On the bimodality in an introductory programming course: an analysis of student performance factors

Lars Josef Höök
Division of Scientific Computing, Department of Information Technology, Uppsala University
josef.hook@it.uu.se
2014-03-18

Abstract
In this work, the student performance in an introductory programming course given at the department of information technology at Uppsala University is analyzed. The results from the final exam are studied and the potential bimodality of the distribution is investigated. A questionnaire, answered by the students at the final exam is analyzed, including study behavior and potential connections to the performance at the final exam.

Keywords: educational research in programming, bimodal distribution

1. Introduction
A few years ago an article was published that claimed it had developed a test that could predict with very high accuracy whether students in an introductory programming course would pass or fail the course before reading it [1]. The test consists of a number of questions on basic programming like assignment of values to variables etc. and the background for the development of this test is as follows. Between 30 and 60 percent of the students who study their first programming course fail the exam according to [1] and a lions share of the group that passes do so with excellent grade. This seemingly bimodal shape, (two humps) in the distribution of the results have motivated educational research for more than 25 years, in search for an explanation of the underlying factors for this effect. A controversial view, which tries to explain this bimodality, is that the class of students consists of people with innate talent for programming and students who lack talent. This standpoint is available in parts of the educational research literature for teaching in programming. According to a previous study [2] and references therein there has been some attempts to correlate the results from the exam against gender, ACT/SAT score and performance in mathematics to name a few examples, but without finding any significant correlation. In a study at Arhus [3] it is reported that they have not found any links between gender, income or age to the results.

The purpose of this study is to investigate whether a course in introductory programming (1TD753) at Uppsala University (UU) have the same bimodal distribution in the results of the exam as in [1]. The goal is to try to identify factors that might be linked to the grade. The outline of the paper is as follows. In Section 2 we present the questionnaire. In section 3 the results from the survey are presented together with a statistical analysis of the grade, which is followed by a discussion in section 4.

2. The questionnaire

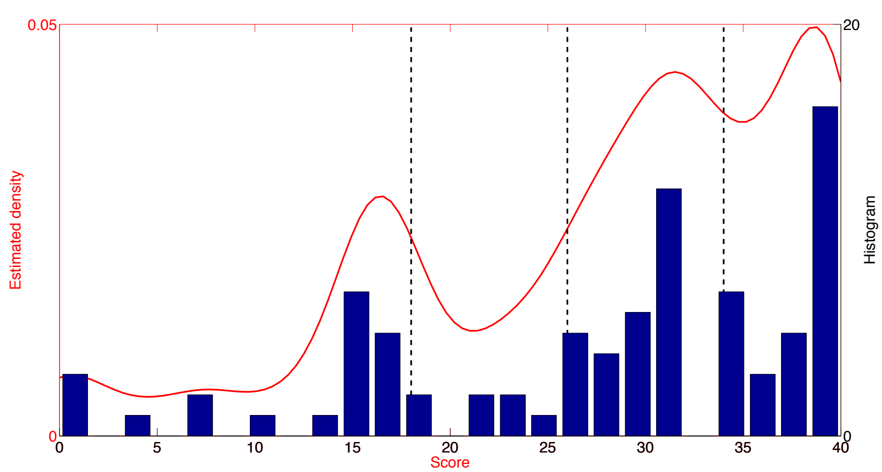
We have survey responses from approximately 230 students answered at the final exam. In this study, we have chosen to focus on students with no previous experience in programming. The questionnaire contains nine questions see Appendix A. The first question is on whether the student had any previous experience in programming. We used this question to filter out the students with no previous experience, which turned out to be 85 students from the original 230. The last two questions were related to the students' goal and how they've studied. Due to the free text nature of these two last questions they were not analyzed in this report. Instead we have focused our analysis on question 2-7 summarized below:

2. Estimate the amount of time you sat at the computer and wrote code yourself (1-5)
3. Estimate the amount of time you sat next to the computer when someone else wrote the code (1-5)
4. How m any of the lectures (max 8) did you participate in?
5. Estimate how many hours you have read the course book. (1-5)
6. What is your view on programming, (1-5)?
a. Fun.
b. Difficult.
c. Intimidating.
d. Useful for my future career.
e. I would like to work with programming in the future.
7. How much more would you like to learn about programming (1-5)

We have translated the answers from the questionnaire to numbers starting from 1 and used this representation in the analysis. The original answers can be obtained by plugging in these numbers into the questionnaire, appendix A.

