Seed Science and Biotechnology ***©2008 Global Science Books***

Does Number of Kernels per Spike Follow a Normal Distribution in Spring Barley?

Marcin Kozak[1*] • Dariusz Gozdowski[1] • Zdzisław Wyszyński[2]

[1] Department of Experimental Design and Bioinformatics, Warsaw University of Life Sciences – SGGW, Nowoursynowska 159, 02-776 Warsaw, Poland
[2] Department of Agronomy, Warsaw University of Life Sciences – SGGW, Nowoursynowska 159, 02-776 Warsaw, Poland

***Corresponding author*: * nyggus@gmail.com**

ABSTRACT

In this paper we investigate whether distribution of number of kernels per spike, which is a component of small grain cereal yield per spike, can be approximated by a normal distribution in spring barley. We emphasize that this trait is not the same as mean number of kernels per spike, which is normally considered in the classical yield component analysis (conducted for grain yield per unit area or per plant), even though the latter is referred to by the same name as the former; we consider this a mistake in terminology. Hence we suggest that in classical yield component analysis one should use the correct name of mean number of kernels per spike. Further, our study shows that the distribution of number of kernels per spike in spring barley is usually not normal, although in some situations it follows the pattern of this distribution. The Box-Cox transformation seldom led to approximate normal distribution of this trait.

Keywords: count variable, grain yield per spike, normal distribution, transformation
Abbreviations: MNKS, mean number of kernel per spike; **NKS**, number of kernels per spike

INTRODUCTION

The number of kernels per spike (NKS) is one of two components of grain yield per spike in small grain cereal investigations, the other being mean kernel weight; such investigations may constitute an element of investigations on grain yield per plant or per unit area (Gozdowski *et al.* 2007). From Gozdowski *et al.*'s (2007) study it follows that both these components are important in determining grain yield per spike, although NKS had a slightly greater effect than mean kernel weight. Unlike mean kernel weight and grain yield per spike, NKS is unlikely to follow a normal distribution because it is a count variable and as such is not normally distributed. On the other hand, it is common practice to assume an approximate normal distribution for a count variable that has a reasonable number of values (e.g., it is not a variable on a 1 to 5 scale, or the like).

Very often the variable being studied is not the number of kernels per spike *per se*, but rather the mean number of kernels per spike (MNKS). This is the case in most classical cereal yield component analyses for data from field trials in which the number of kernels per spike is determined for a sample of spikes from each plot; MNKS for the plot is then estimated as a mean of these values (see Kozak and Mądry 2006 for discussion on this topic). In these situations, MNKS is *not* a count but a continuous variable, and as the mean it follows a normal distribution. For this reason, the assumption that this trait follows a normal distribution in classical yield component analyses is justified.

Nevertheless, when grain yield per spike is of interest, NKS is considered instead of MNKS. Unlike MNKS, NKS *is* a count variable, and thus its distribution needs special attention. To study whether NKS's distribution can be approximated by a normal distribution, one would need to have at one's disposal a large sample of spikes from which NKS would be measured. As we have such data for two spring genotypes in 24 environments (3 years × 4 nitrogen rates × 2 sowing dates), sample sizes for these 48 cases ranging from 180 to 758, we decided to check the assumption in question.

Thus, the aim of this paper is to study whether the assumption that the number of kernels per spike follows a normal distribution is acceptable in spring barley.

