UDC 519.237.4

Dispersion analysis in colorimetric control

S. Yefymenko, I. Hryhorenko, S. Hryhorenko

National Technical University "Kharkiv Polytechnic Institute", Kirpicheva Str., 2, 61002, Kharkiv, Ukraine
sefimenko64@gmail.com

Abstract

In the paper, solutions to the scientific and practical task of applying covariance analysis to determine the factor influence on the functional transformation of the control parameter in colorimetric research are considered. The research involves determining the factor influence on the additive and multiplicative components of the measurement error of the colorimetric control parameter to assess the reliability of conclusions about the factor influence on the transformation of the control parameter. Limits on the amount of the main level (control parameter) and factors affecting the result of colorimetric control are determined. During the study, the equations to assess the reliability of statistical conclusions about the informational significance of indicators of colorimetric control for a simplified model of cross-classification were obtained.

The need to carry out the research is related to the fact that during colorimetric control of grain crops, the uncertainty of the measurement results of the values of the controlled indicators is large enough, and so it must be evaluated. However, the proposed approach is not limited to a few colorimetric controls, but can be extended to a wider range of problems, where it is necessary to consider the factor influence with further ranking of factors by levels of significance.

Keywords: dispersion analysis; yellowness of wheat; factor influence; probability distribution density; colorimetry.

Received: 16.05.2023 Edited: 15.06.2023 Approved for publication: 21.06.2023

Introduction

Everyone knows that wheat is one of the most important and popular grain crops in the world [1], as about 725 million tons of grain are harvested worldwide every year. Durum wheat varieties are used in various food products, including couscous, pasta and bulgur, but mainly in the form of flour for the production of pasta [1, 2]. In this century, attention to food quality rather than to quantity has shifted the focus of research to improving the nutritional quality of wheat, which is measured by various parameters such as raw material moisture, protein content, and flour colour. The latter is related to carotenoid pigments, which are of great importance for human nutrition [3]. Yellow pigment content (YPC) is one of the most important indicators of the quality of durum wheat grain [4] due to the existing correlation between yellowness and the quality of flour products [1, 4]. For this reason, accurate determination and, if possible, improvement of YPC is the main goal of most programmes of both durum wheat selection and its quality processing, especially considering the fact that competition in the market of flour and its products has made this trait even more important [1]. This became especially relevant after the legislative ban on the use of artificial colours in the production of pasta in some European countries [5], which strengthened the role of durum wheat cultivation programmes to improve YPC [1]. Based on the above, the task of considering the factor influence on the determination of the YPC level using colorimetric methods becomes relevant. In work [6], the model of factor influence during colorimetric control is considered. It is noted that it is technically very difficult to provide the required sufficiently large number of non-standard samples while maintaining the uniformity of the measurement experiment. A simplified model is proposed to avoid such complexity of the experiment, leaving only main deviations and deviations caused by pairwise interactions of the factors.

Study of a simplified model of the factor influence on the result of colorimetric control of the YPC level with restrictions on the volume of measurements

In some cases of research of non-deterministic objects, the influence of some values on the initial value of the object cannot be estimated quantitatively. At the same time, a researcher may be interested in how significant the influence of one or another factor on the dispersion of the results of observations of the initial value is. To study the influence of factors that interfere with the initial value (feedback), their overall estimation, ranking and selecting the most significant

ones among them, all methods of screening quantitative factors and regression analysis methods are obviously unsuitable since they involve measuring the levels of the factors being studied [7, 8].

To estimate the influence of each factor on the response and to compare the influence of different factors, it is necessary to establish some quantitative indicator of this influence. To solve this kind of problems, researchers use different techniques and ways of organizing sample data, the essence of which consists in applying different plans for classifying observations according to possible sources of dispersion. The processing of such classified data is carried out using the methods of dispersion analysis, which allow decomposing general dispersion into components caused by the influence of the levels of factors of interest to the researcher [9].

Let us assume that Z is a colorimetric control parameter that needs to be estimated (yellowness of wheat), and $K_1, \dots K_n$ are control indicators determined during the measurement process (for example, grain temperature, humidity, illumination, and others). The result of observing the value of each of the control indicators can be written in the form of a mathematical model, in which the influencing factors are Z and $(n-1)$ factors caused by the variability of the remaining control indicators. This statement is because the remaining indicators quantitatively characterize $(n-1)$ physical properties of the object of control and differ from the colorimetric control parameter Z in that it is possible to directly measure their levels.