3. Results

The distribution of the examination result is seen in figure 1 where the grade levels are marked with dashed lines. The grade limits for the exam were: 18 for grade 3, 26 for grade 4 and 34 for the highest grade 5. The maximum score in the exam was 40 points.



Figur 1 Distribution of the exam result for the 85 students. Grade limits are marked with dashed lines.

An immediate observation in figure 1 shows a multimodal distribution, and not a bimodal. To verify that this is not an under sampled unimodal distribution, we have tested the null hypothesis that the exam result is normally distributed with a Monte Carlo based Lilliefors test. The null hypothesis that the outcome is normally distributed was rejected at 5% significance level, which indicates that the observations are not likely from a normal distribution but have a more complicated structure. Note that the boundaries of the grades seem to affect the shape of the distribution, especially for students that failed the exam (<18) where the local maximum is located on the left side of the grade limit. Could this be a consequence of the correction procedure? The distribution in the result is clearly not of a bimodal form, but rather have several local maxima however, if we study the distribution of the grades then a clear bimodal distribution emerges, figure 2.

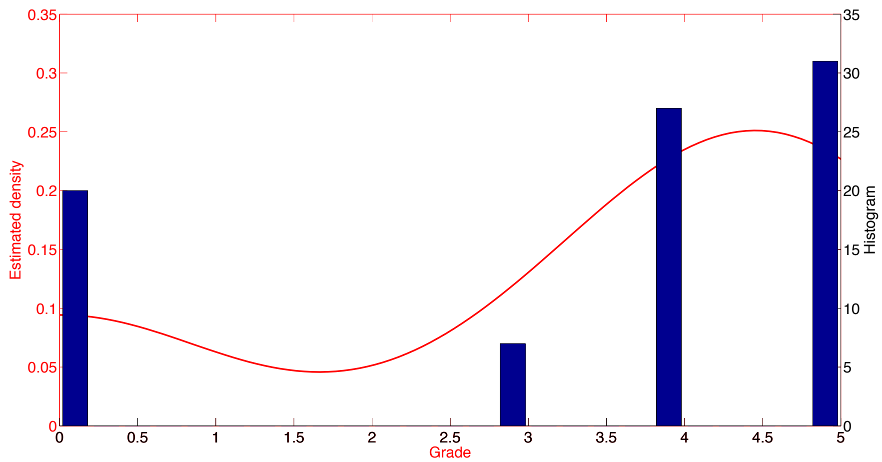


Figure 2 Distribution of scores (5,4,3, U) where U (failed) has been replaced by 0. We have a clear bi-modal distribution of the grades.

We have tested if the distribution of the grades has a significant bimodal structure with the Hartigans dip test [3]. The Hartigans test gives dip = 0.1588, p = 0 which is in the range for a bimodal distribution. We can clearly see in figure 2 and figure 1 that the grade intervals play a significant role for the shape of the distribution of the grade. If you move down the requirement for passing from 18 to 15 we would not observe this bimodality. From these simple observations it is not clear to us that one can claim that the bimodal shape is an effect of talent/lack of talent since the effect of other course related processes like the correction procedures of the final exam can easily modify the shape.

An analysis of the survey response

We have conducted an analysis of the survey responses, as summarized in table 1. In the analysis we divided the responses in to three groups, U, 3, and (4.5) according to the grades, dashed lines in figure 1.

Group	2	3	4	5	6a*	6b*	6c*	6d*	6e*	7*	Results
4,5	6 ± 1	2 ± 1	4 ± 2	2 ± 1	4	3	2	4	3	4	34 ± 5
3	4 ± 2	4 ± 2	3 ± 2	2 ± 1	4	4	2	4	3	4	22 ± 2
U	4 ± 2	3 ± 2	4 ± 2	3 ± 1	4	4	2	4	3	5	12 ± 6

Table 1, Mean, standard deviation and median (star marked) for the responses in each group.

