The 7th International Conference on
Computer Science & Education (ICCSE 2012)
July 14-17, 2012. Melbourne, Australia

Analysis on the relationship between student grades and computer programming time in learning the C programming language*

Zhixin Tie, Hong Zhuang, Qingyang Zhang, Zhaoqing Wang
School of Information Science and Technology
Zhejiang Sci-Tech University
Hangzhou, Zhejiang, 310018, P. R. China
E-mail: {Tiezx, Lisa, zqy, warls}@zstu.edu.cn

***Abstract*—The C Programming Language is one of the most popular teaching programming languages. It is generally considered that the computer programming time that a student spent on his homework directly affects his grade of the C Programming Language. There is little research regarding the correlation of students' computer programming time and their test grades. The computer programming time, grades of final test and the Zhejiang Computer Grade Test Band II (ZCGTBII) of students from a teaching class are collected. The students are divided into two groups based on their programming time. The T-Test is employed to analyze grades of final test and the ZCGTBII of the two groups. Analysis results show that the mean and the passing rates of final test and the ZCGTBII of Group I are greater than those of Group II. Therefore, we conclude that the amount of time students spend practicing programming on computers plays a very important role towards mastering the C Programming Language.**

***Index Terms*— C Programming language, Grade, Hypothesis test, Computer programming time, Homework**

I. INTRODUCTION

The C programming language is one of the most widely used programming languages in industry [1][2]. It is also one of the most popular teaching programming languages [3]-[7]. Harry H. Cheng listed ten good reasons to teach and learn the C programming language [7]. To learn the C programming language, a student needs to practice programming on computers. It is generally considered that the computer programming time of a student directly affects his performance on the course of the C programming language. There is little research regarding the effects of computer programming time and a student's progress learning the C programming language. Some experimental teaching systems and reforms were presented in [8][9]. Educational data mining methods were reviewed in [10]-[12].

In this paper, the computer programming time on homework, final grades and grades of the Zhejiang Computer Grade Test Band II (ZCGTBII) of students from a teaching class are collected. The Chi-Square Goodness-of-Fit Test is used to verify that final grades and ZCGTBII grades of this teaching class subject to the normal distribution. This is what we expect to be. Students of this teaching class are divided into two groups based on students' computer programming time on homework and minimum computer programming time on homework that the course of the C programming language syllabus prescribed for a student to learn this course. Group I is the students whose total computer programming time on homework is greater than or equal to the minimum computer programming time on homework. Group II is the students whose total computer programming time on homework is less than the minimum total computer programming time on homework. The Two-Sample Pooled T-Test is employed to analyze the final grades and the ZCGTBII grades of the two groups of students.

The rest of the paper is organized as follows. Section II gives an overview of some hypothesis tests that will be used in following sections. Section III introduces how to collect the analysis data. Section IV presents students' grades analysis methods. Finally, conclusions are drawn in Section V.

II. HYPOTHESIS TEST OVERVIEW

A. Chi-Square Goodness-of-Fit Test

The Chi-square Goodness-of-Fit Test is used to test if a sample of data comes from a population with a specific distribution [13]. Given a random samples of measurements, $X_1, X_2, \cdots X_n$, The chi-square test is defined for the hypothesis:

H0: The data follow a specified distribution.

* This work is partially supported by Zhejiang new century higher education reform project under Grant No. yb2010023, the Bilingual Education Model Curriculum of Ministry of Education of China.

Ha: The data do not follow the specified distribution.

Test Statistic: For the chi-square goodness-of-fit computation, the data are divided into k bins and the test statistic is defined as

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$$

Where O_i is the observed frequency for bin i and E_i is the expected frequency for bin i. The expected frequency is calculated by

$$E_i = N(F(Y_u) - F(Y_l))$$

Where F is the cumulative distribution function for the distribution being tested, Y_u is the upper limit for class i, Y_l is the lower limit for class i and N is the sample size.

The test statistic follows, approximately, a chi-square distribution with $(k-r-1)$ degrees of freedom where k is the number of non-empty cells and r is the number of estimated parameters.

Therefore, the hypothesis that the data are from a population with the specified distribution is rejected if

$$\chi^2 > \chi^2_{(\alpha,k-r-1)}$$

Where $\chi^2_{(\alpha,k-r-1)}$ is the chi-square percent point function with $(k-r-1)$ degrees of freedom and a significance level of α.