The model of the influence of factors on the measurement results of a single control indicator K (grain moisture) is characterized by two influencing factors (grain temperature, illumination). It has the following form [9]:

$$K_{abcv} = \bar{K} + \delta_a + \alpha_b + \beta_c + (\delta\alpha)_{ab} + (\delta\beta)_{ac} + h_{abcv}, \quad (1)$$

where a, b, c are the numbers of levels of influencing factors;

δ_a is the deviation of the measurement result of the indicator K from its average value $\bar{K}$ due to the influence of the parameter Z (yellowness of wheat – YPC) i.e. different characteristics of selected wheat samples;

α_b, β_c are the deviation of the measurement result of K_{abcv} from $\bar{K}$ due to two factors;

$(\delta\alpha)_{ab}$, $(\delta\beta)_{ac}$ are the deviation due to paired interactions of the influencing factors;

h_{abcv} is a random remainder.

The simplified model (1) can be reduced to two two-factor cross-classification models [9]:

$$K_{abv} = \bar{K} + \delta_a + \alpha_b + (\delta\alpha)_{ab} + h_{(\alpha)abv};\ c = q = v, \quad (2)$$

$$K_{acv} = \bar{K} + \delta_a + \beta_c + (\delta\beta)_{ac} + h_{(\beta)acv};\ b = q = v, \quad (3)$$

where v is the number of multiple measurements of the indicator K in a cell of the model output data tables (2) and (3); $h_{(\alpha)abv}$, $h_{(\beta)abv}$ are random remainders caused by two factors: grain temperature and illumination, respectively.

We will conduct a study of the condition that allows synthesizing model (1) using (2) and (3).

The first condition is to ensure that main deviations in models (2) and (3) are equal to each other. For this, the number of groups of the results of observing the values of the control indicator K should be the same for all models. This corresponds to the same number of terms in the source data table [9]. At the same time, the number of values of the control indicator K in each of the groups should be the same g. For this, the number of groups of the results of observing the values of the control indicator K should be the same for all models.

We will use the following actions to model the additional factor influence (for the columns of the initial data) on the indicator K:

– we rank the internal group values of the control indicator K_i in ascending order, which corresponds to the selected additional influencing factor;

– we divide the ranked (by all g groups) series of the values of the indicator K_i into l subgroups;

– in each subgroup, we select n values of the information indicator K that correspond to n values of the indicator K_i and enter them into the input data cell.

The obtained table $g \times l$ of the results of observations of the values of the control index K with n multiple observations in each of the cells $g \times l$ can be used for variance analysis of any of the models (2) or (3) of cross-classifications corresponding to the given additional factor affecting K_i, $i = \overline{1,2}$. We denote these factors as K_α, K_β. Generally, any of these factors will be denoted as $K_{(i)}$, and any of the models (2) or (3) can be represented as [9]:

$$K_{adv} = \bar{K} + \delta_a + i_d + (\delta_i)_{ad} + h_{adv}. \quad (4)$$

Complete decomposition of the sum of squares of the deviations of K_{adv} values from $\bar{K}$, with the fulfilment of initial conditions and restrictions, model (1) has the following form:

$$\Sigma = \Sigma_\delta + \Sigma_a + \Sigma_{\delta a} + \Sigma_{ea}, \quad (5)$$

where Σ_δ, Σ_a, $\Sigma_{\delta a}$, Σ_{ea} are the sum of the squared deviations of the two influencing factors and their interactions, respectively.

The results of the dispersion analysis of model (5) are given in Table 1 [9], where $\bar{K}_a$, $\bar{K}_d$, $\bar{K}_{ad}$ are the averages for the rows, columns, and cells.

We will consider that for models (2) and (3) $\bar{K}$ and the sums Σ and Σ_δ are the same, we will present the sum of squared deviations K_{abc} from K of mo-

Table 1

The results of the dispersion decomposition of the sum (5)

A source of variability	The number of degrees of freedom	Sum of squared deviations
The main factor Z	$k = g - 1$	$\Sigma_\delta = nl\sum_{a=1}^{g}\left(\overline{K}_a - \overline{K}\right)^2$
An additional factor K_i	$k = l - 1$	$\Sigma_a = ng\sum_{d=1}^{l}\left(\overline{K}_d - \overline{K}\right)^2$
Interaction between Z and K_i	$k = (g-1)(l-1)$	$\Sigma_{\delta a} = n\sum_{a=1}^{a}\sum_{d=1}^{l}\left(\overline{K}_{\delta a} - \overline{K}_\delta - \overline{K}_a + \overline{K}\right)^2$
Remainder (in the middle of the cell)	$k = gl\,(n-1)$	$\Sigma_{ea} = \sum_{a=1}^{g}\sum_{d=1}^{l}\sum_{v=1}^{n}\left(K_{adv} - \overline{K}_{ad}\right)^2$
General	$k = N - 1$	$\Sigma = \sum_{a=1}^{g}\sum_{d=1}^{l}\sum_{v=1}^{n}\left(K_{adv} - \overline{K}\right)^2$