The table shows that the groups have responded almost identical to the questions 6-7, which are questions related to motivation. Thus there is no significant difference in motivation between students who failed the exam and students who received higher grades. The greatest variations were in the questions 2-4 excluding the result on the exam. To investigate the importance of the questions for each group we have performed a principal component analysis on each subgroup, see figure 3. The principal component analysis indicates that the answers to question 3, 6b and 6c are varying in the same direction as the result of the exam for the group that failed the exam. For this group the answers to these questions can give an indication to why they failed the exam. Similar pattern is seen in the group that received grade 3. In this group there is a connection to the result of the exam and the answers to question 5, 6b and 6c. The group that received the highest grade did not have any question grouped together with the result on the exam, which indicate that the answers from this group cannot be related to their performance.

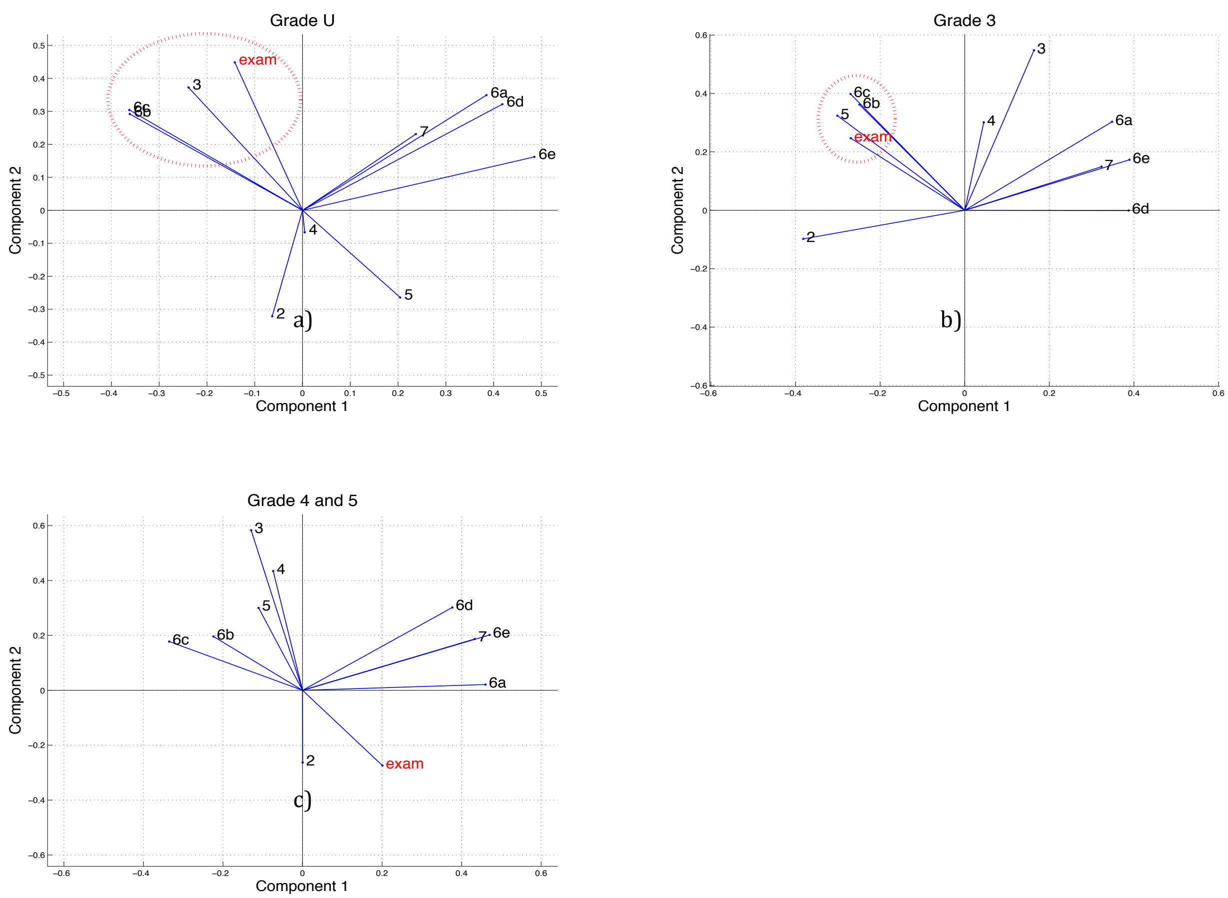


Figure 3 Principal component analysis for each subgroup, a is the failed group, b grade 3 group and c the higher grade group.

To summarize the questionnaire: students who failed the exam did also spend less number of hours working with the computer. The average answer was between 20-30 hours writing code themselves and 9 of 20 spent between 0-5 hours next to someone who was

programming with an average value of 10-20 hours. The individual variation in this group is large. In total this group spent less time with or next to a computer. The majority of students in the failed group thought the topic were difficult (question 6b) but not scary (6c). They thought the topic was useful for their future career (6d) but were slightly negative towards working with programming (6e). Half of the group would like to to learn more about programming, 10 out of 20 replied 5 to question 7 and half of the group spent between 1-5 hours reading the book.
In the group, which received the highest grade the average time spent working with the computer was between 40 to 50 hours almost twice as much as the group that failed. 25 out of 58 students spent between 0-5 hours next to someone else who programmed and the average value sitting beside someone programming was 5-10 hours. Only one person in this group thought that the subject was difficult and 28 out of 58 responded that it was not difficult which is better than the group that failed. The remaining students answered somewhere in between. There is not a clear pattern that can be drawn from this analysis, except that there are indications that low results on the exam are connected to the amount of time spent on programming. The opposite conclusion that a large amount time working with the computer will also give a high degree cannot be drawn due to the results from the principal component analysis for the high-grade students in figure 3c.

