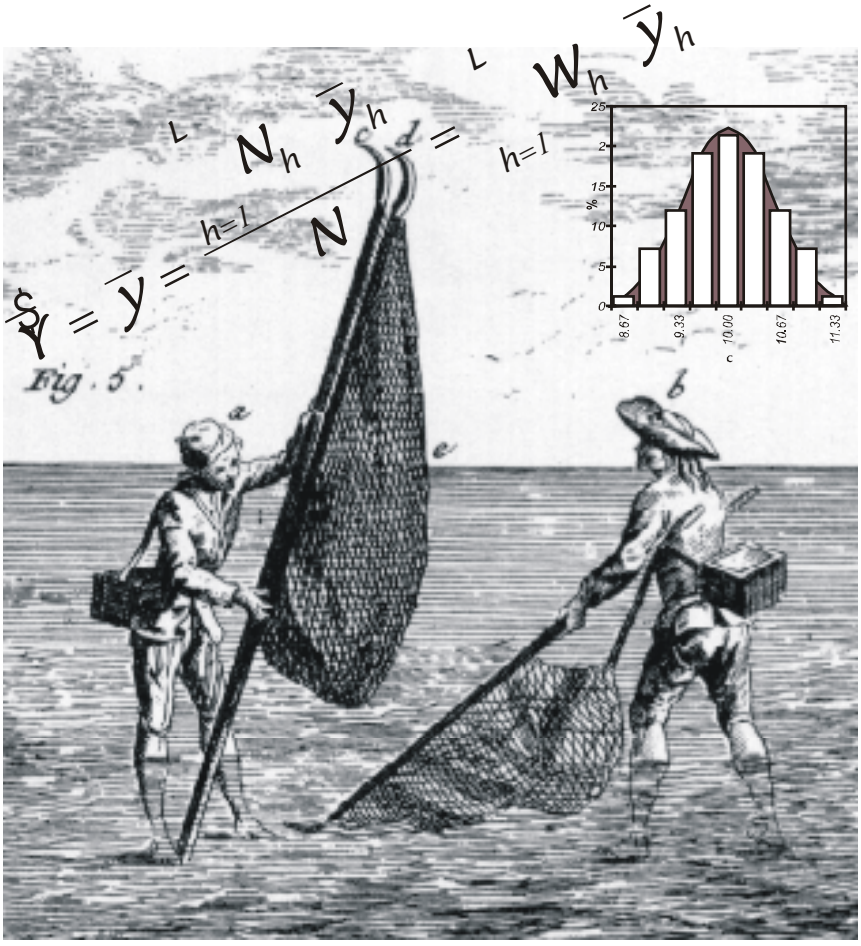
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
Τμήμα Βιολογίας

ΔΕΙΓΜΑΤΟΛΗΨΙΑ

Βασικές έννοιες
και
Εφαρμογές στην Οικολογία



Κωνσταντίνος ΚΟΥΤΣΙΚΟΠΟΥΛΟΣ
Αναπληρωτής Καθηγητής

Πάτρα, 2002

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΔΕΙΓΜΑΤΟΛΗΨΙΑ: ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

1.1. ΠΕΡΙ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

Ο ορισμός της αλιεύσιμης ποσότητας συγκεκριμένου είδους ψαριού σε μια περιοχή, η τοποθέτηση και ο χρονισμός των φωτεινών σηματοδοτών στο οδικό δίκτυο, η εμφάνιση νέου προϊόντος στην αγορά, η μελέτη της συγκέντρωσης χοληστερίνης στον άρρενα πληθυσμό ενός νομού, βασίζονται σε εκτιμήσεις της τρέχουσας κατάστασης των συγκεκριμένων παραμέτρων. Οι εκτιμήσεις αυτές βασίζονται ή καλύτερα προέρχονται από την ανάλυση δειγμάτων. Το **δείγμα** είναι ένα μέρος ενός συνόλου και η μελέτη των χαρακτηριστικών του μας πληροφορεί για τις ιδιότητες του συνόλου. Είναι φανερό από τη σπουδαιότητα των παραπάνω παραδειγμάτων ότι το δείγμα πρέπει να μας πληροφορήσει όσο το δυνατό καλύτερα για την πραγματική κατάσταση που επικρατεί στο σύνολο. Άρα η ποιότητα του δείγματος είναι καθοριστική για το αποτέλεσμα των μελετών.

Για την εκτίμηση μιας παραμέτρου (ενός χαρακτηριστικού) υπάρχουν δυο δυνατότητες: η αναλυτική μελέτη όλων των ατόμων που αποτελούν το σύνολο (μιλάμε πλέον για απαρίθμηση) ή η επιλογή ενός μέρος του πληθυσμού (δείγμα) και η ανάλυση των χαρακτηριστικών του. Στην πρώτη περίπτωση είναι προφανές ότι δεν υπάρχει πιθανότητα λάθους στην εκτίμηση. Γνωρίζουμε την ακριβή πραγματικότητα. Από την άλλη πλευρά η διαδικασία της δειγματοληψίας έχει ένα σοβαρό πλεονέκτημα: τη σημαντική μείωση του κόστους. Η εκτίμηση μπορεί να έχει κάπως μειωμένη ακρίβεια, όμως ο λόγος ποιότητα/κόστος της εκτίμησης μέσω δείγματος είναι συχνότατα πολύ ελκυστικός. Υπάρχουν όμως και άλλα πλεονεκτήματα στη διαδικασία της δειγματοληψίας και αυτά είναι η ταχύτητα διεξαγωγής της μελέτης λόγω του περιορισμένου μεγέθους του δείγματος αλλά και η συχνά μεγαλύτερη ακρίβεια των επί μέρους

μετρήσεων λόγω της χρησιμοποίησης εξειδικευμένου προσωπικού και μηχανημάτων που είναι διαθέσιμα μόνο σε μικρή κλίμακα. Τέλος η δειγματοληψία είναι υποχρεωτική στην περίπτωση που τα υπό μελέτη άτομα καταστρέφονται όπως στην περίπτωση της αλιείας, του κυνηγίου κ.λ.π..

1.2. Ο ΠΛΗΘΥΣΜΟΣ, ΤΟ ΔΕΙΓΜΑ, Ο ΕΚΤΙΜΗΤΗΣ: ΟΡΙΣΜΟΙ - ΣΥΜΒΟΛΙΣΜΟΙ

Στη διαδικασία της δειγματοληψίας επιλέγουμε ένα μέρος από τον πληθυσμό του οποίου κάποιο χαρακτηριστικό θέλουμε να εκτιμήσουμε. Το χαρακτηριστικό αυτό είναι μια **παράμετρος** η οποία χαρακτηρίζει *όλα ανεξαιρέτως τα άτομα* του πληθυσμού. Στη θεωρία της δειγματοληψίας ο όρος **πληθυσμός** δεν υποδηλώνει υποχρεωτικά ένα βιολογικό πληθυσμό. Είναι απλά ένα καθορισμένο εξ αρχής σύνολο ατόμων, ο δε όρος *άτομο* δεν υποδηλώνει υποχρεωτικά ένα ζώντα οργανισμό. Έτσι η εκτίμηση της μέσης ετήσιας δαπάνης για ιατροφαρμακευτική περίθαλψη ανά οικογένεια του νομού Αχαΐας μπορεί να είναι το αντικείμενο της δειγματοληψίας. Σ'αυτή την περίπτωση το άτομο είναι η οικογένεια, ο πληθυσμός είναι το σύνολο των οικογενειών και η παράμετρος είναι η ετήσια δαπάνη.

Οι τιμές της παραμέτρου στα N άτομα του πληθυσμού συμβολίζονται $y_1, y_2,...,y_N$. Οι αντίστοιχες τιμές της παραμέτρου για τα n άτομα που αποτελούν το δείγμα συμβολίζονται $y_1, y_2,...,y_n$. Χρησιμοποιούμε το συμβολισμό y_i (i=1, 2,...,n) όταν αναφερόμαστε σε ένα τυπικό άτομο του δείγματος. Προσοχή το δείγμα δεν περιέχει υποχρεωτικά τα n πρώτα άτομα του πληθυσμού αλλά n τυχαία ή με κάποιο άλλο τρόπο επιλεγμένα μέλη του πληθυσμού. Ισχύει λοιπόν $0 < n \leq N$.

Συνήθως δύο είναι τα κύρια χαρακτηριστικά που προσπαθούμε να εκτιμήσουμε σε ένα πληθυσμό: η μέση τιμή $\overline{Y}$ και το σύνολο των ατόμων ή αφθονία του πληθυσμού Y. Παράδειγμα είναι η εκτίμηση της μέσης κατανάλωσης συγκεκριμένης τροφής ανά άτομο ή η συνολική ποσότητα τροφής που καταναλίσκεται από ένα πληθυσμό. Άλλα χαρακτηριστικά που μπορεί να βρίσκονται στο επίκεντρο της εκτίμησης είναι ο λόγος R δύο παραμέτρων π.χ. ο λόγος ενηλίκων ατόμων προς το σύνολο των ατόμων ενός πληθυσμού ή ακόμα η αναλογία ατόμων που φέρουν συγκεκριμένο παράσιτο.

Μια δεύτερη βασική παράμετρος που χαρακτηρίζει ένα πληθυσμό είναι η διασπορά ατόμων y_i γύρω από τη μέση τιμή $\overline{Y}$ του πληθυσμού. Με άλλα λόγια είναι ενδιαφέρον να γνωρίζουμε κατά πόσο τα άτομα μοιάζουν μεταξύ τους (ως προς την υπό μελέτη παράμετρο) ή υπάρχει ετερογένεια στον πληθυσμό. Το μέγεθος αυτής της ετερογένειας, η **διασπορά** (variance) **του πληθυσμού** ορίζεται ως

$$\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N} \qquad (1.1)$$

Είναι φανερό από τον τύπο ότι η διασπορά δεν είναι τίποτε άλλο παρά η μέση απόκλιση των ατόμων από τη μέση τιμή (πιο συγκεκριμένα το τετράγωνο της). Επειδή είναι σπάνια η περίπτωση που όλα τα άτομα του πληθυσμού αναλύονται αλλά συνήθως η γνώση μας προέρχεται από δείγμα, η διασπορά του δείγματος δίνεται από τον ακόλουθο τύπο.

$$\boxed{s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \qquad (1.2)$$

Οι δύο τύποι διαφέρουν στον παρονομαστή του κλάσματος. Η **διασπορά του δείγματος** (s^2)είναι μια εκτίμηση της διασποράς του πληθυσμού (δηλαδή από το δείγμα μπορούμε να γνωρίζουμε πόσο ετερογενής είναι ο πληθυσμός) και αποδεικνύεται ότι αν εφαρμοσθεί ο τύπος 1.1 για το δείγμα τότε μας δίνει μια εκτίμηση της διασποράς του πληθυσμού που είναι συστηματικά μικρότερη από την πραγματική (σ^2). Άρα διαιρώντας δια n-1 έχουμε μια τιμή μεγαλύτερη που πλησιάζει περισσότερο στην πραγματική τιμή της πραγματικής διασποράς του πληθυσμού (περισσότερα σχόλεια σε επόμενα κεφάλαια). Επειδή η διασπορά χρησιμοποιείται συχνά παρουσιάζεται και η ακόλουθη ισότιμη μορφή της 1.2 που διευκολύνει τον υπολογισμό της παραμέτρου

$$s^2 = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} y_i}{n}}{n-1} \qquad (1.3)$$

Η **τυπική απόκλιση** σ και για το δείγμα s (standard deviation) είναι η τετραγωνική ρίζα της διασποράς του πληθυσμού και του δείγματος αντίστοιχα.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \qquad (1.4)$$

Στη συνέχεια με κεφαλαία γράμματα θα συμβολίζονται χαρακτηριστικά του πληθυσμού και με μικρά τα χαρακτηριστικά του δείγματος. Το σύμβολο ^ υποδεικνύει εκτίμηση του χαρακτηριστικού του πληθυσμού που προέρχεται από την ανάλυση δείγματος. Η μαθηματική έκφραση που μας επιτρέπει να υπολογίσουμε παράμετρο του πληθυσμού από τα δεδομένα του δείγματος ονομάζεται **εκτιμητής**. Τα ακόλουθα σύμβολα λοιπόν αντιπροσωπεύουν

αντίστοιχα την εκτίμηση της μέσης τιμής, του συνόλου και του λόγου δυο χαρακτηριστικών του πληθυσμού $\hat{\bar{Y}}$, $\hat{Y}$, $\hat{R}$. Έτσι ο τύπος $\bar{y}=\sum_{i=1}^{n} y_i/n$ που δίνει τη μέση τιμή του δείγματος, είναι ο εκτιμητής $\hat{\bar{Y}}$ της μέσης τιμής $\bar{Y}$ του πληθυσμού που συνήθως είναι άγνωστη.