B. Two-Sample Pooled T-Test[13][14]

Given two random samples of measurements, $Y_1, Y_2, \cdots Y_{n_1}$ and $Z_1, Z_2, \cdots Z_{n_2}$, which have samples of size n_1 and n_2, respectively, are drawn from two populations with means μ_1 and μ_2, and variances $\sigma_1{}^2$ and $\sigma_2{}^2$.

The following hypothesis test can be used to test that the variances $\sigma_1{}^2$ and $\sigma_2{}^2$ are equivalent or not.

H0: The variances $\sigma_1{}^2$ and $\sigma_2{}^2$ are equivalent.

Ha: The variances $\sigma_1{}^2$ and $\sigma_2{}^2$ are not equivalent.

The test statistic is

$$F = S_1{}^2 / S_2{}^2$$

Where

$$S_1 = \sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}(Y_i - \overline{Y})^2},\ S_2 = \sqrt{\frac{1}{n_2}\sum_{i=1}^{n_2}(Z_i - \overline{Z})^2}$$

The test statistic F follows F-distribution with the numerator degrees of freedom $v_1 = n_1 - 1$ and denominator degrees of freedom $v_2 = n_2 - 1$. Therefore, the hypothesis H0 is rejected if:

$$F \geq F_{\alpha/2,v1,v2} \text{ or } F \leq F_{1-\alpha/2,v1,v2}$$

Where $F_{\alpha/2,v1,v2}$, $F_{1-\alpha/2,v1,v2}$ are the F-value with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom and a significance level of α.

If the variances $\sigma_1{}^2$ and $\sigma_2{}^2$ are equivalent, The Two-Sample Pooled T-Test can be used to test the relation of the means μ_1 and μ_2.

H0: $\mu_1 - \mu_2 \geq \delta$.

Ha: $\mu_1 - \mu_2 < \delta$.

The test statistic is

$$t = \frac{(\overline{Y} - \overline{Z}) - \delta}{S\sqrt{1/n_1 + 1/n_2}}$$

Where

$$S^2 = \frac{S_1{}^2(n_1 - 1) + S_2{}^2(n_2 - 1)}{n_1 + n_2 - 2}$$

$$\overline{Y} = \frac{1}{n1}\sum_{i=1}^{n_1} Y_i,\ \overline{Z} = \frac{1}{n2}\sum_{i=1}^{n_2} Z_i$$

The test statistic t follows t-distribution with $v = n_1 + n_2 - 2$ degrees of freedom. Therefore, the hypothesis H0 is rejected if:

$$t \leq -t_{\alpha,v}$$

Where $t_{\alpha,v}$ is the t-value with $v = n_1 + n_2 - 2$ degrees of freedom and a significance level of α.

III. DATA COLLECTION

According to teaching syllabus of the course of the C Programming Language revised in 2008 [15], teaching activities include 16 lectures and 16 computer programming classes. Each lecture has two class hours and each computer programming class has three class hours. There is homework after each computer programming class. Students are asked to spend at least three hours working on a programming assignment in our public computer centers.

In order to analyze the relation between students' grades and the time that students spent on programming after classes, one of C programming teaching classes is selected. This teaching class has 84 students. Students' final grades and grades of the Zhejiang Computer Grade Test Band II (ZCGTBII) are collected, as shown in column two and three of TABLE I. The statistics graph of students' final grades and the ZCGTBII grades are made, as shown in Figure 1 and Figure 2, respectively.

Students programming time can be divided into two parts. One is the time of computer programming classes. The other, which is called the computer programming time on homework, is the time that students used to do their homework after classes. The former is the same to each student, thus when analyzing students programming time, this part will be subtracted.

TABLE I. GRADES AND HOMEWORK PROGRAMMING TIME OF A TEACHING CLASS.

No.	Final Grade	ZCGTBII Grade	Homework Programming Time (Min.)
1	88	62	1170
2	87	80	2566
…	…	…	…
84	…	…	…

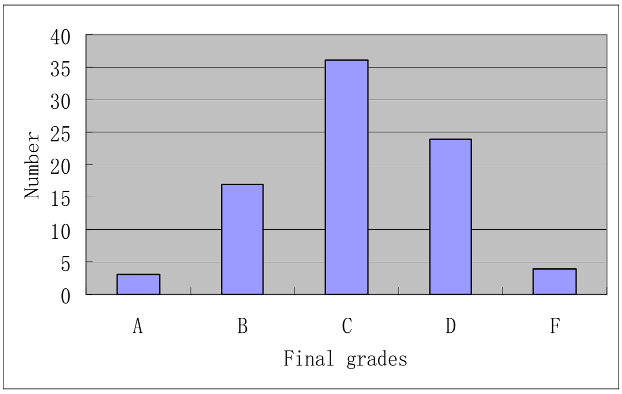

Figure 1. Figure 1 the students' final grade statistics graph.

