3.ΑΠΛΗ ΤΥΧΑΙΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑ (SIMPLE RANDOM SAMPLING)

Η πιο απλή τουλάχιστον στην φιλοσοφία της και στην ανάλυση των δεδομένων της μέθοδος δειγματοληψίας είναι αυτή κατά την οποία ένας αριθμός n ατόμων (πρόκειται για μονάδες δειγματοληψίας) επιλέγονται τυχαία από τα N άτομα του πληθυσμού. Η τυχαία επιλογή γίνεται με τη βοήθεια πινάκων τυχαίων αριθμών (παράρτημα 1), ρουτίνας τυχαίων αριθμών σε υπολογιστή ή με οποιαδήποτε άλλο τρόπο η αμεροληψία του οποίου είναι εξακριβωμένη.

3.1. ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΟΥ ΤΟΥ ΠΛΗΘΥΣΜΟΥ

Το ενδιαφέρον συχνά περιορίζεται σε δύο χαρακτηριστικά του πληθυσμού: Τη **μέση τιμή** (π.χ. αριθμός μελών ανά οικογένεια) και την **αφθονία** ή το **σύνολο** των ατόμων του πληθυσμού(αριθμός δένδρων σε μια συγκεκριμένη περιοχή).

ΕΚΤΙΜΗΤΕΣ

Δεδομένα

$$y_1, y_2, y_3, \ldots, y_i, \ldots, y_n$$

Παράμετροι του δείγματος

$$f = \frac{n}{N} \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

Εκτίμηση της μέσης τιμής ($\hat{\bar{Y}}$) του πληθυσμού

$$\hat{\bar{Y}} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad v(\hat{\bar{Y}}) = s_{\bar{y}}^2 = \frac{s_y^2}{n}(1-f)$$

$$P\{\bar{y} - t_{\alpha/2} s_{\bar{y}} < \bar{Y} < \bar{y} + t_{\alpha/2} s_{\bar{y}}\} = 1 - \alpha$$

Εκτίμηση του συνόλου ($\hat{Y}$) του πληθυσμού

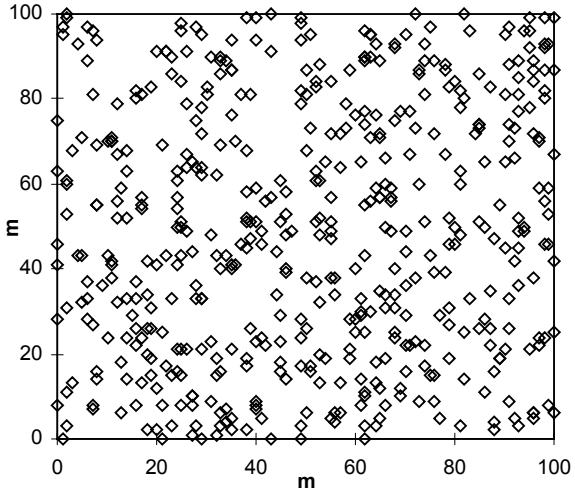
$$\hat{Y} = N\bar{y}c \qquad v(\hat{Y}) = N^2 s_{\bar{y}}^2 c$$

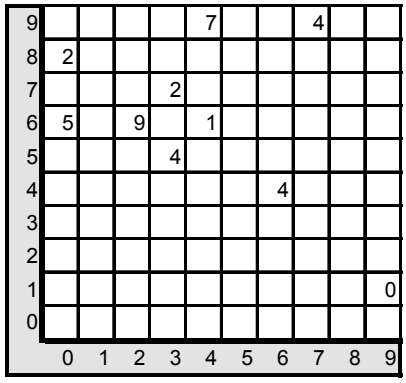
$$P\{\hat{Y} - t_{\alpha/2}\sqrt{v(\hat{Y})} < Y < \hat{Y} + t_{\alpha/2}\sqrt{v(\hat{Y})}\} = 1 - \alpha$$

⚠ Για τον ορισμό του διαστήματος εμπιστοσύνης και συγκεκριμένα για την επιλογή των τιμών του t, z ή άλλων ακολουθούμε τις επιλογές που παρουσιάζονται στο διάγραμμα 1.8.

