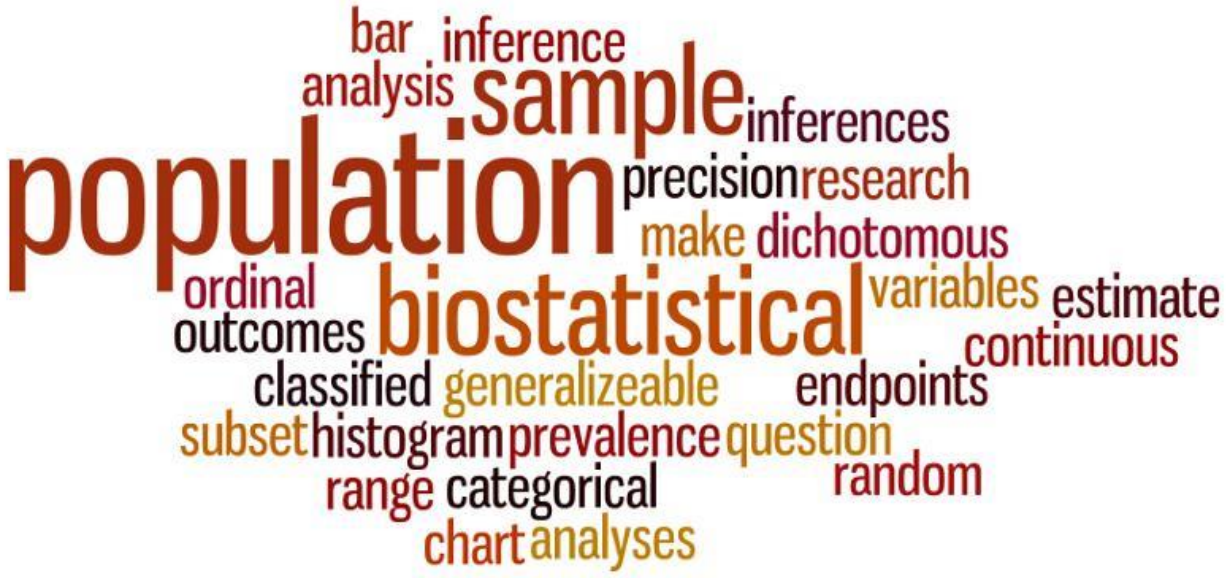
Άρθρο: Βασικές Έννοιες της Βιοστατιστικής

Basic Concepts for Biostatistics

Author:
Lisa Sullivan, PhD, Professor of Biostatistics, Boston University School of Public Health



Εισαγωγή

Βιοστατιστική είναι η εφαρμογή στατιστικών αρχών σε ερωτήματα και προβλήματα στην ιατρική, τη δημόσια υγεία ή τη βιολογία. Μπορεί κανείς να φανταστεί ότι έχει ενδιαφέρον να μελετηθεί ένας συγκεκριμένος πληθυσμός (π.χ. ενήλικες στη Βοστώνη ή όλα τα παιδιά στις Ηνωμένες Πολιτείες) σε σχέση με το ποσοστό των υπέρβαρων ατόμων ή το ποσοστό που έχουν άσθμα, και θα ήταν επίσης σημαντικό να εκτιμηθεί το μέγεθος αυτών των προβλημάτων (παθήσεων) στην πορεία του χρόνου ή ίσως σε διαφορετικές γεωγραφικές περιοχές.