MATERIALS AND METHODS

Plant material

A detailed description of the experiment is provided by Gozdowski *et al.* (2007). Here we simply provide the information needed to follow our analysis. The field experiment was carried out at the Chylice Experimental Station of the Warsaw University of Life Sciences (52° 05′ N, 20°32′ E) in 1999, 2001 and 2002. Soils of the experiment were classified as black earth formed of loamy sand of glacial origin. Besides the year, the following factors were studied: cultivar ('Rasbet' and 'Rastik', the former having a hulled grain while the latter a hulless grain), sowing date (early and delayed about three weeks) and nitrogen fertilization level (0, 30, 60 and 90 kg/ha). Doses of 30 and 60 kg N per ha were applied before sowing in the form of ammonium sulphate; a dose of 90 kg N per ha was split-applied, 60 kg N per ha being applied before sowing and 30 kg N per ha during shoot elongation. The experiment was arranged in a split-plot randomized complete block design with four replications within years; the cultivar × sowing date combination constituted main plots, whereas nitrogen rates the subplots. The plot area was 30 m^2. At harvest, plants were taken for measurements from a 0.22 m^2 (two 1-meter length rows) area per plot, and among other traits, number of kernels per spike was counted for each spike.

Probing into the NKS distribution

We based the distribution checking mainly (a) histograms of NKS in the environments, and (b) the Shapiro-Wilk test (Shapiro and Wilk 1965), with the help of which we verified the hypothesis on lack of a normal distribution of NKS. Note that the number of classes for the histograms was equal to the number of exclusive values NKS taken in the particular case considered (cultivar ×

Received: 27 April, 2008. Accepted: 9 June, 2008.

Table 1 Actual *P*-values for the Shapiro-Wilk test for the original and transformed via the Box-Cox transformation (in round brackets) number of kernels per spike of two spring barley genotypes for the combinations of N dose, sowing date and year.

N	Sowing date	'Rasbet'			'Rastik'		
		1999	2001	2002	1999	2001	2002
0	early	<0.0001 (0.0005)	0.0001 (0.0005)	0.0001 (0.0002)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	<0.0001 (<0.0001)
	delayed	<0.0001 (<0.0001)	<0.0001 (0.0010)	0.0049 (0.0077)	0.0026 (0.0057)	<0.0001 (0.0001)	0.0012 (0.0013)
30	early	<0.0001 (0.0001)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	<0.0001 (<0.0001)
	delayed	<0.0001 (<0.0001)	<0.0001 (0.0022)	<0.0001 (<0.0001)	<0.0001 (0.0026)	<0.0001 (0.0073)	0.0016 (0.0019)
60	early	<0.0001 (<0.0001)	<0.0001 (<0.0001)	0.0002 (0.0573)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	<0.0001 (<0.0001)
	delayed	<0.0001 (0.0004)	<0.0001 (0.0005)	<0.0001 (0.0001)	<0.0001 (<0.0001)	<0.0001 (0.0051)	<0.0001 (<0.0001)
90	early	<0.0001 (0.0003)	0.0064 (0.0332)	0.0002 (0.0003)	<0.0001 (<0.0001)	<0.0001 (<0.0001)	<0.0001 (0.0006)
	delayed	<0.0001 (<0.0001)	0.0003 (0.0071)	0.0021 (0.0099)	<0.0001 (<0.0001)	<0.0001 (0.0045)	<0.0001 (<0.0001)

