongside describing and interpreting single distributions and exploring differences between them we wanted the participants to make conclusions about a wider universe and try to make generalizations of their findings. A typical task in connection with group comparisons was: "Is there a difference regarding a variable between two groups or could that difference have occurred at random due to the selection of our sample?" At the end, a statistical project work concluded the course. In the next part of this article we describe a possible method to answer such a question posed above with the help of a randomization test (see for example Rossman, 2008) facilitated with TinkerPlots. We focus on subject matter and knowledge of our students and do not discuss pedagogical content knowledge.

THEORETICAL FRAMEWORK

In general it is reported that pupils, students and pre-service teachers have a lot of misconceptions in testing hypotheses and using p-values (see e.g., Garfield and Ben-

Zvi 2008, p.270). There we also get to know that most of the students have problems with questions concerning generalizing of a result found in a sample. An opportunity for emergent inferential reasoning, especially in connection with results from group comparisons, is a so called randomization test. For a detailed and formal introduction in randomization tests see Ernst (2004). Rossman (2008) recommends starting inferential reasoning with randomization tests. He introduces randomization test with the example "dolphin therapy". For details see Rossman (2008). A big advantage according to Rossman (2008) and an important argument for a first step into informal inferential reasoning via randomization tests is that "…this procedure for introducing introductory students to the reasoning process of statistical inference is that it makes clear the connection between the random assignment in the design of the study and the inference procedure" (p.10). Rossman (2008) further points out that a randomization test "…also helps to emphasize the interpretation of a p-value as the longterm proportion of times that a result at least as extreme as in the actual data would have occurred by chance alone under the null model" (p.10). Furthermore, Rossman points to the problem of the "final" conclusion and the possibility that the null model is correct (although this is "unlikely" given a small p-value). Cobb (2007) emphasizes that, with randomization tests, students in introductory courses have a better opportunity to understand the "core logic of inference" converse to an approach based on calculations from normal-based probability distributions. This was also proposed by R.A. Fisher "but […] was not realistic in his day due to the absence of computers". Cobb (2007) also emphasizes his 3R´s: Randomize data production, Repeat by simulation to see what´s typical (and what´s not) and Reject any model that puts your data in its tail. The randomization of the data production (Cobb´s first "R") is an important condition. However, we have decided to use randomization tests also in the context of observational studies comparing data to the hypothetical situation that a random assignment process has produced the data (see also Konold, 1994 for such an approach and Konold & Pollatsek, 2002 for the "process approach"). A bootstrap method would have been preferable if we deal with random samples, however, resampling with replacement seemed us too difficult for explaining it to students. If the data were produced by such a process (null hypothesis), it has to be judged whether a difference is likely to be due to random variation under the null model. Rossman (2008) claims that teachers could use randomization tests to connect the randomness that students perceive in the process of collecting data to the inference to be drawn. He provides examples of how such a randomization-based approach might be implemented at tertiary level, while Scheaffer and Tabor (2008) propose such an approach for the secondary curriculum and provide relevant examples. Which misconceptions of students occur when doing a hypothesis test? Vallecillos (1994) report that many students who thought (like in a deductive process) that the correct application of a test with a significant result implies the truth of the alternative hypothesis. Another misconception regarding p-values is that the p-value is supposed to be the probability that the null hypothesis is true, given the

observed data (Garfield & Ben-Zvi, 2008, p. 270). In the following section we describe how we used Tinkerplots for doing randomization tests. We see this as a refinement of group comparisons in concrete informal terms being aware that our approach cannot solve all the problems students have with hypothesis testing.

RANDOMIZATION TESTS FACILITATED BY TINKERPLOTS

The sampler of TinkerPlots 2.0 can be used as a useful tool for simulations. At first we will describe a typical task handed out to the participants working with the so-called Muffins data (Biehler, Kombrink & Schweynoch, 2003), which is a complex data set with 538 cases and about fifty variables of a questionnaire concerning media use and leisure time of eleven graders. "How much time do the girls read more than the boys on average?". Exploring the Muffins data reveals that the girls on average read approx. 0.82 hours per week more than the boys. This motivates the question: "Is there a difference regarding the variable "Time_reading" between boys and girls or could that difference be due to the selection of our sample?" The 538 students are neither a random sample of a clear-cut population, nor do we do random "treatment assignments" However, we can just imagine a process, where reading time is independent from gender. If the data were produced by such a process (null hypothesis), we have to judge whether a difference of 0.82 hours can be due to random variation under the assumption of no difference. We imagine that we divide the group of 533 students randomly into a group of 301 pseudo-females and 232 pseudo-males. Then we can calculate the mean difference of reading time in these two random subgroups. When we repeat this process many times, we can estimate the probability to get a mean difference greater or equal to 0.82 just by random group selections. The advantage of Tinkerplots is that such a random selection model can directly be implemented in the software. We formulate our null hypothesis "The data were produced by a process where there is no difference regarding the variable *time reading* between boys and girls." We estimate the probability that the difference between boys and girls is 0.82 hours or even higher under the assumption that the null hypothesis is true. This can be done in the following way by a simulation in TP. Figure 1 shows a screenshot with several steps of the randomization test in TP.