Σε άλλες περιπτώσεις, θα ήταν σημαντικό να γίνουν συγκρίσεις μεταξύ ομάδων ατόμων προκειμένου να καθοριστεί εάν ορισμένες συμπεριφορές (π.χ. κάπνισμα, άσκηση κ.λπ.) σχετίζονται με μεγαλύτερο κίνδυνο ορισμένων διαταραχών υγείας. Θα ήταν, φυσικά, αδύνατο να απαντηθούν όλα αυτά τα ερωτήματα συλλέγοντας πληροφορίες (δεδομένα) ***από όλα τα μέλη*** στους πληθυσμούς που μελετώνται. Μια πιο ρεαλιστική προσέγγιση είναι η μελέτη ***δειγμάτων*** (= ***υποσυνόλων)*** ενός πληθυσμού. Ο κλάδος της βιοστατιστικής παρέχει εργαλεία και τεχνικές για τη συλλογή δεδομένων και στη συνέχεια τη σύνοψη, την ανάλυση και την ερμηνεία τους. Εάν τα δείγματα που λαμβάνονται είναι αντιπροσωπευτικά του πληθυσμού ενδιαφέροντος, θα παρέχουν καλές εκτιμήσεις σχετικά με τον συνολικό πληθυσμό. Κατά συνέπεια, στη βιοστατιστική αναλύει κανείς δείγματα για να βγάλει συμπεράσματα για τον πληθυσμό. Αυτή η ενότητα εισάγει θεμελιώδεις έννοιες και ορισμούς για τη βιοστατιστική.

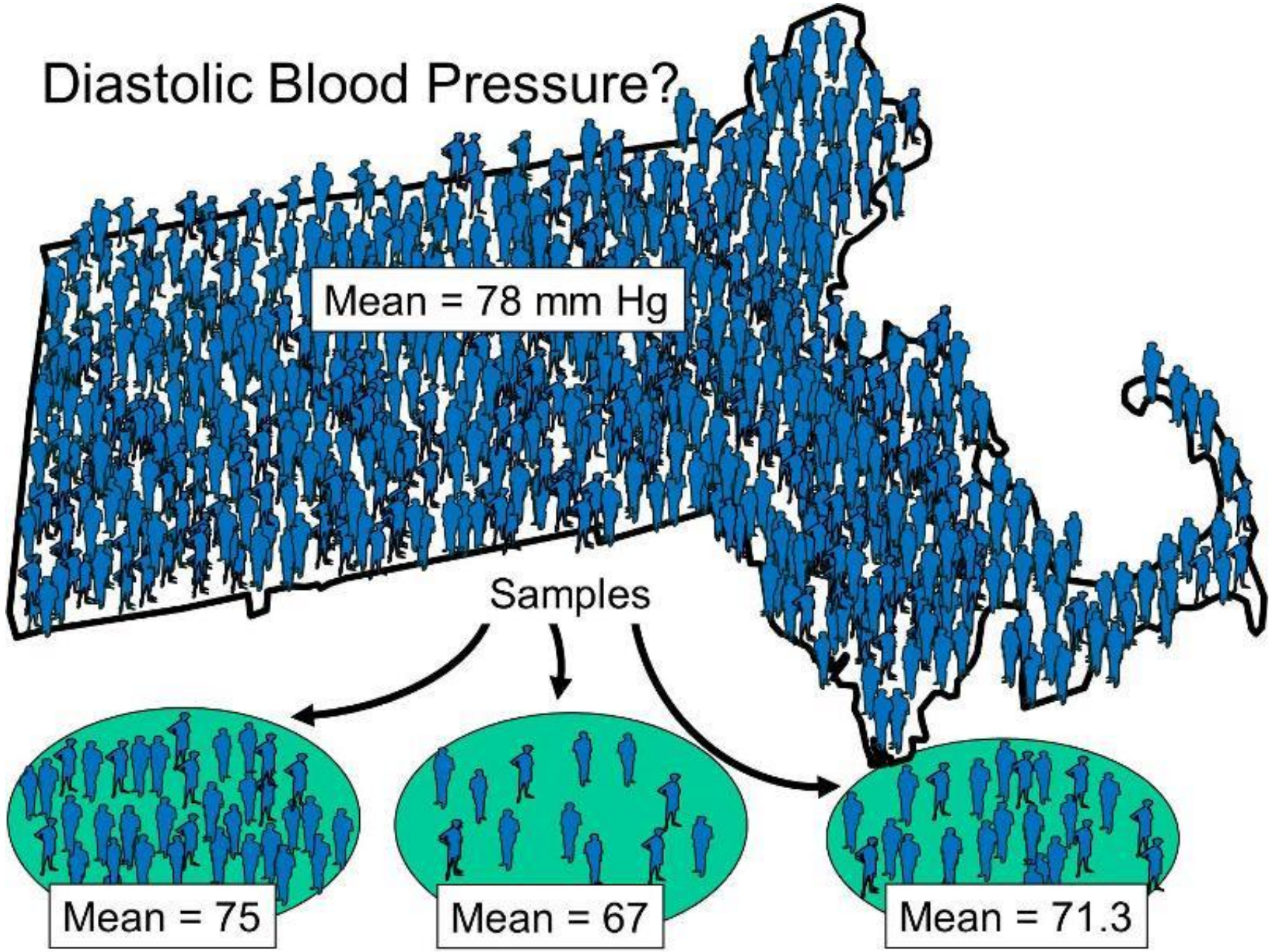
Παράμετροι πληθυσμού έναντι στατιστικών δειγμάτων

