

ОБЕМ НА ПИЛОТНАТА ПРОБА В СТАТИСТИЧЕСКИЯ АНАЛИЗ

К. Куманов

Институт по овощарство, Остромила 12, Пловдив 4004

THE PILOT-SAMPLE SIZE IN STATISTICAL ANALYSIS

K. Koumanov

Fruitgrowing Institute, 4004 Plovdiv, Bulgaria

РЕЗЮМЕ

В резултат на проведени анализи и математически преобразувания са изведени емпирични зависимости на обема на пробата N от отношението $(CV/\Delta_{p,k})^2$, където CV е вариационният коефициент на пилотна проба. Предмет на настоящата публикация е метод за определяне обема на пилотната проба в съответствие с предварително зададено вариране на CV . В направения анализ са използвани данни за масата на плодовете при малина, череша и ябълка; на ябълката са анализирани също диаметърът и височината на плодовете. Всеки плод е претеглен/измерен отделно. Формирани са седем редици от по 100 измервания на малиновите плодове, три редици от по 400 измервания за черешовите плодове и девет редици от по 50 измервания за ябълковите плодове. Чрез аналогия на т.н. "bootstrap methods" във всяка редица от измервания при всеки обем на извадката от съответно два, три, четири плода и т.н. е определено варирането на коефициента на вариация (т.е. CV на CV). Получени са зависимости на CV на CV от обема на пилотната проба, които са подложени на корелационен и регресионен

SUMMARY

Series of analyses and mathematical transformations resulted in deriving of empirical relationships between the sample size and the term $(CV/\Delta_{p,k})^2$, where CV is the variation coefficient of a pilot sample.

The present paper objective is a method for evaluating the pilot-sample size according to a predetermined variation of CV .

The analysis was based on data of the mass of raspberry, cherry and apple fruits; with apple, subject of analysis were also the diameter and the height of the fruits. Each fruit was weighed/measured separately.

Experimental data formed seven series of 100 raspberry fruit measurements each, three series of 400 cherry fruit measurements each, and nine series of 50 apple fruit measurements each.

Using an analogy of the bootstrap method, the variation of the variation coefficient (CV of CV) was evaluated in all series, for samples of respectively two, three, four fruits and so on.

This resulted in relationships between CV of CV and the pilot-sample

анализ. Установена е много силна корелация между обема на пилотната проба и CV на CV на редицата от измервания като зависимостта може да бъде изразена чрез нелинейното трипараметрично уравнение на Hill (1910). Доказано е, че при всички експерименти в условията на нормално разпределение, независимо от изучавания обект, обемът на пилотната проба може да бъде изчислен със зададена точност (стойност на CV на CV) чрез едно нелинейно непараметрично уравнение.

Ключови думи: извадка, генерална съвкупност, коефициент на вариация, bootstrap methods

УВОД

Обемът на пробата е важен за представителността на научните изследвания и валидността на направените заключения. За да представя правилно генералната съвкупност, пробата трябва да бъде достатъчно голяма. От друга страна увеличаването на обема на пробата води до по-голям разход на труд, време и като цяло оскъпява изследването. Проблемът има и етична страна, доколкото може да бъде свързан с изпитването на агресивни въздействия и медикаменти върху хора и животни. За подобряване методиката на експерименталната работа и привеждането ѝ в съответствие с нормите на статистиката, през 2006 г. в Института по овощарство – Пловдив, започна изследване върху редица

size, which were subjected to correlation and regression analyses. There was very strong correlation found between the pilot-sample size and the CV of CV, and the relationship was very well fitted by the nonlinear three-parametric equation of Hill (1910). It was proven that, under normal distribution, the pilot-sample size might be evaluated with a predetermined precision (CV of CV value) through a nonlinear nonparametric equation. That equation may be used in all experiments under normal distribution, irrespective of the studied subject.

Key words: sampling, general population, coefficient of variation, bootstrap methods

INTRODUCTION

The size of a sample is important for the representativeness of the research and the validity of the drawn conclusions.

The sample has to be large enough in order to characterize in a proper way the general population.

On the other hand, the increase in the sample size is associated with increased labor and time consumption. Moreover, it generally raises the cost of the investigation.