Πληθυσμός	Δείγμα
Αριθμός ατόμων : N	n
Σύνολο : $Y = \sum_{i=1}^{N} y_i = y_1 + y_2 + \cdots + y_N$	$\sum_{i=1}^{n} y_i = y_1 + y_2 + \cdots + y_n$
Μέση τιμή : $\mu = \bar{Y} = \frac{y_1 + y_2 + \cdots + y_N}{N} = \frac{\sum_{i=1}^{N} y_i}{N}$	$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n} = \hat{\bar{Y}}$
Διασπορά : $\sigma^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{Y})^2}{n-1} = \hat{\sigma}^2$
Τυπική απόκλιση : σ	s

1.3. Η ΘΕΩΡΙΑ ΤΗΣ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΚΑΙ Η ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ

Η θεωρία της δειγματοληψίας καθορίζει τη μέθοδο επιλογής του δείγματος και τον τρόπο υπολογισμού των χαρακτηριστικών του πληθυσμού από τα δεδομένα του δείγματος

Η θεωρία της δειγματοληψίας έχει δύο στόχους. Πρώτον να καθορίσει επακριβώς τη μεθοδολογία για την επιλογή του δείγματος (σχεδιασμός) και δεύτερον να ορίσει με ποιο τρόπο τα χαρακτηριστικά του δείγματος θα μας πληροφορήσουν για τις ιδιότητες του συνόλου (συμπερασματολογία). Σε κάθε περίπτωση ο στόχος είναι πως θα αξιοποιηθεί όλη η υπάρχουσα πληροφορία ώστε να μειωθεί το κόστος της μελέτης χωρίς να επηρεάζεται η ακρίβεια των εκτιμήσεων.

Ειδικά στη μελέτη της φύσης όπου τα φαινόμενα χαρακτηρίζονται από έντονες διακυμάνσεις (εποχικές αλλαγές, ημερονύκτιοι ρυθμοί, τάσεις, κ.λ.π.) και η παρατήρηση είναι δύσκολη και συχνά υψηλού κόστους, η ποιότητα του δείγματος είναι καθοριστική για την εκτίμηση της τρέχουσας κατάστασης. Περιέργως όμως στην πλειοψηφία τους τα προγράμματα σπουδών αλλά και τα διδακτικά βιβλία ασχολούνται με την ανάλυση δεδομένων και τη στατιστική συμπερασματολογία, ασχολούνται δηλαδή με τα στάδια που ακολουθούν τη συλλογή και ανάλυση του δείγματος. Πολύ λίγα παρατίθενται για τη διαδικασία καθορισμού και συλλογής του δείγματος. άμως η αλήθεια είναι ότι ένα κακό δείγμα δίνει πάντα αναξιόπιστες ή και ανακριβείς πληροφορίες όσο εξεζητημένη κι αν είναι η μέθοδος στατιστικής ανάλυσης των αποτελεσμάτων. Και κάτι που διαφεύγει της προσοχής. Η πανοπλία των στατιστικών μεθόδων

που συχνά χρησιμοποιούμε είναι αναγκαία μόνο και μόνο γιατί οι μελέτες μας βασίζονται σε δείγματα. Για παράδειγμα, εάν όλα τα άτομα δύο πληθυσμών είχαν μετρηθεί ένα προς ένα δεν θα υπήρχε λόγος στατιστικών συγκρίσεων (test) για ν'αποφανθούμε για τις τυχόν διαφορές ανάμεσα στους πληθυσμούς. Η απάντηση θα ήταν ξεκάθαρη χωρίς πιθανότητα λάθους.

1.4. ΤΟ ΑΝΤΙΠΡΟΣΩΠΕΥΤΙΚΟ ΔΕΙΓΜΑ

Από την εποχή που οι δημοσκοπήσεις της κοινής γνώμης έγιναν πια καθημερινή υπόθεση στα μέσα μαζικής ενημέρωσης, η φράση "η μελέτη έγινε σε δείγμα n ατόμων αντιπροσωπευτικό του πληθυσμού" περνά απαρατήρητη. Στην προηγούμενη παράγραφο έγινε λόγος για το "καλό δείγμα". Αυθόρμητα θα λέγαμε ότι ένα "αντιπροσωπευτικό" ή "καλό" δείγμα είναι αυτό που μας δίνει αποτελέσματα που βρίσκονται πολύ κοντά στην πραγματικότητα π.χ. το μέσο ύψος των ατόμων του δείγματος είναι το ίδιο με το μέσο ύψος των ατόμων του πληθυσμού. Το πρόβλημα όμως είναι ότι την πραγματική τιμή του πληθυσμού δεν την γνωρίζουμε αφού για να τη μάθουμε γίνεται η μελέτη. Έτσι περιοριζόμαστε στο να φροντίζουμε ώστε η σύνθεση του δείγματος, αναφορικά με χαρακτηριστικά που μπορούν να επηρεάσουν την τιμή της προς μελέτη παραμέτρου, να είναι ίδια με αυτή του πληθυσμού. Εάν για παράδειγμα το φύλλο επηρεάζει το ύψος των ατόμων και κατά μέσο όρο τα αγόρια είναι ψηλότερα από τα κορίτσια, τότε για την εκτίμηση του μέσου ύψους των φοιτητών του Πανεπιστημίου το δείγμα μας πρέπει να περιέχει την ίδια αναλογία των δύο φύλλων με αυτή του πληθυσμού των φοιτητών του συγκεκριμένου Πανεπιστημίου. Τα άτομα δε του δείγματος πρέπει να έχουν επιλεγεί με τέτοιο τρόπο ώστε η πιθανότητα των φοιτητών να εμφανιστούν στο δείγμα να είναι ίση για όλους. Αυτό είναι ένα αντιπροσωπευτικό δείγμα. Κατά συνέπεια το αντιπροσωπευτικό δείγμα των κοινών δημοσκοπήσεων είναι αυτό του οποίου η ποσοστιαία σύνθεση όσον αφορά στο επίπεδο εκπαίδευσης, εισοδημάτων, ηλικίας καθώς και στην επαγγελματική σύνθεση είναι ίδια με αυτή του πληθυσμού "στόχου".

1.5. ΤΟ ΔΕΙΓΜΑ ΚΑΙ ΟΙ ΕΠΙΠΤΩΣΕΙΣ ΤΟΥ

ΠΑΡΑΔΕΙΓΜΑ 1.1

Για να γίνουν κατανοητές οι συνέπειες της δειγματοληψίας και η διαδικασία εκτίμησης ας θεωρήσουμε το μικρό σύνολο ατόμων του πίνακα 1.1 που θα το θεωρήσουμε ως τον πληθυσμό. Οι αριθμοί αυτοί μπορεί να αντιπροσωπεύουν την ηλικία ατόμων, τον αριθμό καρπών ανά φυτό ή ακόμα και τον αριθμό αγροκτημάτων ανά τετραγωνικό χιλιόμετρο. Η μέση τιμή του πληθυσμού αυτού είναι $\overline{Y}=\sum_{i=1}^{N} y_i/N=10.0$ και η διασπορά των τιμών γύρω από τη μέση τιμή $\sigma^2=1.333$. Ας υποθέσουμε ότι από τον πληθυσμό αυτό επιλέγουμε τυχαία 3 άτομα που αποτελούν το δείγμα. Τα άτομα αυτά είναι 9, 11, 11 (α/α:47,

ΠΙΝΑΚΑΣ 1.1 *Πληθυσμός αποτελούμενος από 9 άτομα που χαρακτηριζονται από τις τιμές y συγκεκριμένης παραμέτρου.*

y_i
8,9,9,10,10,10,11,11,12
N=9
$\bar{Y}$=10.0
σ^2=1.333

πίνακας 1.2). Καλούμαστε από το δείγμα αυτό να βγάλουμε συμπεράσματα για τη μέση τιμή του πληθυσμού. Χρησιμοποιούμε γι'αυτό τη μέση τιμή του δείγματος που είναι $\bar{y}$=10.33. Αν όμως το δείγμα μας το αποτελούσαν τα άτομα 8, 10, 10, τότε η μέση τιμή θα ήταν 9.33 (α/α: 19). Ποια από τις δύο τιμές είναι η καλύτερη (η σωστότερη); Η πραγματική μέση τιμή του πληθυσμού είναι κατά κανόνα άγνωστη κι έτσι δυστυχώς δεν υπάρχει κριτήριο επιλογής. Ο πληθυσμός αυτός των 9 ατόμων μπορεί να δώσει 84 διαφορετικούς συνδυασμούς 3 ατόμων, δηλαδή 84 διαφορετικά δείγματα[1].

Κάθε ένας από τους συνδυασμούς αυτούς έχει την ίδια με τους άλλους πιθανότητα να αποτελεί το δείγμα μας. Κάθε ένας δε από τους συνδυασμούς αυτούς έχει διαφορετική μέση τιμή άρα δίνει και διαφορετική εκτίμηση για την πραγματική μέση τιμή που μας ενδιαφέρει, αυτή του πληθυσμού. Έτσι ανακαλύπτουμε ότι αφού η εκτίμηση κυμαίνεται από το ένα δείγμα στο άλλο ο εκτιμητής είναι μια τυχαία μεταβλητή η κατανομή της οποίας ονομάζεται **κατανομή δειγματοληψίας** (sampling distribution). Ο πίνακας 1.2 παρουσιάζει αναλυτικά τις μέσες τιμές των 84 αυτών δειγμάτων.