Όπως σημειώθηκε στην εισαγωγή, μία θεμελιώδης αποστολή της βιοστατιστικής είναι η *ανάλυση – μελέτη δειγμάτων* προκειμένου ***να εξάγει συμπεράσματα*** σχετικά με τον πληθυσμό από τον οποίο αντλήθηκαν τα δείγματα.

Για να το αναδείξουμε αυτό, λάβετε υπόψη τον πληθυσμό της Μασαχουσέτης το 2010, ο οποίος αποτελούταν από 6.547.629 άτομα. Ένα χαρακτηριστικό (ή μεταβλητή) δυνητικού ενδιαφέροντος μπορεί να είναι ***η διαστολική αρτηριακή πίεση*** του πληθυσμού της Μασαχουσέτης .

Υπάρχουν διάφοροι τρόποι αναφοράς και ανάλυσης αυτού, οι οποίοι θα εξεταστούν στην ενότητα για τη σύνοψη δεδομένων. Ωστόσο, προς το παρόν, θα επικεντρωθούμε στη ***μέση διαστολική αρτηριακή πίεση*** όλων των ανθρώπων που ζουν στη Μασαχουσέτη.

Προφανώς δεν είναι εφικτό να μετρηθούν και να καταγραφούν οι αρτηριακές πιέσεις για όλους τους κατοίκους, αλλά θα μπορούσε κανείς να πάρει δείγματα του πληθυσμού για να εκτιμήσει τη μέση διαστολική αρτηριακή πίεση του πληθυσμού.


Diastolic Blood Pressure?
Mean = 78 mm Hg
Samples
Mean = 75
Mean = 67
Mean = 71.3

Παρά την απλότητα αυτού του παραδείγματος, εγείρει μια σειρά από έννοιες και όρους που πρέπει να οριστούν. Οι όροι ***πληθυσμός***, ***υποκείμενα***, ***δείγμα***, ***μεταβλητή*** και ***στοιχεία δεδομένων*** ορίζονται σε παρακάτω πίνακα.

Είναι δυνατό να επιλεγούν πολλά δείγματα από έναν δεδομένο πληθυσμό και θα δούμε σε άλλες μαθησιακές ενότητες ότι υπάρχουν διάφορες μέθοδοι που μπορούν να χρησιμοποιηθούν για την επιλογή *υποκειμένων* από έναν πληθυσμό σε δείγμα.