ΠΑΡΑΔΕΙΓΜΑ 3.1

Το αριστερό διάγραμμα παρουσιάζει την κατανομή είδους φυτού σε ένα λιβάδι. Η έκταση αυτή χωρίστηκε σε τετράγωνες περιοχές εμβαδού $100m^2$ από τις οποίες 10 επιλέχθηκαν τυχαία και μελετήθηκαν. Ο αριθμός των ατόμων του φυτού που βρέθηκαν σε κάθε μία από αυτές τις δειγματοληπτικές μονάδες φαίνεται στο δεξί διάγραμμα. Ποια είναι η μέση πυκνότητα του συγκεκριμένου φυτού σε άτομα/$100m^2$ και ποιος ο συνολικός αριθμός φυτών στην περιοχή;





Δεδομένα:

$(y_1, ..., y_n)$= 2, 5, 9, 4, 2, 7, 1, 4, 4, 0

Παράμετροι του δείγματος:

Μέγεθος δείγματος (προσοχή πρόκειται για τον αριθμό των δειγματοληπτικών μονάδων όχι των αριθμό των ατόμων) [n=]

Μέγεθος του πληθυσμού [N=]

Δειγματοληπτικό κλάσμα [f=]

(βασικοί υπολογισμοί) [$\sum_{i=1}^{n} y_i$=] [$\sum_{i=1}^{n} y_i^2$=]

Μέση τιμή του δείγματος [$\bar{y}$=] (άτομα ανά 100 m^2)

Διασπορά του δείγματος [s^2=]

Εκτίμηση της μέσης τιμής του πληθυσμού:

[$\hat{\bar{Y}}$=] (άτομα ανά 100 m^2)

[$s_{\bar{y}}^2$=]

[t=] (πιθανότητα=5% και [d.f=])

Διάστημα εμπιστοσύνης στο επίπεδο : 95%

[<$\bar{Y}$<]

Εκτίμηση του συνόλου του πληθυσμού:

[$\hat{Y}$=]

[v($\hat{Y}$)=]

Διάστημα εμπιστοσύνης στο επίπεδο 95% :

[<Y<]

ΠΑΡΑΔΕΙΓΜΑ 3.1

3.2. ΕΚΤΙΜΗΣΗ ΛΟΓΟΥ ΔΥΟ ΠΑΡΑΜΕΤΡΩΝ

Σε αρκετές περιπτώσεις σε κάθε δειγματοληπτική μονάδα (που επιλέγεται τυχαία από το σύνολο των μονάδων του πληθυσμού)

μετρούμε δύο διαφορετικές παραμέτρους και το ενδιαφέρον μας στρέφεται στο λόγο των δυο αυτών χαρακτηριστικών. Για παράδειγμα σε μια οικογένεια που αποτελεί τη βασική δειγματοληπτική μονάδα της μελέτης καταγράφουμε την συνολική κατανάλωση τροφής και τον αριθμό των ατόμων και το ενδιαφέρον μας στρέφεται στη μέση κατανάλωση τροφής ανά άτομο και στην συνολική κατανάλωση τροφής στον πληθυσμό.

ΕΚΤΙΜΗΤΕΣ

Δεδομένα

$$y_1, y_2, y_3, \ldots, y_i, \ldots, y_n$$

$$x_1, x_2, x_3, \ldots, x_i, \ldots, x_n$$

οι τιμές y_i, x_i προέρχονται από την ίδια δειγματοληπτική μονάδα i, αποτελούν δηλαδή ζεύγος.

Παράμετροι του δείγματος

$$f = \frac{n}{N}^{(*)} \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(*) ή όποιος άλλος δείκτης δίνει το λόγο του μεγέθους του δείγματος προς το συνολικό μέγεθος του πληθυσμού όπως στο παράδειγμα που ακολουθεί

Εκτίμηση του λόγου ($\hat{R}$) στον πληθυσμό

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \qquad s_{\hat{R}}^2 = \frac{1-f}{n\bar{x}^2} \cdot \frac{\sum_{i=1}^{n} y_i^2 - 2\hat{R}\sum_{i=1}^{n} x_i y_i + \hat{R}^2 \sum_{i=1}^{n} x_i^2}{n-1}$$

$$P\{\hat{R} - t_{\alpha/2} s_{\hat{R}} < R < \hat{R} + t_{\alpha/2} s_{\hat{R}}\} = 1 - \alpha$$