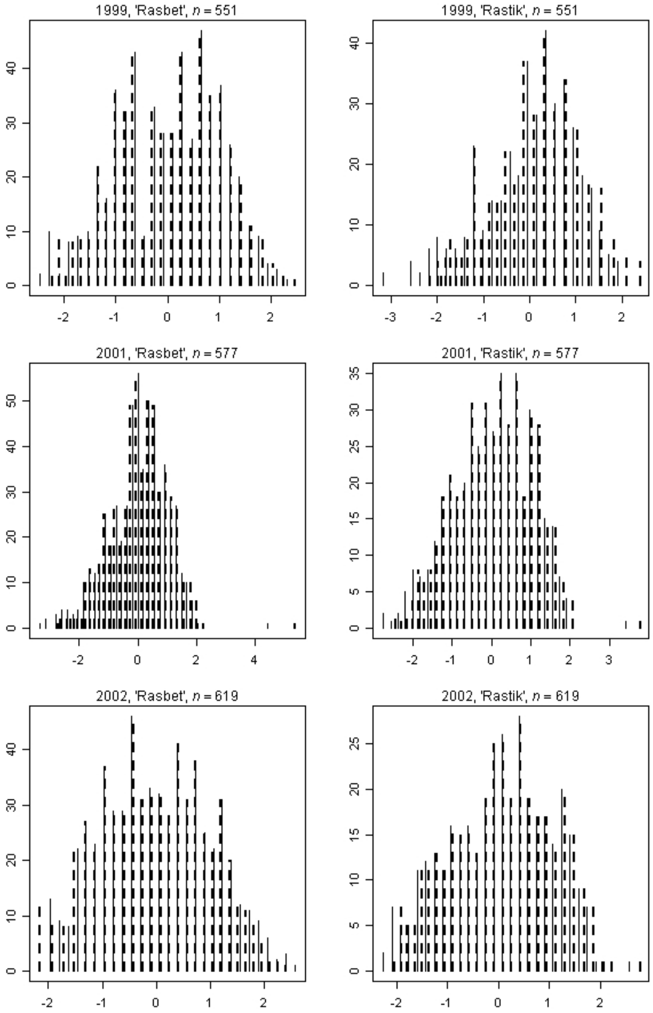


Fig. 1 Distribution of original (solid line) and transformed (bold dashed line) number of kernels per spike of two spring barley cultivars at the early sowing data and N dose of 0 kg/ha, in 1999, 2001 and 2002 (*n* = sample size). The variables are standardized to zero mean and unit variance.

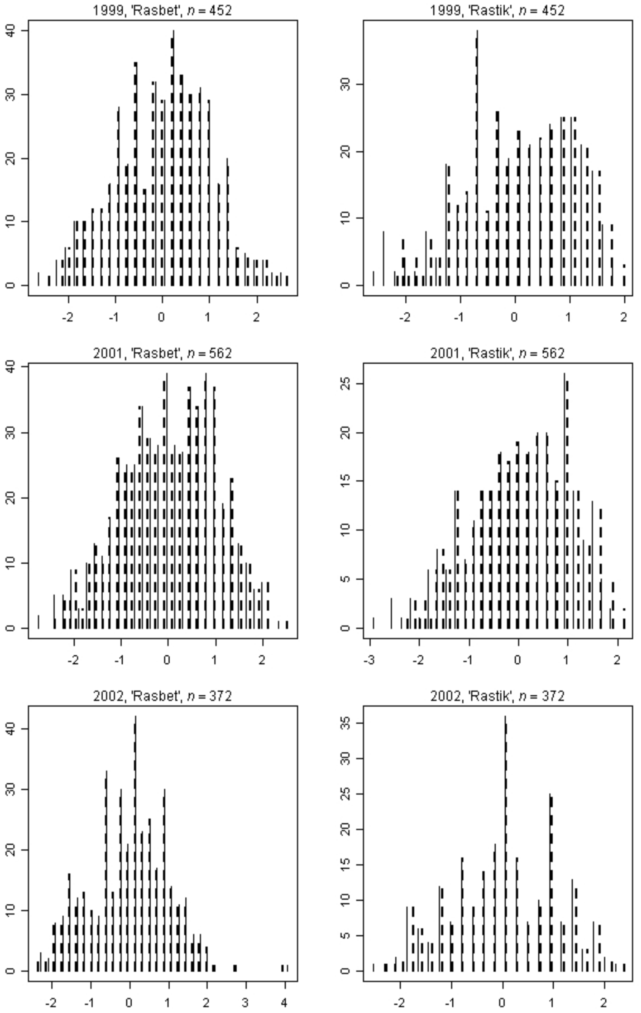


Fig. 2 Distribution of original (solid line) and transformed (bold dashed line) number of kernels per spike of two spring barley cultivars at the late sowing data and N dose of 0 kg/ha, in 1999, 2001 and 2002 (*n* = sample size). The variables are standardized to zero mean and unit variance.