The problem has also ethical aspects, as far as it may be related to testing of aggressive treatments and drugs on humans or animals. In order to improve the methodology of the experiments and to set it in accordance with the standards of the statistical

биометрични показатели, включително масата и размерите на плодовете. Първите публикувани резултати (Koumanov and Kolev, 2007; 2008) са за минималния брой плодове, необходими за определяне средната маса на малиновите, черешовите и ябълковите плодове, както и на средната височина и диаметър на ябълковите плодове. Доказано е (Koumanov and Kolev, 2009), че при нормално разпределение обемът на пробата N може лесно и с голяма точност да бъде изчислен чрез линейно уравнение от вида:

$$N = a \left(\frac{CV}{\Delta} \right)^2 + b; \quad (1)$$

като параметрите a и b зависят от избраната доверителна вероятност p , а CV е коефициентът на вариация на пилотна проба с относително малък обем (Митков и Минков, 1989).

Както е известно, CV на всяка една извадка е толкова близко до CV на генералната съвкупност, колкото по-голяма е извадката. При това варирането на CV е най-голямо при минимален обем на пробата и намалява с нарастването на размера y . Варирането зависи обаче и от мястото на отделните

analysis, an investigation on some biometrical characteristics, including the mass and the size of the fruit, has been started and carried out since 2006 in the Fruitgrowing Institute in Plovdiv. The first published results (Koumanov and Kolev, 2007; 2008) concerned the minimum fruit number needed to evaluate the average mass of one raspberry, cherry or apple fruit, as well as the average diameter and the height of apple fruit. It was proven that, under normal distribution, the sample size could be evaluated using a linear equation of the type:

$$N = a \left(\frac{CV}{\Delta} \right)^2 + b; \quad (1)$$

the parameters a and b being related to a predetermined confidence probability p ; CV is the variation coefficient of a pilot sample with relatively small size (Mitkov and Minkov, 1989).

As it is well known, the sample CV approaches the population CV as the sample size increases, the sample CV variation being reciprocal to the sample size.

However, the CV variation depends also on the position of the individual measurements in the series.

измервания при случайното им подреждане в общата редица. Възможно е при малък брой измервания да се получи нисък вариационен коефициент, ако в началото се случат измервания с близки стойности. От друга страна CV може да се повиши при увеличаване броя на измерванията и включване на членове със значително отклоняващи се стойности. Доколкото CV е единствената променлива в уравнение (1), обемът на пилотната проба се оказва решаващ за точността на решението.

Предмет на настоящата публикация е метод за определяне обема на пилотната проба в съответствие с предварително зададено вариране на CV или с други думи в съответствие с предварително зададен CV на CV на редицата от измервания.

МАТЕРИАЛ И МЕТОДИ

В направения анализ са използвани данни за масата на плодовете при малина – седем проби от по 100 случайно избрани плода от сорта "Люлин", череша – три проби с обем от по 400 случайно подбрани плода от сортовете "Хъдзън", "Налина", и "Съмит", които са съответно с дребни, средни и едри плодове, и ябълка – три проби с обем от по 50 плода от сортовете "Карастоянка", "Флорина" и Джърси ред", съответно с

Low number of measurements may yield low coefficient of variation, too, if the values of the first measurements in a series are close to each other.

On the other side, a larger number of measurements may increase the CV values if significantly deviating values were included.

As far as CV is the only variable in eq. (1), the pilot-sample size is critical for the precision of the solution.

The present paper objective is a method for evaluating the pilot-sample size according to a predetermined variation of CV or, in other words, according to a predetermined CV of CV of the measurement series.

MATERIAL AND METHODS

Present analysis was based on data of the fruit mass of raspberry – seven samples of 100 randomly collected fruits each of the 'Lyulin' cultivar; cherry – three samples of 400 randomly collected fruits each of 'Hudson', 'Nalina' and 'Summit' cultivars, bearing small, medium and large fruit, respectively; and apple – three samples of 50 fruit each of 'Karastoyanka', 'Florina' and 'Jersey Red' cultivars, respectively with small, middle and large fruit.

дребни, средни и едри плодове. При ябълката същите проби са анализирани по отношение на диаметъра и височината на плодовете. Всеки плод от тези проби е претеглен/измерен отделно. Така са формирани седем редици от по 100 измервания на малиновите плодове, три редици от по 400 измервания за черешовите плодове и девет редици от по 50 измервания за ябълковите плодове.