Εκτίμηση του συνόλου ($\hat{Y}$) του πληθυσμού

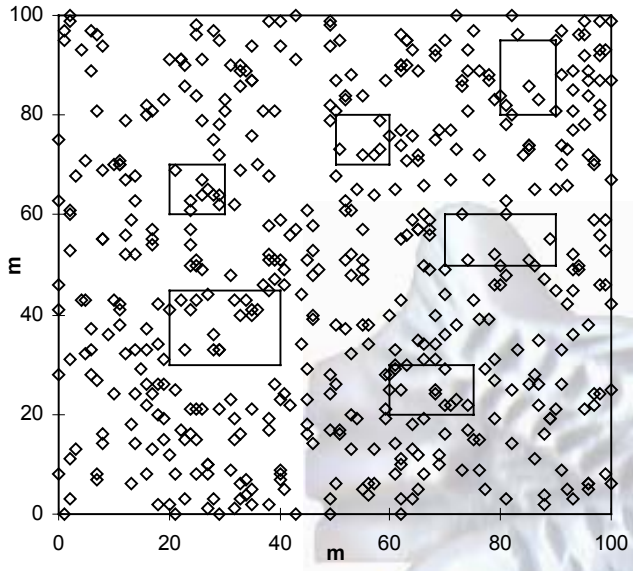
$$\hat{Y} = RX \qquad v(\hat{Y}) = s_{\hat{Y}}^2 = X^2 s_{\bar{R}}^2$$

$$P\{\hat{Y} - t_{\alpha/2} s_{\hat{Y}} < Y < \hat{Y} + t_{\alpha/2} s_{\hat{Y}}\} = 1 - \alpha$$

⚠ Ο εκτιμητής αυτός δεν είναι αμερόληπτος και για $n<30$ η μεροληψία (bias) γίνεται μεγάλη.

ΠΑΡΑΔΕΙΓΜΑ 3.2

Το αριστερό διάγραμμα παρουσιάζει την κατανομή των ατόμων μικρού τρωκτικού σε ένα λιβάδι. Περιορισμένος αριθμός ζώων συλλέχθηκε με τη βοήθεια διχτυού. Οι περιοχές στις οποίες έγιναν οι συλλήψεις φαίνονται στο ίδιο διάγραμμα. Η επιφάνεια που κάλυπτε το δίχτυ ήταν έντονα κυμαινόμενη λόγω τεχνικών δυσκολιών. Σε κάθε δείγμα γινόταν και μέτρηση της επιφάνειας που κάλυπτε το δίχτυ (πίνακας). Με αυτά τα δεδομένα να εκτιμηθεί η μέση πυκνότητα του συγκεκριμένου είδους σε άτομα/m^2 και ο συνολικός αριθμός ζώων στην περιοχή;



α/α	αριθμός ατόμων	επιφάνεια δείγματος (m^2)
1	9	100
2	11	150
3	16	300
4	5	100
5	6	200
6	5	150

Δεδομένα:

$(y_1, ..., y_n)$= 9, 11, 16, 5, 6, 5

$(x_1, ..., x_n)$= 100, 150, 300, 100, 200, 150

Παράμετροι του δείγματος:

Μέγεθος δείγματος (προσοχή πρόκειται για τον αριθμό των δειγματοληπτικών μονάδων όχι των αριθμό των ατόμων) [n=]

Μέγεθος του πληθυσμού (συνολική επιφάνεια) [N=]

Δειγματοληπτικό κλάσμα (συνολική επιφάνεια κάλυψης των διχτυών προς τη συνολική επιφάνεια αναφοράς) [f=]

(βασικοί υπολογισμοί) [$\sum_{i=1}^{n} y_i$=] [$\sum_{i=1}^{n} y_i^2$=]

[$\sum_{i=1}^{n} x_i$=] [$\sum_{i=1}^{n} x_i^2$=] [$\sum_{i=1}^{n} x_i y_i$=]

Μέση τιμή του δείγματος [$\bar{y}$=]

Διασπορά του δείγματος [s^2=]

Εκτίμηση του λόγου R=y/x **του πληθυσμού:**

[$\hat{R}$=] (άτομα ανά m^2)

[v($\hat{R}$)=]

[t=] (πιθανότητα=5% και [d.f=])