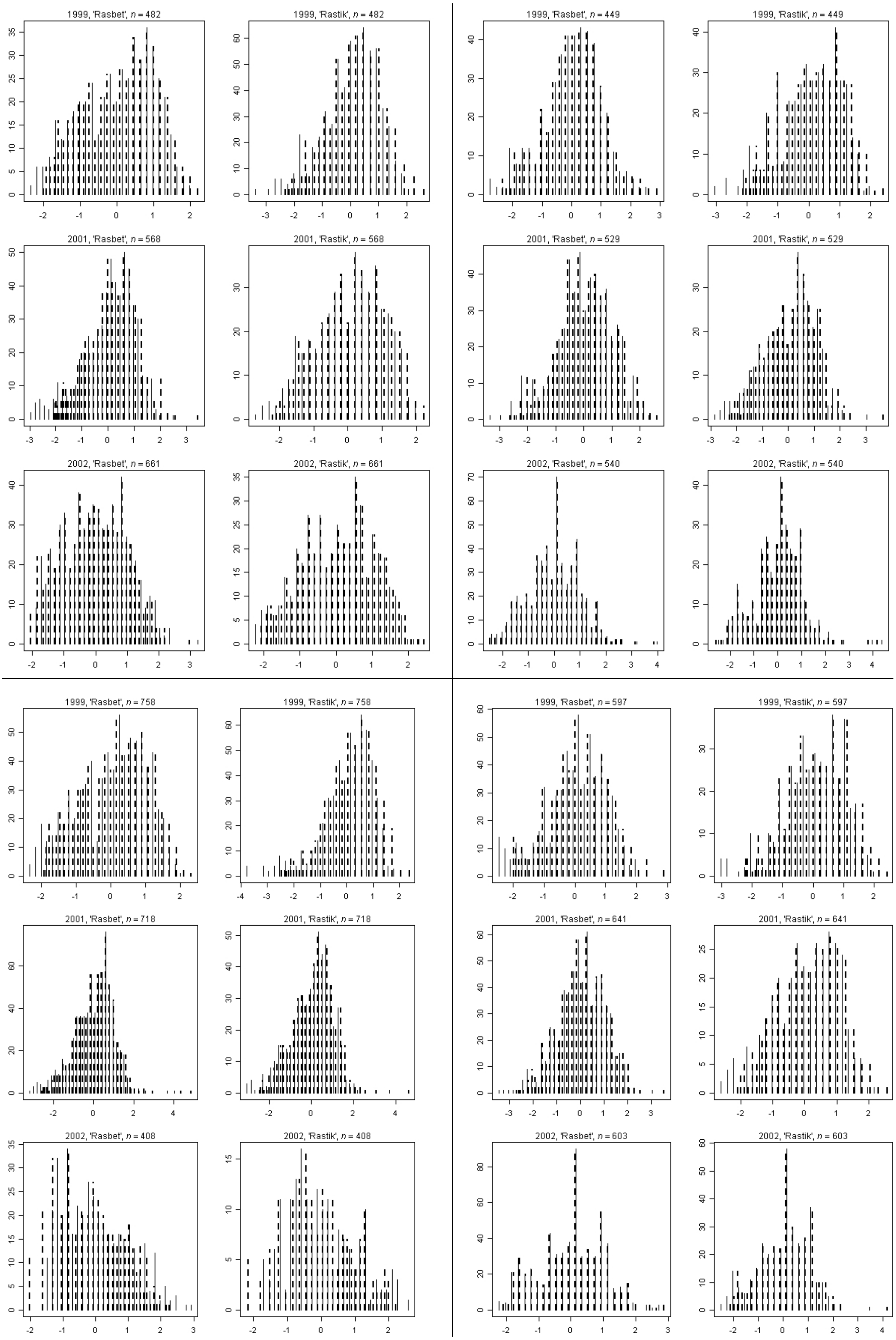
environment). When the Shapiro-Wilk test rejected the hypothesis at the 0.05 probability of type I error (which actually was the case for each of the 48 cases), we applied the Box-Cox transformation (Box and Cox 1964) to normalize the data. The histogram of a transformed variable was added to the corresponding histogram for the original variable. Since we were interested in the shape of the variables' distribution, central tendency and variability of the variables could be discounted. This is why standardized variables with zero mean and unit variance were used to plot the histograms.

