Zpravodaj Československého sdružení uživatelů TeXu

Vít Novotný

The Trends in the Usage of TeX for the Preparation of Theses and Dissertations at the Masaryk University in Brno

Zpravodaj Československého sdružení uživatelů TeXu, Vol. 25 (2015), No. 1-2, 80–85

Persistent URL: http://dml.cz/dmlcz/150229

The Trends in the Usage of TEX for the Preparation of Theses and Dissertations at the Masaryk University in Brno

Vít Novotný

This article is a statistical analysis of theses and dissertations written and defended during the period 2010–2015 at the Masaryk University in Brno (MU). The author assesses the trends in the usage of TEX and tests the hypothesis that theses and dissertations written using TEX received significantly better rating than those not written using TEX.

Key words:

thesis, dissertation, statistics

Introduction

A statistical analysis of theses defended at the Masaryk University in Brno (MU) between years 2010 and 2015 was carried out by the author of this article. The sample data for the analysis were kindly provided by doc. Ing. Michal Brandejs, CSc., the head of the Computer Systems Unit at the Faculty of Informatics at MU.

Table 1 shows the distribution of theses written and defended during the years 2010–2015 across the faculties of MU and Table 2 illustrates how many of these theses were written using TEX. Table 3 then presents the trends in the usage of TEX by the students of bachelor's, master's and doctoral degree programmes at the Faculty of Informatics (FI) and the Faculty of Science (Sci). Other faculties of MU were not considered since the number of theses written at these faculties using TEX was statistically insignificant (see Table 2). Theses written by students of lifelong education programmes were likewise ignored since none of them were written using TEX.

Analysis

A thesis was considered to be written using TEX if one or more files submitted with it satisfied one or more of the following conditions:

- The suffix was `tex`.
- The magic number was that of a DVI file.

 doi: 10.5300/2015-1-2/80

Faculty	#	%
Arts	10 000	21.98
Education	8 219	18.07
Social Studies	5 599	12.31
Science	5 275	11.60
Law	4 824	10.60
Economics & Administration	4 591	10.09
Informatics	2 904	6.38
Sports Studies	2 062	4.53
Medicine	2 014	4.43
Total	**45 488**	**100.00**

Table 1: The distribution of theses defended during 2010–2015 across the faculties of MU

Faculty	**With TEX**	**Total**	%
Informatics	1 716	2 904	59.09
Science	786	5 275	14.90
Economics & Administration	64	4 591	1.39
Arts	69	10 000	0.69
Medicine	8	2 014	0.40
Law	15	4 824	0.31
Education	19	8 219	0.23
Social Studies	12	5 599	0.21
Sports Studies	3	2 062	0.15
Total	**2 692**	**45 488**	**5.92**

Table 2: The distribution of theses written using TEX, which were defended during 2010–2015 across the faculties of MU

Degree	Fac.	2010	2011	2012	2013	2014	R
Bachelor's	FI	58.92	59.44	49.54	53.77	59.06	−0.195
	Sci	11.55	13.00	15.90	19.79	15.16	+0.703
	All	5.08	6.19	6.00	6.08	6.24	+0.731
Master's	FI	60.61	59.91	60.08	64.50	57.96	−0.046
	Sci	19.38	13.54	13.75	13.78	17.71	−0.180
	All	6.02	4.88	5.22	6.59	6.29	+0.490
Doctoral	FI	100.00	76.67	71.88	83.87	90.91	−0.155
	Sci	18.09	10.71	12.75	10.19	8.85	−0.830
	All	8.83	8.23	8.41	9.38	7.43	−0.361
All	**FI**	**60.83**	**60.53**	**54.92**	**60.57**	**59.34**	**−0.188**
	Sci	**14.86**	**12.96**	**14.74**	**16.55**	**15.45**	**+0.577**
	All	**5.67**	**5.70**	**5.73**	**6.41**	**6.28**	**+0.855**

Table 3: The percentage of theses written using TEX which were defended in each year during the years 2010–2014 and the sample correlation coefficient R between the percentage and the years with remarkably strong correlations emphasized

- The MIME type was `application/postscript` and the file contained the `TeXDict` substring suggesting that the file was a PostScript document which had been created using the dvips utility.
- The MIME type was `application/pdf` and either the `Creator` or the `Producer` PDF header contained the `TeX` substring suggesting that the file had been created using either the dvipdfm utility or a TEX engine which supports PDF output.

Provided the heuristic is sound, there was a marked and steady increase in the use of TEX for the typesetting of theses at MU during the period 2010–2014 (see Table 3). This, however, does not necessarily hold true for individual faculties and degree study programmes with some of them showing barely any correlation between the years and the use of TEX others showing a strong negative correlation. A particularly striking example of the latter case is the pronounced downward trend in the use of TEX for the typesetting of doctoral theses at the Faculty of Science.

