ISSN 1733-8670

ZESZYTY NAUKOWE NR 10(82)
AKADEMII MORSKIEJ
W SZCZECINIE

IV MIĘDZYNARODOWA KONFERENCJA NAUKOWO-TECHNICZNA
E X P L O - S H I P 2 0 0 6

Leszek Chybowski, Zbigniew Matuszak

Wybrane sposoby określania liczności próby

Słowa kluczowe: liczność próby, pobieranie próby, losowanie próby, losowanie z warstw

W analizie statystycznej wydarzeń w siłowni okrętowej, i nie tylko, istnieje konieczność określania liczności próby. Nie jest ona wymagana, jeżeli zebrane wyniki nie są obrabiane aparatem statystycznym, a służą tylko do zaprezentowania zaobserwowanych zjawisk lub przebiegów. W prezentowanym materiale przedstawiono niektóre sposoby określania liczności próby oraz pobierania próby. Scharakteryzowano losowanie próby ze zbioru nieznanego i z tzw. warstw, wielostopniowe i systematyczne oraz wiązkami. Przedstawiono również uwagi o kosztach pobierania próby i proporcjach w próbie.

Selected Methods of Sample Size Determination

Key words: sample size, sampling, drawing a sample, sample drawing from layers

There is a need for specifying a sample size in statistical analysis of, among others, marine power plant events. Sample size is not required when the collected results are not processed with statistical tools, but only are used to present observed phenomena. This work points out some of the methods of sample size determination and drawing a sample. The authors describe the drawing of a sample from an unknown set, from layers, multistage, systematic and in bundles. Remarks on costs of sampling and on proportions in a sample are also given.

Wprowadzenie

Sposoby określania liczności próby oraz pobierania próby mają szczególne znaczenie dla wiarygodności badań statystycznych obserwowanych zjawisk lub przebiegów. Szczególne znaczenie ma to dla długotrwałych i obszernych badań, gdzie występuje wiele elementów o bardzo różnych cechach charakterystycznych, mających znaczenie dla wnioskowania o przebiegu badanych zjawisk lub procesów. W dalszej części prezentowanego materiału przedstawiono, z konieczności w bardzo uproszczony i skrócony sposób, niektóre sposoby pobierania prób losowych oraz uwagi na temat ich kosztów i proporcji losowanych elementów w badanych zbiorach.

1. Pobieranie próby

W celu wyznaczenia jednostek najbardziej typowych dla badanej zbiorowości urządzeń technicznych, elementów, zbiorów elementów pobiera się określoną liczbę elementów ze zbioru, nazywaną powszechnie próbą losową

Na podstawie tego badania wyciąga się wnioski o całej zbiorowości – całej partii wyrobów lub eksploatowanych urządzeń czy elementów. Podstawowym problemem jest określenie sposobu przeprowadzenia losowania, tj. czy wybrane elementy należycie reprezentują całą populację zbioru. Przy pobieraniu próbek należy zapewnić jednakową szansę trafienia do próby każdemu z elementów badanej populacji. Najprostsza metoda polega na ponumerowaniu elementów zbioru i wylosowaniu odpowiedniej liczby kartek z numerami, które wskazują elementy do analizy. Czasem stosuje się w tym celu specjalne tabele liczb losowych. Przy losowaniu ustala się np. stronę, z której odczytuje się kolejno ilość liczb, które będą numerami elementów wybranych do badania, albo też na określonej stronie wybiera się liczby z pewnej kolumny itp.

W dalszej części przedstawiono bardziej złożone sposoby pobierania prób losowych oraz zagadnienia charakteryzujące koszt pobierania tych prób.

1.1. Losowanie ze zbioru nieznanego

Należy np. ustalić, ile paliwa zużywają silniki zespołów prądotwórczych tego samego typu na serii masowców w określonym czasie τ przy porównywalnych mocach prądnic. W tym celu losuje się n zespołów i dokładnie ustala zużycie przez nie paliwa w czasie τ. Zespoły prądotwórcze wybiera się przy użyciu tablic liczb losowych; liczby te wyznaczają numery zespołów. Należy podkreślić, że wszystkie zespoły powinny być tego samego typu i o tych samych parametrach technicznych, aby był to zbiór jednorodny.

1.2. Losowanie z warstw

Metoda losowania z warstw polega na podziale całego zbioru na warstwy i następnie losowaniu z każdej warstwy próby, w której każda warstwa będzie reprezentowana proporcjonalnie do jej liczebności w całym zbiorze [3]. Można to przedstawić zależnością:

$$N_1 : N_2 : N_3 : ... = n_1 : n_2 : n_3 ... \tag{1}$$

gdzie:
N_i – liczebność warstwy,
n_i – liczebność próby.