Το παραπάνω απλό παράδειγμα δείχνει τρία μικρά δείγματα που ελήφθησαν για να εκτιμηθεί η μέση διαστολική αρτηριακή πίεση των κατοίκων της Μασαχουσέτης, αν και δεν διευκρινίζει πώς λήφθηκαν τα δείγματα. Σημειώστε επίσης ότι καθένα από τα δείγματα παρείχε *διαφορετική* ***εκτίμηση της μέσης τιμής για τον πληθυσμό*** και καμία από τις εκτιμήσεις δεν ήταν ίδια με την πραγματική μέση τιμή για το συνολικό πληθυσμό (78 mm Hg σε αυτό το υποθετικό παράδειγμα). Στην πραγματικότητα, κανείς γενικά δεν γνωρίζει τις πραγματικές μέσες τιμές των χαρακτηριστικών του πληθυσμού, γι' αυτό φυσικά προσπαθούμε να τις εκτιμήσουμε από δείγματα. Ως εκ τούτου, είναι σημαντικό να ορίσουμε και να διακρίνουμε μεταξύ:

- μέγεθος πληθυσμού σε σχέση με το μέγεθος του δείγματος
- *παράμετρος του πληθυσμού* σε σχέση με την *παράμετρο του δείγματος* (που υπολογίστηκε από το δείγμα).

Sample Statistics

In order to illustrate the computation of sample statistics, we selected a small subset (n=10) of participants in the Framingham Heart Study. The data values for these ten individuals are shown in the table below. The rightmost column contains the body mass index (BMI) computed using the height and weight measurements. We will come back to this example in the module on Summarizing Data, but it provides a useful illustration of some of the terms that have been introduced and will also serve to illustrate the computation of some sample statistics.

Δειγματικές Στατιστικές

Προκειμένου να επεξηγήσουμε τον υπολογισμό των στατιστικών στοιχείων από δείγμα, επιλέξαμε ένα μικρό υποσύνολο (n=10) συμμετεχόντων από τη μελέτη Framingham. Οι τιμές δεδομένων για αυτά τα δέκα άτομα φαίνονται στον παρακάτω πίνακα. Η δεξιά στήλη περιέχει τον δείκτη μάζας σώματος (ΔΜΣ) που υπολογίζεται χρησιμοποιώντας τις μετρήσεις ύψους και βάρους. Αυτό το παράδειγμα παρέχει μια χρήσιμη απεικόνιση ορισμένων από τους όρους που έχουν ήδη αναφερθεί και θα χρησιμεύσει επίσης για την απεικόνιση του υπολογισμού ορισμένων στατιστικών.

Στοιχεία από Μικρό Δείγμα

Participant ID **Συμμέτοχος**	Systolic Blood Pressure **Συστολική Πίεση**	Diastolic Blood Pressure **Διαστολική Πίεση**	Total Serum Cholesterol **Ολική Χοληστερίνη**	Weight **Βάρος**	Height **Ύψος**	Body Mass Index **Δείκτης Μάζας Σώματος**
1	141	76	199	138	63.00	24.4
2	119	64	150	183	69.75	26.4
3	122	62	227	153	65.75	24.9
4	127	81	227	178	70.00	25.5
5	125	70	163	161	70.50	22.8
6	123	72	210	206	70.00	29.6
7	105	81	205	235	72.00	31.9
8	113	63	275	151	60.75	28.8
9	106	67	208	213	69.00	31.5
10	131	77	159	142	61.00	26.8

Το πρώτο στοιχείο που είναι σημαντικό να αναφερθεί είναι το μέγεθος του δείγματος. Σε αυτό το παράδειγμα το μέγεθος του δείγματος είναι n=10. Επειδή αυτό το δείγμα είναι μικρό (n=10), είναι εύκολο να συνοψίσουμε το δείγμα ελέγχοντας τις παρατηρούμενες τιμές, για παράδειγμα, αναφέροντας τις διαστολικές πιέσεις του αίματος σε αύξουσα σειρά::