Във всяка редица от измервания коефициентът на вариация е определен за извадки от съответно два плода, три плода, четири плода и т.н. до 100, 400 или 50 плода в зависимост от дължината на дадената редица. Това е направено 70 пъти като преди всяко определяне стойностите на измерванията в редицата са пренареждани на случаен принцип, аналогия на т.н. "bootstrap methods" (Efron, 1979; 1982), с помощта на онлайн приложение за генериране на случайни числа (<http://www.random.org>). Във всяка редица от измервания при всеки обем на извадката от съответно два, три, четири плода и т.н. е определено средното аритметично и варирането на коефициента на вариация (т.е. CV на CV на редицата от измервания) от получените при пренареждането 70 стойности. Получени са

With apple, subject of analysis were also the diameter and the height of the fruits.

Each fruit of those samples was weighed/measured separately. Finally, the analysis comprised seven series of 100 measurements of raspberry fruits each, three series of 400 measurements of cherry fruits each, and nine series of 50 measurements of apple fruits each.

In each series of measurements, the coefficient of variation was evaluated for samples of respectively two fruits, three fruits, four fruits and so on up to 100, 400, and 50 fruits, according to the size of each series.

That was done 70 times, the measurements in the series being randomly rearranged (resampled) before each evaluation, an analogy of the bootstrap methods (Efron, 1979; 1982), using an on-line application for random number generation (<http://www.random.org>). In each series of measurements and for each sample size of respectively two, three, four fruits and so on, the average CV and the variation coefficient of CV (i.e. CV of CV) were calculated out of the 70 values obtained after resampling.

The obtained relationships between CV of CV and the number

зависимости на CV на CV от броя на измерванията, т.е. от обема на пилотната проба, които са подложени на корелационен и регресионен анализ. За целта е използвано компютърното приложение за анализ на данни и графични изображения OriginPro 8.5.0 SR1 на OriginLab Corporation.

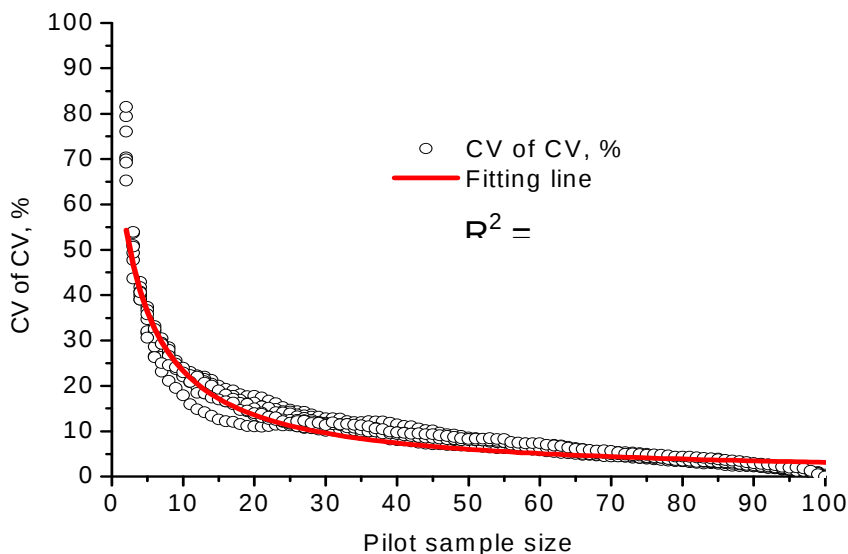
of measurements (the sample size) were subjected to correlation and regression analyses using OriginPro 8.5.0 SR1 data analysis and graphing software of the OriginLab Corporation.

РЕЗУЛТАТИ И ОБСЪЖДАНЕ

Зависимостите на CV на CV от обема на пробата (броя измервания) са онагледени на Фигури 1, 2 и 3 съответно за малиновите, черешовите и ябълковите плодове.