Odchylenie standardowe charakteryzuje dyspersję każdej warstwy; im większe odchylenie standardowe, tym bardziej jest różnorodna warstwa.

Pobieranie losowej próby z każdej warstwy proporcjonalnie do odchylenia standardowego tych warstw można wyrazić jako:

$$\sigma_1 : \sigma_2 : \sigma_3 : ... = n_1 : n_2 : n_3 ... \tag{2}$$

We wzorze (2) liczebność elementów z każdej warstwy nie uwzględnia liczebności tych warstw. Aby można było przeprowadzić losowanie, biorąc pod uwagę liczebność warstw i odchylenia standardowe każdej warstwy, stosuje się zależność:

$$N_1\,\sigma_1 : N_2\,\sigma_2 : N_3\,\sigma_3 : ... = n_1 : n_2 : n_3 ... \tag{3}$$

Zastosowanie losowania z warstw powoduje, że maksymalne różnice między średnią próby a średnią arytmetyczną całego zbioru są mniejsze, aniżeli w wypadkach, gdy losuje się z całego zbioru bez uwzględniania warstw. Losowanie metodą z warstw zabezpiecza przed możliwością otrzymania próby bardzo mało reprezentatywnej.

1.3. Losowanie wielostopniowe

Dotyczy ono określenia jakości jednorodnych elementów (podzespołów) otrzymywanych w partiach w jednorodnych zbiorczych opakowaniach. Przeprowadza się najpierw losowanie kilku największych opakowań; następnie z nich losuje się kilka mniejszych opakowań, by z nich pobrać losowo kilka elementów (podzespołów) i sprawdzić ich jakość.

Przy losowaniu wielostopniowym można zastosować losowanie z warstw, jeżeli można wyraźnie te warstwy rozróżnić [3, 4].

1.4. Losowanie systematyczne

Polega na cyklicznym pobieraniu (losowaniu) elementów do oceny. Ten sposób losowania systematycznego można stosować również przy losowaniu z każdej warstwy lub przy losowaniu wielostopniowym, co ma często zastosowanie przy kontroli jakości wyrobów, kiedy bada się np. co dwudziestą czy co czterdziestą sztukę wyprodukowanego wyrobu. Należy przestrzegać, by pobierane kolejno do oceny sztuki nie były „zgodne" z rytmem produkcji, np. nie następowały po zmianie narzędzi lub regulacji maszyn [3].

1.5. Losowanie wiązkami

Stosuje się je dla określonych grup, w specyficznych środowiskach. Dotyczy losowania grup pracowników lub skupisk ludności i w zagadnieniach typowo technicznych nie ma zastosowania, chyba, że w badaniach socjologicznych grup pracowniczych.

2. Koszty pobierania próby

Istotnym zagadnieniem przy pobieraniu próby jest minimalizacja wariancji średnich, gdyż wtedy wnioski wysnute na podstawie próby będą możliwie najbardziej trafne. Aby zmniejszyć wariancję średnich prób, należy niemal zawsze zwiększyć liczebność próby. Jednak przy każdym badaniu, zwiększenie liczebności próby pociąga za sobą wzrost kosztów badań. Dysponując kwotą c, stanowiącą koszt badań, można ją podzielić na: c_o – koszty stałe, niezależne od liczebności próby; c_h – koszt pobrania jednego elementu do próby. Jeżeli próba liczy n elementów, to:

$$c = c_o + c_h \cdot n \tag{4}$$

a zatem liczebność próby przy założonym koszcie badań wynosi:

$$n = \frac{c - c_o}{c_h} \tag{5}$$

Bardziej skomplikowane jest określenie liczebności próby, kiedy prowadzi się losowanie z warstw. Jeżeli liczebność próby z każdej warstwy wynosi n_h wówczas równanie funkcji kosztów ma postać [4]:

$$c = c_o + \sum c_h n_h \tag{6}$$

Jeżeli przeprowadzane jest losowanie z warstw tak, że liczebność próby jest proporcjonalna do iloczynu liczebności warstwy i odchylenia standardowego warstwy, wtedy liczebność próby z każdej warstwy ustala się na podstawie zależności [4]:

$$n_h = \frac{\frac{N_h \cdot \sigma_h}{\sqrt{c_h}}}{\sum \frac{N_h \cdot \sigma_h}{\sqrt{c_h}}} \cdot n \tag{7}$$

gdzie:

N_h – liczebność warstwy,
σ_h – odchylenie standardowe warstwy,
c_h – koszt pobrania jednego elementu z warstwy.

