ΝΙΚΟΛΑΟΥ ΦΑΡΜΑΚΗ

ΤΟ ΜΕΓΕΘΟΣ ΤΩΝ ΔΕΙΓΜΑΤΩΝ ΣΤΗΝ ΑΠΛΗ ΤΥΧΑΙΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑ

ΔΙΑΓΡΑΜΜΑ

ΠΕΡΙΛΗΨΗ

Η περίπτωση της Απλής Τυχαίας Δειγματοληψίας (ΑΤΔ), κατά την οποία παίρνουμε δείγμα δ, μεγέθους n, από πληθυσμό Π, μεγέθους N (N>n), είναι η σπουδαιότερη θεωρητικά και η συχνότερα εφαρμοζόμενη μορφή Δειγματοληψίας. Στην προκείμενη περίπτωση εξετάζεται εκτενέστερα το μέγεθος n, που πρέπει να έχει το δείγμα, ώστε, με πιθανότητα $100(1-a)\% \geq 90\%$, να έχουμε μέγιστο σφάλμα εκτίμησης μικρότερο από h ή λόγο σφάλματος προς την Τυπική Απόκλιση s, $\frac{h}{s} = \lambda$ γνωστό εκ των προτέρων.

1. ΕΙΣΑΓΩΓΗ

Έχουμε πληθυσμό Π με Ν στοιχεία (μονάδες) και παρακολουθούμε, με τη βοήθεια μιας τυχαίας μεταβλητής (τ.μ.) Υ, μία χαρακτηριστική ιδιότητα των μονάδων του πληθυσμού. Ο σκοπός μας είναι να πετύχουμε εκτιμήσεις για τη μέση τιμή του πληθυσμού $\overline{Y}$ καθώς και για το άθροισμα των τιμών του πληθυσμού $Y = N \cdot \overline{Y} = \sum_{i=1}^{N} Y_i$, όπου Y_i είναι οι τιμές της τ.μ. Υ, μία για κάθε στοιχείο του πληθυσμού. Τις εκτιμήσεις αυτές τις έχουμε με τη βοήθεια δείγματος δ, μεγέθους n, από τον πληθυσμό Π και είναι

$$\hat{\overline{Y}} = \overline{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i \quad \text{και} \quad \hat{Y} = y = \sum_{i=1}^{n} y_i \tag{1.1}$$

οι αντίστοιχοι εκτιμητές που προέρχονται από το δείγμα αυτό. Οι ποσότητες

$$\left|\hat{\overline{Y}} - \overline{Y}\right| \quad \text{και} \quad \left|\hat{Y} - Y\right| \tag{1.2}$$

ονομάζονται σφάλματα εκτίμησης της μέσης τιμής και του αθροίσματος αντίστοιχα.

Με τη βοήθεια του Νόμου των Μεγάλων Αριθμών (ΝΜΑ) του Kolmogorov και του Κεντρικού Οριακού Θεωρήματος (ΚΟΘ) (βλ. Κούνιας Σ. κ.ά., 1985, κεφ. 4ο και Φαρμάκης Ν., 1992, κεφ. 2ο), έχουμε ότι:

$$\overline{y} \sim N\left(\overline{Y}, \mathrm{Var}\overline{y}\right), \quad \mathrm{Var}\overline{y} = \frac{1-f}{n} \cdot S^2, \quad f = \frac{n}{N}, \quad S^2 = \frac{1}{N-1} \cdot \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2 \tag{1.3}$$

Τα παραπάνω ισχύουν όταν $n \geq 30$. Στην παρούσα εργασία ενδιαφερόμαστε για τιμές του n πολύ μεγαλύτερες του 30. Αποδεικνύεται ότι για να έχουμε (με βεβαιότητα 100 · (1-a)%) $\left|\hat{\overline{Y}} - \overline{Y}\right| \leq h$, πρέπει

$$n \geq \frac{N}{1 + \left(\dfrac{h}{z_{a/2} \cdot s}\right)^2 \cdot N} = n_1(h, s, a, N), \quad s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 \tag{1.4}$$

και για να έχουμε $\left|\hat{Y} - Y\right| \leq h$ (με βεβαιότητα 100 · (1-a)%), πρέπει

$$n \geq \frac{N}{1+\left(\frac{h}{z_{a/2}\cdot s}\right)^2 \cdot \frac{1}{N}} = n_2(h,s,a,N) \tag{1.5}$$

Θα εξετάσουμε χωριστά τη σχέση (1.4) για τη μέση τιμή και χωριστά την (1.5) για το άθροισμα στις επόμενες παραγράφους.