For all computation and graphs R language was used (R Development Core Team 2008).

RESULTS

In each case (cultivar × environment) the Shapiro-Wilk test rejected the null hypothesis that NKS follows a normal distribution. It also rejected the hypothesis for most of the transformed variables (see **Table 1** for the actual *P*-values) – only in two cases ('Rasbet' 2001, 90 kg N per ha, early sowing date; and 'Rasbet' 2002, 60 kg N per ha, early sowing date) the actual *P*-values were not very small, in the former being 0.0573 while in the latter 0.0332, indicating that the distribution might be approximately normal. Although the results of the testing should not constitute the basis for acceptance/rejection of normality (in our study also because of such a large sample, which gave rise to large statistical power of the tests, so it was rather easy to reject the hypotheses), similar conclusions can be drawn from **Figs. 1-8**, which represent the distributions of both original and transformed variables for all 48 cases.

Clearly the Box-Cox transformation was seldom of help in normalizing the variables even though sometimes it did provide an approximately normal distribution; in general this approximation could not be thought of as fair enough


1999, 'Rasbet', n = 482
1999, 'Rastik', n = 482
1999, 'Rasbet', n = 449
1999, 'Rastik', n = 449
2001, 'Rasbet', n = 568
2001, 'Rastik', n = 568
2001, 'Rasbet', n = 529
2001, 'Rastik', n = 529
2002, 'Rasbet', n = 661
2002, 'Rastik', n = 661
2002, 'Rasbet', n = 540
2002, 'Rastik', n = 540
1999, 'Rasbet', n = 758
1999, 'Rastik', n = 758
1999, 'Rasbet', n = 597
1999, 'Rastik', n = 597
2001, 'Rasbet', n = 718
2001, 'Rastik', n = 718
2001, 'Rasbet', n = 641
2001, 'Rastik', n = 641
2002, 'Rasbet', n = 408
2002, 'Rastik', n = 408
2002, 'Rasbet', n = 603
2002, 'Rastik', n = 603

Previous page: Top left = Fig. 3; Top right = Fig. 4; Bottom left = Fig. 5; Bottom right = Fig. 6

Fig. 3 Distribution of original (solid line) and transformed (bold dashed line) number of kernels per spike of two spring barley cultivars at the early sowing data and N dose of 30 kg/ha, in 1999, 2001 and 2002 (*n* = sample size). The variables are standardized to zero mean and unit variance.

Fig. 4 Distribution of original (solid line) and transformed (bold dashed line) number of kernels per spike of two spring barley cultivars at the late sowing data and N dose of 30 kg/ha, in 1999, 2001 and 2002 (*n* = sample size). The variables are standardized to zero mean and unit variance.

Fig. 5 Distribution of original (solid line) and transformed (bold dashed line) number of kernels per spike of two spring barley cultivars at the early sowing data and N dose of 60 kg/ha, in 1999, 2001 and 2002 (*n* = sample size). The variables are standardized to zero mean and unit variance.

Fig. 6 Distribution of original (solid line) and transformed (bold dashed line) number of kernels per spike of two spring barley cultivars at the late sowing data and N dose of 60 kg/ha, in 1999, 2001 and 2002 (*n* = sample size). The variables are standardized to zero mean and unit variance.