Wariancję średnich można obliczyć z zależności [1 – 4]:

$$\sigma_{\bar{x}}^2 = \frac{1}{N^2} \cdot \frac{1}{n} \left(\sum N_h \cdot \sigma_h \cdot \sqrt{c_h} \right) \left(\sum \frac{N_h \cdot \sigma_h}{\sqrt{c_h}} \right) - \frac{1}{N^2} \sum N_h \sigma_h^2 \tag{8}$$

Mając określony zbiór, który podzielono na warstwy, oraz pewną kwotę, którą rozporządza się dla przeprowadzenia badań, można ustalić optymalną liczebność próby z każdej warstwy, aby otrzymać przy tych założeniach możliwie najmniejszą wariancję średnich prób.

Bardzo często nie jest znane odchylenie standardowe warstw, wówczas szacuje się je za pomocą odchylenia standardowego, obliczonego na podstawie prób.

Mając jednak określoną z góry wariancję średnich, należy tak przeprowadzić badania, aby przy badaniu pewnych wartości, których średnia jest określona, wariancja średnich nie była większa od zadanej wartości. Poniżej przedstawiono to na przykładzie badania ciężaru nakrętek, których średnia waga wynosi 100 N, a wariancja średnich ma być nie większa niż 2 N.

Dla rozkładu normalnego ustalono, że w granicach od $\bar{x} - 3\sigma$ do $\bar{x} + 3\sigma$ mieści się 0,0074 wszystkich elementów zbioru, mającego rozkład normalny. Wobec tego prawdopodobieństwo, że *x* przybierze wartości mniejsze od dolnej granicy tego przedziału ($\bar{x} - 3\sigma$) lub większe od górnej granicy ($\bar{x} + 3\sigma$) wynosi 1,0000 – 0,9974 = 0,0026.

Badania należy przeprowadzić tak, aby wariancja średnich prób przy badaniu nakrętek była nie większa niż 2 N tj. d = 2 N, czyli prawdopodobieństwo otrzymania takiej właśnie próby było 0,9974, a więc, by średnia próby w 9974 wypadkach na 10 000 mieściła się w żądanej granicy ± 2 N. Żąda się więc, aby $d = 3\sigma_{\bar{x}}$ lub $d^2 = 3^2 \sigma_{\bar{x}}^2$.

Stąd wariancja średnich prób:

$$\sigma_{\bar{x}}^2 = \frac{d^2}{3^2} = D^2 \tag{9}$$

W analizowanym przykładzie, przy badaniu ciężaru nakrętek, zgodnie ze wzorem (9) $\sigma_{\bar{x}}^2 = \frac{2^2}{3^2} = \frac{4}{9}$.

Wariancję średnich określa zależność:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \tag{10}$$

czyli żąda się, aby wariancja średnich prób wynosiła $\sigma_{\bar{x}}^2 = \frac{4}{9}$, a skoro wariancja całego zbioru (całej partii produkowanych nakrętek) wynosi $\sigma^2 = 4$, a odchylenie standardowe ciężaru nakrętek wynosi $\sigma = 2$ N, to liczność próby wynosi $n = \frac{\sigma^2}{\sigma_{\bar{x}}^2} = \frac{4}{\frac{4}{9}} = 9$ sztuk nakrętek.

Jeżeli przeprowadza się losowanie z warstw, to losując proporcjonalnie do liczebności warstwy, wariancję średnich prób określa zależność [1, 2, 3, 4]:

$$\sigma_{\bar{x}}^2 = \frac{1}{N}\sum \frac{N_h \sigma_h^2}{n} - \frac{1}{N^2}\sum N_h \sigma_h^2 \tag{11}$$

gdzie:

N – liczebność całego zbioru,

N_h – liczebność warstwy,

σ_h – wariancja warstwy.

