International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13802-13806

On the Theoretically Achievable Accuracy of a Selective Assessment of Grain Quality

Urii Igorevich Minkin[1], Aleksei Vladimirovich Panchenko[2] and Ivan Andreevich Konovalenko[3], Dmitri Valerevich Polevoy[4]

[1,2,3] JSC Cognitive, 60-letya Oktyabrya Avenue, 9, Moscow, 117312, Russia.
[4]FRC CSC RAS, 60-letya Oktyabrya Avenue, 9, Moscow, 117312, Russia.

Orcid: 0000-0002-5148-5795

Abstract

This article provides the analysis of the statistical properties of a random grain quality assessment for the best case. In addition, the distribution of the true unknown grain quality value is derived under the condition of a fixed sample and its main characteristics. The work is designed to select a sample size that can theoretically provide the required quality of the assessment, as well as to set algorithms for assessing grain quality and comparing their experimentally measured quality of work with theoretically the maximum achievable performance.

Keywords: precision agriculture, grain quality, statistical training.

INTRODUCTION

The work of agricultural enterprises can't be imagined without grain quality assessment technologies. The quality of grain is a multidimensional concept, which includes various characteristics. For example, in the paper [1] the problem of screening out dead, chalky, cracked and immature grains of rice is considered. The paper [2] provides an overview of sensors that provide information on moisture content and straw yield. And the paper [3] describes a system that estimates the fraction of cracked grain and impurities based on the photographic image of the grain. Evaluation of the above parameters of grain quality is made both automatically and manually by sorting grains into quality and substandard.

In all the cases, not the whole batch of grain is analyzed, but some small sample [4]. It is obvious that the estimation of the fraction of substandard grains in the batch, made on the basis of a random sample, is also random, and therefore limited in precision. In this paper, a quantitative analysis of the maximum possible accuracy of such an estimate is made. This will make it possible to compare the practically measured accuracy of the methods of estimating the fraction of poor-quality grain with the greatest possible accuracy [5].

In order to maximize the accuracy studied by us, first of all we need to make the assumption that we have a way to accurately determine whether the grain is qualitative or not. In practice, this would mean that we have a detector of substandard grains, whose false positive and false negative errors kind are zero.

Let us pass directly to the analysis of the greatest possible accuracy.

MATHEMATICAL MODEL

In this case, the general total of grains is a batch of grain, in which it is required to estimate the proportion of poor-quality grains. We will model an unknown fraction of poor-quality grains by a random variable P having an a priori distribution density $f_p(p)$ with support [0,1]: $P \sim f_p(p)$, $f_p(p) = 0$ for $p \notin [0,1]$. In the absence of a priori representations of P, a uniform a priori distribution should be used $f_p(p) = 1_{[0,1]}(p)$.

To assess the quality of grain from the batch, a random sample is taken in the amount of n of grains. Such a grain sample will be considered a representative sample from the general total of grains. Let us introduce Bernoulli random variables X_i, such that $X_i = 0$, if i grain turns out to be qualitative, $X_i = 1$ - otherwise.

Since the sample from the general total was taken randomly, the quantities X_i, are conditionally independent for each P and are equally distributed as follows

$$\mathbb{P}[X_i = 1|P = p] = 1 - p$$

$$\mathbb{P}[X_i = 1|P = p] = p = \mathbb{E}[X_i|P = p]$$

Let us introduce the conditional probability function for each sample element

$$f_{X_i|P}(x_i|p) = \mathbb{P}[X_i = x_i|P = p]$$

Then otherwise it can be written as follows

$$f_{X_i|P}(x_i|p) = (1-p)^{1-x_i}p^{x_i}$$

Due to the conditional independence of X_i, the conditional probability function of the entire sample will look like this

$$f_{X|P}(x_i|p) = \Pi_{i=1}^{n} f_i(x_i|p) = \Pi_{i=1}^{n}((1-p)^{1-x_i}p^{x_i}) = (\Pi_{i=1}^{n}(1-p)^{1-x_1})(\Pi_{i=1}^{n}p^{x_1}) = (1-p)^{n-1^T x}p^{1^T x}$$

where $1^T = |1 \quad 1 \dots 1|$.