At first the null hypothesis h_1 was supposed that the grades awarded to theses written using and not using TEX, respectively, have the same distribution on the significance level $\alpha = 0.05$. The one-tailed Pearson's χ^2 test (Pearson, 1900) of the goodness of fit was applied to the observations of awarded grades (see Table

	Without TEX	E(**With TEX**)	O(**With TEX**)	$(E-O)^2/E$
A	15 476	987.635	1 181	37.858
B	9999	638.108	587	4.093
C	7 926	505.815	381	30.799
D	4 020	256.545	194	15.248
E	2 783	177.603	128	13.853
F	1 979	126.294	145	2.771
Total	**42 183**	**2 692**	**2 692**	**104.623**

Table 4: The contingency table of the numbers of marks awarded to theses written and defended during 2010–2015 with Pearson's goodness-of-fit measure $(E-O)^2/E$ between the expected (E) and the observed (O) numbers of marks awarded to theses written using TEX

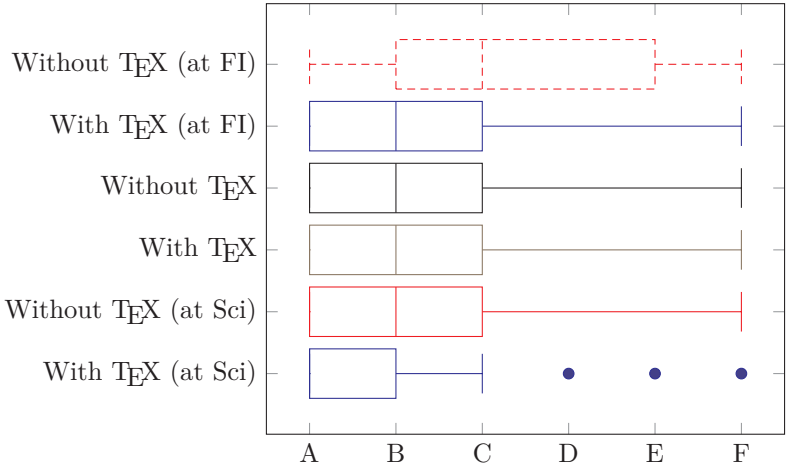

Figure 1: A box plot of the grades of theses written and defended during the period 2010–2015 at FI, Sci and all the faculties of MU with and without TEX

4). Since

$$\sum_{\mathrm{A,B,\ldots,F}} (\mathrm{E}-\mathrm{O})^2/\mathrm{E} = 104.623 \gg 11.07 = \chi^2_{1-\alpha}(5) \tag{1}$$

the null hypothesis h_1 was refused and it was concluded that the grades are indeed differently distributed on the significance level α.

Having shown that the distribution of grades awarded to theses written using and not using TEX is different, the author proceeded to test if this holds for individual grades. The null hypothesis h_A was supposed that the distribution of grade A being awarded to theses written using and not using TEX is equivalent. The two-tailed Mann-Whitney U test (Mann and Whitney, 1947) was applied to the observations of grade A being and not being awarded to theses written using and not using TEX:

$$\begin{aligned}
m_1 &= 15\,476 && \text{(Without TEX (grade A))}\\
m_2 &= 1\,181 && \text{(With TEX (grade A))}\\
n_1 &= 42\,183 && \text{(Without TEX (total))}\\
n_2 &= 2\,692 && \text{(With TEX (total))}\\
U_1 &= m_1(m_2 \cdot 0.5 + (n_2 - m_2)) + (n_1 - m_1)((n_2 - m_2) \cdot 0.5) && (2)\\
&= 52\,699\,952.5\\
U_2 &= m_2(m_1 \cdot 0.5 + (n_1 - m_1)) + (n_2 - m_2)((n_1 - m_1) \cdot 0.5) && (3)\\
&= 60\,856\,683.5\\
U &= \min(U_1, U_2) = U_1 = 52\,699\,952.5 && (4)
\end{aligned}$$

Since $n_1 n_2 \gg 20$, $U \sim \mathrm{N}\left(\frac{n_1 n_2}{2}, \frac{n_1 n_2 (n_1+n_2+1)}{12}\right)$. After normalization to

$$\mathrm{N}(0,1) \sim z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \approx -\frac{4\,078\,365.5}{651\,662.46} \approx -6.258 \tag{5}$$

the two-tailed p-value β was computed as follows:

$$\begin{aligned}
&\arg\min_\beta P(\Phi^{-1}_{\beta/2} \le z \le \Phi^{-1}_{1-\beta/2}) = \beta && (6)\\
&\iff \Phi^{-1}_{\beta/2} = -6.258 \iff \beta/2 = 1 - \Phi(6.258) \iff \beta \approx 0
\end{aligned}$$

Since $\beta < \alpha$ the null hypothesis h_A on the significance level α was refused. Following a similar procedure for grades from B to F, the following conclusions on the significance level α could be made:

- Theses written using TEX had been awarded grade A significantly more often than those not written using TEX.

- Theses written using TEX had been awarded grades C and D significantly less often than those not written using TEX.
- No significant difference was observed in the distributions of grades B, E and F being awarded to theses written using and not using TEX.

A box plot of the grades is shown in Figure 1.

Conclusion

The author has shown that there was a marked and steady increase in the usage of TEX at MU the period during 2010–2014 and that the usage of TEX correlated with statistically significantly better grades during the period 2010–2015.

Literatura

Mann, H. B., Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 1947, Vol. 18, No. 1, s. 50–60.

Pearson, C. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 1900, Vol. 50, Iss. 302, s. 157–175.

Vít Novotný
witiko@mail.muni.cz