62 63 64 67 70 72 76 77 81 81

Η απλή επισκόπηση αυτού του μικρού δείγματος μας δίνει μια αίσθηση του *«κέντρου»* των παρατηρούμενων διαστολικών πιέσεων και επίσης μας δίνει μια αίσθηση του πόση *«μεταβλητότητα»* υπάρχει. Ωστόσο, για ένα μεγάλο δείγμα, η επισκόπηση των μεμονωμένων τιμών δεδομένων δεν παρέχει μια ουσιαστική περίληψη και είναι απαραίτητη μια στατιστική επεξεργασία. Τα δύο βασικά στοιχεία μιας χρήσιμης επεξεργασίας για μια συνεχή μεταβλητή είναι:

- μια περιγραφή του κέντρου / κεντρικής τιμής ή του «μέσου» όρου των δεδομένων (δηλαδή, ποια είναι μια τυπική κεντρική τιμή;)
- Και μια ένδειξη της «μεταβλητότητας» των δεδομένων.

Δειγματικός Μέσος

Υπάρχουν πολλά στατιστικά στοιχεία που περιγράφουν το «κέντρο» των δεδομένων, αλλά προς το παρόν θα επικεντρωθούμε στον ***μέσο όρο*** του δείγματος, ο οποίος υπολογίζεται αθροίζοντας όλες τις τιμές για μια συγκεκριμένη μεταβλητή στο δείγμα και διαιρώντας με το μέγεθος του δείγματος. Για το δείγμα της διαστολικής αρτηριακής πίεσης στον παραπάνω πίνακα, η μέση τιμή του δείγματος υπολογίζεται ως εξής:

$$\text{Sample Mean} = \frac{62+63+64+67+70+72+76+77+81+81}{10} = 71.3$$

Για να απλοποιήσουμε τους τύπους για τα στατιστικά στοιχεία του δείγματος (και για τις παραμέτρους πληθυσμού), συνήθως συμβολίζουμε τη μεταβλητή ενδιαφέροντος ως "X". Το X είναι απλώς ένα σύμβολο κράτησης θέσης για τη μεταβλητή που αναλύεται. Εδώ X=διαστολική αρτηριακή πίεση.

Ο γενικός τύπος για τον δειγματικό μέσο είναι:

$$\overline{X} = \frac{\Sigma X}{n}$$

Το X με τη γραμμή (μπάρα) πάνω του αντιπροσωπεύει τη μέση τιμή του δείγματος και διαβάζεται ως " X μπάρα" .

Το Σ υποδηλώνει άθροιση (δηλαδή, άθροισμα των X ή άθροισμα των διαστολικών πιέσεων του αίματος σε αυτό το παράδειγμα).

Διακύμανση των στοιχείων του Δείγματος και Τυπική Απόκλιση

Εάν δεν υπάρχουν ακραίες ή «απομακρυσμένες» τιμές της μεταβλητής σε ένα δείγμα, ο ***μέσος όρος*** είναι η πιο κατάλληλη «σύνοψη» / «αντιπροσώπευση» ως τυπική τιμή

και για να εκφράσουμε τη *μεταβλητότητα* στα δεδομένα εκτιμούμε συγκεκριμένα τη μεταβλητότητα των στοιχείων του δείγματος *γύρω από τη μέση τιμή του δείγματος*. Εάν όλες οι παρατηρούμενες τιμές σε ένα δείγμα είναι κοντά στη μέση τιμή του δείγματος, η ***τυπική απόκλιση*** θα είναι μικρή (δηλαδή κοντά στο μηδέν) και εάν οι παρατηρούμενες τιμές ποικίλλουν *ευρέως* γύρω από τη μέση τιμή του δείγματος, η τυπική απόκλιση θα είναι μεγάλη.

Εάν όλες οι τιμές στο δείγμα είναι πανομοιότυπες, η τυπική απόκλιση του δείγματος θα είναι μηδέν.

Κατά τη συζήτηση του μέσου όρου του δείγματος, διαπιστώσαμε ότι ο μέσος όρος του δείγματος για τη διαστολική αρτηριακή πίεση = 71,3.