del (1) as a combination of sum (5), $i = \overline{1,2}$ with a residual sum Σ_e^f, which characterizes the influence of factors not considered in the model:

$$\Sigma = \Sigma_\delta + \Sigma_\alpha + \Sigma_\beta + \Sigma_{\delta\alpha} + \Sigma_{\delta\beta} + \Sigma_e^f. \quad (6)$$

The obtained ratio (6) allows simplifying model (1) to the following form:

$$K_{abcv} = \overline{K} + \delta_a + \alpha_b + \beta_c + (\delta\alpha)_{ab} + (\delta\beta)_{ac} + h_{abcv}^f, \quad (7)$$

where h_{abcv}^f is a remainder caused by two factors: grain temperature and illumination, respectively.

The calculation of the sums of expression (7) is carried out according to the equations of the sums of squares (Table 1), except for Σ_e^f, by replacing the factor $K_{(i)}$ with a specific additional factor affecting K_α, K_β. The sum Σ_e^f can be calculated from equations (8) or (9) [9]:

$$\begin{cases} \Sigma_e^f = \Sigma_{e\alpha} - \Sigma_\beta - \Sigma_{\delta\beta}; \\ \Sigma_e^f = \Sigma_{e\beta} - \Sigma_\alpha - \Sigma_{\delta\alpha}; \end{cases} \quad (8)$$

$$\Sigma_e^f = \Sigma_{e\alpha} + \Sigma_{e\beta} + 2\Sigma_\delta - 2\Sigma. \quad (9)$$

Table 2 presents the results of the variance analysis for the simplified model (7) [9].

Let us consider the properties of the simplified model and how exactly it affects the colorimetric control indicator. Model (1), used for the results of observations of the values of K_{abcv} of the information parameter K, occupies an intermediate position between the full model [9] and models (2) and (3). From (8), it becomes clear that the residual sum of the simplified model (9) is less than the residual sums of models (2) and (3). This may indicate the increased accuracy of the model compared to cross-classification models (2) and (3). Based on Table 2, we can conclude that the number of degrees of freedom of the residual sum Σ_e^f decreases with an increase in the number g of groups (according to the levels of the colorimetric control parameter Z) and the number l of subgroups (according to the level of additional factors affecting K_α and K_β). Let us rewrite k_e from the Table 2 as follows:

Table 2

Results of the variance analysis for the simplified model (7)

A source of variability	The number of degrees of freedom	Sum of squared deviations
The main factor Z	$k_Z = g - 1$	$\overline{\Sigma}_\delta = \Sigma_\delta / k_Z$
An additional factor K_α	$k_\alpha = l - 1$	$\overline{\Sigma}_\alpha = \Sigma_\alpha / k_\alpha$
An additional factor K_β	$k_\beta = l - 1$	$\overline{\Sigma}_\beta = \Sigma_\beta / k_\beta$
Interaction ZK_α	$k_{\delta\alpha} = (g-1)(l-1)$	$\overline{\Sigma}_{\delta\alpha} = \Sigma_{\delta\alpha} / k_{\delta\alpha}$
Interaction ZK_β	$k_{\delta\beta} = (g-1)(l-1)$	$\overline{\Sigma}_\delta = \Sigma_{\delta\beta} / k_{\delta\beta}$
Remainder	$k_e = N - g(2l-1)$	$\overline{\Sigma}_e = \Sigma_e^f / k_e$
General	$k = N - 1$	$\overline{\Sigma} = \Sigma / k$