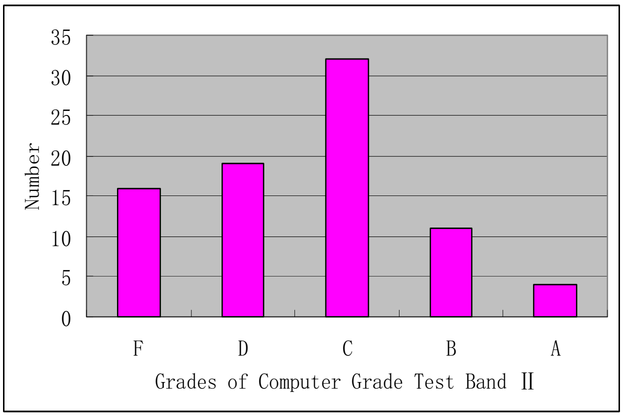

Figure 2. Figure 2 Statistics graph of the students' grades of the Zhejiang Computer Grade Test Band II (ZCGTBII).

In our university, students do their C Programming Language homework in our public computer centers. Thus the computer programming time on homework of a student can be extracted from the Homework System that is developed to collect students' homework online, and the Computer Laboratory Management System (CLMS) which is installed in our public computer centers to manage the usage of the computers. Because there has a record when a student commits his homework in the Homework System, and there has a record when a student starts using a computer and how long does it last in the CLMS. The computer programming time on homework of a student is summed from each time period that he (or she) uses a computer in public computer centers between the ends of each computer programming class and the moment that he (or she) commits his (or her) homework.

The total computer programming time on homework is a student spending on all homework assignments of the course of the C Programming Language. Because there have 16 homework assignments for the C Programming Language, thus, the total computer programming time on homework is a sum of each computer programming time on homework. The total computer programming time on homework of each student is listed in column four of TABLE II.

IV. STUDENTS' GRADES ANALYSIS METHOD

Different students have different final grades and ZCGTBII grades, because they have different efforts on the course of the C Programming Language. Thus students' final grades and ZCGTBII grades are considered random variables. The hypothesis tests that mentioned in section II can be employed to analyze students' grades. Firstly the Chi-Square goodness-of-fit test is employed to test if students' final grades and ZCGTBII grades are following the normal distribution, which is expected to be. Then the Two-Sample Pooled T-Test is used to test if students who spend more computer programming time on homework get better scores than those of who spend less computer programming time on homework.

A. Grades distribution hypothesis testing

1) Students' final grades distribution hypothesis testing

The hypothesis is:

H0: the students' final grades $X \sim N(\mu, \sigma^2)$.

Ha: the students' final grades X do not follow the distribution $N(\mu, \sigma^2)$.

Where μ, σ is unknown, denote the mean and standard deviation of X, respectively.

The test statistic is calculated in TABLE II. Here, $n = 84$, $k = 4$, $r = 2$. The significance level α is select to $\alpha = 0.01$. Thus, if H0 is true, the test statistic follows, approximately, a chi-square distribution with $(k - r - 1) = 1$ degrees of freedom.

As shown in TABLE II:

$$\chi^2 = 1.630 < \chi^2{}_{0.01}(1) = 6.63$$

Thus, hypothesis H0 is accepted. The students' final grades X follow normal distribution $N(\mu, \sigma^2)$ with $\mu = 73.47$, $\sigma = 9.70$, and a significance level of $\alpha = 0.01$

2) Students' ZCGTBII grades distribution hypothesis testing

Using the same method as shown in last subsection, the test statistic is calculated in TABLE III.

$$\chi^2 = 3.689 < \chi^2{}_{0.01}(1) = 6.63$$

Thus, The students' ZCGTBII grades follow normal distribution $N(\mu, \sigma^2)$ with $\mu = 69.62$, $\sigma = 11.32$, and a significance level of $\alpha = 0.01$

B. Spending more computer programming time on homework versus spending less computer programming time on homework: "Effects" on score

Do students who spend more computer programming time on homework receive higher scores than those who spending less computer programming time on homework? This question arises when instructors consider encouraging students to spend more programming time on the course of C programming Language. It is supposed to get higher grades if you spend more time to practice C programming. Does it true? The Two-Sample Pooled T-Test is employed to test it.