Now we find the density of the conditional distribution of the proportion of poor-quality grains under the condition of sampling

$$f_{P|X}(p|x) = f_{X,P}(x,p)/f_X(x) = f_{X,P}(x,p)/\int_{\mathbb{R}} f_{X,P}(x,t)d = f_{X,P}(x|p)f_P(p)/\int_{\mathbb{R}} f_{X|P}(x|t)f_P(t)dt = f_{X,P}(x|p)f_P(p)/\int_0^1 f_{X|P}(x|t)f_P(t)dt$$

MAXIMUM LIKELIHOOD METHOD

Let us apply the maximum likelihood method

$$\hat{P} = \arg\max_{p\in[0,1]}\left[f_{P|X}(p|X)\right] = \arg\max_{p\in[0,1]}\left[f_{X|P}(X|p)f_p(p)/\int_0^1 f_{X|P}(X|t)f_P(t)dt\right]$$

As an example, take $f_P(p) = 1_{[0,1]}(p)$. Then

$$\hat{P} = \arg\max_{p\in[0,1]}\left[f_{P|X}(p|X)\right] = \arg\max_{p\in[0,1]} f_{X|P}(X|p)f_p(p)/\int_0^1 f_{X|P}(X|t)f_P(t)dt = \arg\max_{p\in[0,1]}\left[f_{X|P}(X|p)/\int_0^1 f_{X|P}(X|t)dt\right] = \arg\max_{p\in[0,1]}\left[f_{X|P}(X|p)\right] = \arg\max_{p\in[0,1]}\left[(1-p)^{n-1^T X}p^{1^T X}\right] = \arg\max_{p\in[0,1]}\left[e^{(n-1^T X)\log(1-p)}e^{1^T X\log p}\right] = \arg\max_{p\in[0,1]}\left[(n-1^T X)\log(1-p) + 1^T\log p\right]$$

To solve this optimization problem, we find the zero derivative of the optimized function

$$\frac{d}{dP}\left[(n-1^T X)\log(1-\hat{P}) + 1^T\log\hat{P}\right] = 0$$

$$\frac{n-1^T X}{1-\hat{P}} = \frac{1^T X}{\hat{P}}$$

$$n\hat{P} - 1^T X\hat{P} = 1^T X - 1^T X\hat{P}$$

$$n\hat{P} = 1^T X$$

$$\hat{P} = \frac{1}{n}1^T X$$

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Thus, the maximum likelihood estimate for the fraction of poor-quality grain in the general total is the fraction of poor-quality grain in the sample. As it will be shown below, this estimate is unbiassed and, in addition, has the smallest variance among all the unbiassed estimates, i.e. is optimal.

Analysis of the accuracy of the estimate. Evaluation distribution

So, the optimal estimate looks like this

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Let's find out how it is distributed if the fraction of poor-quality grains in the general total P is actually equal to the true fraction p. First, we find the error mean square X_i for a fixed $P = p$

$$\mathbb{D}[X_i|P=p] = \mathbb{E}[(X_i - \mathbb{E}[X_i|P=p])^2|P=p] = \mathbb{E}[X_i^2|P=p] - (\mathbb{E}[X_i|P=p])^2 = \mathbb{E}[X_i^2|P=p] - p^2 = \mathbb{E}[X_i|P=p] - p^2 = p - p^2 = p(1-p)$$

The sum of identically distributed independent Bernoulli quantities $Y = n\hat{P} = \sum_{i=1}^{n} X_i$ is distributed binomially, so we can find the following

The distribution density $[Y|P=p]$

$$f_{Y|P}(y|p) = \mathbb{P}[Y=y|P=p] = C_n^y p^y(1-p)^{n-y}.$$

Mode $[Y|P=p]$

$$mode[Y|P=p] = arg\max_{y\in[0,1,2,\dots,n]}\left[f_{Y|P}(y|P=p)\right] = (n+1)p$$