Διάστημα εμπιστοσύνης στο επίπεδο : 95%

[<R<]

Εκτίμηση του συνόλου του πληθυσμού:

[$\hat{Y}$=]

[v($\hat{Y}$)=]

Διάστημα εμπιστοσύνης στο επίπεδο 95% :

[<Y<]

ΠΑΡΑΔΕΙΓΜΑ 3.2

3.3. ΕΚΤΙΜΗΣΗ ΠΟΣΟΣΤΟΥ

Συχνά το ενδιαφέρον της μελέτης στρέφεται στην εκτίμηση του ποσοστού των ατόμων που φέρουν συγκεκριμένο χαρακτηριστικό π.χ. κάποια ασθένεια, συγκεκριμένο χρωματισμό ή βρίσκονται σε συγκεκριμένο στάδιο ανάπτυξης. Σε τέτοιες μελέτες δύο είναι συνήθως οι παράμετροι του πληθυσμού που μας ενδιαφέρουν: το ποσοστό των ατόμων που φέρουν το χαρακτηριστικό και το σύνολο των ατόμων του πληθυσμού που ανήκουν σ'αυτή την κατηγορία.

ΕΚΤΙΜΗΤΕΣ

Δεδομένα

$$y_1, y_2, y_3, \ldots, y_i, \ldots, y_n$$

Η δειγματοληπτική μονάδα y_i παίρνει την τιμή 0 ή 1 ανάλογα με το αν έχει το προς μελέτη χαρακτηριστικό ή ιδιότητα

Παράμετροι του δείγματος

$$a = \sum_{i=1}^{n} y_i \qquad p = \frac{a}{n} \qquad q = \frac{n-a}{n} \qquad p + q = 1$$

p είναι το κλάσμα, η αναλογία (ποσοστό) των στοιχείων του δείγματος που έχουν τον συγκεκριμένο χαρακτήρα.

Εκτίμηση της μέσης αναλογίας του πληθυσμού

$$\hat{P} = p = \frac{a}{n} \qquad v(\hat{P}) = s_p^2 = \frac{pq}{n-1}(1-f)$$

$$P\{p - z_{\alpha/2}s_p - \frac{1}{2n} < P < p + z_{\alpha/2}s_p + \frac{1}{2n}\} = 1 - \alpha$$

Το ½n είναι μια διόρθωση για την ασυνέχεια των δεδομένων[1]

Εκτίμηση του συνόλου των ατόμων του πληθυσμού που φέρουν το συγκεκριμένο χαρακτηριστικό

$$\hat{A} = Npc \qquad v(\hat{A}) = \frac{pq}{n-1}N(N-n)$$

$$P\{\hat{A} - z_{\alpha/2}\sqrt{v(\hat{A})} - \frac{N}{2n} < A < \hat{A} + z_{\alpha/2}\sqrt{v(\hat{A})}\} + \frac{N}{2n} = 1 - \alpha$$

⚠ Το z ακολουθεί την κανονική κατανομή. Η δημιουργία του διαστήματος εμπιστοσύνης με τη βοήθεια του z είναι μία προσέγγιση του προβλήματος και ισχύει μόνο αν για την καταγραφήσα τιμή της αναλογίας p το μέγεθος του δείγματος είναι ίσο ή μεγαλύτερο από τα ακόλουθα όρια (πηγή Cochran 1977).

p	n
0.5	30
0.4	50
0.3	80
0.2	200
0.1	600
0.05	1400

Στην περίπτωση που το δείγμα δεν ικανοποιεί αυτές τις συνθήκες τότε για την εκτίμηση του διαστήματος εμπιστοσύνης χρησιμοποιούνται οι πίνακες της υπεργεωμετρικής κατανομής. Μια προσέγγιση της είναι η δυονιμική κατανομή[2].

[1] Η διόρθωση οφείλεται στο γεγονός ότι ο αριθμός των ατόμων που φέρουν το συγκεκριμένο χαρακτηριστικό είναι ακέραιος π.χ. 15 ή 16 αλλά δεν μπορεί να είναι 15.3 άτομα.. Χωρίς αυτή τη διόρθωση η κανονική προσέγγιση δίνει συστηματικά στενότερο διάστημα εμπιστοσύνης.