ΠΙΝΑΚΑΣ 1.2 - Όλα τα δυνατά δείγματα μεγέθους 3 ατόμων που μπορούν να δημιουργηθούν από συνδυασμό των ατόμων του πληθυσμού του παραδείγματος 1.1. Για κάθε δείγμα παρουσιάζεται ο αύξων αριθμός, τα άτομα του δείγματος καθώς και η μέση τιμή και η διασπορά των δειγμάτων

α/α	δείγμα (y_1, y_2, y_3)			μέση τιμή $\bar{y}c$	διασπορά s^2	α/α	δείγμα (y_1, y_2, y_3)			μέση τιμή $\bar{y}c$	διασπορά s^2	α/α	δείγμα (y_1, y_2, y_3)			μέση τιμή $\bar{y}c$	διασπορά s^2
1	8	9	9	8.667	0.333	29	9	9	10	9.333	0.333	57	10	10	11	10.000	1.000
2	8	9	10	9.000	1.000	30	9	9	10	9.333	0.333	58	10	10	12	10.333	2.333
3	8	9	10	9.000	1.000	31	9	9	10	9.333	0.333	59	10	10	11	10.000	1.000
4	8	9	10	9.000	1.000	32	9	9	11	9.667	1.333	60	10	10	11	10.000	1.000
5	8	9	11	9.333	2.333	33	9	9	11	9.667	1.333	61	10	10	12	10.333	2.333
6	8	9	11	9.333	2.333	34	9	9	12	10.000	3.000	62	10	11	11	10.333	1.333
7	8	9	12	9.667	4.333	35	9	10	10	9.667	0.333	63	10	11	12	10.667	2.333
8	8	9	10	9.000	1.000	36	9	10	10	9.667	0.333	64	10	11	12	10.667	2.333
9	8	9	10	9.000	1.000	37	9	10	11	10.000	1.000	65	10	10	10	10.000	0.000
10	8	9	10	9.000	1.000	38	9	10	11	10.000	1.000	66	10	10	11	10.333	0.333
11	8	9	11	9.333	2.333	39	9	10	12	10.333	2.333	67	10	10	11	10.333	0.333
12	8	9	11	9.333	2.333	40	9	10	10	9.667	0.333	68	10	10	12	10.667	1.333
13	8	9	12	9.667	4.333	41	9	10	11	10.000	1.000	69	10	10	11	10.333	0.333
14	8	10	10	9.333	1.333	42	9	10	11	10.000	1.000	70	10	10	11	10.333	0.333
15	8	10	10	9.333	1.333	43	9	10	12	10.333	2.333	71	10	10	12	10.667	1.333
16	8	10	11	9.667	2.333	44	9	10	11	10.000	1.000	72	10	11	11	10.667	0.333
17	8	10	11	9.667	2.333	45	9	10	11	10.000	1.000	73	10	11	12	11.000	1.000
18	8	10	12	10.000	4.000	46	9	10	12	10.333	2.333	74	10	11	12	11.000	1.000
19	8	10	10	9.333	1.333	47	9	11	11	10.333	1.333	75	10	10	11	10.333	0.333
20	8	10	11	9.667	2.333	48	9	11	12	10.667	2.333	76	10	10	11	10.333	0.333
21	8	10	11	9.667	2.333	49	9	11	12	10.667	2.333	77	10	10	12	10.667	1.333
22	8	10	12	10.000	4.000	50	10	10	10	9.667	0.333	78	10	11	11	10.667	0.333
23	8	10	11	9.667	2.333	51	10	10	10	9.667	0.333	79	10	11	12	11.000	1.000
24	8	10	11	9.667	2.333	52	10	10	11	10.000	1.000	80	10	11	12	11.000	1.000
25	8	10	12	10.000	4.000	53	10	10	11	10.000	1.000	81	10	11	11	10.667	0.333
26	8	11	11	10.000	3.000	54	10	10	12	10.333	2.333	82	10	11	12	11.000	1.000
27	8	11	12	10.333	4.333	55	10	10	10	9.667	0.333	83	10	11	12	11.000	1.000
28	8	11	12	10.333	4.333	56	10	10	11	10.000	1.000	84	11	11	12	11.333	0.333

μέση τιμή των μέσων τιμών $\bar{Y}$ = 10.0, διασπορά μέσων τιμών $\sigma_{\bar{y}}^2$ = 0.333

Παρατηρούμε ότι:

[1] Οι συνδυασμοί n από N είναι $_N C_n = \frac{N!}{n!(N-n)!}$

- αν και οι μέσες τιμές κυμαίνονται από το ένα δείγμα στο άλλο, η μέση τιμή των μέσων τιμών είναι 10, η ίδια δηλαδή με αυτή του πληθυσμού
- η διασπορά των μέσων τιμών γύρο από την πραγματική μέση τιμή είναι μικρότερη από αυτή των ατόμων του πληθυσμού γύρο από τη μέση τους τιμή (0.333 και 1.333 αντίστοιχα).

Αν αντί για δείγμα 3 ατόμων επιλέγαμε δείγμα 6 ατόμων ποιες θα ήταν οι συνέπειες; Αυθόρμητα θα λέγαμε ότι όσο μεγαλύτερο είναι το δείγμα τόσο καλύτερη, ακριβέστερη, σωστότερη θα είναι η εκτίμηση. Ο πίνακας 1.3 παρουσιάζει τις μέσες τιμές όλων των δυνατών δειγμάτων μεγέθους 6 ατόμων από τον αρχικό πληθυσμό των 9. Οι συνδυασμοί 6 από 9 είναι πάλι 84. Παρατηρούμε ότι και σε αυτή την περίπτωση η μέση τιμή των μέσων τιμών είναι 10 αλλά η διασπορά τους (η διασπορά του εκτιμητή αφού η μέση τιμή του δείγματος είναι ο εκτιμητής της πραγματικής μέσης τιμής του πληθυσμού) είναι 0.083, μικρότερη δηλαδή από τη διασπορά των μέσων τιμών των δειγμάτων μεγέθους 3 ατόμων.

ΠΙΝΑΚΑΣ 1.3 - Όλα τα δυνατά δείγματα μεγέθους 6 ατόμων που μπορούν να δημιουργηθούν από συνδυασμό των ατόμων του πληθυσμού του παραδείγματος 1.1. Για κάθε δείγμα παρουσιάζεται ο αύξων αριθμός, τα άτομα του δείγματος καθώς και η μέση τιμή και η διασπορά των δειγμάτων

α/α	μέση τιμή $\bar{y}c$	διασπορά s^2	α/α	μέση τιμή $\bar{y}$	διασπορά s^2	α/α	μέση τιμή $\bar{y}$	διασπορά s^2	α/α	μέση τιμή $\bar{y}$	διασπορά s^2
1	9.333	0.667	22	9.667	1.067	43	10.000	2.000	64	10.167	1.367
2	9.500	1.100	23	9.833	1.767	44	10.000	2.000	65	10.167	1.367
3	9.500	1.100	24	9.833	1.367	45	10.167	2.167	66	10.333	1.467
4	9.667	1.867	25	10.000	2.000	46	9.833	1.367	67	10.000	0.800
5	9.500	1.100	26	10.000	2.000	47	10.000	2.000	68	10.167	1.367
6	9.500	1.100	27	9.833	1.367	48	10.000	2.000	69	10.167	1.367
7	9.667	1.867	28	10.000	2.000	49	10.167	2.167	70	10.333	1.467
8	9.667	1.467	29	10.000	2.000	50	10.167	2.167	71	10.333	1.467
9	9.833	2.167	30	10.167	2.167	51	10.000	1.200	72	10.167	0.567
10	9.833	2.167	31	9.833	1.367	52	10.167	1.767	73	10.333	1.067
11	9.500	1.100	32	10.000	2.000	53	10.167	1.767	74	10.333	1.067
12	9.500	1.100	33	10.000	2.000	54	10.333	1.867	75	10.500	1.100
13	9.667	1.867	34	10.167	2.167	55	10.333	1.867	76	10.500	1.100
14	9.667	1.467	35	10.167	2.167	56	10.333	1.867	77	10.500	1.100
15	9.833	2.167	36	9.667	1.067	57	9.833	0.567	78	10.167	0.567
16	9.833	2.167	37	9.667	1.067	58	9.833	0.567	79	10.333	1.067
17	9.667	1.467	38	9.833	1.767	59	10.000	1.200	80	10.333	1.067
18	9.833	2.167	39	9.833	1.367	60	10.000	0.800	81	10.500	1.100
19	9.833	2.167	40	10.000	2.000	61	10.167	1.367	82	10.500	1.100
20	10.000	2.400	41	10.000	2.000	62	10.167	1.367	83	10.500	1.100
21	9.667	1.067	42	9.833	1.367	63	10.000	0.800	84	10.667	0.667

μέση τιμή των μέσων τιμών $\bar{Y} = 10.0$, διασπορά μέσων τιμών $\sigma^2_{\bar{y}} = 0.0833$

Πράγματι οι διαφορές φαίνονται και στο εύρος διακύμανσης των μέσων τιμών των δειγμάτων γύρο από τη μέση τιμή τους. Στην πρώτη περίπτωση των δειγμάτων μεγέθους 3 οι τιμές κυμαίνονται από 8.67 έως 11.33 ενώ για το δείγμα μεγέθους 6 από 9.33 έως 10.67. Οι διαφορές αυτές εμφανίζονται αντίστοιχα και στο μέγεθος της διασποράς των μέσων τιμών που είναι 0.333 και 0.083 αντίστοιχα.

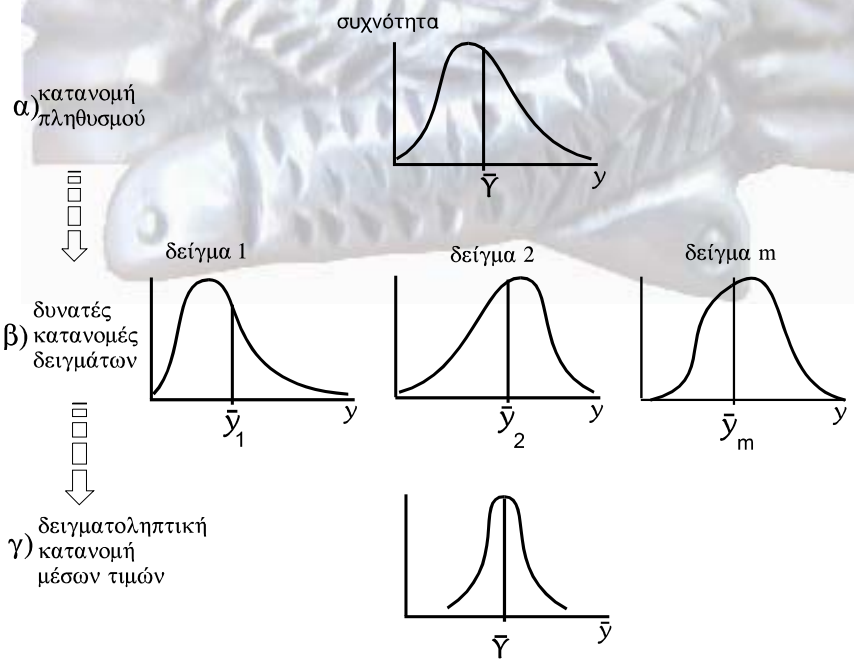
άσο λοιπόν το δείγμα μεγαλώνει τόσο οι εκτιμήσεις πλησιάζουν προς την πραγματική τιμή της παραμέτρου στον πληθυσμό.

ΠΑΡΑΔΕΙΓΜΑ 1.1

1.6. Η ΚΑΤΑΝΟΜΗ ΤΩΝ ΜΕΣΩΝ ΤΙΜΩΝ ΚΑΙ ΤΩΝ ΔΙΑΣΠΟΡΩΝ ΤΩΝ ΔΕΙΓΜΑΤΩΝ

Η έννοια της κατανομής δειγματοληψίας μπορεί να δημιουργήσει προβλήματα στην κατανόηση της. Μερικές διευκρινήσεις χρειάζονται. Η κατανομή των τιμών μιας παραμέτρου που χαρακτηρίζει τα διάφορα άτομα ενός πληθυσμού μπορεί να έχει διάφορες μορφές. Για παράδειγμα η κατανομή της κατ'άτομο κατανάλωσης τροφής ενός συνόλου ζώων μιας παραγωγικής μονάδας που αποτελούν τον πληθυσμό (σχήμα 1.1 α).