Natomiast gdy losuje się z warstw proporcjonalnie do iloczynu liczebności warstwy i odchylenia standardowego tej warstwy, wariancja średnich prób wynosi jak w zależności (8):

$$\sigma_{\bar{x}}^2 = \frac{1}{N^2} \cdot \frac{1}{n}\left(\sum N_h \cdot \sigma_\mathrm{h} \cdot \sqrt{c_h}\right)\left(\sum \frac{N_h \cdot \sigma_h}{\sqrt{c_h}}\right) - \frac{1}{N^2}\sum N_h \sigma_h^2 \tag{12}$$

Mając więc założoną pewną wariancję średnich prób i żądaną tzw. dokładność szacunku (d), można określić liczebność próby, aby spełniała ona powyższe założenia [3, 4]:

– przy losowaniu z warstw proporcjonalnie do ich liczebności:

$$n = \frac{N\sum N_h\sigma_h^2}{N^2D^2 + \sum N_h\sigma_h^2} \tag{13}$$

– przy losowaniu z warstw proporcjonalnie do iloczynu liczebności i odchylenia standardowego warstw:

$$n = \frac{\left(\sum N_h\sigma_h^2\sqrt{c_h}\right)\left(\frac{\sum N_h\sigma_h^2}{\sqrt{c_h}}\right)}{N^2D^2 + \sum N_h\sigma_h^2} \tag{14}$$

Obie powyższe zależności (13) i (14) wymagają znajomości wariancji całych warstw σ_h^2, a w praktyce bardzo często nie są one znane. Szacuje się je wtedy na podstawie np. uprzednio badanych prób, albo też pobiera się próby z każdej warstwy i oblicza ich wariancje, które następnie służą do szacowania wariancji warstw.

3. Proporcje w próbie

Poprzednio rozważano różne rodzaje losowania w celu ustalenia średniej lub też odchylenia standardowego. Często jednak zadaniem jest określenie udziału pewnego rodzaju elementów w całym zbiorze na podstawie próby. Chcąc na przykład ustalić, jaki udział (procent) stanowią zużyte łożyska trzech typów A, B i C, albo też jaki jest udział tych łożysk w całej populacji zamawianych łożysk. Zbadanie powyższych procesów w dłuższym okresie, np. w ciągu roku, czy też zbadanie całej populacji jest bardzo pracochłonne i pociąga za sobą znaczne koszty. Można tego dokonać w pewnym krótszym okresie (tydzień, miesiąc) i na podstawie tego badania (próby) oszacować proporcję poszczególnych typów łożysk w całym roku.

Jeżeli elementy mające pewną cechę oznaczy się przez 1, to elementy nie mające tej cechy będą oznaczone przez 0. Badając na przykład zużycie łożysk, typ A oznaczono przez 1, a pozostałe typy łożysk przez 0. Tak więc, jeżeli zaobserwowano zużycie np. 100 łożysk (N), a wśród nich było 30 łożysk typu A, to proporcja zużytych łożysk typu A dla całej badanej populacji wynosi: $p = 30/100 = 0{,}3$.

Podczas badań trwających dwa tygodnie, w pierwszym tygodniu na 100 zużytych łożysk 30 było typu A, natomiast w drugim tygodniu na 200 zużytych łożysk 50 było typu A, tak więc proporcje są następujące: $p_1 = 30/100 = 0{,}3$; $p_2 = 50/200 = 0{,}25$.

Udział zużytych łożysk typu A w okresie badanym (dwa tygodnie) wynosi:

$$P=\frac{30+50}{100+200}=0{,}267 \quad \text{lub} \quad P=\frac{100\cdot 0{,}3+200\cdot 0{,}25}{100+200}=0{,}267$$

Powyższe obliczenia określa wzór:

$$P=\frac{\sum N_h \cdot p_h}{N} \tag{15}$$

gdzie:
N_h – liczebność warstwy,
N – liczebność całego zbioru $\left(N=\sum N_h\right)$.

3.1. Obliczanie wariancji

Oznaczając przez P udział elementów w zbiorze mającym określoną cechę, a przez Q udział elementów w zbiorze, które tej cechy nie mają, to wariancja proporcji zbioru jest równa [1, 2]:

$$\sigma_p^2=\frac{N-n}{N-1}\cdot\frac{PQ}{n} \tag{16}$$

gdzie:
N – liczebność zbioru,
n – liczebność próby,
P – proporcja (udział) jednostek mających określoną cechę,
Q – proporcja (udział) jednostek nie mających określonej cechy.

Natomiast wariancję proporcji na podstawie próby oblicza się z zależności [1, 2]:

$$S_p^2=n\frac{pq}{n-1} \tag{17}$$

gdzie:
S_p^2 – wariancja proporcji obliczona na podstawie próby,
n – liczebność próby,
p – proporcja (udział) jednostek mających określoną cechę,
q – proporcja (udział) jednostek nie mających określonej cechy.