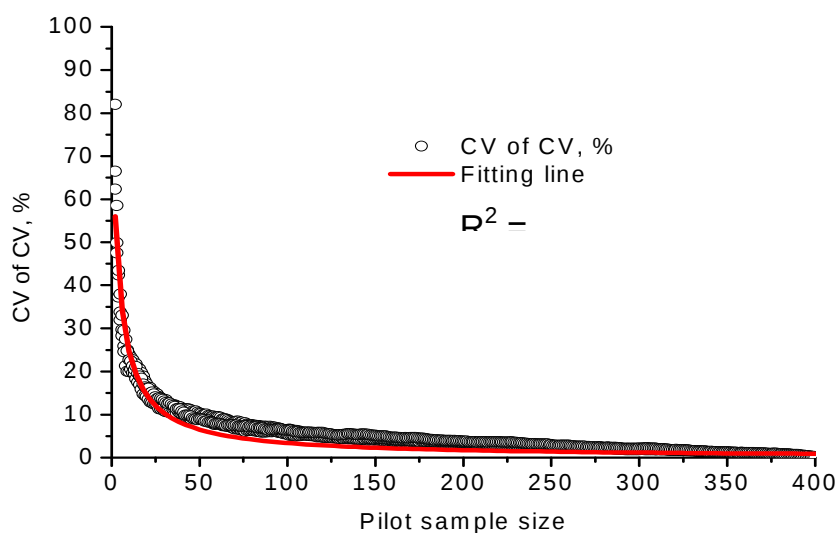
RESULTS AND DISCUSSION

The relationships between CV of CV and the sample size (number of measurements) are shown on Figures. 1, 2 and 3 for raspberry, cherry and apple fruit, respectively.



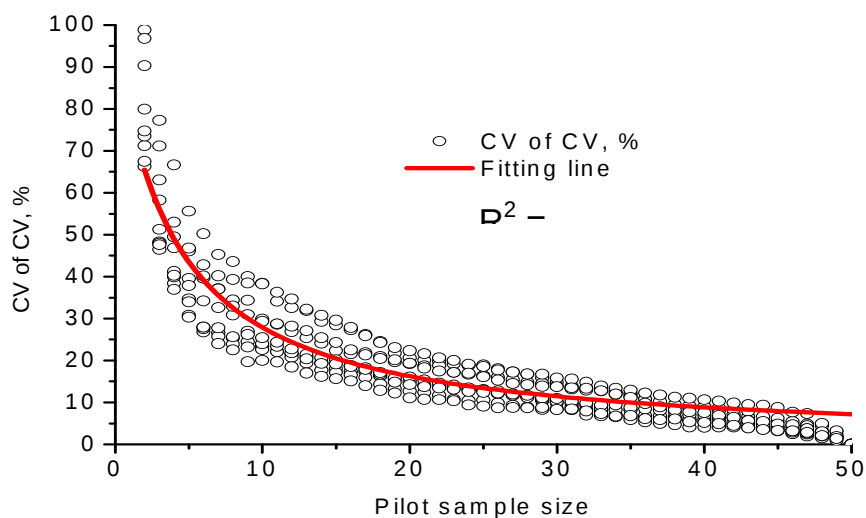
Фиг. 1. Размер на пилотната проба в зависимост от CV на CV по отношение масата на малиновите плодове

Fig. 1. Pilot-sample size in relation to CV of CV regarding the mass of raspberry fruits



Фиг. 2. Размер на пилотната проба в зависимост от CV на CV по отношение масата на черешовите плодове

Fig. 2. Pilot-sample size in relation to CV of CV regarding the mass of cherry fruits