Απ'αυτόν τον πληθυσμό μπορούν να επιλεγούν δείγματα n ατόμων. Για κάθε δείγμα η κατ'άτομο κατανάλωση έχει μια κατανομή που μπορεί να πάρει διάφορες μορφές ανάλογα με τα άτομα που επιλέγηκαν (σχήμα 1.1 β). Κάθε δείγμα έχει τη δική του μέση τιμή και τυπική απόκλιση. Φυσικά υπάρχουν τόσες κατανομές δειγμάτων όσα και τα δυνατά δείγματα. Για παράδειγμα εάν η μονάδα είχε 50 ζώα και επιλέγαμε δείγματα 8 ατόμων, θα υπήρχαν περίπου 537 εκατομμύρια διαφορετικά δείγματα. Θα υπήρχαν λοιπόν 537 εκατ. δυνατές κατανομές δειγμάτων και 537 εκατ. μέσες τιμές.

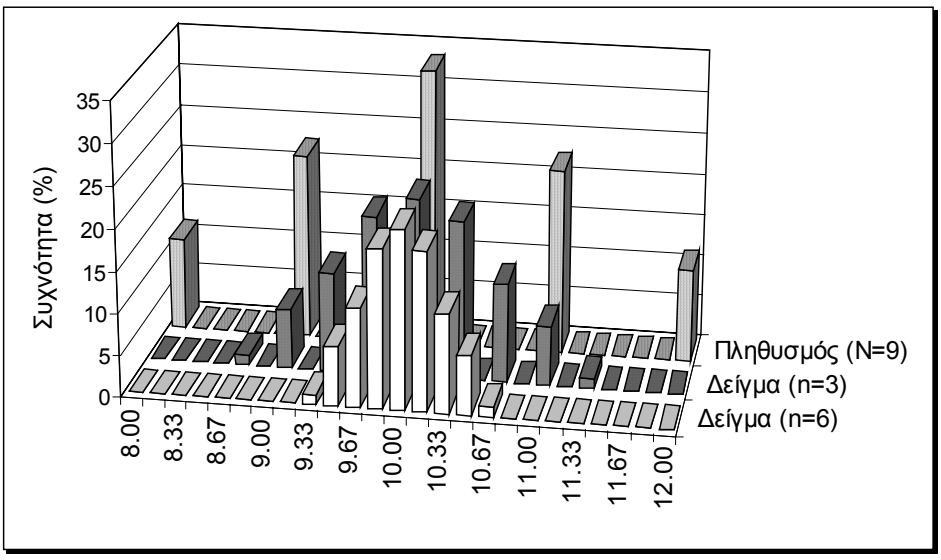


Σχημα 1.1 - *Διαγραμματική παρουσιαση της σχέσης ανάμεσα στην κατανομή των ατομών στον πληθυσμό (α), στα δείγματα (β) και στη δειγματοληπτική κατανομή (γ)*

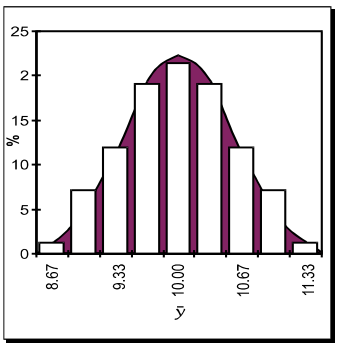
Η κατανομή των μέσων τιμών αυτών που αποτελεί την δειγματοληπτική κατανομή είναι *μία και μοναδική* σε αντίθεση με τις κατανομές των δειγμάτων. Στην δειγματοληπτική κατανομή οι δυνατές μέσες τιμές δειγμάτων συγκεκριμένου μεγέθους κατανέμονται γύρο από τη μέση τιμή της δειγματοληπτικής

κατανομής που όπως είδαμε συμπίπτει με την πραγματική μέση τιμή της παραμέτρου στον πληθυσμό (σχήμα 1.1 γ). Υπολογίζεται αθροίζοντας όλες τις δυνατές μέσες τιμές και διαιρώντας δια του συνολικού αριθμού δυνατών δειγμάτων όπως φαίνεται και από τους πίνακες 1.2 και 1.3 . Η τυπική απόκλιση της δειγματοληπτικής κατανομής μετρά την απόκλιση των μέσων τιμών των δειγμάτων γύρο από τη μέση τιμή της κατανομής. Από το παράδειγμα της προηγουμένης παραγράφου φαίνεται ότι το εύρος της κατανομής των μέσων τιμών των δειγμάτων μεγέθους 6 ατόμων είναι πιο μικρό από το εύρος της κατανομής των δειγμάτων 3 ατόμων. Με άλλα λόγια οι μέσες τιμές των μεγάλων δειγμάτων βρίσκονται πιο κοντά στην πραγματική μέση τιμή. Αυθόρμητα σκεφτόμαστε ότι ένα μεγάλο δείγμα είναι καλύτερο από ένα μικρό.

Το διάγραμμα 1.2 δείχνει την κατανομή των τιμών του αρχικού πληθυσμού καθώς και τις κατανομές των μέσων τιμών όλων των δυνατών δειγμάτων μεγέθους 3 και 6 ατόμων. Είναι φανερό ότι αν παίρναμε μόνο ένα άτομο από τον αρχικό πληθυσμό για να μας πληροφορήσει για τα χαρακτηριστικά του πληθυσμού, τότε η πιθανότητα να βρεθούμε κοντά στην πραγματική μέση τιμή (10.0) είναι σχετικά περιορισμένη.



ΣΧΗΜΑ 1.2 Κατανομή των ατόμων του πληθυσμού του παραδείγματος 1, καθώς και οι κατανομές των μέσων τιμών όλων των δυνατών δειγμάτων μεγέθους 3 και 6 ατόμων από τον πληθυσμό αυτόν (δειγματοληπτικές κατανομές).



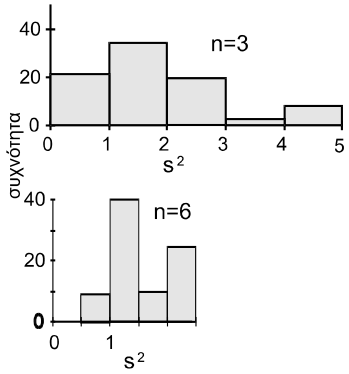
Σχημα 1.3 Σύγκριση της δειγματοληπτικής κατανομής δειγμάτων μεγέθους 3 ατόμων από τον πληθυσμό του παραδειγματος 1.1 με την αντιστοιχη κανονική κατανομή.

Αυξάνοντας το μέγεθος του δείγματος, η πιθανότητα αυτή αυξάνει και φυσικά γίνεται 1 (100%) όταν το δείγμα περιέχει όλα τα άτομα του πληθυσμού (n=N). Φυσικά σ'αυτή την περίπτωση δεν μιλάμε για δειγματοληψία αλλά για απαρίθμηση και η υπολογισθείσα μέση τιμή έχει διασπορά 0 αφού μόνο ένα τέτοιο δείγμα υπάρχει. Πως όμως συνδέεται η διασπορά των ατόμων του πληθυσμού με τη διασπορά των μέσων τιμών των δειγμάτων; Η σχέση αυτή είναι

$$\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \qquad (1.5)$$

Η τετραγωνική ρίζα της διασποράς αυτής είναι η τυπική απόκλιση των μέσων τιμών των δειγμάτων από την πραγματική μέση τιμή του πληθυσμού και ονομάζεται **τυπικό σφάλμα** ($\sigma_{\bar{y}}$, standard error).

$$\sigma_{\bar{y}} = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} \qquad (1.6)$$



ΣΧΗΜΑ 1.4 *Κατανομή των διασπορών των 84 δυνατών δειγμάτων μεγέθους 3 και 6 ατόμων από τον πληθυσμό του παραδείγματος* 1.1.

άταν το n τείνει προς το N τότε η τιμή της παρένθεσης τείνει στο 0 και έτσι ικανοποιείται η μηδενική διασπορά του εκτιμητή για την περίπτωση της απαρίθμησης (n=N) που αναφέρθηκε προηγουμένως. άταν δε το n=1 τότε η διασπορά των μέσων τιμών είναι ίση με αυτή του πληθυσμού αφού ουσιαστικά έχουμε δείγματα 1 ατόμου. Η έκφραση της παρένθεσης (N-n)/(N-1) είναι γνωστή σα **διόρθωση πεπερασμένου πληθυσμού** (finite population correction).

Η σχέση 1.5 συχνά εμφανίζεται κυρίως στα βιβλία της στατιστικής απλά σαν $\sigma_{\bar{y}}^2=\sigma^2/n$. Αυτό συμβαίνει διότι οι συγγραφείς αναφέρονται σε πληθυσμό με άπειρο μέγεθος (ή στην πράξη πολύ μεγαλύτερο από το n) και κατά συνέπεια η τιμή της διόρθωσης του πεπερασμένου πληθυσμού είναι ίση με 1 οπότε και παραλείπεται. Από τη σχέση 1.5 είναι φανερό ότι όσο μεγαλώνει το δείγμα (n) τόσο μειώνεται η διασπορά των μέσων τιμών.

Ένα δεύτερο στοιχείο εξάγεται από την παρατήρηση του διαγράμματος 1.2 . Η κατανομή των μέσων τιμών εμφανίζεται συμμετρική γύρω από την πραγματική μέση τιμή και επί πλέον η εμφάνιση της μοιάζει με την κανονική κατανομή (normal distribution). Αυτό είναι καθαρότερο στο διάγραμμα 1.3 όπου η δειγματοληπτική κατανομή συγκρίνεται με την αντίστοιχη κανονική κατανομή. Σύμφωνα με το **κεντρικό οριακό θεώρημα** (central limit theoreme) όσο το n και το N-n μεγαλώνουν, τόσο η κατανομή των μέσων τιμών των δειγμάτων προσεγγίζει την κανονική κατανομή όποια κι αν είναι η κατανομή των τιμών στον αρχικό πληθυσμό. Στην περίπτωση δε που η κατανομή του αρχικού πληθυσμού πλησιάζει την κανονική τότε και ένα μικρό δείγμα είναι αρκετό για να ισχύει η προσέγγιση της κανονικής κατανομής από τις μέσες τιμές των δειγμάτων[2]. Το θεώρημα αυτό έχει τη βάση του στις εργασίες του Laplace στις αρχές του 19ου αιώνα και θεμελιώθηκε από τον Ούγγρο μαθηματικό G. Polya στις αρχές του 20ου. Αποτελεί ένα ύψιστης

[2] Το θεώρημα ισχύει εάν η διασπορά του αρχικού πληθυσμού είναι πεπερασμένη και το δείγμα επιλέγεται τυχαία από τον πληθυσμό.

σημασίας θεώρημα για ολόκληρη τη στατιστική συμπερασματολογία. Η σημασία του θα φανεί στις επόμενες παραγράφους.

Αντίθετα με την κατανομή των μέσων τιμών, η κατανομή των διασπορών των δειγμάτων κατά κανόνα διαφέρει από την κανονική κατανομή όπως φαίνεται και από το διάγραμμα 1.4 που παρουσιάζει τα δεδομένα του παραδείγματος 1.1.