4. Discussion

The picture that emerges from the analysis is that the group that failed the exam consists of students who are trying to learn a new subject, find it interesting and understand the usage of what is learned but also find it difficult and “do not understand the thinking”. This is something that has been observed in previous studies [5]. At the same time less effort is spent reading and programming in this group, which can bee seen as slightly contradictive since they understand the purpose and importance of the subject.
The group with the higher grades spent in general more time working with the computer and did not perceive the subject as difficult. We cannot draw a conclusion that the more time spent is a factor for a higher grade.

The difficulty in learning programming may be caused by the many things that have to be learned at the same time, from semantics, how computers work to abstraction of problems. The topic can be thought of as a closed system with inherent rules and laws, which may be difficult to relate to if the students have not previously encountered programming. It is a "new world with new rules of the game" and in such a scenario people who learn by associating to past experiences and people with a deep oriented learning perspective are greatly challenged. Here a teaching tool that could potentially increase the support for such students would be to introduce a couple of lab-based supplemental instructions (SI) in the course where the students with the highest grade from the previous year act as teachers. The link between the number of hours that the student spent working with the computer (in the fail and grade 3 groups) and the result on the exam indicate that it can be beneficial if students work individually on the lab sessions rather than in groups of two or three people. This is something that has been introduced in this year's edition of the course. According to Dunn and Dunns Learning Style Inventory, [6], 28 percent of students in elementary school prefer to work in groups. In [7] it is argued that it may be reasonable to assume that this also applies to higher education. If we assume that about a quarter of the students work best in groups, one could test the so-called pair programming technique, [8], which is a concept where students sit in pairs and program with one student performing the actual programming and the other student is watching, but they often change who performs the actual programming. This could be a good educational technique for students who prefer to work in groups and at the same time force everyone to spend more time programming. It is questionable whether classical examination is the best way to examine the programming. In [2] the students were examined in the form of work samples in which the students had to solve a number of programming tasks in a limited time. In this study they did not see a bimodal distribution in the grade.