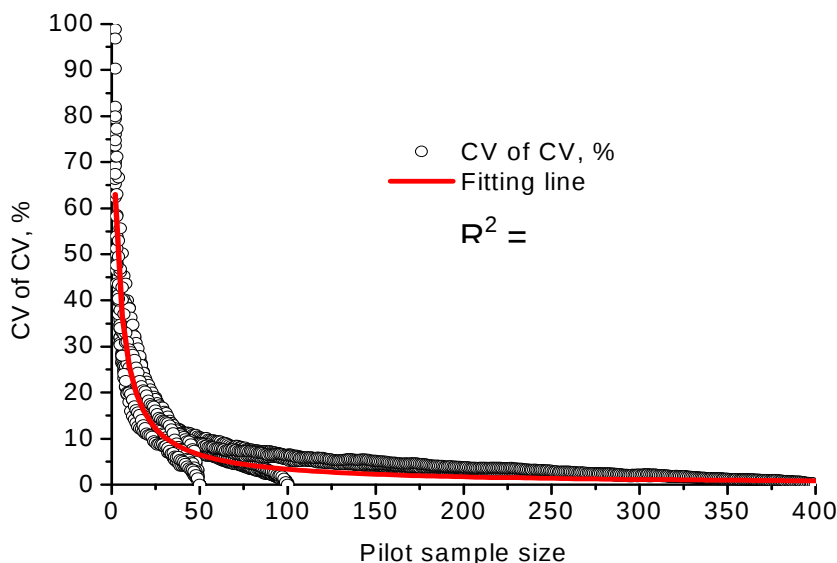


Фиг. 3. Размер на пилотната проба в зависимост от CV на CV по отношение масата, диаметра и височината на ябълковите плодове

Fig. 3. Pilot-sample size in relation to CV of CV regarding the mass, the diameter and the height of apple fruits

Получената въз основа на всички опитни данни зависимост (2350 двойки числа общо за малиновите, черешовите и ябълковите плодове) е показана на Фигура 4.

The based on all experimental data together relationship (2350 pairs of values, total for raspberry, cherry and apple fruit) is shown on Figure 4.



Фиг. 4. Размер на пилотната проба в зависимост от CV на CV по отношение масата на малиновите, черешовите и ябълковите плодове, както и диаметъра и височината на ябълковите плодове (2350 двойки стойности)

Fig. 4. Pilot-sample size in relation to CV of CV regarding the mass of raspberry, cherry and apple fruits, as well as the diameter and the height of apple fruits (2350 pairs of values)

На всички фигури са показани и регресионните линии, изравняващи емпиричните точки, както са дадени и съответните коефициенти на детерминация, R^2 . Всички зависимости могат да бъдат изразени чрез нелинейното трипараметрично уравнение на

The regression lines, fitting the empirical data, and the corresponding coefficient of determination, R^2 , are shown on the figures, as well. All relationships may be expressed by the nonlinear three-parametric equation of Hill (Hill, 1910; Coutelle et al., 2008):

Hill (Hill, 1910; Coutelle et al., 2008):

$$y = V \frac{x^\alpha}{k^\alpha + x^\alpha};$$

(2)

където y е CV на CV на съответната редица от измервания, изразен в проценти; x – обемът на пилотната проба (броят измервания); а V , k и α са параметри на уравнението.

Стойностите на параметрите V , k и α , както и на коефициента на детерминация, R^2 , за всяка една от четирите зависимости са дадени в Таблица 1.

$$y = V \frac{x^\alpha}{k^\alpha + x^\alpha}; \quad (2)$$

where y is CV of CV of the corresponding sample size in percent; x is the pilot-sample size (number of measurements); while V , k , and α are equation parameters.

The values of the parameters V , k , and α , and the determination coefficients, R^2 , are given in Table 1 for each of the four relationships.

Таблица 1. Стойности на параметрите на уравнение (2) и на коефициента на детерминация, R^2 , при малина, череша, ябълка и общо за трите култури
Table 1. Values of the parameters of equation (2) and of the coefficient of determination, R^2 , for raspberry, cherry, apple and for all together

Култура Crop	V	k	α	R^2
Малина Raspberry	81.510	4.000	- 1	0.93
Череша Cherry	82.026	4.309	- 1	0.89
Ябълка Apple	98.816	3.933	- 1	0.87
Общо All together	98.816	3.520	- 1	0.92

Стойности за обема на пробата, изчислени по уравнение (2) за характерни стойности на CV на CV, са дадени в Таблица 2 по отношение на малината, черешата и ябълката, както и обобщено за всички опитни данни.

Получените резултати

Table 2 represents sample sizes, obtained from eq. (2) for common values of CV of CV for raspberry, cherry and apple, respectively, as well as for all experimental data together.

The results show an extremely strong correlation

показват изключително силна корелация между обема на пробата и CV на CV.

between the sample size and the CV of CV.