1.7. Η ΕΠΙΔΡΑΣΗ ΤΟΥ ΜΕΓΕΘΟΥΣ ΤΟΥ ΔΕΙΓΜΑΤΟΣ

αυξανομένου του μεγέθους του δείγματος η πιθανότητα να περιέχει μόνο άτομα με ακραία χαρακτηριστικά, που θα έδιναν και λανθασμένη εκτίμηση μειώνεται

Κοιτάζοντας το διάγραμμα 1.2 βλέπουμε ότι αν το δείγμα περιέχει μόνο ένα άτομο, η πιθανότητα η τιμή του ατόμου αυτού να είναι 8 είναι περίπου 11%. Άρα η πιθανότητα να έχουμε τιμή που να είναι πολύ μικρότερη από την πραγματική μέση τιμή (10.0) είναι σχετικά μεγάλη. Στην περίπτωση δείγματος 2 ατόμων η πιθανότητα και τα δύο άτομα να έχουν τιμές πολύ μικρότερες της μέσης, δηλαδή να είναι 8 και 9 είναι 2/36=5.6%. Πράγματι οι συνδυασμοί 2 από 9 είναι 36 και επειδή υπάρχουν 2 άτομα με τιμή 9 υπάρχουν και δύο δείγματα 8, 9 στον πληθυσμό. Αλλά και σ'αυτή την περίπτωση η μέση τιμή του που είναι 8.5 είναι πιο κοντά στην πραγματική μέση τιμή (10.0) απ'ότι στην περίπτωση εκτίμησης από μεμονωμένο άτομο. Δηλαδή όχι μόνο η πιθανότητα ακραίας εκτίμησης είναι μικρότερη αλλά και σ'αυτή την περίπτωση η τιμή βρίσκεται πλησιέστερα στην πραγματική. Στην περίπτωση δείγματος 3 ατόμων η πιθανότητα και τα τρία άτομα να είναι μικρότερα από τη μέση τιμή είναι 1/84 (ο συνδυασμός 8, 9, 9) και η μέση τιμή τους είναι 8.67, πλησιάζει δηλαδή προς το 10.0. Για δείγματα 4 και πλέον ατόμων δεν υπάρχει καμία πιθανότητα όλα τα άτομα του δείγματος να είναι μικρότερα από τη μέση τιμή του πληθυσμού.

Η στρατηγική της δειγματοληψίας φροντίζει να καθορίζει το μέγεθος του δείγματος και τον τρόπο επιλογής των ατόμων που θα το αποτελέσουν έτσι ώστε να αποφεύγονται οι ακραίες (exreme) καταστάσεις. Ακραίες δε, είναι οι καταστάσεις στις οποίες όλα τα άτομα του δείγματος έχουν κοινά χαρακτηριστικά τα οποία όμως απέχουν πολύ από την πραγματική μέση τιμή του πληθυσμού. Από το προηγούμενο παράδειγμα φαίνεται καθαρά ότι η πιθανότητα "ακραίου" δείγματος μειώνεται με την αύξηση του μεγέθους του δείγματος. Στη φύση για παράδειγμα είναι εξαιρετικά σπάνιο όλα τα άτομα ενός μεγάλου δείγματος να πάσχουν από συγκεκριμένη και όχι κοινή νόσο εάν η επιλογή των ατόμων έχει γίνει τυχαία και δεν συντρέχουν ειδικοί λόγοι (π.χ. συγκέντρωση των φορέων σε συγκεκριμένη περιορισμένη περιοχή ή αδυναμία τους να αποφύγουν τη σύλληψη). Αυτό το γεγονός μεταφράζει εξ'άλλου και ο τύπος 1.5 όπου φαίνεται καθαρά ότι με την αύξηση του n η διασπορά των μέσων τιμών γύρο από την πραγματική μέση τιμή μειώνεται. Στα επόμενα κεφάλαια θα παρουσιασθούν και διάφορες τεχνικές για την επιλογή των ατόμων που στοχεύουν στη μείωση της πιθανότητας

εμφάνισης "ακραίων" δειγμάτων και κατά συνέπεια οδηγούν σε καλύτερες (ακριβέστερες και αξιόπιστες) εκτιμήσεις.

1.8. Οι επιπτωσεις της διορθωσης του πεπερασμενου πληθυσμου (fpc)

Από τον τύπο 1.5 φαινεται ότι η διασπορά των μέσων τιμών μειώνεται όταν το μέγεθος του δειγματος αυξάνει. Σ'αυτή την περιπτωση η εκτιμηση ειναι ακριβέστερη. Πως επιδρά όμως η διόρθωση του πεπερασμένου πληθυσμού (N-n)/(N-1);

Ας υποθέσουμε ότι έχουμε δύο πληθυσμούς, ο ένας

1000 και ο άλλος 20000 ατόμων. Οι πληθυσμοι αυτοι έχουν την ιδια διασπορά $\sigma_1^2=\sigma_2^2=100$. Από κάθε πληθυσμό επιλέγουμε δειγμα 50 ατόμων. Σύμφωνα με τον τύπο 1.5 η διασπορά των μέσων τιμών ειναι 1.9 για τον πληθυσμό 1 και 1.995 για τον 2. Παρατηρούμε ότι οι διαφορές στη διασπορά των μέσων τιμών ειναι σχεδόν αμελητέες. Αυθόρμητα θα θεωρούσαμε ότι το δειγμα 1 θα έδινε πολύ ακριβέστερες εκτιμήσεις (μικρότερη διασπορά) διότι αντιπροσωπεύει ένα πολύ μεγαλύτερο μέρος του πληθυσμού 1. Στην πράξη όμως φαινεται ότι το $\sigma_{\bar{y}}^2$ μειώνεται μόνο κατά 4.76% από το δειγμα 2 στο δειγμα 1. Συνήθως όταν οι τιμές του δειγματοληπτικού κλάσματος n/N ειναι μικρότερες από 5% αγνοούμε την fpc και ο τύπος της διασποράς των μέσων τιμών απλοποιειται σε $\sigma_{\bar{y}}^2=\sigma^2/n$. Αυτό συμβαινει συχνά στη μελέτη της φύσης όπου το δειγμα αντιπροσωπεύει ένα μικρό μέρος του φυσικού πληθυσμού.

1.9. Η πιθανοτητα σωστης εκτιμησης

Τοποθέτηση του προβλήματος

Όπως ήδη αναφέρθηκε ο εκτιμητής ειναι μια τυχαια μεταβλητή που η τιμή της εξαρτάται από τα χαρακτηριστικά του πληθυσμού "στόχου" αλλά και από το δειγμα. ,τσι αν ο σκοπός μας ειναι η εκτιμηση της μέσης τιμής του πληθυσμού του παραδειγματος 1.1 μέσω ενός δειγματος 3 ατόμων, τότε η εκτιμηση θα προέλθει μέσα από ένα από τα 84 δυνατά δειγματα (συνδυασμοι 3 απο 9) του πινακα 1.2 . Όπως ήδη παρατηρήσαμε οι μέσες τιμές των δειγμάτων κυμαινονται γύρο από την πραγματική μέση τιμή. Ας υποθέσουμε ότι επιλέξαμε το δειγμα 47 (9, 11, 11) με μέση τιμή 10.33 και διασπορά $s^2=1.33$. Η τιμή 10.33 ειναι εκτιμηση της πραγματικής μέσης τιμής του πληθυσμού την οποια όμως λογικά δεν γνωριζουμε αφου αυτήν προσπαθούμε να εκτιμήσουμε με τη δειγματοληψια. Δεν γνωριζουμε ούτε καν αν ειναι μεγαλύτερη ή μικρότερη από 10.33 που μας έδωσε το δείγμα. Τα ερωτήματα που γεννιώνται ειναι: Πόσο κοντα βρισκόμαστε στην πραγματική μέση τιμή; Ποιά ειναι η πιθανότητα να την έχουμε βρει ακριβώς ή έστω να κάνουμε ένα πολύ μικρό λάθος της τάξης του 5 ή 10%; Το μόνο που γνωριζουμε από τις προηγούμενες παραγράφους ειναι ότι βρισκεται σχετικά κοντά στην

πραγματική μέση τιμή. Το πρόβλημα πλέον ειναι να αποφασισουμε τι σημαινει "κοντά".

Για να ξεπεράσουμε αυτή την αβεβαιότητα η λύση ειναι να ορισουμε ένα διάστημα τιμών και να προτεινουμε ότι το πραγματικό $\overline{Y}$ που ψάχνουμε βρισκεται μέσα σ'αυτό το διάστημα. Μπορούμε να προτεινουμε ότι η πραγματική μέση τιμή που μας ενδιαφέρει βρισκεται στο διάστημα 10.0 - 10.66. Μήπως όμως ειναι στενό; Πόσο βέβαιοι μπορούμε να ειμαστε ότι όντως το πραγματικό $\overline{Y}$ περικλειεται σ'αυτό το διάστημα; Σιγουρα ένα μεγαλύτερο διάστημα π.χ. 9.0 - 11.66 ειναι καλύτερο. Όμως αν αυξήσουμε πολύ το διάστημα τότε η εκτιμηση (πρόβλεψη) παύει πλέον να έχει έννοια. Σκεφτήτε μια μελέτη που θα λέει ότι το μέσο ύψος του ,λληνα φοιτητή ειναι ανάμεσα στο 1 και τα 2 μέτρα. Ουσιαστικά δεν προσφέρει πληροφορια.

Στην ουσια χρειαζόμαστε ένα μικρό διάστημα (για να ειναι χρήσιμο) το οποιο να έχει αρκετές πιθανότητες να περιέχει την πραγματική μέση τιμή. Μια πιθανότητα 90% ή 95% ειναι συνήθως αρκετή. Αυτή η πιθανότητα ονομάζεται **επιπεδο ή συντελεστής εμπιστοσύνης** (confidence level) και το αντιστοιχο διάστημα ονομάζεται **διάστημα εμπιστοσύνης** (confidence interval). Για κάθε διάστημα αντιστοιχει και ένα επιπέδο εμπιστοσύνης. Το διάστημα εμπιστοσύνης γύρο από τη μέση τιμή του δειγματος που αντιστοιχει στο επιπεδο 95% σημαινει ότι υπάρχουν 95 περιπτώσεις στις 100 που το διάστημα θα περιέχει την πραγματική μέση τιμή του πληθυσμού, που επαναλαμβάνουμε μας ειναι άγνωστη. Αυτό σημαινει ότι αν από ένα πληθυσμό επιλέξουμε 100 διαφορετικά τυχαια δειγματα συγκεκριμένου μεγέθους και ορισουμε το διάστημα εμπιστοσύνης γύρο από τη μέση τιμή του κάθε δειγματος τότε τα 95 από αυτά θα περιέχουν την πραγματική μέση τιμή. Για να γίνει κατανοητή η έννοια του διαστήματος εμπιστοσύνης είναι προτιμότερο να εξετάσουμε πρώτα μια ιδανική περίπτωση που φυσικά δεν παρουσιάζεται στην πράξη.

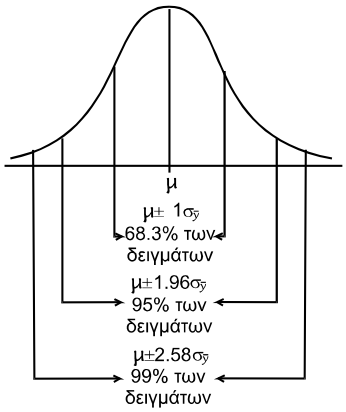
Ιδανική προσέγγιση

Ας πάρουμε την περίπτωση του πληθυσμού του παραδείγματος 1.1 του οποίου η μέση τιμή $\overline{Y}$ και η τυπική απόκλιση σ είναι γνωστές. Αυτό φυσικά δεν συμβαίνει (ή εξαιρετικά) σπάνια στην πράξη. Συνήθως από το δείγμα προσπαθούμε να εκτιμήσουμε αυτές τις παραμέτρους με τη βοήθεια δειγμάτων.

Το $\overline{Y}$ και το σ είναι γνωστά και παίρνουμε ένα δείγμα 3 ατόμων από τον πληθυσμό. Η μέση τιμή 10.33 του δειγματος που επιλέξαμε ειναι μια από τις 84 δυνατές μέσες τιμές που μπορούν να δώσουν δειγματα 3 ατόμων από τα 9 του πληθυσμού. Ειναι μια τυχαια τιμή από την κατανομή των μέσων τιμών δειγμάτων μεγέθους 3 από τον αρχικό πληθυσμό. Σύμφωνα με το κεντρικό οριακό θεώρημα η κατανομή των

μέσων τιμών έχει μέση τιμή ιση με την πραγματική μέση τιμή του πληθυσμού και διασπορά ιση με σ^2/n (αγνοώντας προς το παρόν την fpc). Η κατανομή των μέσων τιμών ειναι επίσης κανονική αφού η κατανομή των ατόμων του πληθυσμού ειναι σχεδόν κανονική (σχήμα 1.3).