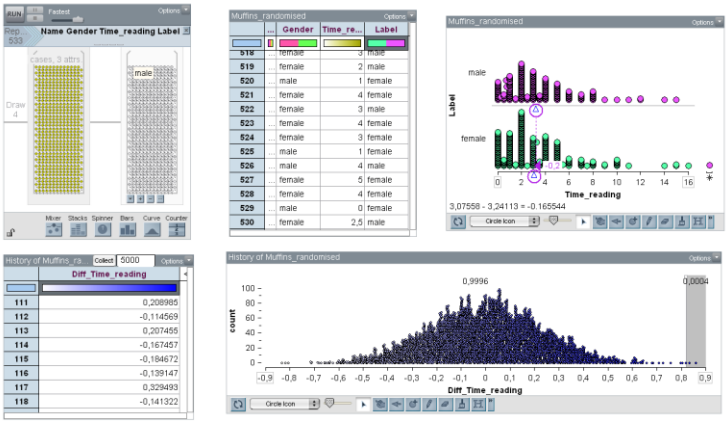

Figure 1: Screenshot of doing a randomization test with TinkerPlots

We place 533 balls labelled with the 533 times of the variable "Time_reading" in our group of respondents in box 1 (see figure 1, upper-left corner, left urn) and construct another box 2 (see figure 1, upper-left corner, right urn) with 232 balls labelled with (pseudo-) "male" and 301 balls labelled with (pseudo-) "female". In the next step a ball from each box is drawn without replacement. This is repeated 533 times (because of 533 cases): The 533 values of the variable "gender" are randomly assigned to the 533 values of the variable "Time_reading". The results can be seen in the table and the plot "Muffins_randomised" (figure 1, middle and upper-right corner, see also the column "label" that contains the "pseudo-gender"). This whole process is repeated 5000 times and we collect the measure "difference of the means of the two groups" with the "History"-function in TinkerPlots. The table (see figure 1, bottom-left corner) contains the collected measures of 5000 simulated random assignments. In only three of the 5000 random assignments (see figure 1, bottom-right corner), the simulated result turns out to be as extreme or even more extreme than the observed difference in the Muffins data. With the divider-tool we can determine the number and the proportion of cases (measures) that are 0.82 hours and higher. This proportion is 0.0004, which is our p-value. The "result" of the randomization test is an estimated probability (estimated by relative frequencies) of 0.0004 that the difference between the means of the two groups equals 0.82 or is even higher under the assumption that there is no difference in the process producing gender and reading time. What can we conclude from these results? Due to this very small p-value there is a very strong evidence against the null hypothesis of no difference.

DESIGN OF THE LEARNING TRAJECTORY

In this paragraph we want to describe the learning trajectory we created for the introduction into randomization tests on the one hand and a so-called randomization test-plan we handed out to the participants while doing a randomization test on the other hand. The major goal was that the participants learn to perform randomization tests with TinkerPlots as we demonstrated in the paragraph above (although we have paid just few attention to teaching the logic of inference at this stage). At first the participants were introduced in the sampler of TinkerPlots by modelling and performing some simple chance-experiments. Our introduction into randomization tests began with the "Extrasensory perception (ESP)" - task (Rossman et al., 2001, pp. 376), which the students had to do on their own in teams of two. In the working phase the first and second author gave support and feedback when problems occurred. Afterwards the results were discussed in the whole group. In the next step, our intention was to draw parallels between the simulation of ESP-task and the performance of a randomization test regarding the Muffins task: "Is there a difference regarding the variable "Time_reading" between boys and girls or did that difference occur at random due to the selection of our sample?". The participants were handed out a randomization test-plan and then worked in groups of two on the muffins task described above. Before continuing with the description of the learning trajectory we want to give some arguments for handing out a randomization test-plan: Sweller

(1999) have found out that the exploration of a complex learning trajectory such as the randomization test in our example, tasks the cognitive load of the learner very much. Due to the limitation of cognitive load he proposes to give the learner some help in form of structural aspects. With the randomization test-plan we give the participants a possibility to structure their thoughts and steps. With similar ideas to Biehler and Maxara (2007) we created the plan seen below (fig. 2).