Based on teaching syllabus of the course of C Programming Language revised in 2008, students are asked to spend no less than three class hours to do each of their homework assignments. Thus the total computer programming time on homework of a student is supposed to be no less than 48 class hours, which is 2160 minutes. However, some students spend much more than the suggested time on homework assignments, and some students spend much less than the suggested time working on their assignments. Using this time, students are divided into two groups. One is the group of students whose total computer programming time on homework is greater than or equal to this time, denoted as Group I. The other is the group of students whose total computer programming time on homework is less than this time, denoted as Group II. Based on this classification principle, there are 68 students in Group I and 16 students in Group II. Students' final grades and ZCGTBII grades are used to test if the Group I receives higher grades than the Group II.

1) Comparing the passing rate of two groups of students

We define the passing rate of a group in an exam as number of students who pass the exam divide by total number of students in that group. The passing rates of final test and ZCGTBII of the two groups of students are shown in TABLE iv. From TABLE iv, we know that the passing rates of final test and ZCGTBII of Group I are greater than those of Group II by 18.75 and 19.54 percentage points, respectively.

TABLE II. COMPUTING THE TEST STATISTIC OF STUDENTS' FINAL GRADES.

Bin NO	**Grades Bin**	O_i	$F(Y_u) - F(Y_l)$	E_i	$(O_i - E_i)^2 / E_i$
1	A or B	20	0.2504	21.0342	0.051
2	C	36	0.3894	32.7091	0.331
3	D	24	0.2779	23.3420	0.019
4	F	4	0.0823	6.9147	1.229
Sum	/	84	/	/	1.630

TABLE III. COMPUTING THE TEST STATISTIC OF STUDENTS' ZCGTBII GRADES.

Bin NO	**Grades Bin**	O_i	$F(Y_u) - F(Y_l)$	E_i	$(O_i - E_i)^2 / E_i$
1	A or B	15	0.1796	14.7261	0.005
2	C	32	0.3071	25.1783	1.848
3	D	19	0.3157	25.8881	1.833
4	F	16	0.1977	16.2075	0.003
Sum	/	82	/	/	3.689

TABLE IV. THE PASSING RATE OF FINAL TEST AND ZCGTBII OF THE TWO GROUPS OF STUDENTS.

Group NO.	passing rate of Final test	passing rate of ZCGTBII
I	100.00%	83.82%
II	81.25%	64.29%

2) Hypothesis test on two groups of students' final grades

In order to use the Two-Sample Pooled T-Test to test if two groups of students' final grades have the same mean, the variances of two groups of students' final grades are equivalent should be tested firstly. We use μ_I and μ_{II} , and variances σ_I^2 and σ_{II}^2 to denote the mean and the variances of the two group respectively. Thus the first hypothesis is:

H0: The variances σ_I^2 and σ_{II}^2 are equivalent

Ha: The variances σ_I^2 and σ_{II}^2 are not equivalent.

Relevant statistics of two groups of students' final grades are calculated in TABLE V. The test statistic F is

$$F = 0.324$$

Because $F_{0.01/2,67,15} = 3.26$ and $F_{1-0.01/2,67,15} = 0.307$, Thus, the hypothesis H0 is accepted with a significance level of $\alpha = 0.01$.

Because the variances of two groups of students' final grades are equivalent, the Two-Sample Pooled T-Test can be used to test if the mean of students' final grades of Group I is greater than that of Group II. Thus, the second hypothesis is:

H0: $\mu_I - \mu_{II} \geq 9.0$

Ha: $\mu_I - \mu_{II} < 9.0$

From TABLE V, the test statistic t is

$$t = 0.288 > -t_{0.01,82} = -2.375$$

Thus, the hypothesis H0 is accepted with a significance level of $\alpha = 0.01$, i.e. the mean of group I of students' final grades is greater than or equal to that of group II by 9.0 points.

3) Hypothesis test on two groups of students' ZCGTBII grades

Because two students did not take the ZCGTBII, the total sample size is 82 students in this hypothesis test.

Using the same methodology as shown in last subsection, relevant statistics of two groups of students' ZCGTBII grades are calculated in TABLE VI.

TABLE V. RELEVANT STATISTICS OF THE TWO-SAMPLE POOLED T-TEST ON TWO GROUPS OF STUDENTS' FINAL GRADES.

Attributes	Group I	Group II
Mean	75.412	65.688
Standard deviation	7.710	13.549
Number of. measurements	68	16
Degrees of freedom	67	15
F	0.324	/
t	0.288	/

TABLE VI. RELEVANT STATISTICS OF THE TWO-SAMPLE POOLED T-TEST ON TWO GROUPS OF STUDENTS' ZCGTBII GRADES.