Αφού η κατανομή αυτή είναι κανονική, τότε 95% από τις μέσες τιμές όλων των δυνατών δειγμάτων απέχουν λιγότερο από $1.96\sigma_{\bar{y}}$ από τη μέση τιμή του πληθυσμού (σχήμα 1.5). Κατά συνέπεια η πραγματική μέση τιμή δεν θα απέχει πάνω από 1.96 $\sigma_{\bar{y}}$ από το 95% των δυνατών μέσων τιμών των δειγμάτων 3 ατόμων. Για να γίνει κατανοητό αυτό ας πάρουμε ένα απλοϊκό παράδειγμα. Ας θεωρήσουμε τις κατοικίες 1000 φοιτητών που βρίσκονται σε διάφορες αποστάσεις από το Πανεπιστήμιο. Αν 95% από τις κατοικίες βρίσκονται σε απόσταση μικρότερη των 10 km από το Πανεπιστήμιο τότε σίγουρα το Πανεπιστήμιο δεν θα απέχει περισσότερο από 10 km από την κάθε μια από αυτές τις κατοικίες. Εάν επιλέξουμε τυχαία ένα μεγάλο αριθμό κατοικιών από τις 1000 τότε το Πανεπιστήμιο θα απέχει λιγότερο από 10 km από το 95% των κατοικιών που επελέγησαν.



ΣΧΗΜΑ 1.5 *Ποσοστά του πληθυσμού που περιέχονται σε διάφορες περιοχές της κανονικής κατανομής.*

Από τα χαρακτηριστικά της κανονικής κατανομής (σχήμα 1.5) γνωρίζουμε ότι 64%, 95% και 99% των τιμών βρίσκονται σε μια απόσταση το πολύ 1, 1.96 και 2.58σ αντίστοιχα γύρο από τη μέση τιμή της κατανομής (χαρακτηριστικά της κανονικής κατανομής παρουσιάζονται στο παράρτημα 1). Η απόσταση της μέσης τιμής του δείγματος από την πραγματική μέση τιμή είναι $\bar{y}$-$\bar{Y}$. Μπορούμε να εκφράσουμε την απόσταση αυτή σε μονάδες τυπικού σφάλματος διαιρώντας με $\sigma_{\bar{y}}$. Έτσι η ποσότητα

$$z = \frac{\bar{y} - \bar{Y}}{\sigma_{\bar{y}}} \qquad (1.7)$$

κατανέμεται κανονικά γύρο από το 0 με τυπική απόκλιση 1. Τιμές του z 1 ή 1.96 σημαίνουν ότι η μέση τιμή του δείγματος απέχει 1 ή αντίστοιχα 1.96 τυπικά σφάλματα από την πραγματική μέση τιμή. Σε κάθε διάστημα που ορίζεται από την τιμή του z βρίσκεται και ένα συγκεκριμένο και γνωστό ποσοστό των τιμών της κατανομής (σχήμα 1.5). Άρα μπορούμε να γράψουμε

$$P\{-1.96 < \frac{\bar{y} - \bar{Y}}{\sigma_{\bar{y}}} < 1.96\} = 0.95$$

όπου P είναι η πιθανότητα να συμβαίνει το περιεχόμενο της αγκύλης. Αυτό γράφεται και $P\{-1.96\ \sigma_{\bar{y}} < \bar{y} - \bar{Y} < 1.96\ \sigma_{\bar{y}}\} = 0.95$ το οποίο είναι ίσο με

$$P\{\bar{y} - 1.96\sigma_{\bar{y}} < \bar{Y} < \bar{y} + 1.96\sigma_{\bar{y}}\} = 0.95$$

Αυτό σημαίνει ότι η πιθανότητα P είναι 95% η πραγματική μέση τιμή να περικλείεται από τις τιμές $\bar{y}$-1.96$\sigma_{\bar{y}}$ και $\bar{y}$+1.96$\sigma_{\bar{y}}$c Αποκαλούμε τις δύο αυτές τιμές κατώτερο (L_l) και ανώτερο (L_u) όριο εμπιστοσύνης στο επίπεδο 95%. Αν η πιθανότητα 95% δεν είναι ικανοποιητική τότε αντικαθιστώντας το 1.96 με 2.576 αυξάνουμε το επίπεδο εμπιστοσύνης στο 99%.

Γενικότερα το διάστημα εμπιστοσύνης που ορίζεται από τη μέση τιμή του δείγματος $\bar{y}$ και με την προϋπόθεση ότι η τυπική απόκλιση σ του πληθυσμού είναι γνωστή είναι

$$\boxed{P\{\bar{y} - z_\alpha \sigma_{\bar{y}} < \bar{Y} < \bar{y} + z_\alpha \sigma_{\bar{y}}\} = 1 - \alpha} \qquad (1.8)$$

όπου η τιμή του z εξαρτάται επίπεδο εμπιστοσύνης 1-α και βρίσκονται στους πίνακες της κανονικής κατανομής[3] (παράρτημα 1).

1.10. ΤΟ ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΣΤΗΝ ΠΡΑΞΗ

Η προηγούμενη περίπτωση στην οποία η τυπική απόκλιση σ του πληθυσμού είναι εκ των προτέρω γνωστή είναι σπάνια. Συνήθως τα χαρακτηριστικά του πληθυσμού είναι άγνωστα και προσπαθούμε να τα εκτιμήσουμε μέσα από την ανάλυση του δείγματος. Έτσι η μέση τιμή $\bar{y}$ και η τυπική απόκλιση s του δείγματος είναι οι μόνες εκτιμήσεις που διαθέτουμε για τις αντίστοιχες παραμέτρους $\bar{Y}$ και σ του πληθυσμού. Μπορούμε να χρησιμοποιήσουμε αυτές τις τιμές για να ορίσουμε ένα διάστημα εμπιστοσύνης όπως το ορίσαμε στην προηγούμενη παράγραφο;

Δυστυχώς όχι κι αυτό διότι όπως ήδη αναφέρθηκε το $\bar{y}$ και το s του δείγματος είναι τυχαίες μεταβλητές η τιμή των οποίων αλλάζει από δείγμα σε δείγμα. Για ένα δείγμα μπορεί η τιμή $\bar{y}$ να είναι κοντά στην πραγματική μέση τιμή $\bar{Y}$ του πληθυσμού αλλά η τυπική απόκλιση s του δείγματος να είναι πολύ μικρότερη από το σ του πληθυσμού. Κατά συνέπεια ο λόγος ($\bar{y}$-$\bar{Y}$)/$s_{\bar{y}}$ θα είναι μεγάλος παρά το γεγονός ότι το η απόσταση του $\bar{y}$ από το $\bar{Y}$ είναι μικρή (θυμίζουμε ότι $s_{\bar{y}}$=s/$\sqrt{n}$ και αφού το s είναι μικρό και το τυπικό σφάλμα θα είναι μικρό). Αντίθετα αν το $\bar{y}$-$\bar{Y}$ είναι μεγάλο αλλά και το s επίσης μεγαλύτερο από το σ τότε ο λόγος τους μπορεί να είναι σχετικά μικρός. Για παράδειγμα ο λόγος αυτός είναι 0.32 για το δείγμα 27 του πίνακα 1.2 και 1.155 για το δείγμα 66 αν και τα δύο δείγματα έχουν την ίδια μέση τιμή. Κατά συνέπεια η κατανομή του λόγου αυτού θα είναι ευρύτερη και πιο επίπεδη από την κανονική κατανομή της προηγουμένης παραγράφου.

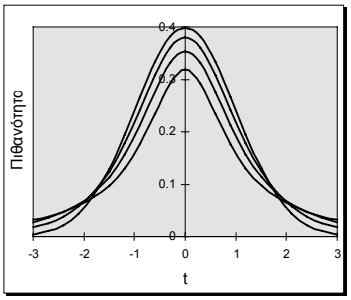
[3] Οι πίνακες του z δίνουν έμφαση στην πιθανότητα σφάλματος α, δηλαδή στην πιθανότητα να μην περικλείει το διάστημα εμπιστοσύνης την πραγματική τιμή της παραμέτρου. Συνεπώς η πιθανότητα εμπιστοσύνης είναι 1-α.

Αφού λοιπόν η κατανομή διαφέρει από την κανονική δεν μπορούμε να χρησιμοποιήσουμε τον τύπο 1.8 για να ορίσουμε το διάστημα εμπιστοσύνης.

Μετά από εργασιες του W.S Gosset στις αρχές του αιώνα (ο οποιος για λόγους εργασιακούς υπέγραφε τις δημοσιεύσεις του με το ψευδόνυμο Student) που συμπληρώθηκαν από τον R.A. Fisher απεδειχθη ότι η ποσότητα

$$t = \frac{\bar{y} - \bar{Y}}{s/\sqrt{n}} \qquad (1.9)$$

ακολουθει την περίφημη κατανομή t του Student (που αναφέρεται σε Ελληνικά βιβλια στατιστικής σαν t του σπουδαστή).



Σχημα 1.6 *Κατανομή του t για 1, 2, 5 βαθμούς ελευθερίας (df) (από κάτω προς τα πάνω) και η κανονική κατανομή (ανώτερη καμπύλη) που συμπίπτει με την κατανομή του t για άπειρους df. (η σχετική θέση των καμπυλών εξετάζεται στο κέντρο του διαγράμματος).*

Όπως φαινεται και από τον τύπο του η κατανομή t εξαρτάται από το n (διάγραμμα 1.6). Οι πινακες της κατανομής t που βρισκονται σε όλα τα εγχειριδια στατιστικής δινουν για κάθε τιμή των βαθμών ελευθεριας (degrees of freedom, df) που ισούνται με n-1 και για συγκεκριμένη πιθανότητα (α) την αντιστοιχούσα τιμή t. Ο πινακας 1.4 ειναι απόσπασμα από λεπτομερή πινακα τιμών της κατανομής t του παραρτήματος 2.

Η ποσότητα αυτή t ειναι όπως και στην προηγούμενη περίπτωση η απόκλιση της μέσης τιμής του δειγματος από την πραγματική μέση τιμή του πληθυσμού εκφρασμένη σε μονάδες τυπικού σφάλματος αφου η ποσότητα $s/\sqrt{n}$ ειναι το τυπικό σφάλμα. ¸τσι μια τιμή t=1.5 σημαινει ότι η μέση τιμή του δειγματος απέχει 1.5 φορά το τυπικό σφάλμα από την πραγματική μέση τιμή. Από τη στιγμή που η κατανομή t ειναι γνωστή σημαινει ότι γνωριζουμε ακριβώς ποιά ειναι η πιθανότητα η διαφορά $\bar{y}$-c$\bar{Y}$ να ειναι ακριβώς 0.6, 1.9 ή 2.5 τυπικά σφάλματα. Ειναι επισης γνωστή η πιθανότητα η διαφορά αυτή να ειναι σε απόλυτη τιμή μικρότερη, για παράδειγμα, από 1.6 τυπικά σφάλματα. Αυτή η τελευταια παρατήρηση μεταφράζεται ως εξής

$$P\{-t < \frac{\bar{y} - \bar{Y}}{s/\sqrt{n}} < t\} = 1 - \alpha$$

που οδηγεί όπως προηγουμένως στην

$$\boxed{P\{\bar{y} - t_{\alpha}\frac{s}{\sqrt{n}} < \bar{Y} < \bar{y} + t_{\alpha}\frac{s}{\sqrt{n}}\} = 1 - \alpha} \qquad (1.10)$$

και είναι η σχέση που χρησιμοποιούμε για να ορίσουμε το διάστημα εμπιστοσύνης χρησιμοποιώντας παραμέτρους του δείγματος.[4] Εφ'όσον η τιμή t εξαρτάται από το επίπεδο εμπιστοσύνης (1-α) και από τους βαθμούς ελευθερίας, ο πλήρης συμβολισμός του είναι $t_{\alpha[n-1]}$ για λόγους όμως συντομίας παρουσιάζεται σαν t_{α}.