Plan: How to do a Randomization Test

No	Step	ESP	Muffins
1	**Observation** Which difference do you observe between the means of the two groups in the dataset?	Number of correct answers = 20	*Mean of Time_Reading of boys = 2.685* *Mean of Time_Reading of girls = 3.503* *Difference = 0.818*
2	**Hypothesis H_0** As said in the task, the difference of the means of the two groups could have occurred at random. Generate an adequate N-hypothesis for your investigation.	The person does not have any extrasensory perception (ESP). He/she guesses with a success rate p = 0.25.	*The difference of the means of Time_Reading of boys and girls has occurred at random.*
3	**Simulation of H_0** How can you investigate the n-hypothesis with a simulation? Explain your procedure	Drawing 40times with replacement from an urn which is filled with 4 balls: 1 ball is labeled „r" (right) and 3 balls are labeled „f" (false)	*Place the 533 cases of Time_Reading in urn1. Construct urn2 with 232 balls labeled with "m" (male) and 301 balls labeled with "w" (female). Draw 533 times without replacement*
4	**Test statistic** Define the test statistic	X = Number of correct predictions	$X = \bar{x}_{Group1} - \bar{x}_{Group2}$
5	**p-value** Read of the p-value	P (X ≥ 20) = 0.0004 = 0.04%	*P (X ≥ 0.818) = 0.0006 = 0.06 %*
6	**Conclusions** Which conclusions can you make regarding your n-hypothesis?	The p-value (=0.0004) is very small, so we have a strong evidence against our n-hypothesis. We assume that the fortune teller has not guessed. Another possibility is: he could have guessed but that would have been very unlikely.	*The p-value (=0.0006) is very small. So we have a strong evidence against our hypothesis "The difference of the means of Time_Reading of boys and girls has occurred at random." Another possibility is: the difference occurred at random, but that is very unlikely.*

Figure 2: Randomization test-plan (with entries in "Muffins" – column, a task that our students had to do themselves)

In this "randomization test-plan", the participants have a structure for the simulation-process on the one hand and can write down their findings on the other hand. On the left side of the plan, the participants get an overview of the structure of the

randomization test, short instructions and leading questions for each (the forth column is without entries when it is handed out to the participants). A special feature of the plan is the third column "ESP". The "ESP"-task is the exemplary task we used when introducing our students in randomization tests. To support the participants in step 6 to draw conclusions from p-values, we gave them further material in form of a hand-out which was supposed to give them hints how to evaluate possible p-values, as follows:

> Hand-out: *We have a very strong evidence against H_0, if $p < 0.1\%$. *We have a strong evidence against H_0, if $p < 1\%$. *We have a medium evidence against H_0, if $p < 5\%$. *We have a small evidence against H_0, if $p < 10\%$. (Hand-out for the participants)

So after doing the "ESP"-task on their own with feedback of the first and second author as we have described above, the participants were handed out the "Muffins"-task which they had to do in teams of two. The results were discussed in the whole group afterwards. In reflecting on our "randomization test" sessions we found that there were two neuralgic points when doing a randomization test: First the correct formulation of the null hypothesis, second an adequate conclusion drawing from the resulting p-value. Finally the participants had to do a randomization test in their statistical project work, which was a requirement for completing the course.

RESEARCH QUESTIONS

In this paper, we will focus on the final randomization tests in the students' statistical project reports. Three main research questions emerge: 1. How did the participants finally perform a randomization test with using of TinkerPlots in their project works? 2. Were they able to fulfil the six steps of the randomization-test plan? 3. At which stages/steps did problems (which?) occur?

DATA, PARTICIPANTS & METHODOLOGY

We have had a look at eleven statistical project reports that the participants had to do at the end of the course in teams of two. They were allowed to choose their own questions related to the data sets we provided. Doing at least one randomization test in their report was a requirement. 23 participants attended the course, 22 of them worked on the project reports in teams of two, so we have 11 project works in total. The number of students'' semesters varies from 4 to 11, most (11 from 23) students were in their fifth semester. For a deeper analysis, we analysed all written extracts of the statistical project reports that dealt with the randomization test-task while focussing on the successful execution of the six steps of the randomization test-plan and typical problems that occurred when going along these steps. So we have had a global view (cf. research question 1 and 2) and a local view (cf. research question 3) on the randomization tests of the project works. In the "global observation" we checked how well the participants performed the 6 steps of the simulation plan generally. If they accomplished a step as described in the example above, we called it

“Step x successfully done”. We analysed the several steps with the background of our theoretical framework. We defined categories of typical problems with a focus on step 5 & 6. Furthermore we will also give typical examples for the categories in form of written extracts of students´ project works.