In conclusion, the study opens up for questions that need to be studied more thoroughly for example, does it matter if you sit down yourself and program or is it sufficient to watch some that program? Is the total amount of time programming the key factor? Another question that might be interesting to examine is whether deep-learning students and students who prefer to work in groups have more difficulty in learning because of the individual nature of programming?

Acknowledgment

The author greatly acknowledges the support from Anna Eckerdal for providing with the raw material for this study and for giving feedback and comments.

References

[1] S. Dehnadi; R. Bornat: The camel has two humps (working title), February 6, 2006
[2] ME Caspersen et al. Mental Models and Programming Aptitude, manuscript.
[3] J. Bennedsen and ME Casperson. An investigation of potential success factors for an Introductory Model-Driven Programming Course. In Proceedings for the Fifth Koli Calling Conference on Computer Science Education, p 166-169, Koli, Finland, 18 to 20 November 2005.
[4] Hartigan, JA; Hartigan, PM The Dip Test of Unimodality. The Annals of Statistics 13 (1985), no. 1 70 - 84.
[5] Berglund, A. Eckerdal, "What do CS Student Strive to Learn?" A. Computer Science Education, 16 (3), 2006
[6] Dunn, Rita, Nu fattar jag!: Att hitta och använda sin inlärningsstil, Jönköping: Barin Books, 2001
[7] M. Elmgren et al. University of Pedagogy, 2010, Collins.
[8] C. McDowel et al. The impact of pair programming on student performance, perception and persistence. In Proc. of ICSE'03, pages 602-607, 2003.

Appendix A - the questionnaire (in Swedish)

FRIVILLIG UPPGIFT.
För att vi bättre ska förstå hur man lär sig programmering, och för att vi ska kunna förbättra undervisningen ber vi dig att fylla i nedanstående frågor. Det är frivilligt och kommer inte att påverka resultatet på tentamen. Det tar bara någon minut. Frågorna är i första hand avsedda för dem som läst kursen nu i period 3.

DIN TENTAKOD (eller namn):________________________

1. Har du någon programmeringserfarenhet mer än Beräkningsvetenskap 1 och programmeringsteknik 1?
□Nej □ Ja.
Om Ja, vad, i vilket sammanhang och hur mycket? (T ex "50 poäng på gymnasiet, Java", "mycket hemma sedan jag var 14 år, Basic och C++", etc)

2. Uppskatta hur lång tid du suttit vid datorn och ***själv skrivit kod***, i labbsal eller någon annanstans, ***under kursen***. Kursen har haft 29 schemalagda tillfällen, dvs c:a 58 timmar sammanlagt i labbsal.
□ 0-5 h □5-10 h □10-20 h □20-30 h □30-40 h □40-50 h □50-60 h □ >60 h

3. Uppskatta hur lång tid du ***suttit bredvid datorn*** och tittat på när någon annan skrivit kod ***under kursen***.
□ 0-5 h □5-10 h □10-20 h □20-30 h □30-40 h □40-50 h □50-60h □ >60 h

4. Hur många av kursens 8 föreläsningar deltog du i? _____ föreläsningar

5. Uppskatta hur många timmar du läst boken, föreläsningsanteckningar, programmeringsteori på internet etc:
□0 h □1-5 h □5-10 h □10-15 h □15-20 h □20-25h □> 25 h

6. Hur ser du på programmering? Skala 1-5 där 1=håller inte alls med, och 5=håller med helt och hållet. Kryssa för det alternativ som passar bäst.

	1	2	3	4	5
Roligt	□	□	□	□	□
Svårt	□	□	□	□	□
Skrämmande	□	□	□	□	□
Nyttigt för min framtida arbetskarriär	□	□	□	□	□
Jag vill hålla på att programmera i framtiden	□	□	□	□	□

7. Hur mycket mer vill du lära dig om programmering (om du fick välja)? Svara på skalan där 1 betyder inget mer, och 5 betyder mycket mer.

1	2	3	4	5
□	□	□	□	□

8. Vad tycker du har varit mest värdefullt av det du lärt dig under kursen?

9. Fria kommentarer om *hur du studerat* under kursen och om *dina mål* med kursen: