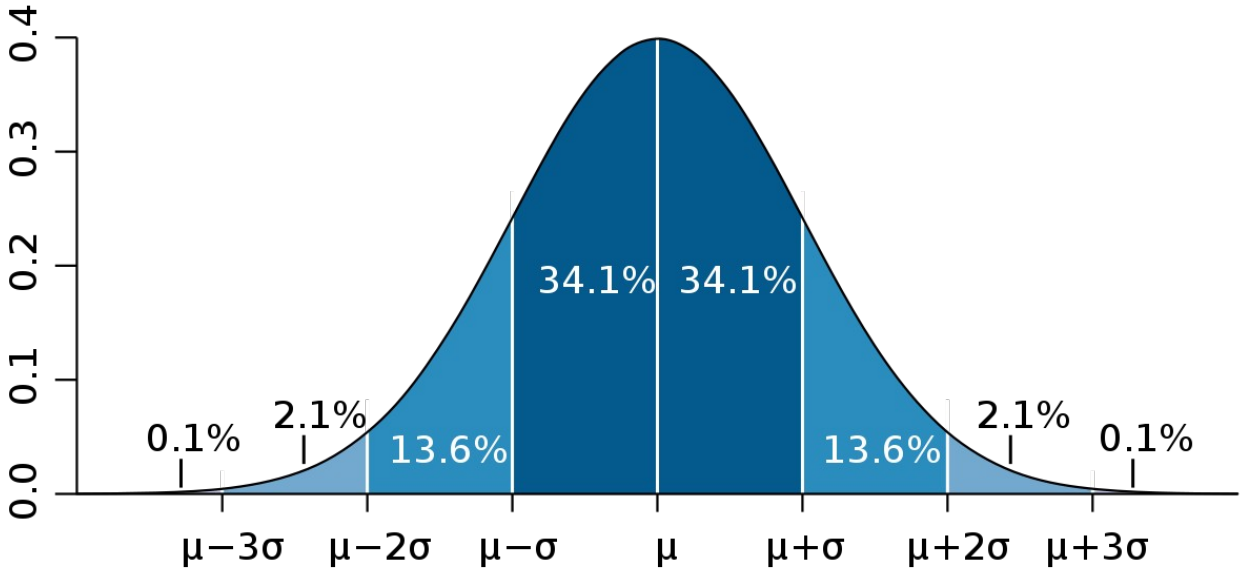
Θεωρία Πιθανοτήτων και Στατιστική



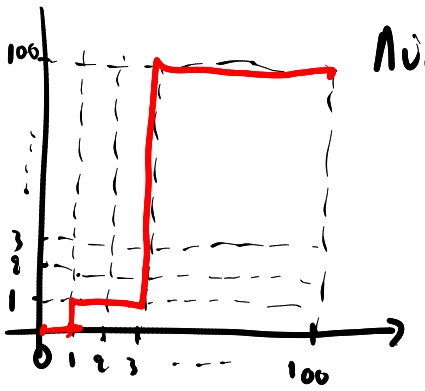
Διδάσκων: Επαμεινώνδας Διαμαντόπουλος
Επικοινωνία: epdiaman@ee.duth.gr

Ασκήσεις στο σύνολο της ύλης

1. (α) Να βρεθεί το πλήθος των δυνατών τρόπων με τους οποίους μπορούμε να πάμε από το σημείο (0, 0) στο (100, 100) πηγαίνοντας πάνω ή δεξιά με βήμα 1 σε ένα καρτεσιανό σύστημα αξόνων.

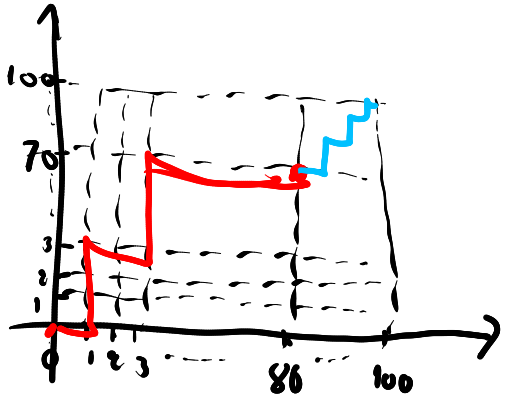
(β) Να βρεθεί το πλήθος των δυνατών τρόπων με τους οποίους μπορούμε να πάμε από το σημείο (0, 0) στο (100, 100) πηγαίνοντας πάνω ή δεξιά με βήμα 1 σε ένα καρτεσιανό σύστημα αξόνων, με την επιπλέον υποχρέωση να περάσουμε από το σημείο (80, 70).