RESULTS

Let us have a look (table 1) how well the teams did in the several steps when doing a randomization test with TinkerPlots (Note, that every team have had an different topic, so some of them were confronted with p-values larger than 10%, others with p-values smaller than 0.1%, for example).

Steps successfully done	**Number of teams (of 11)**	**%**
Step1	11	100.00
Step2	8	72.73
Step3	10	90.91
Step4	10	90.91
Step5	5	45.45
Step6	5	45.45

Table 1: Overview-Randomization tests in project works

Almost every team was able to conduct and fulfil the simulation of the randomization test with TinkerPlots in form of accomplishing steps 1,3 and 4. Step 2 (formulating null hypothesis), step 5 (identifying and reading of the p-value) and step 6 (drawing conclusions from p-value) seemed to be problematic points as we mentioned above. So we want to have a closer look on that now.

Step 1 – Reading of the difference between groups in the dataset

As seen in the table every team accomplished step 1 successfully.

Step 2 – Formulation of null hypothesis

Regarding step 2 we can say that the majority of the teams (8 of 11) gave a correct formulation of the null hypothesis when doing the randomization test. Two teams formulated the alternative hypothesis instead of the null hypothesis. For example when comparing the reading habits of boys and girls in the muffins data and investigating the variable “Time_reading” the hypothesis of one of the teams was:

H. & P.: The Girls tend to spend more time on reading than boys.

Two other teams (three teams in total) showed the same problem when formulating an adequate null hypothesis.

Step 3 & Step 4 – Modelling the simulation process in TinkerPlots

As seen in the table nearly every team (10 out of 11) managed to model the simulation process of the randomization test in TinkerPlots.

Step 5 – Reading of the p-value

A notable problem which occurred in step 5 was a false identification of p-value in form of the mean of the collected measures. Two out of eleven teams identified the mean of the measures as p-value. Let us have a look on the case of Laura & Sarah. They wrote, when investigating the hypothesis "the gender-difference on the means of the variable `Time_phone_20min` (Number of phone calls per week that last longer than 20 minutes) did happen by chance":

> L. & S.: The mean of all means is approximately 0.000238873 after 5000 simulations und therefore smaller than 0.1%, which shows a strong evidence against the null hypothesis.

For them the mean of the 5000 collected differences is a very small value and seemed to turn out for them as a p-value.

Step 6 – Drawing conclusions from the observed p-value

In step 6 we found two phenomena: on the one hand "drawing premature conclusions from the p-value" such as "the p-value is smaller than 5%, therefore the null hypothesis can be rejected." and on the other hand "drawing false conclusions from an observed p-value". We will give an example for "premature conclusions" first. Alex and Kathrin concluded under the null hypothesis "the difference of the means of the variable "age" concerning the marital status of students is due to random effects":

> A. & K.: The statement can be rejected with a p-value of 4% (which is smaller than 5%). Therefore the null hypothesis […] can be rejected.

We consider this as premature because we taught the students not to take a definite decision but express the uncertainty when a small p-value occurs as an amount of evidence. An example of drawing false conclusions from the observed p-value is the following: The null hypothesis is seen as true, because the p-value is significantly high (> 10%). Victoria and Corinna were investigating whether a gender-difference of time spent on working (in hours per week) occurred at random and concluded:

> V. & C.: "The result of the randomization test shown here (0.1033) is a probability. Here we have a relative frequency of 0.1033 or 10.33%. This value corresponds to our p-value. [...] The p-value is bigger than 10% which means that the evidence is not so strong and therefore the null hypothesis must be true. "

They made a typical mistake, which is also reported in Garfield & Ben-Zvi (2008, p.270). Having a "large" p-value, they concluded, that the null hypothesis must be true. It is noticeable that this problem occurred at every (precisely: 3 out of 11) team, who conducted a randomization test in which the p-value turned out to be larger than