2. ΜΕΛΕΤΗ ΓΙΑ ΤΗ ΜΕΣΗ ΤΙΜΗ

Από τη σχέση (1.4) έχουμε:

$$n \geq \frac{N}{1+\left(\frac{h}{z_{a/2}\cdot s}\right)^2 \cdot N} = \frac{1}{\frac{1}{N}+\left(\frac{h}{z_{a/2}\cdot s}\right)^2} = n_1(h,s,a,N) \tag{2.1}$$

Συμβολίζουμε με $\lambda = \frac{h}{s}$ και η (2.1) γίνεται

$$n \geq \frac{1}{\frac{1}{N}+\left(\frac{\lambda}{z_{a/2}}\right)^2} = n_1(\lambda,a,N) \tag{2.2}$$

και είναι

$$\lim_{N\to\infty} n_1(\lambda,a,N) = \left(\frac{z_{a/2}}{\lambda}\right)^2 = n_1(\lambda,a) \tag{2.3}$$

Συνηθισμένες τιμές για το λ είναι οι μικρότερες του 0.60. Εάν $\lambda \geq 0.60$, τότε δεν είναι δυνατό να οδηγηθούμε σε μεγάλα δείγματα. Είναι άλλωστε απαίτησή μας για την προκείμενη περίπτωση να πάρουμε το $n_1(\lambda, a) \geq 30$. Άρα:

$$\lambda \leq \frac{z_{a/2}}{\sqrt{30}} = \frac{z_{a/2}}{30}\cdot\sqrt{30} = 0.18\cdot z_{a/2}. \tag{2.4}$$

Μερικές περιπτώσεις εφαρμογής του τύπου (2.4) είναι οι παρακάτω:

a	$\lambda \leq$
0.05	0.353
0.01	0.464
0.005	0.513

Εξετάζεται στον παρακάτω πίνακα 2.1, για a=0.05 και $\lambda \leq 0.1$, η συμπεριφορά του $n_1(\lambda, 0.05)$, ενώ στην τελευταία στήλη γράφονται οι ελάχιστες τιμές. Βασιζόμαστε στη σχέση

$$n_1(\lambda, 0.05) = \frac{1.96^2}{\lambda^2} = \frac{3.8416}{\lambda^2}.$$

ΠΙΝΑΚΑΣ 2.1

λ	$n_1(\lambda, 0.05)$	$n \geq$
0.10	384.16	385
0.09	474.27	475
0.08	600.25	601
0.05	1536.64	1537
0.02	9604.00	9604
0.01	38416.00	38416

Στον παραπάνω πίνακα 2.1 υποτίθεται ότι $N \to \infty$.

3. ΜΕΛΕΤΗ ΓΙΑ ΤΟ ΑΘΡΟΙΣΜΑ

Η σχέση (1.5) γίνεται

$$n \geq \frac{N}{1+\left(\frac{h}{z_{a/2}\cdot s}\right)^2 \cdot \frac{1}{N}} = \frac{N^2}{N+\left(\frac{\lambda}{z_{a/2}}\right)^2} = n_2(\lambda, a, N) \qquad (3.1)$$

Η τιμή του h όμως —άρα και του λ— είναι τώρα πολύ μεγαλύτερη. Οι παράμετροι h και λ έχουν τώρα τάξη μεγέθους N φορές μεγαλύτερη από την αντίστοιχη τάξη μεγέθους των παραμέτρων της προηγούμενης παραγράφου. Γενικά είναι $\lambda \in [0.01 \cdot N, 0.5 \cdot N]$. Συνηθισμένες τιμές για το λ είναι στην πράξη από 100 μέχρι 1.000.000.

Η σχέση (3.1) δίνει

$$\lim_{\lambda \to 0} n_2(\lambda, a, N) = \lim_{\lambda \to 0} \frac{N^2 \cdot z_{a/2}^2}{N \cdot z_{a/2}^2 + \lambda^2} = N \tag{3.2}$$

και η παράγωγος

$$\frac{\partial n_2}{\partial \lambda} = \frac{-2\lambda \cdot N^2 \cdot z_{a/2}^2}{\left(N \cdot z_{a/2}^2 + \lambda^2\right)^2} < 0 \tag{3.3}$$

εξασφαλίζει το συμπέρασμα ότι η συνάρτηση n_2 (λ, a, N) είναι φθίνουσα συνάρτηση του λ, όταν τα N, a παραμένουν σταθερά. Πιο ενδιαφέρον είναι να εξετάσουμε τη συμπεριφορά του λ σε σχέση με το δειγματοληπτικό κλάσμα $f = \frac{n}{N} \in [0,1]$. Συνηθισμένος περιορισμός: $f \le b \in [0,1]$. Από τις σχέσεις (3.1) και (3.2) έχουμε

$$\frac{N^2 \cdot z_{a/2}^2}{N \cdot z_{a/2}^2 + \lambda^2} \le f \cdot N \iff \frac{N \cdot z_{a/2}^2}{N \cdot z_{a/2}^2 + \lambda^2} \le f \iff \lambda^2 \ge \frac{1-f}{f} \cdot N \cdot z_{a/2}^2$$

δηλαδή

$$\lambda \ge z_{a/2} \cdot \sqrt{\frac{1-f}{f} \cdot N} \tag{3.4}$$

Στην παραπάνω σχέση χρησιμοποιούμε συνήθως το ίσον στις πράξεις.