Attributes	Group I	Group II
Mean	70.853	63.643
Standard deviation	11.215	10.172
Number of. measurements	68	14
Degrees of freedom	67	13
F	1.215	/
t	0.065	/

From TABLE VI, the test statistic F is

$$F = 1.215$$

Because $F_{0.01/2,67,13} = 3.65$ and $F_{1-0.01/2,67,13} = 0.274$, we can conclude that the variances of two groups of students' ZCGTBII grades are equivalent with a significance level of $\alpha = 0.01$.

The following hypothesis test is used to test the relation of the two groups of students' ZCGTBII grades.

H0: $\mu_I - \mu_{II} \geq 7.0$

Ha: $\mu_I - \mu_{II} < 7.0$

From TABLE VI, the test statistic t is

$$t = 0.201 > -t_{0.01,80} = -2.373$$

Thus, the hypothesis H0 is accepted with a significance level of $\alpha = 0.01$, i.e. the mean of group I of students' ZCGTBII grades is greater than or equal to that of group II by 7.0 points.

V. CONCLUSION

Based on students' computer programming time on homework and the minimum computer programming time on homework that the C programming language syllabus prescribed for a student to learn this course, students of a teaching class are divided into two groups. Group I is the students whose total computer programming time on homework is greater than or equal to the minimum computer programming time on homework. Group II is the students whose total computer programming time on homework is less than the minimum computer programming time on homework. Hypothesis tests are employed to analyze the final grades and the ZCGTBII grades of the two groups of students. Analysis results show that the mean of the final grades and the mean of ZCGTBII grades of Group I are greater than those of Group II by 9.0 and 7.0 points, respectively. Analysis results also show that the passing rates of final test and ZCGTBII of Group I are greater than those of Group II by 18.75 and 19.54 percentage points, respectively. Thus we can conclude that the time that a student practicing programming on computers plays a very important role in their C Programming Language learning. Therefore, in teaching process of the course of the C Programming Language, In addition to normal classroom teaching, teachers/instructors should encourage students to use computers to do an appropriate amount of homework.

REFERENCES

[1] Programming Language Popularity, http://www.langpop.com/

[2] TIOBE Software, "TIOBE programming community index for June 2011," http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html

[3] Harry H. Cheng, "C for the course," Mechanical Engineering Magazine, September, 2009, pp 50-52

[4] Qinming He，et al, "National program of web-delivery for elaborate courses - The C programming language, " http://jpkc.zju.edu.cn/k/409/ml15.htm

[5] Kuanquan Wang，et al, "National program of web-delivery for elaborate courses - The C programming language, " http://cms.hit.edu.cn/elite/

[6] Zhaoqing Wang et al, "Ministry of Education Bilingual Education Program - The C programming language," http://ecs.zstu.edu.cn/ce/KCGK/kcgk.htm.

[7] Harry H. Cheng, "Ten Reasons to Teach and Learn Computer Programming in C," http://iel.ucdavis.edu/publication/WhyC.html.

[8] Ding Haiyan，Zou Jiang，Qiu Sha, "Experiment teaching system and reform measures in C language," Experim ental Technology and Management, Vol 27, No. 11, pp. 179–181, Nov. 2010 (in Chinese)

[9] Chen Ting, "Research on the teaching reform of C Language Programming," Vol 27, No. 10, pp. 182–184, Oct. 2010 (in Chinese)

[10] C. Romero and S. Ventura, "Educational data mining: a survey from 1995 to 2005," Expert Systems with Applications, Vol. 33, Issue 1, pp. 135-146, July 2007

[11] R. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," J. Educ. DataMining, vol. 1, no. 1, pp. 3–17, 2009.

[12] Crist´obal Romero and Sebasti´an Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE transactions on systems, man, and cybernetics -- part c: applications and reviews, vol. 40, no. 6, pp601-618, November 2010.

[13] NIST/SEMATECH, " Engineering Statistics Handbook," http://www.itl.nist.gov/div898/handbook.

[14] Ribakd E. Walpole, Raymond H. Myers, Sharon L. Myers and Keying Ye, "Probability & statistics for engineers & scientists," 7th edition, Prentice Hall, 2002.

[15] Teaching committee of Instructional Division of Computer Technology, Zhejiang Sci-Tech University, "Syllabus Of the C Programming Language," unpublished.