[2] Στη δυωνιμική κατανομή η πιθανότητα το δείγμα να περιέχει α άτομα με το χαρακτηριστικό είναι $P(\alpha) = \frac{n!}{\alpha!(n-\alpha)!}p^{\alpha}q^{n-\alpha}$

ΠΑΡΑΔΕΙΓΜΑ 3.3

Από το σύνολο των 400 δένδρων αγροκτήματος ένα τυχαίο δείγμα 64 εξετάσθηκε και 24 από αυτά βρέθηκαν να φέρουν συγκεκριμένο παράσιτο. Ποια είναι η αναλογία των δένδρων που έχουν προσβληθεί και ποιος ο συνολικός αριθμός των δένδρων που φέρουν το παράσιτο;

Παράμετροι του δείγματος:

Μέγεθος δείγματος [n=]

Μέγεθος του πληθυσμού [N=]

Αριθμός ατόμων που φέρουν συγκεκριμένο χαρακτηριστικό [α=]

Δειγματοληπτικό κλάσμα [f=]

Εκτίμηση της αναλογίας του πληθυσμού:

Αναλογία ατόμων που φέρουν το παράσιτο [p=]

Αναλογία ατόμων που δεν το φέρουν [q=]

Διασπορά της αναλογίας [s^2_p=]

[z=] (πιθανότητα=5%)

Διάστημα εμπιστοσύνης στο επίπεδο : 95% [<P<]

Εκτίμηση του συνόλου των ατόμων του πληθυσμού που φέρουν το παράσιτο:

[$\hat{A}$=]

[v($\hat{A}$)=]

Διάστημα εμπιστοσύνης στο επίπεδο 95% : [<$\hat{A}$<]

ΠΑΡΑΔΕΙΓΜΑ 3.3

3.4. ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΤΗΣ ΑΠΛΗΣ ΤΥΧΑΙΑΣ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

⇧ Πλεονεκτήματα ⇧

- Είναι γνωστή και κοινά αποδεκτή.
- Αποτελεί τη βάση της στατιστικής συμπερασματολογίας και το μεγαλύτερο μέρος των παραμετρικών ή μη παραμετρικών αναλύσεων εφαρμόζονται στα δεδομένα των δειγμάτων.
- Οι εκτιμητές είναι αμερόληπτοι (εκτός από τον εκτιμητή λόγου).
- Εάν υπάρχουν βάσιμες υποψίες ότι η κατανομή των ατόμων στον πληθυσμό δεν έχει φανερά συναθροιστικό χαρακτήρα τότε η μέθοδος επιλογής των ατόμων του δείγματος είναι σχετικά απλοϊκή π.χ. σε μια περιοχή με τοπογραφική και υδρολογική ομοιογένια ενά δείγμα πλαγκτού σε δεδομένη τοποθεσία μπορεί να θεωρηθεί ένα τυχαίο δείγμα. Επίσης η επιλογή κάποιων σπόρων από ένα δοχείο το περιεχόμενο του οποίου έχει μηχανικά ομεγενοποιηθεί (ανακάτεμα) αποτελεί μια τυχαία δειγματοληψία.

⇩ Μειονεκτήματα ⇩C

- Η δημιουργία ενός πλήρους καταλόγου με όλες τις μονάδες (στατιστικές μονάδες και όχι τα φυσικά άτομα) που αποτελούν τον πληθυσμό είναι μια υπόθεση δύσκολη έως αδύνατη και συχνά με υπέρογκο κόστος.
- Η συλλογή των μονάδων που θα αποτελέσουν το δείγμα είναι συχνά δύσκολη (προβλήματα πρόσβασης, μεγάλες αποστάσεις).
- Τα δεδομένα προσφερονται μόνο για τις εκτιμήσεις των παραμέτρων - στόχων αλλά προσφέρουν περιορισμένη πληροφορία για βασικές επεξεργασίες που λόγω της σπανιότητας των δειγμάτων είναι επιθυμητές κυρίως στις μελέτες της φύσης (χαρτογράφηση, ανάλυση βιοκοινωνιών, συσχέτηση με περιβαλλοντικές παραμέτρους.
- Η ακρίβεια των εκτιμήσεων είναι περιορισμένη σε σχέση με άλλες στρατηγικές και αυτό διότι το συγκεκριμένο σχέδιο δεν χρησιμοποιεί καμμία διαθέσιμη πληροφορία.