Παράδειγμα 3.1: Ο πληθυσμός αποτελείται από N=10.000 μονάδες και θέλουμε να εξετάσουμε τις τιμές του λόγου $\lambda = h/s$, όταν η δειγματοληψία εκτείνεται στο $100 \cdot f\% \in \{1\%, 2\%, 3\%, \ldots 10\%\}$ του πληθυσμού.

Απάντηση: Ο επόμενος πίνακας 3.1 δίνει την απάντηση στο ερώτημά μας και κατασκευάστηκε με βάση τη σχέση (3.4) και το γεγονός ότι N=10.000 και a=0.05:

ΠΙΝΑΚΑΣ 3.1

f	λ	n
0.01	1950.175	100
0.02	1372.000	200
0.03	1114.503	300
0.04	960.200	400
0.05	854.344	500
0.06	755.790	600
0.07	714.412	700
0.08	664.668	800
0.09	623.240	900
0.10	588.000	1000

Είναι a=0.05 και N=10.000.

Παράδειγμα 3.2: Αν στο παράδειγμα 3.1 η εξεταζόμενη τ.μ. Y ακολουθεί κανονική κατανομή και από το δείγμα έχουμε $|y_{max} - y_{min}| = 800$, να βρεθεί μία εκτίμηση για το απόλυτο σφάλμα h του αθροίσματος του πληθυσμού και μία εκτίμηση για το σχετικό (ως προς τη μέση τιμή $\overline{Y}$) σφάλμα του αθροίσματος του πληθυσμού, το $h/\overline{Y}$, αν $\overline{Y} = 1.600$.

Απάντηση: Επειδή δεν έχουμε εκτίμηση για τη διασπορά και γνωρίζοντας ότι η κατανομή της τ.μ. Y είναι η κανονική, παίρνουμε:

$$|y_{max} - y_{min}| = 6 \cdot S$$

και για μεγαλύτερη βεβαιότητα

$$|y_{max} - y_{min}| = 5 \cdot s = 800 \quad \Leftrightarrow \quad s = 160$$

οπότε είναι

$$h = s \cdot \lambda \quad \text{και} \quad h/\overline{Y} = \frac{s \cdot \lambda}{\overline{Y}} = 0.1 \cdot \lambda \tag{3.5}$$

Τα αποτελέσματα για $f \in \{0.01, 0.02, \ldots 0.10\}$ φαίνονται στον παρακάτω πίνακα 3.2.

ΠΙΝΑΚΑΣ 3.2

f	λ	n	h	$h/\bar{Y}$
0.01	1950.175	100	312028.0	195.02
0.02	1372.000	200	219520.0	137.20
0.03	1114.503	300	178320.5	111.45
0.04	960.200	400	153632.0	96.02
0.05	854.344	500	136695.0	85.43
0.06	755.790	600	120926.4	75.58
0.07	714.412	700	114305.9	71.44
0.08	664.668	800	106346.9	66.47
0.09	623.240	900	99718.4	62.32
0.10	588.000	1000	94080.0	58.80

Είναι a=0.05 και N=10.000.

Σημείωση: Πήραμε $5 \cdot s$ αντί του $6 \cdot s$, διότι το εύρος εκτιμήθηκε από το δείγμα και στον πληθυσμό θα είναι λίγο μεγαλύτερο. Η μέθοδος αυτή είναι εμπειρική και χρησιμοποιείται πολύ συχνά για γρήγορη εκτίμηση της τυπικής απόκλισης s, που θα χρησιμοποιηθεί στον τύπο (2.1). Πολλές φορές, ανάλογα με την εμπειρία μας, μπορεί να πάρουμε εύρος $= 4 \cdot s$, για περισσότερη σιγουριά στον υπολογισμό του n (βλ. Φαρμάκης Ν. (1992), ασκήσεις 2.2, 2.3, σελ. 66 και 67). Σημειώνεται τέλος ότι στην (3.5) συνήθως χρησιμοποιείται η δειγματική τιμή $\bar{y}$, της μέσης τιμής, αντί της πληθυσμιακής $\bar{Y}$.

SUMMARY

The size of sample in the Simple Random Sampling Process is examined as a function of the sample variance, s^2, (or of the sample standar deviation, s) and of the permited error limmit, h, for the mean value estimation and the population sum estimation. Very large values for population size N are assumed. We also assume that the sample size n is $n \geq 30$, in this paper. The level of significance in most of cases is taken as $a=0.05$. In order to avoid a great deal of work and for uniform presentation reasons we introduce the ratio parameter $\lambda=\frac{h}{s}$ in all the cases.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Κούνιας Σ., Κολυβά Φ., Μπαγιάτης Κ., Μπόρα Ε., *Εισαγωγή στη Στατιστική,* Υπηρεσία Δημοσιευμάτων του Α.Π.Θ., Θεσσαλονίκη 1985.

Φαρμάκης Ν., *Εισαγωγή στη Δειγματοληψία,* Εκδόσεις Επιστημονικών Βιβλίων Κ. Χριστοδουλίδη, Θεσσαλονίκη 1992.