10%. The problem may have occurred due to paying not enough attention to a p-value larger than 10% in our learning trajectory. We can conclude that the participants have several problems to make conclusions on their own. Summary: two of eleven teams have identified the p-value as the mean of the collected measures. Four out of eleven teams rejected their null-hypothesis (particularly due to a small p-value) in form of "drawing premature conclusions from a given p-value". Three of eleven teams fell into the category "drawing false conclusions from a given p-value" concluded that the null hypothesis must be true, because of a large p-value (> 10%). All in all we can say that almost every team was able to deal with the technical process of the simulation in TinkerPlots, but they had partly problems with steps 5 and 6. They have acquired procedural knowledge of performing randomization tests in TinkerPlots, but some still fail to formulate an adequate null hypothesis or to identify a p-value or fail to draw adequate conclusions from it. When evaluating a p-value in the project reports, the participants seem to have the attitude either to accept or to reject a null hypothesis, instead of saying something like "…there is a small/medium/strong/very strong evidence against the null hypothesis". Living with uncertainty obviously is something uncomfortable.

LIMITATIONS AND FURTHER RESEARCH

For evaluating the learning trajectory in the sense of design research approaches (see Cobb, Confrey, diSessa, Lehrer & Schauble, 2003) we plan a retrospective analysis and a redesign of the learning trajectory. Obviously there is a need to revise our learning trajectory in order to improve students' conceptual understanding of randomization tests. How can they be better supported at major problems (for example the formulation of a correct null hypothesis; drawing conclusions from a given p-value)? We are currently (end of summer term 2012) conducting a qualitative interview study with the same participants, which is two-phased. In phase 1 they have to work on a group comparison (including a randomization test) exercise in teams of two. In phase 2 we interview them in form of a stimulated recall-method to elicit the thoughts and strategies of them when working on the task. This will hopefully give further insights into the cognitive processes of the students while working with randomization test with TinkerPlots. The interviews are informed by our previous analyses of project reports that we presented in this paper and are also directed towards the levels of conceptual understanding of the null model that they implemented with TinkerPlots sampler. The difficulties we observed with formulating null hypotheses may have a deeper origin in understanding a "null model" of no difference in a process approach with no random treatment assignments in the data production process.

REFERENCES

Biehler, R., Kombrink, K., & Schweynoch, S. (2003). MUFFINS – Statistik mit komplexen Datensätzen – Freizeitgestaltung und Mediennutzung von Jugendlichen. *Stochastik in der Schule, 23*(1), 11-25.

Biehler, R. & Maxara, C. (2007). Integration von stochastischer Simulation in den Stochastikunterricht mit Hilfe von Werkzeugsoftware. *Der Mathematikunterricht 53* (3): 45-62.

Cobb, P., Confrey, J., deSessa, A., Lehrer, R. & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13.

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

Ernst, M. (2004). Permutation methods: A basis for exact inference. *Statistical science, 19*, 676-685.

Frischemeier, D. & Biehler, R. (2012). Statistisch denken und forschen lernen mit der Software TinkerPlots. In Kleine, M. und Ludwig, M. (Eds.): *Beiträge zum Mathematikunterricht 2012*, WTM: Münster.

Garfield, J. & Ben-Zvi, D. (2008). *Developing students´statistical reasoning: Connecting research and teaching pratice*. Berlin: Springer.

Konold, C. (1994). Understanding probability and statistical inference through resampling. In Brunelli, L. & Cicchitelli, G (Eds.): *Proceedings of the first scientific meeting of the IASE*. Perugia, Italy: University of Perugia, 199-21.1

Konold, C., & Pollatsek, A. (2002). Data Analysis as the Search for Signals in Noisy Processes. *Journal for Research in Mathematics Education, 33*(4), 259-289.

Konold, C. & Miller, C. (2011). TinkerPlots TM Version 2 [computer software]. Emeryville, CA: Key Curriculum Press.

Rossman, A. Chance, B., Lock, R.H. (2001). *Workshop Statistics: Discovery with Data*. New York, Key College Publishing.

Rossman, A. (2008). "Reasoning about Informal Statistical Inference: A Statistician´s View." *Statistics Education Research Journal 7*(2): 5-19.

Scheaffer, R. & Tabor, J. (2008). Statistics in the high school mathematics curriculum: Building sound reasoning under uncertainty. *Mathematics Teacher, 102*(1), 56-61.

Sweller, J. (1999). Instructional design in technical areas. Melbourne: ACER Press.

Vallecidos, A. (1994). *Theoretical and experimental study on errors and conceptions about hypothesis testing in university students*. Unpublished Ph.D., University of Granada, Spain.

Wild, C.J. & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-265.