Table 3

Estimated ratios

Factor influences	$\sigma^2_{\Delta K}$
An additional factor K_α	$\Sigma_\beta+\Sigma_{\delta\alpha}+\Sigma_{\delta\beta}+\Sigma_e^f / k_\beta+k_{\delta\alpha}+k_{\delta\beta}+k_e$
An additional factor K_β	$\Sigma_\alpha+\Sigma_{\delta\alpha}+\Sigma_{\delta\beta}+\Sigma_e^f / k_\alpha+k_{\delta\alpha}+k_{\delta\beta}+k_e$
K_α, K_β	$\Sigma_{\delta\alpha}+\Sigma_{\delta\beta}+\Sigma_e^f / k_{\delta\alpha}+k_{\delta\beta}+k_e$
K_α, K_β, ZK_α	$\Sigma_{\delta\beta}+\Sigma_e^f / k_{\delta\beta}+k_e$
K_α, K_β, ZK_β	$\Sigma_{\delta\alpha}+\Sigma_e^f / k_{\delta\alpha}+k_e$
K_α, K_β, ZK_α, ZK_β	$\overline{\Sigma}_e$

$$k_e = N - gl\left(2-\frac{g}{l}\right). \qquad (10)$$

It can be seen from formula (10) that the number of degrees of freedom is greater, with $g \cdot l = const$, the greater the ratio g/l. Based on the conclusion, it is possible to plan the number of groups and subgroups in the tables of output data of models (2) and (3). It is desirable to increase the number of g groups (subranges of measurements of the parameter K of the colorimetric control Z), and reduce the number of subgroups by reducing l to a minimum. This will make it possible to increase the number of degrees of freedom of the residual sum Σ_e^f. The advantage of the proposed simplified model is the possibility of simultaneously testing the hypothesis H_0: the influence of factors Z, K_α, K_β on the information indicator K. The latter is absent, therefore H_0: $\delta_h = \ldots \delta_g = 0$ [9]. The components of this basic hypothesis are as follows:

$$H_0^{\alpha}: \alpha_h = \ldots = \alpha_1 = 0; \quad H_0^{\beta}: \beta_h = \ldots = \beta_1 = 0;$$

$$H_0^{\delta\alpha}: (\delta\alpha)_h = \ldots = (\delta\alpha)_{gl} = 0; \; H_0^{\delta\beta}: (\delta\beta)_h = \ldots = (\delta\beta)_{gl} = 0.$$

Verification of the listed hypotheses by the ratio of the corresponding mean squares ($\overline{\Sigma}_\delta$, $\overline{\Sigma}_\alpha$, $\overline{\Sigma}_\beta$, $\overline{\Sigma}_{\delta\alpha}$, $\overline{\Sigma}_{\delta\beta}$) to the residual mean square $\overline{\Sigma}_e$ with a further comparison of the obtained F – statistics with the corresponding percentage points for F – distributions [9]. This advantage of the simplified model (7) makes it possible to estimate the amount of expected information about the levels of the colorimetric control parameter Z for the information indicator K, considering the levels of both influencing factors and their interactions [9]:

$$I = log\sqrt{1+\left(\frac{\sigma_K}{\sigma_{\Delta K}}\right)^2}, \qquad (11)$$

where $\sigma_K^2 = \overline{\Sigma}_\delta$, a $\sigma_{\Delta K}^2$ – is a function of the sum of squared deviations ($\overline{\Sigma}_\delta$, $\overline{\Sigma}_\alpha$, $\overline{\Sigma}_\beta$, $\overline{\Sigma}_{\delta\alpha}$, $\overline{\Sigma}_{\delta\beta}$, $\overline{\Sigma}_e^f$) in Table 1 and Table 2.

Table 3 summarizes the equation to calculate the $\sigma_{\Delta K}^2$ for the simplified model with various combinations of factors affecting the information indicator K.

In this way, the equations for estimating the reliability of statistical conclusions about the informational significance of colorimetric control indicators for the simplified cross-classification model were obtained.

Conclusions

1. An analysis was made, and a simplified cross-classification model was proposed, which considers the effects of the simultaneous interaction of two factors (moisture of raw materials, illumination) on the result of measuring the colorimetric control indicator (yellowness of wheat – *YPC*), and studied.

2. Expressions for estimating the reliability of statistical conclusions characterizing the informational significance of colorimetric control indicators for a simplified cross-classification model were obtained.

Дисперсійний аналіз у колориметричному контролі

С.А. Єфименко, І.В. Григоренко, С.М. Григоренко

Національний технічний університет "Харківський політехнічний інститут", вул. Кирпичова, 2, 61002, Харків, Україна
sefimenko64@gmail.com