Ο παρακάτω πίνακας δείχνει καθεμία από τις παρατηρούμενες τιμές μαζί με την αντίστοιχη απόκλιση από τη μέση τιμή του δείγματος.

$$\text{Deviation from Sample Mean} = X - \bar{X}$$

Table - Diastolic Blood Pressures and Deviations from the Sample Mean

X=Diastolic Blood Pressure	Deviation from the Mean
76	4.7
64	-7.3
62	-9.3
81	9.7
70	-1.3
72	0.7
81	9.7
63	-8.3
67	-4.3
77	5.7
$\Sigma X = 713$	$\Sigma\left(X - \bar{X}\right) = 0$

Οι αποκλίσεις από τον μέσο αντανακλούν *πόσο απέχει* η διαστολική αρτηριακή πίεση κάθε ατόμου από τη μέση διαστολική αρτηριακή πίεση. Η διαστολική αρτηριακή πίεση του πρώτου συμμετέχοντα είναι 4,7 μονάδες πάνω από τη μέση τιμή ενώ η διαστολική αρτηριακή πίεση του δεύτερου συμμετέχοντα είναι 7,3 μονάδες κάτω από τη μέση τιμή.

- *Αυτό που χρειαζόμαστε είναι μια «σύνοψη» αυτών των αποκλίσεων από τη μέση τιμή, ειδικότερα **ένα μέτρο** του πόσο απέχει, κατά μέσο όρο, κάθε συμμετέχων από τη μέση διαστολική αρτηριακή πίεση*. Αν υπολογίσουμε τον μέσο όρο των αποκλίσεων αθροίζοντας τις αποκλίσεις και διαιρώντας με το μέγεθος του δείγματος, αντιμετωπίζουμε πρόβλημα. Το άθροισμα των αποκλίσεων από τον μέσο όρο είναι μηδέν. Αυτό θα συμβαίνει πάντα καθώς είναι μια ιδιότητα του μέσου όρου του δείγματος, *δηλαδή, το άθροισμα των αποκλίσεων κάτω από το μέσο όρο θα ισούται πάντα με το άθροισμα των αποκλίσεων πάνω από το μέσο όρο*. Ωστόσο, ο στόχος είναι να αποτυπωθεί το μέγεθος αυτών των αποκλίσεων σε ένα «*συνοπτικό» μέτρο*. Για να αντιμετωπίσουμε αυτό το πρόβλημα των αποκλίσεων που αθροίζονται στο μηδέν, θα μπορούσαμε να πάρουμε απόλυτες τιμές ή να τετραγωνίσουμε κάθε απόκλιση από τον μέσο όρο. Και οι δύο μέθοδοι θα λύσουν το πρόβλημα. Η πιο δημοφιλής μέθοδος για τη σύνοψη των αποκλίσεων από τον μέσο όρο περιλαμβάνει τον τετραγωνισμό των αποκλίσεων (οι απόλυτες τιμές είναι δύσκολες στις μαθηματικές αποδείξεις)

Ο παρακάτω πίνακας εμφανίζει καθεμία από τις παρατηρούμενες τιμές, τις αντίστοιχες αποκλίσεις από τη μέση τιμή του δείγματος και τις τετραγωνικές αποκλίσεις από τη μέση τιμή.

X=Diastolic Blood Pressure	Deviation from the Mean $(X-\overline{X})$	Squared Deviation from the Mean $(X-\overline{X})^2$
76	4.7	22.09
64	-7.3	53.29
62	-9.3	86.49
81	9.7	94.09
70	-1.3	1.69
72	0.7	0.49
81	9.7	94.09
63	-8.3	68.89
67	-4.3	18.49
77	5.7	32.49
$\Sigma X = 713$	$\Sigma(X-\overline{X}) = 0$	$\Sigma(X-\overline{X})^2 = 472.10$

Οι τετραγωνικές αποκλίσεις ερμηνεύονται ως εξής: ***Η τετραγωνική απόκλιση*** του πρώτου συμμετέχοντα είναι 22,09 που σημαίνει ότι η διαστολική αρτηριακή του πίεση είναι 22,09 μονάδες στο τετράγωνο μακριά από τη μέση διαστολική αρτηριακή πίεση και η διαστολική αρτηριακή πίεση του δεύτερου συμμετέχοντα είναι 53,29 μονάδες στο τετράγωνο μακριά από τη μέση διαστολική αρτηριακή πίεση.