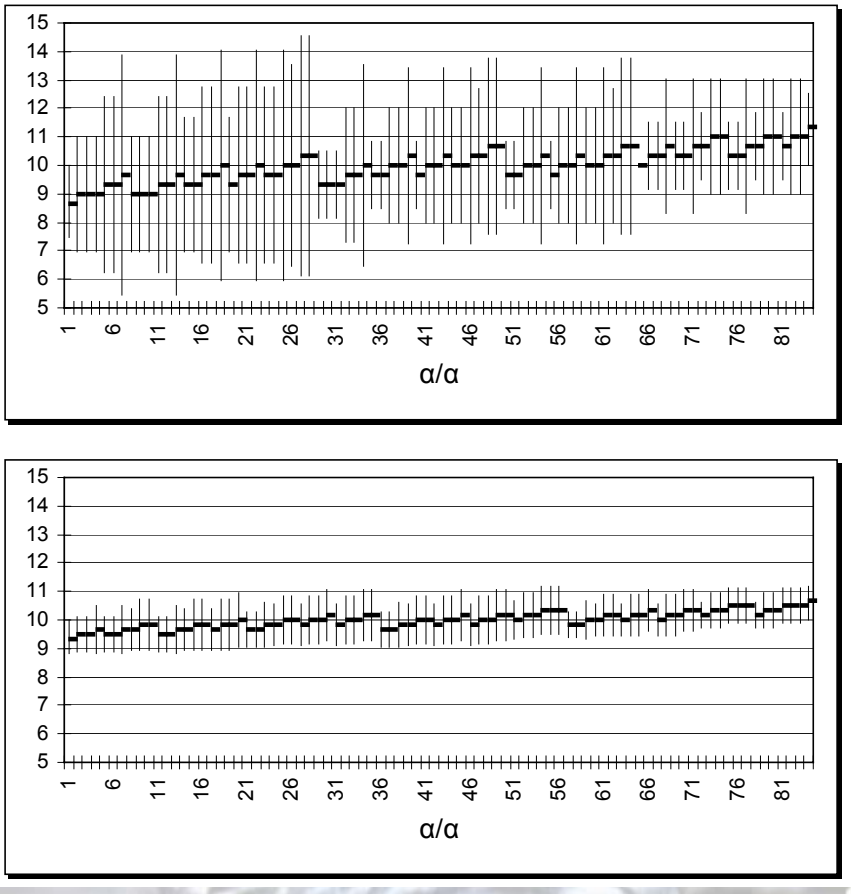
Η τιμή του t=4.30 που αντιστοιχει σε 2 βαθμούς ελευθεριας και πιθανότητα 0.05 υποδηλώνει ότι για ένα δειγμα 3 ατόμων (*df*=n-1) η πιθανότητα η μέση τιμή του δειγματος να απέχει ανεξαρτήτως κατεύθυνσης (μικρότερη δηλαδή ή μεγαλύτερη) από την πραγματική μέση τιμή του πληθυσμού $\bar{Y}$ πάνω από 4.30 τυπικά σφάλματα ειναι ιση με 5%. Με άλλα λόγια η πραγματική μέση τιμή του πληθυσμού έχει 95% (1-*α*) πιθανότητες να βρισκεται σε ένα διάστημα 4.30 τυπικών σφαλμάτων γύρο από τη μέση τιμή του δειγματος. Αυτή η έκφραση θυμιζει τον ορισμό του διαστήματος εμπιστοσύνης που αναφέρθηκε ανωτέρω.

ΠΙΝΑΚΑΣ 1.4 *Απόσπασμα από πίνακα της κατανομής t. α το επίπεδο εμπιστοσύνης, df οι βαθμοί ελευθερίας*

df\α	0.1	0.05	0.01
1	6.31	12.70	63.66
2	2.92	4.30	9.93
3	2.35	3.18	5.84

Ας γυρισουμε στο δειγμα μας. Το δειγμα αυτό έχει n=3, μέση τιμή $\bar{y}$=10.33, διασπορά s^2=1.33. Κατά συνέπεια το τυπικό σφάλμα ειναι $s_{\bar{y}}$=0.577 και οι βαθμοι ελευθεριας 3-1=2. Η αντιστοιχούσα τιμή του tc για το επιπεδο εμπιστοσύνης 95% ειναι 4.30. ϕρα η πραγματική μέση τιμή του πληθυσμού έχει 95% πιθανότητες να περικλειεται στο διάστημα $\bar{y}$-t$s_{\bar{y}}$=7.85 και $\bar{y}$+t$s_{\bar{y}}$=12.81. Πράγματι το $\bar{Y}$ που ειναι 10 περικλειεται στο διάστημα αυτό. Εάν με τον ιδιο τρόπο υπολογισουμε το διάστημα εμπιστοσύνης για τα 84 δυνατά δειγματα 3 ατόμων από τον αρχικό πληθυσμό των 9 του παραδειγματος 1 (πίνακας 1.2)παρατηρούμε ότι 2 από τα διαστήματα δεν περιέχουν την πραγματική μέση τιμή (τα δειγματα 1 και 84 του διαγράμματος 1.7). Το ποσοστό 2/84=2.38% ειναι κοντά στο επιπεδο του 5% (1-α) που επιλέξαμε. Η διαφορά οφειλεται στο γεγονός ότι το δειγμα και ο πληθυσμός ειναι μικρά. Σε μεγάλα δειγματα από μεγάλους πληθυσμούς (n και N μεγάλα) το πραγματικό ποσοστό των δειγμάτων που ικανοποιούν τη συνθήκη ειναι πολύ κοντά στο επιπεδο εμπιστοσύνης που καθοριζει το διάστημα.

[4] Οι πίνακες του t του Student δίνουν έμφαση στην πιθανότητα σφάλματος α, δηλαδή στην πιθανότητα να μην περικλείει το διάστημα εμπιστοσύνης την πραγματική τιμή της παραμέτρου. Συνεπώς η πιθανότητα εμπιστοσύνης είναι 1-α.



ΣΧΗΜΑ 1.7 *Μέσες τιμές και διαστήματα εμπιστοσύνης στο επίπεδο 95% για όλα τα δυνατά δείγματα μεγέθους 3 (άνω) και 6 (κάτω) ατόμων (πίνακες* 1.2 , 1.3 *από τον πληθυσμό του παραδείγματος 1.* α/α: αύξων αριθμός του δείγματος στους πίνακες.

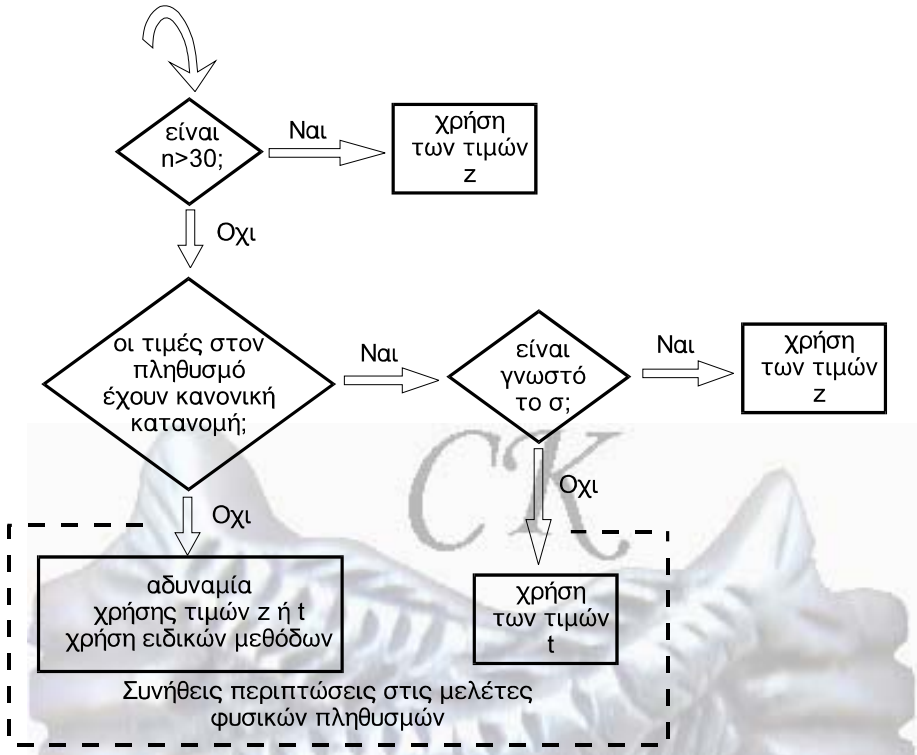
Από τον ορισμό του διαστήματος εμπιστοσύνης φαινεται ότι

- όσο η πιθανότητα σωστής εκτιμησης (πρόβλεψης) αυξάνει τόσο το διάστημα γινεται ευρύτερο (για το προηγούμενο παράδειγμα το διάστημα εμπιστοσύνης είναι 8.65-12.02 για το επίπεδο εμπιστοσύνης 90% και 4.60-16.06 για το επίπεδο 99%)
- για ένα δεδομένο επιπεδο εμπιστοσύνης το εύρος του διαστήματος μειώνεται όσο το μέγεθος του δειγματος αυξάνει (διότι και το t και το $s_{\bar{y}}$ που εξαρτώνται από το n μικραινουν).

Μερικές παρατηρήσεις ειναι αναγκαιες:

- οι τύποι για το διάστημα εμπιστοσύνης ισχύουν όταν η αρχική κατανομή των ατόμων του πληθυσμού δεν απέχει πολύ από την κανονική κατανομή (για να ισχύει το κεντρικό οριακό θεώρημα δηλαδή η κατανομή των μέσων τιμών να ειναι κανονική)
- εάν το n ειναι μεγάλο ($n>30$) τότε η κατανομή του t του Student πλησιάζει πολύ την κανονική κατανομή με $\mu=0.0$ και $\sigma=1.0$ και κατά συνέπεια οι τιμές του t εξαρτώνται πλέον μόνο από το επιπεδο πιθανότητας

- αν το n ειναι μεγάλο τότε και στην περιπτωση που η αρχική κατανομή διαφέρει σημαντικά από την κανονική, οι μέσες τιμές των δειγμάτων έχουν σχεδόν κανονική κατανομή και όλα τα παραπάνω ισχύουν.



Σχημα 1.8 *Κριτήρια και τις επιλογές των συντελεστών για τη δημιουργία των διαστημάτων εμπιστοσύνης*

Οι τιμές του t και κατά συνέπεια ο τύπος 1.10 ισχύει μόνο όταν η κατανομή των τιμών y_i της παραμέτρου στον πληθυσμό δεν απέχουν πολύ από την κανονική κατανομή και το n ειναι μεγάλο. Μικρές αποκλισεις από την κανονικότητα δεν έχουν σημαντικές επιπτώσεις. Για μικρά όμως δειγματα με ιδιαιτερα ασύμετρες κατανομές ειδικές μέθοδοι πρέπει να χρησιμοποιούνται. Το ακόλουθο διάγραμμα συνοψίζει τα διάφορα κριτήρια και τις επιλογές των συντελεστών για τη δημιουργία των διαστημάτων εμπιστοσύνης.