Average $[Y|P=p]$

$$\mathbb{E}[Y|P=p] = \mathbb{E}[\sum_{i=1}^{n} X_i|P=p] = \sum_{i=1}^{n}\mathbb{E}[X_i|P=p] = \sum_{i=1}^{n} p = np.$$

Variance $[Y|P=p]$

$$\mathbb{D}[Y|P=p] = \mathbb{D}\left[\sum_{i=1}^{n} X_i|P=p\right] = \sum_{i=1}^{n}\mathbb{D}[X_i|P=p] = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

Finally, on this basis, we clarify the distribution properties of the maximum likelihood estimate $\left[\hat{P}\middle|P=p\right]$

Density of distribution $\left[\hat{P}\middle|P=p\right]$

$$f_{\hat{P}|P}(\tau|p) = \mathbb{P}\left[\hat{P}=\tau\middle|P=p\right] = \mathbb{P}\left[\frac{1}{n}Y=\tau|P=p\right] = \mathbb{P}[Y=n\tau|P=p] = f_{Y|P}(n\tau|p) = C_n^{n\tau}p^{n\tau}(1-p)^{n-n\tau} = C_n^{n\tau}(p^\tau(1-p)^{1-\tau})^n$$

Mode $\left[\hat{P}\middle|P=p\right]$

$$mode\left[\hat{P}\middle|P=p\right] = arg \max_{\tau\in\left[0,\frac{1}{n},\frac{2}{n}\ldots,1\right]}\left[f_{\hat{P}|P}(\tau|P=p)\right]$$
$$= mode\left[\frac{Y}{n}|P=p\right] = \frac{1}{n}mode[Y|P=p]$$
$$= \frac{1}{n}[(n+1)p]$$

Average $\left[\hat{P}\middle|P=p\right]$

$$\mathbb{E}\left[\hat{P}|P=p\right] = \mathbb{E}\left[\frac{Y}{n}|P=p\right] = \frac{1}{n}\,\mathbb{E}[Y|P=p] = \frac{n}{n}p = p$$

thus, the maximum likelihood estimate is unbiased $\mathbb{E}\left[\hat{P}\middle|P=p\right] = p$.

Variance $\hat{P}\middle|P=p$

$$\mathbb{D}\left[\hat{P}|P=p\right] = \mathbb{D}\left[\frac{Y}{n}|P=p\right] = \frac{1}{n^2}\mathbb{D}[Y|P=p]$$
$$= \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Mean square error $\sigma_{\hat{P}|P=p}$

$$\sigma_{\hat{P}|P=p} = \sqrt{\mathbb{D}\left[\hat{P}\middle|P=p\right]} = \sqrt{\frac{p(1-p)}{n}}$$

The figure 1 shows the examples of the distributions $f_{\hat{P}|P}(\tau|p)$

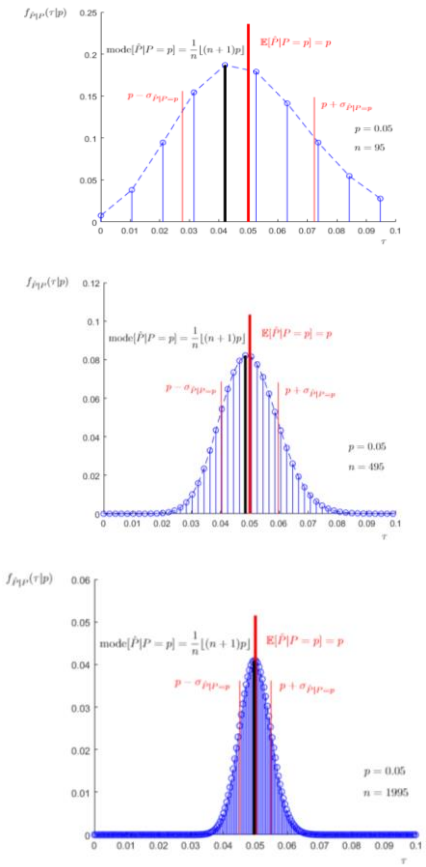

Figure 1: Maximum likelihood estimate distributions for different n (a (n=95), b (n=495), c (n=1995))