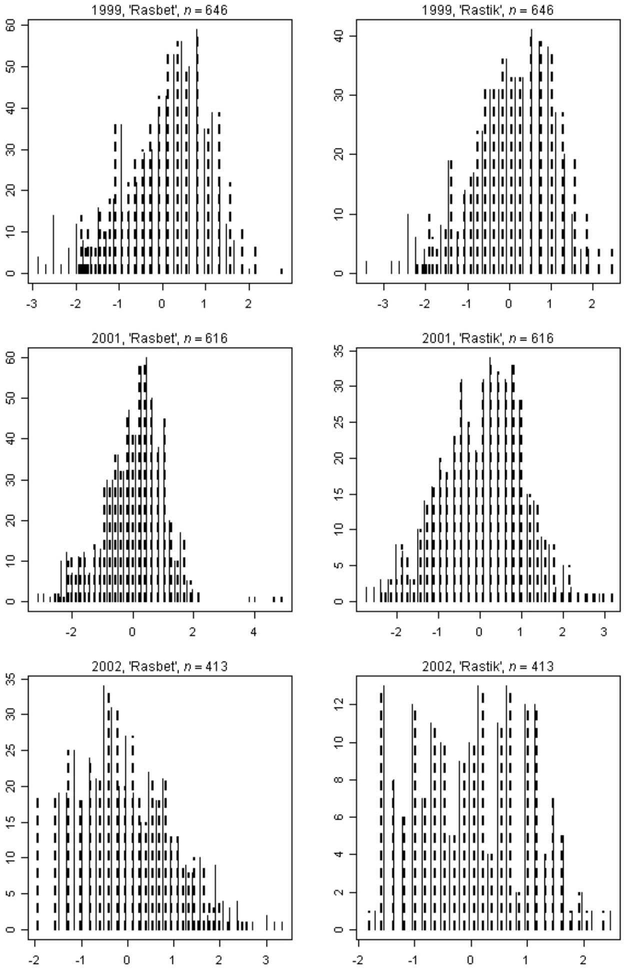


Fig. 7 Distribution of original (solid line) and transformed (bold dashed line) number of kernels per spike of two spring barley cultivars at the early sowing data and N dose of 90 kg/ha, in 1999, 2001 and 2002 (*n* = sample size). The variables are standardized to zero mean and unit variance.