Ένα μέγεθος που χρησιμοποιείται συχνά για τη μέτρηση της μεταβλητότητας σε ένα δείγμα ονομάζεται ***διακύμανση δείγματος*** και είναι ουσιαστικά *ο μέσος όρος των τετραγωνικών αποκλίσεων*. Η διακύμανση του δείγματος συμβολίζεται με s^2 και υπολογίζεται ως εξής:

$$\text{Sample variance} = s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}$$

Γιατί διαιρούμε με (n-1) αντί για n;

Η διακύμανση του δείγματος *δεν είναι στην πραγματικότητα ο μέσος όρος των τετραγωνικών αποκλίσεων, επειδή διαιρούμε με (n-1) αντί για n.*

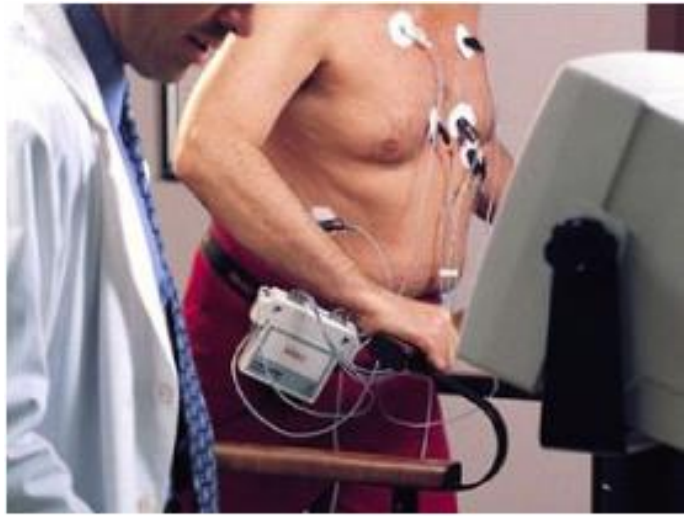
Στα στατιστικά συμπεράσματα κάνουμε γενικεύσεις ή εκτιμήσεις των παραμέτρων του πληθυσμού βάσει δειγματοληπτικών στατιστικών. Εάν υπολογίζαμε τη διακύμανση του δείγματος παίρνοντας τον μέσο όρο των τετραγωνικών αποκλίσεων και διαιρώντας με το n, θα υποτιμούσαμε σταθερά την πραγματική διακύμανση του πληθυσμού. Η διαίρεση με το (n-1) παράγει μια καλύτερη εκτίμηση της διακύμανσης του πληθυσμού. Ωστόσο, η διακύμανση αυτή του δείγματος συνήθως αναφέρεται ως η ***μέση*** τετραγωνική απόκλιση από τον μέσο όρο.

Σε αυτό το δείγμα n=10 διαστολικών πιέσεων αίματος, η διακύμανση του δείγματος είναι $s^2 = 472,10/9 = 52,46$. Έτσι, κατά μέσο όρο οι διαστολικές πιέσεις είναι 52,46 μονάδες στο τετράγωνο από τη μέση διαστολική αρτηριακή πίεση.