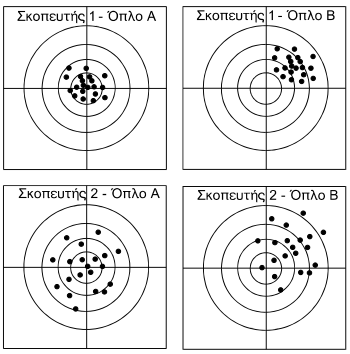
1.11. ΑΜΕΡΟΛΗΨΙΑ ΚΑΙ ΑΚΡΙΒΕΙΑ

Από το προηγούμενο παράδειγμα φαίνεται ότι μπορεί η μέση τιμή ενός δείγματος να είναι διαφορετική από την πραγματική τιμή της παραμέτρου στον πληθυσμό αλλά η μέση τιμή των μέσων τιμών όλων των δυνατών δειγμάτων συμπίπτει. Αυτό σημαίνει ότι κατά μέσο όρο τα δείγματα δίνουν τη σωστή εικόνα της κατάστασης που επικρατεί στον πληθυσμό. Η μέση τιμή των εκτιμήσεων όλων των δυνατών δειγμάτων είναι η *μαθηματική ελπίδα* του εκτιμητή. άταν λοιπόν η

τιμή αυτή συμπίπτει με την πραγματική τιμή του πληθυσμού τότε ο εκτιμητής χαρακτηρίζεται **αμερόληπτος** (unbiased estimator). Η έννοια αυτή της αμεροληψίας είναι σημαντική και στις περισσότερες περιπτώσεις μελετών, δημοσκοπήσεων κ.λ.π. ψάχνουμε έναν αμερόληπτο εκτιμητή.

Το δεύτερο στοιχείο που παρατηρούμε στο παράδειγμα 1 είναι ότι οι μέσες τιμές των διαφόρων δειγμάτων πλησιάζουν όλο και περισσότερο προς την πραγματική μέση τιμή του πληθυσμού όσο το μέγεθος του δείγματος μεγαλώνει. Είναι φανερό ότι στην περίπτωση δείγματος 6 ατόμων οι εκτιμήσεις βρίσκονται πιο κοντά στην πραγματική τιμή άρα η πιθανότητα μεγάλου λάθους από την εκτίμηση ενός τυχαίου δείγματος (ενός από τα 84) είναι περιορισμένη. Ο εκτιμητής λοιπόν των δειγμάτων μεγέθους 6 είναι πιο **ακριβής** (precise estimator) από τον εκτιμητή δειγμάτων μεγέθους 3 ατόμων.

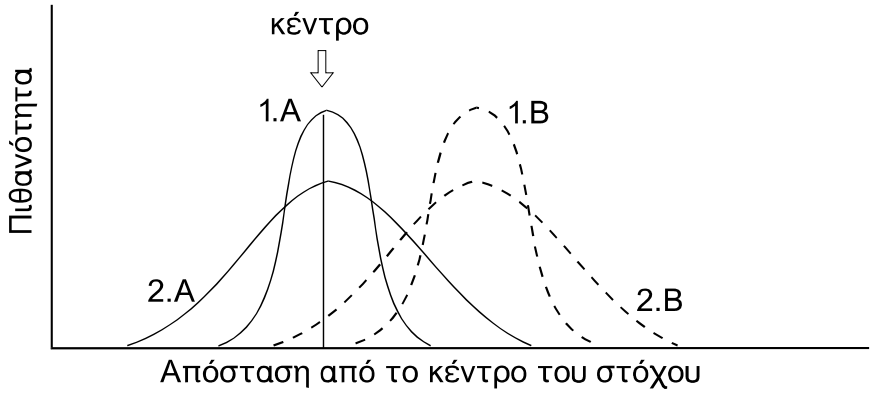
Οι έννοιες της ακρίβειας και της αμεροληψίας γίνονται κατανοητές μέσα από το ακόλουθο παράδειγμα. Δύο σκοπευτές εξετάζουν τη συμπεριφορά δύο διαφορετικών όπλων. Τα αποτελέσματα των βολών τους φαίνονται στο διάγραμμα 1.9. Παρατηρούμε ότι με το όπλο Α ο σκοπευτής 1 βρίσκεται κατά μέσο όρο πιο κοντά στο κέντρο του στόχου από το σκοπευτή 2 αλλά οι βολές και των δύο κατανέμονται γύρο από το κέντρο. Με το όπλο Β οι βολές του σκοπευτή 1 παραμένουν πιο συγκεντρωμένες απ'αυτές του σκοπευτή 2 όμως και των δύο οι βολές φαίνεται ότι κατανέμονται γύρο από ένα σημείο που βρίσκεται δεξιά και πάνω από το κέντρο του στόχου. Αυτό μπορεί να οφείλεται σε ελάττωμα του στοχάστρου του συγκεκριμένου όπλου. Ο σκοπευτής σκοπεύει σωστά αλλά η τροχιά του βλήματος είναι διαφορετική από αυτή που επιθυμεί. Η κατανομή των βολών τους μεταφέρεται στο διάγραμμα 1.10. Από τα αποτελέσματα αυτά παρατηρούμε ότι:



ΣΧΗΜΑ 1.9 *Αποτελέσματα άσκησης σκοποβολής με χρήση δυο διαφορετικών όπλων από δύο σκοπευτές.*

- ο σκοπευτής 1 είναι πιο ακριβής από το σκοπευτή 2 αφού η διασπορά των βολών του είναι συστηματικά μικρότερη
- το όπλο Β είναι αναξιόπιστο αφού η μέση θέση των βολών είναι καθαρά διαφορετική από το επιθυμητό σημείο του κέντρου του στόχου.

Το παράδειγμα αυτό παρουσιάζει απλοϊκά τις έννοιες της μεροληψίας και της ακρίβειας. Ας υποθέσουμε ότι το κέντρο του στόχου δεν μας ήταν γνωστό αλλά θα έπρεπε να το βρούμε μελετώντας τις βολές των σκοπευτών. Είναι ξεκάθαρο ότι θα επιλέγαμε πάνω απ'όλα έναν αμερόληπτο εκτιμητή, άρα το όπλο Α και στη συνέχεια έναν ακριβή εκτιμητή άρα το σκοπευτή 1. Τώρα στην περίπτωση που μόνο το όπλο Β ήταν διαθέσιμο και για σοβαρούς λόγους θα έπρεπε κάποιο από τα λίγα διαθέσιμα βλήματα να φθάσει στο κέντρο του στόχου, τότε θα επιλέγαμε το σκοπευτή 2.



Σχημα 1.10 *Κατανομή των βολών των σκοπευτών γύρο από το κέντρο του στόχου.*

Σ'αυτή την περίπτωση ο συνδυασμός της μεροληψίας με την περιορισμένη ακρίβεια όπως φαίνεται και από το διάγραμμα 1.10 αφήνει κάποιες πιθανότητες το βλήμα να φθάσει στο κέντρο, πράγμα εντελώς απίθανο για το σκοπευτή 1. Μπορεί λοιπόν να υπάρξουν κάποιες σπάνιες περιπτώσεις που μας οδηγούν στην επιλογή ενός μεροληπτικού εκτιμητή γι'αυτό και στην αρχή της παραγράφου αναφέρεται ότι "στις περισσότερες (και όχι σε όλες) περιπτώσεις μελετών, δημοσκοπήσεων κ.λ.π. ψάχνουμε έναν αμερόληπτο εκτιμητή". Πράγματι ένας ελαφρά μεροληπτικός εκτιμητής αλλά πολύ ακριβής μπορεί να είναι καλύτερος από έναν αμερόληπτο αλλά πολύ ανακριβή.

Ως **μεροληψία** (bias) μπορεί να θεωρηθεί μια πληροφορία που εισάγεται αρχικά στα δεδομένα από την ίδια την πράξη της δειγματοληψίας ή από τη μέθοδο μέτρησης και την οποία βρίσκουμε στο τέλος κατά τη διάρκεια της ανάλυσης

1.12. Το κοστος και η ακριβεια των εκτιμησεων

Στο προηγούμενο παράδειγμα μπορούμε να πούμε ότι ο μέσος αριθμός καρπών ανά φυτό του πληθυσμού κυμαίνεται από 7.85 ως 12.81 με πιθανότητα 95% η πρόβλεψη αυτή να είναι σωστή. Μια άλλη έκφραση του ιδίου αποτελέσματος είναι: ο μέσος αριθμός καρπών είναι 10.33±24%. Αυτό το ποσοστό προκύπτει ως εξής: $ts_{\bar{y}}/\bar{y}*100$.

άταν μια δειγματοληπτική μελέτη ξεκινά ο υπεύθυνος μπορεί να ορίσει το ολικό κόστος της μελέτης και να προσπαθήσει στη συνέχεια αριστοποιώντας τον τρόπο επιλογής και το μέγεθος του δείγματος να πετύχει τη μεγαλύτερη δυνατή ακρίβεια. Σε άλλες περιπτώσεις όμως μπορεί να ορίσει το επίπεδο ακρίβειας της μελέτης καθορίζοντας το εύρος του διαστήματος εμπιστοσύνης, π.χ. ±5% άνθη ανά φυτό ή ±3% του ολικού πληθυσμού των ασθενών σε μια περιοχή. Με άλλα λόγια η πραγματική τιμή της παραμέτρου θα έχει συγκεκριμένες πιθανότητες π.χ. 95% να βρίσκεται σε ένα διάστημα το εύρος του οποίου εκφράζεται σαν ποσοστό της εκτιμηθήσας τιμής. Σ'αυτή την τελευταία κατηγορία βρίσκονται συχνά οι μελέτες που είναι συνδεδεμένες με την υγεία ή τη δημόσια ασφάλεια.

Από τη στιγμή που η επιθυμητή ακρίβεια της εκτίμησης έχει ορισθεί ο υπεύθυνος της μελέτης πρέπει να φροντίσει ώστε οι παράμετροι που καθορίζουν το τελικό αποτέλεσμα να ικανοποιούν τους όρους.

Βλέποντας τον τύπο 1.10 για το διάστημα εμπιστοσύνης και σύμφωνα με αυτά που έχουν λεχθεί, είναι φανερό ότι το εύρος του διαστήματος μειώνεται όταν αυξάνει το μέγεθος του δείγματος n και αυτό διότι το τυπικό σφάλμα μειώνεται (το n είναι στον παρονομαστή του τύπου 1.10) και παράλληλα και η τιμή του t για ένα συγκεκριμένο επίπεδο εμπιστοσύνης μειώνεται όπως φαίνεται από τον πίνακα 1.4.

Η αύξηση όμως του μεγέθους του δείγματος έχει σαν αποτέλεσμα την αύξηση του κόστους της μελέτης. Το ερώτημα λοιπόν γεννιέται. Υπάρχει τρόπος αύξησης της ακρίβειας της εκτίμησης χωρίς σοβαρές επιπτώσεις στο τελικό κόστος της μελέτης; Η θεωρία της δειγματοληψίας και οι εφαρμογές της δίνουν λύσεις σ'αυτό το πρόβλημα. Ο σχεδιασμός μιας διαδικασίας επιλογής του δείγματος που χρησιμοποιεί όσο το δυνατό περισσότερη πληροφορία για τα χαρακτηριστικά του πληθυσμού, τη συμπεριφορά των ατόμων, την κατανομή τους κ.λ.π. έχει σαν αποτέλεσμα τη μείωση του τυπικού σφάλματος (της διασποράς των μέσων τιμών γύρο από την πραγματική μέση τιμή) και κατά συνέπεια αύξηση της ακρίβειας της εκτίμησης. Αυτός ο σχεδιασμός είναι μέρος της στρατηγικής της δειγματοληψίας που θα αναλυθεί στα επόμενα κεφάλαια.