The distribution of the unknow true fraction

In the previous section, the statistical properties of the best estimate of the fraction of poor-quality grains were derived for a fixed true value of the fraction of poor-quality grains. This result is important for the theory, but it is not applicable in practice, since in practice the value of the fraction of poor-quality grains is unknown. Therefore, in this section, based on the previous section, we will elucidate the statistical properties of the true unknown fraction of poor-quality grains with the value of the estimate obtained. In other words, we will study the properties of a posteriori distribution of the true fraction of poor-quality grains.

Here we continue to work in the framework of the example, when the a priori distribution of the true fraction is uniform

$$f_P(p) = 1_{[0,1]}(p).$$

So, the problem is to find the a posteriori distribution of the unknown true fraction of a poor-quality grain $f_{P|\hat{P}}(p|\tau)$. We do this according to the Bayes theorem

$$f_{P|\hat{P}}(p|\tau) = \frac{f_{P,\hat{P}}(p,\tau)}{f_{\hat{P}}(\tau)} = \frac{f_{P,\hat{P}}(p,\tau)}{\int_{\mathbb{R}} f_{P,\hat{P}}(t,\tau)dt} = \frac{f_{\hat{P}|P}(\tau|p)f_P(p)}{\int_{\mathbb{R}} f_{\hat{P}|P}(\tau|t)f_P(p)dt} =$$
$$\frac{f_{\hat{P}|P}(\tau|p)}{\int_0^1 f_{\hat{P}|P}(\tau|p)dt} = \frac{C_n^{n\tau}(p^\tau(1-p)^{1-\tau})^n}{\int_0^1 C_n^{n\tau}(t^\tau(1-t)^{1-\tau})^n dt} = \frac{(p^\tau(1-p)^{1-\tau})^n}{\int_0^1 (t^\tau(1-t)^{1-\tau})^n dt} =$$
$$\frac{p^{n\tau}(1-p)^{n(1-\tau)}}{\int_0^1 t^{\tau}(1-t)^{n(1-\tau)}dt}.$$

Let us introduce the notation $\alpha = 1+n\tau$, $\beta = 1+n(1-\tau)$

$$f_{P|\hat{P}}(p|\tau) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt} = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)},$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$ - beta function, $\Gamma(\alpha)$ - gamma function.

Now we have come to the conclusion that $f_{P|\hat{P}}(p|\tau)$ is the beta density, i.e. a posteriori distribution of the true unknown fraction of grains is a beta distribution

$$\left[P|\hat{P}=\tau\right] \sim Beta(\alpha,\beta).$$

The most important characteristics of this distribution is

The distribution density $\left[P|\hat{P}=\tau\right]$, as already known, is equal to

$$f_{P|\hat{P}}(p|\tau) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}, \quad \alpha = 1+n\tau, \beta = 1+n(1-\tau).$$

Mode $\left[P|\hat{P}=\tau\right]$

$$mode\left[P|\hat{P}=\tau\right] = \arg\max_{p\in[0,1]}\left[f_{P|\hat{P}}\left(\hat{P}=\tau\right)\right] = \frac{\alpha-1}{\alpha+\beta-2} =$$
$$\frac{n\tau}{n\tau+n(1-\tau)} = \tau.$$

Thus, we see that the a posteriori density of the distribution of the true fraction of poor-quality grain reaches its maximum at the point corresponding to the estimate obtained.