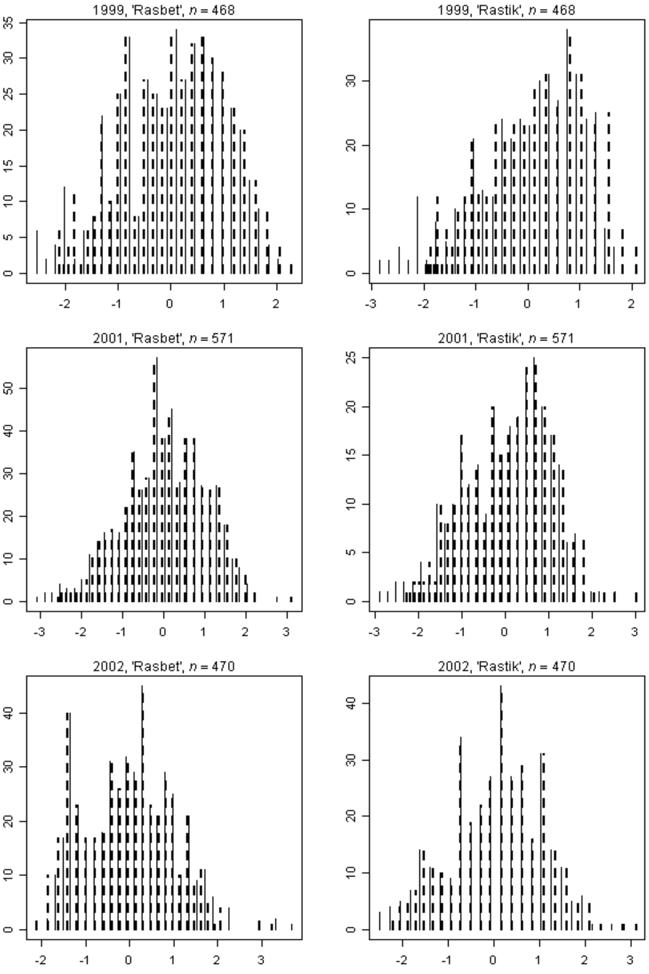


Fig. 8 Distribution of original (solid line) and transformed (bold dashed line) number of kernels per spike of two spring barley cultivars at the late sowing data and N dose of 90 kg/ha, in 1999, 2001 and 2002 (*n* = sample size). The variables are standardized to zero mean and unit variance.

for statistical purposes. Quite often transformed distributions had fairly the same shape as the corresponding original ones (**Figs. 1-8**). Interestingly, the transformation affected mostly the borders of the distributions, usually having no or negligible impact on their central parts.

DISCUSSION AND CONCLUSIONS

With the knowledge provided by this study it is difficult to decide whether spring barley NKS can be generally assumed to be approximated by a normal distribution. The results rather show that this approximation is poor despite applying the Box-Cox transformation, which is generally assumed to normalize data well, also in the case of skewed distributions (Quinn and Keough 2002, p 66). Even though NKS was usually symmetrically distributed, sometimes it was not: see, for example, the plots for 'Rasbet' in 2002, 60 kg N per ha, early sowing date (**Fig. 5**); 'Rasbet' and 'Rastik' in 2002, 90 kg N per ha, early sowing date (**Fig. 7**); and 'Rastik', 1999, 90 kg N per ha, delayed sowing date (**Fig. 8**). It is worth noting that if only the distribution was far from symmetric, the transformation did not help overcome this problem.

If NKS does not follow a normal distribution, despite the use of transformation, discontinuous distributions might be tried such as Poisson or, in case of overdispersion, quasi-Poisson (see Agresti 2002). Nonetheless, if one decides to analyze the variable as though it had been normally distributed, then one should always remember that the inference and interpretation from any statistical method should be considered as approximate, and this fact should be underlined when reporting such analyses. The decision as to which approach one should follow should be taken on a case-by-case basis, taking account of which method one is going to apply and to what kinds of questions one seeks answers.

We believe that the above conclusions apply also to other small grain cereal species. This is due to similarity of grain morphology and biological processes that determine grain of small grain species. But the main reason behind this is of theoretical nature: NKS is a count variable, and as such sometimes it may, though quite often does not have to, follow the pattern similar to that of a normal distribution. Nonetheless, similar studies for other crop species would be valuable.

Finally, let us recall what we have touched upon in the Introduction. In many classical cereal yield component analyses it is MNKS (mean number of kernels per spike) that is measured, analyzed and interpreted, and not NKS (number of kernels per spike), even though it is standard to call the former with the name of the latter. As deep-seated an approach as it is, it is misleading and may provide unfortunate miscomprehension of the choice of a distribution for a variable considered (no matter which of these two traits is studied). Hence, for any cereal crop species, it is crucial to use correct names for both these traits – number of kernels per spike and mean number of kernels per spike.

REFERENCES

Agresti A (2002) *Categorical data analysis* (2nd Edn) John Wiley & Sons, 710 pp

Box GEP, Cox DR (1964) An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association* **77**, 209-210

Gozdowski D, Kozak M, Kang MS, Wyszyński Z (2007) Dependence of grain weight of spring barley genotypes on traits of individual stems. *Journal of Crop Improvement* **21 (1/2)**, 223-233

Kozak M, Mądry W (2006) Note on yield component analysis. *Cereal Research Communications* **34 (2-3)**, 933-940

Quinn GP, Keough MJ (2002) *Experimental Design and Data Analysis for Biologists*, Cambridge University Press, Cambridge, 537 pp

R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online: http://www.R-project.org

Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591-611