Λόγω του τετραγωνισμού, η διακύμανση δεν είναι ιδιαίτερα ερμηνεύσιμη. Το πιο κοινό μέτρο της μεταβλητότητας σε ένα δείγμα είναι ***η τυπική απόκλιση του δείγματος***, *που ορίζεται ως η τετραγωνική ρίζα της διακύμανσης του δείγματος:*

η τυπική απόκλιση ή τυπική απόκλιση τετραγώνου *(του δείγματος)*

$$\text{Sample standard deviation} = s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$$



Παράμετροι του πληθυσμού

Η προηγούμενη σελίδα περιέγραψε τα στατιστικά του δείγματος για τη μέτρηση της διαστολικής αρτηριακής πίεσης στο δείγμα μας. Εάν είχαμε μετρήσεις διαστολικής αρτηριακής πίεσης για όλα τα άτομα στον πληθυσμό, θα μπορούσαμε επίσης να υπολογίσουμε τις παραμέτρους του πληθυσμού ως εξής:

Ο μέσος του πληθυσμού

Συνήθως, ένας μέσος όρος πληθυσμού προσδιορίζεται με το *πεζό ελληνικό γράμμα* **μ** (προφέρεται «mu») και ο τύπος είναι ο ακόλουθος:

$$\mu = \frac{\Sigma X}{N}$$

όπου "N" είναι το μέγεθος του πληθυσμού.

Διακύμανση και τυπική απόκλιση του πληθυσμού

Οι αντίστοιχες εξισώσεις για τη ***διακύμανση*** του πληθυσμού και την ***τυπική απόκλιση*** του πληθυσμού θα είναι οι ακόλουθες (*είναι το πεζό ελληνικό γράμμα σίγμα):*

διακύμανση του πληθυσμού

$$\text{Population variance} = \sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

τυπική απόκλιση του πληθυσμού

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

Στατιστικά Συμπεράσματα *(Επαγωγική Στατιστική)*

Συνήθως δεν έχουμε πληροφορίες για όλα τα θέματα σε έναν πληθυσμό που μελετάμε, επομένως παίρνουμε δείγματα από τον πληθυσμό για να βγάλουμε συμπεράσματα για *άγνωστες παραμέτρους πληθυσμού*. Μια προφανής ανησυχία θα ήταν το πόσο καλά είναι τα στατιστικά «μέτρα» που υπολογίσαμε από τα στοιχεία ενός δεδομένου δείγματος στην εκτίμηση των χαρακτηριστικών του πληθυσμού από τον οποίο προήλθε. Υπάρχουν πολλοί παράγοντες που επηρεάζουν τα επίπεδα της διαστολικής αρτηριακής πίεσης, όπως η ηλικία, το σωματικό βάρος, η φυσική κατάσταση και η κληρονομικότητα. Θα θέλαμε ιδανικά το δείγμα να είναι αντιπροσωπευτικό του πληθυσμού. Διαισθητικά, θα ήταν προτιμότερο να υπάρχει ένα τυχαίο δείγμα, πράγμα που σημαίνει ότι όλα τα υποκείμενα του πληθυσμού έχουν ίσες πιθανότητες να επιλεγούν στο δείγμα. Αυτό θα ελαχιστοποιούσε τα *«συστηματικά»* σφάλματα που προκαλούνται από μεροληπτική δειγματοληψία. Επιπλέον, είναι επίσης διαισθητικό ότι μικρά δείγματα μπορεί να μην είναι αντιπροσωπευτικά του πληθυσμού απλώς τυχαία, και τα μεγάλα δείγματα είναι λιγότερο πιθανό να επηρεαστούν από την «τύχη της κλήρωσης». Αυτό θα μείωνε το λεγόμενο τυχαίο σφάλμα. Δεδομένου ότι συχνά βασιζόμαστε σε ένα μόνο δείγμα για την εκτίμηση των παραμέτρων του πληθυσμού, δεν γνωρίζουμε ποτέ πόσο καλές είναι οι εκτιμήσεις μας. Ωστόσο, μπορεί κανείς να χρησιμοποιήσει μεθόδους δειγματοληψίας που μειώνουν την *μεροληψία* και ο βαθμός του τυχαίου σφάλματος σε ένα δεδομένο δείγμα μπορεί να εκτιμηθεί για να αποκτήσει μια αίσθηση της ακρίβειας των εκτιμήσεών μας.