Mean $\left[P|\hat{P}=\tau\right]$

$$\mathbb{E}\left[P|\hat{P}=\tau\right]=\frac{\alpha}{\alpha+\beta}=\frac{1+n\tau}{2+n}.$$

Variation $\left[P|\hat{P}=\tau\right]$

$$\mathbb{D}\left[P|\hat{P}=\tau\right]=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}=\frac{(1+n\tau)(1+n(1-\tau))}{(2+n)^2(3+n)}.$$

Hence the standard deviation of $\sigma_{P|\hat{P}=\tau}$ is

$$\sigma_{P|\hat{P}=\tau}=\sqrt{\mathbb{D}\left[P|\hat{P}=\tau\right]}=\sqrt{\frac{(1+n\tau)(1+n(1-\tau))}{(2+n)^2(3+n)}}.$$

Calculate the mean square deviation of the true fraction from its estimate

$$MSE\left(P|\hat{P}=\tau\right)=\mathbb{E}\left[\left(P-\hat{P}\right)^2|\hat{P}=\tau\right]=\mathbb{E}\left[\left(P-\hat{P}\right)^2|\hat{P}=\tau\right].$$

Lets use $\mu=\mathbb{E}\left[P|\hat{P}=\tau\right]=\frac{1+n\tau}{2+n}$

$$\begin{aligned}MSE\left(P|\hat{P}=\tau\right)&=\mathbb{E}\left[\left((P-\mu)-(\tau-\mu)\right)^2|\hat{P}=\tau\right]=\\&=\mathbb{E}\left[(P-\mu)^2-2(P-\mu)(\tau-\mu)\right.\\&\left.+(\tau-\mu)^2|\hat{P}=\tau\right]=\\&=\mathbb{E}\left[(P-\mu)^2|\hat{P}=\tau\right]\\&-2\mathbb{E}\left[(P-\mu)(\tau-\mu)+(\tau-\mu)^2|\hat{P}=\tau\right]=\\&=\mathbb{D}\left[P|\hat{P}=\tau\right]\\&-2(\tau-\mu)\mathbb{E}\left[(P-\mu)|\hat{P}=\tau\right]+(\tau-\mu)^2=\\&=\mathbb{D}\left[P|\hat{P}=\tau\right]+(\tau-\mu)^2=\\&=\frac{(1+n\tau)\left(1+n(1-\tau)\right)}{(2+n)^2(3+n)}\\&+\left(\frac{1+n\tau}{2+n}-\tau\right)^2=\frac{2+\tau(1-\tau)(n-6)}{(2+n)^2(3+n)}\end{aligned}$$

The MSE standard deviation $\sigma_{MSEP|\hat{P}=\tau}$ is equal to

$$\sigma_{MSEP|\hat{P}=\tau}=\sqrt{MSEP|\hat{P}=\tau}=\sqrt{\frac{2+\tau(1-\tau)(n-6)}{(2+n)^2(3+n)}}.$$

This value that primarily characterizes the theoretically achievable accuracy of estimating the fraction of poor-quality grain. The figure 2 shows the examples of the distributions $f_{P|\hat{P}}(p|\tau)$.