3.2. Obliczanie odchylenia standardowego

Odchylenie standardowe proporcji obliczone na podstawie próby jest pierwiastkiem z wariancji i określa je zależność [1, 2]:

$$s_p = \sqrt{\frac{npq}{n-1}} \tag{18}$$

W analizowanym wcześniej przykładzie odchylenie standardowe proporcji zużytych łożysk typu A wynosi $s_p = \sqrt{0{,}212} = 0{,}46$.

Uwagi końcowe

Chcąc ustalić liczność próby tak, żeby ścisłość szacunku była z dokładnością do $\pm\sigma_p$, można tego dokonać pobierając odpowiednio liczną próbę. Mając np. $z = 3$ oraz p i q znane z próby, ustala się liczność próby ze wzoru [3]:

$$n = \frac{z^2 pq}{d^2} \tag{19}$$

Obliczenia wariancji całego zbioru przy losowaniu z warstw dokonuje się z zależności [3, 4]:

$$\sigma_p^2 = \frac{1}{N^2} \sum N_h \frac{N_h - n_h}{N_h} \cdot \frac{1}{n_h} \cdot \frac{N_h}{N_h - 1} P_h Q_h \tag{20}$$

Ponieważ zwykle nie jest znana proporcja w zbiorze, szacujemy ją na podstawie próby i wówczas wariancję zbioru można oszacować na podstawie informacji uzyskanych z próby, ze wzoru [3, 4]:

$$\sigma_{p_h}^2 = \frac{1}{N^2} \sum N_h^2 \frac{pq}{n_h} \tag{21}$$

Wzór (21) można stosować, kiedy liczebność próby jest niewielka w stosunku do całego zbioru. We wzorze tym zostało opuszczone wyrażenie $N_h/(N_h - 1)$ ponieważ przy dużej liczebności zbioru jest ono bardzo bliskie jedności.

W przypadku, gdy należy tak zorganizować badania, aby otrzymać wynik z określoną dokładnością, czyli oszacowany z próby udział pewnych elementów w całej zbiorowości nie może być większy od założonej liczby (d), a dodatkowo ma to być szacunek wystarczająco pewny (np. prawdopodobieństwo, że będzie on zawarty w określonym przedziale ($\pm d$), czyli w granicach „trzech sigma" $D = d/3$), można to osiągnąć przez dobór odpowiedniej liczebności próby losowej. Jeżeli losuje się próbę z warstw proporcjonalnie do liczebności warstwy, to liczebność próby z każdej warstwy określana jest na podstawie zależności [3, 4]:

$$n = \frac{N \sum N_h P_h Q_h}{N^2 D^2 + \sum N_h P_h Q_h} \tag{22}$$

Jeżeli losowanie przeprowadza się proporcjonalnie do iloczynu liczebności i odchylenia standardowego warstwy, a należy też uwzględnić koszty losowania z każdej warstwy, to liczebność próby określa się na podstawie zależności [3, 4]:

$$n = \frac{\left(\sum N_h \sqrt{c_h} P_h Q_h\right)\left(\sum N_h \sqrt{\frac{P_h Q_h}{c_h}}\right)}{N^2 D^2 + \sum N_h P_h Q_h} \tag{23}$$

a liczebność próby z każdej warstwy określa się proporcjonalnie do wartości $N_h \sqrt{\frac{P_h Q_h}{c_h}}$.

Zaprezentowane zagadnienia przedstawiają tylko niektóre sposoby określania liczności próby oraz pobierania próby. Literatura szeroko charakteryzuje opisywane problemy, które w prezentowanym materiale tylko zasygnalizowano.

Literatura

1. Bobrowski D., *Probabilistyka w zastosowaniach technicznych*, Wydawnictwo Naukowo-Techniczne, Warszawa 1986.
2. Bobrowski D., Maćkowiak-Łybacka K., *Wybrane metody wnioskowania statystycznego*, Wyd. Politechniki Poznańskiej, Poznań 1988.
3. Greń J., *Statystyka matematyczna. Modele i zadania*, Państwowe Wydawnictwo Naukowe, Warszawa 1984.
4. Pacut A., *Prawdopodobieństwo. Teoria. Modelowanie probabilistyczne w technice*, Wydawnictwo Naukowo-Techniczne, Warszawa 1985.

Wpłynęło do redakcji w lutym 2006 r.

Recenzent

prof. dr hab. inż. Mieczysław Hann

Adresy Autorów

mgr inż. Leszek Chybowski
dr hab. inż. Zbigniew Matuszak, prof. nadzw. AM
Akademia Morska w Szczecinie
Instytut Technicznej Eksploatacji Siłowni Okrętowych
70-500 Szczecin, ul. Wały Chrobrego 1-2
e-mail: zbimat@am.szczecin.pl