Таблица 2 Обем на пилотната проба при различни стойности на CV от CV
Table 2 Size of the pilot sample at different values of CV of CV

CV of CV%	Обем на пилотната проба, бр. / Pilot-sample size, pcs			
	Малина Raspberry	Череша Cherry	Ябълка Apple	Общо All together
1	322	349	385	344
2	159	172	190	170
3	105	114	126	112
4	78	84	93	83
5	61	66	74	66
6	50	55	61	54
7	43	46	52	46
8	37	40	45	40
9	32	35	39	35
10	29	31	35	31
15	18	19	22	20
20	12	13	15	14
25	9	10	12	10
30	7	7	9	8
35	5	6	7	6
40	4	5	6	5
45	3	4	5	4
50	3	3	4	3

Очевидно размерът на плодовете, съответно овощният вид и сортът, както и измерваната величина (маса или дължина) не са от съществено значение за обема на пробата. В такъв случай при всички експерименти, независимо от обекта на изследване, обемът на пилотната проба може да бъде изчислен със зададена точност (стойност на CV на CV на съответната редица от измервания) чрез следното нелинейно непараметрично уравнение:

Apparently, the fruit size (respectively the fruit species and the cultivar) as well as the measured quantity (mass or length) have no significant impact on the sample size.

Hence, in all experiments under normal distribution, irrespective of the studied subject, the size of the pilot sample may be estimated with a predefined precision (CV of CV value of the corresponding series of measurements) from the following nonlinear nonparametric equation:

$$n = \frac{3.52(98.816 - y)}{y};$$

(3)

където n е обемът на пилотната проба, а y е зададената стойност на CV на CV (%).

$$n = \frac{3.52(98.816 - y)}{y};$$

(3)

where n is the pilot-sample size and y is a predefined value of CV of CV (%).

ИЗВОДИ

Налице е много силна корелация между обема на пилотната проба и CV на CV на редицата от измервания като зависимостта може да бъде изразена с висока точност чрез нелинейното трипараметрично уравнение на Hill (1910).

Размерът на плодовете, съответно овощният вид и сортът, както и измерваната величина (маса или дължина) не са от съществено значение за обема на пробата.

При всички експерименти в условията на нормално разпределение, независимо от изследвания обект, обемът на пилотната проба може да бъде изчислен със зададена точност (стойност на CV на CV на редицата от измервания) чрез едно нелинейно непараметрично уравнение.

CONCLUSIONS

There is very strong correlation between the pilot-sample size and the CV of CV of the series of measurements. The relationship may be expressed by the nonlinear three-parametric equation of Hill (1910).

Fruit size (respectively the fruit species and the cultivar) as well as the measured quantity (mass or length) have no significant impact on the sample size.

In all experiments under normal distribution, irrespective of the studied subject, the size of the pilot sample may be evaluated with a predefined precision (CV of CV value of the corresponding series of measurements) through a nonlinear nonparametric equation.

ЛИТЕРАТУРА / REFERENCES

1. **Митков А. и Д. Минков.** 1989. Статистически методи за изследване и оптимизиране на селскостопанската техника – I част. Земиздат, София, 380 стр.
2. **Coutelle S., M. Maurin, F. Rougier, X. Barbaut, L. Bourguignon, M. Ducher, P. Maire.** 2008. The Hill equation: a review of its capabilities in pharmacological modeling. *Fundamentals & Clinical Pharmacology* 22: 633-648.
3. **Efron B.** 1979. Bootstrap methods: another look at the jackknife. *Ann. Statist.* 7: 1-26.

4. **Efron, B.** 1982. The jackknife, the bootstrap and other resampling plans. Soc. Ind. Appl. Math. CBMS - Natl. Sci. Found. Monogr. 38, 92 pp.
5. **Hill, A.V.** 1910. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. Proceedings of the Physiological Society. J Physiol. 40 (Suppl.) iv-vii.
6. **Koumanov K and K. Kolev.** 2007. On the volume of raspberry fruit sample when estimating the average mass of one fruit. Journal of Mountain Agriculture on the Balkans 10(6):1128-1139.
7. **Koumanov K and K. Kolev.** 2008. Volume of the sample for estimating the average mass and the dimensions of one cherry and apple fruit. Journal of Mountain Agriculture on the Balkans 11(7): 1478-1490.
8. **Koumanov K. and K. Kolev.** 2009. Sample size estimation. Journal of Mountain Agriculture on the Balkans 12(6): 1566-1575.