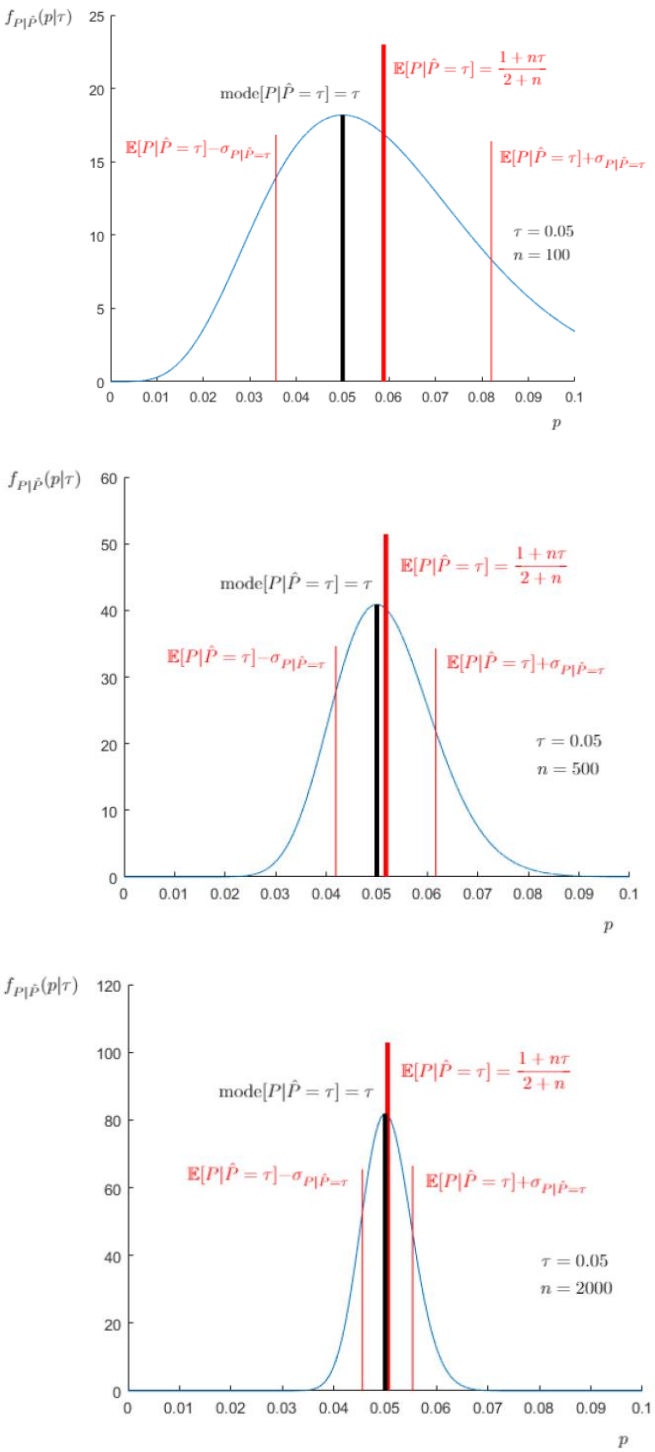

Figure 2. Distributions of the true unknown fraction of poor-quality grain for different n (a (n=100), b (n=500), c (n=2000))

CONCLUSION

The distribution of the best selective evaluation of grain quality and its main characteristics, depending on the true grain quality, was derived in the paper. In addition, the distribution parameters of the true unknown grain quality value were obtained under the condition of a fixed sample.

This work allows, firstly, to select a sample size that can theoretically provide the required quality of the assessment; secondly, it helps in setting up grain quality estimation algorithms; and thirdly, it allows to compare their experimental measured performance with the theoretically maximum achievable performance.

ACKNOWLEDGEMENTS

This work was financially supported by the Ministry of Education and Science of Russian Federation (Agreement No. 14.579.21.0119, Applied Research and Experimental Development individual identifier RFMEFI57915X0119)

REFERENCES

[1] Y-N Wan, C-M Lin, and J-F Chiou. Rice quality classification using an automatic grain quality inspection system. Transactions of the ASAE, 45(2):379, 2002.

[2] P Reyns, B Missotten, H Ramon, and J De Baerdemaeker. A review of combine sensors for precision farming. Precision Agriculture, 3(2):169_182, 2002.

[3] Mathias Escher M.Sc. and Dipl.-Ing. Thilo Krause. Grain quality camera 4th International Conference on Machine Control and Guidance, pages 8-15, 2014.

[4] GOST 13586.3. Grain. Regulations of admission and methods of sampling. 2015

[5] A. V. Panchenko, N. V. Reshetnyak, A. M. Samohin, V. V. Postnikov, "The automated testing facility based on machine vision for optimizing grain quality control technology", Proc. SPIE 10341, Ninth International Conference on Machine Vision (ICMV 2016), 103411Q (17 March 2017).