Анотація

Розглянуто можливість використання дисперсійного аналізу для обробки даних колориметричного контролю та розробки спрощеної моделі перехресних класифікацій, що враховує ефекти одночасної взаємодії двох факторів (вологості сировини та освітленості) на результат вимірювання показника колориметричного контролю (жовтизна пшениці – YPC), та проведено її дослідження. Оскільки вміст жовтого пігменту YPC є одним із найважливіших показників якості зерна твердої пшениці, то саме він буде найбільш наочно визначатися методом колориметричного контролю. Дослідження показало обмеження на кількість рівнів параметра контролю та факторів, що впливають на результат. У ході дослідження були встановлені рівняння для оцінювання достовірності статистичних висновків щодо інформаційної значущості показників колориметричного контролю при використанні спрощеної моделі перехресної класифікації. Спрощена модель дала можливість оцінити кількість очікуваної інформації щодо рівнів параметра колориметричного контролю при обмеженнях на об'єм вимірювань. Доведено, що саме використання дисперсійного аналізу повною мірою дає можливість врахувати факторний вплив на результат колориметричного контролю, необхідність якого на сьогодні пов'язана зі зростанням вимог до якості пшеничного борошна та продуктів, що з нього виготовляються. Це стало особливо актуальним після законодавчої заборони на використання штучних барвників при виготовленні макаронних виробів у деяких країнах Європи, що посилило роль програм селекції твердої пшениці для покращення YPC. Зважаючи на вищесказане, стає актуальним завдання урахування факторного впливу на визначення рівня YPC за допомогою методів колориметрії, як одного з методів експрес-контролю, що швидко дає необхідне значення параметра при обмеженнях на час контролю, але не на його точність.

Ключові слова: дисперсійний аналіз; жовтизна пшениці; факторний вплив; щільність розподілу ймовірності; колориметрія.

References

1. Parada R., Royo C., Gadaleta A., Colasuonno P. et al. Phytoene synthase 1 (Psy-1) and lipoxygenase 1 (Lpx-1) Genes Influence on Semolina Yellowness in Wheat Mediterranean Germplasm. *International Journal of Molecular Sciences*, 2020, vol. 21(13):4669. doi: https://doi.org/10.3390/ijms21134669
2. Randhawa H.S., Asif M., Pozniak C., Clarke J.M. et al. Application of molecular markers to wheat breeding in Canada. *Plant Breeding*, 2013, vol. 132, issue 5, pp. 458–471. doi: https://doi.org/10.1111/pbr.12057
3. Kuldeep Kaur, Achla Sharma, Gurvinder Singh Mavi, Puja Srivastava et al. Biofortified wheat: Harnessing genetic diversity for improved nutritional quality to eradicate hidden hunger. *Crop Science*, 2022, vol. 62, issue 2, pp. 802–819. doi: https://doi.org/10.1002/csc2.20701
4. Hu J., Xiao G., Jiang P. et al. QTL detection for bread wheat processing quality in a nested association mapping population of semi-wild and domesticated wheat varieties. *BMC Plant Biology*, 2022, vol. 129. doi: https://doi.org/10.1186/s12870-022-03523-x
5. Chegdali Y., Ouabbou H., Essamadi A. et al. Distribution of alleles related to grain weight and quality in Moroccan and North American wheat landraces and cultivars. *Euphytica*, 2022, vol. 218 (123). doi: https://doi.org/10.1007/s10681-022-03078-w
6. Hare R. Agronomy of the durum wheats Kamilaroi, Yallaroi, Wollaroi and EGA Bellaroi. Available at: http://www.dpi.nsw.gov.au/__data/assets/pdf_file/0007/63646/Agronomy-of-thedurum-wheats-Primefact-140-final.pdf (accessed on 30.04.2023).
7. Yefymenko S., Hryhorenko I., Khoroshailo Yu., Hryhorenko S. Zastosuvannya kovariatsiynoho analizu dlya vyznachennya faktornoho vplyvu na parametr kontrolyu pry kolorymetrychnomu doslidzhenni [Applying covariance analysis to determine the factor influence on the control parameter in colorimetric study]. *Ukrainian Metrological Journal*, 2022, no. 3, pp. 49–55 (in Ukrainian). doi: https://doi.org/10.24027/2306-7039.3.2022.269783
8. Yefymenko S., Hryhorenko I., Khoroshilo Iu., Hryhorenko S., Petrovska I. Evaluation of Informativeness of Indicators in Colorimetric Control Using Discriminative Analysis Models. *Proceedings of XXXII International Scientific Symposium Metrology and Metrology Assurance (MMA)*, Sozopol, Bulgaria, 2022, pp. 1–4. doi: https://doi.org/10.1109/MMA55579.2022.9992712
9. Yefymenko S.A. Kolorymetrychnyy metod ta zasib dlya ekspres-kontrolyu yakosti zernovykh kultur: dys. dokt. filosofii [Colorimetric method and tool for express quality control of grain crops: PhD diss.]: 152. Kharkiv, 2022. 165 p. (in Ukrainian).