(α)



Λύση: $\binom{200}{100}$

(β)



Λύση: $\binom{150}{80} \cdot \binom{50}{20}$.

Ασκήσεις στο σύνολο της ύλης

2. Σε μία διεργασία παραγωγής ενός μεγάλου αριθμού αντικειμένων, μπορεί να συμβούν δύο ειδών σφάλματα Α και Β, καθένα με πιθανότητα 0.25, και χωρίς να μπορούν να συμβούν μαζί. Αν το σφάλμα Α συμβεί τότε το 50% των αντικειμένων είναι ελαττωματικά, ενώ αν το σφάλμα Β συμβεί τότε το 75% των αντικειμένων είναι ελαττωματικά. Τέλος, αν δεν υπάρχει κανένα σφάλμα τότε το αντικείμενο δεν είναι ελαττωματικό. Επιλέγουμε 5 αντικείμενα από την παραγωγή.
α) Να υπολογιστεί η πιθανότητα ότι 2 από τα αντικείμενα είναι ελαττωματικά.
β) Να υπολογιστεί η πιθανότητα ότι το πολύ ένα από τα αντικείμενα είναι ελαττωματικό.

$P(A) = 0,25$, $P(B) = 0,25$

$P(E|A) = 0,5$

$P(E|B) = 0,75$

$n = 5$. $X = \{\text{πλήθος } E \text{ στα } 5\} \sim B(5, p_E)$.

$p_E = P(E) = P(E|A) \cdot P(A) + P(E|B) \cdot P(B) =$

$= 0,5 \cdot 0,25 + 0,75 \cdot 0,25 = 0,3125$

(α) $P(X=2) = \binom{5}{2} \cdot 0,3125^2 \cdot (1-0,3125)^{5-2} = 0,3173$

(β) $P(X \leq 1) = P(X=0) + P(X=1) = \binom{5}{0} \cdot 0,3125^0 \cdot 0,6875^5 + \binom{5}{1} \cdot 0,3125^1 \cdot 0,6875^4 = 0,503$

Ασκήσεις στο σύνολο της ύλης

3. Η εσωτερική διάμετρος κυλίνδρου μηχανής είναι τυχαία μεταβλητή X (cm) με $X \sim N(10, 9\cdot10^{-4})$.
Να υπολογιστεί το ποσοστό των κυλίνδρων με εσωτερική διάμετρο :
(α) Μικρότερη από 9,95 cm.
(β) Μεταξύ 9,95 και 10,06 cm.
(γ) Ποια είναι η πιθανότητα αν επιλέξουμε 10 κυλίνδρους, οι 3 από αυτούς να έχουν διάμετρο μεταξύ 9,95 και 10,06cm;

$\mu = 10, \quad \sigma^2 = 9\cdot10^{-4} \Rightarrow \sigma = 3\cdot10^{-2} = 0,03.$

(α) $P(X < 9,95) = P\left(Z < \frac{9,95-10}{0,03}\right) = P(Z < -1,67) = 0,047$

$-1,67$ ↗ $-1,6$ γραμμή, ↘ $0,07$ στήλη

(β) $P(9,95 < X < 10,06) = P\left(\frac{9,95-10}{0,03} < Z < \frac{10,06-10}{0,03}\right)$

$= P(-1,67 < Z < 2) = \Phi(2) - \Phi(-1,67)$

$= 0,977 - 0,047 = 0,93$

(γ) X: {πλήθος στους 10 με διάμετρο 9,95-10,06} $\sim B(10, 0.93)$ και $P(X=3) = \binom{10}{3} \cdot 0.93^3 \cdot 0,07^7$

Ασκήσεις στο σύνολο της ύλης

4. Το πλήθος των αιτημάτων σε έναν διακομιστή ιστού ακολουθεί την κατανομή Poisson(λ), όπου λ = 50 ανά λεπτό. Να βρεθεί η πιθανότητα να υπάρξουν τουλάχιστον 90 αιτήματα σε διάστημα 2 λεπτών.

$X \sim Poisson(\lambda)$, $X =$ {πλήθος αιτημάτων σε 2 λεπτά} $\sim Poisson(100)$

$\lambda = 50$ αιτ./min $= 100$ αιτ./2min | $X \sim Poisso(100) \approx N(100, 100)$

$$P(X \geq 90) = P_N(X \geq 89.5) = P_N\left(Z \geq \frac{89.5 - 100}{\sqrt{100}}\right) =$$

$$= P_N(Z \geq -1,15) = 1 - P_N(Z < -1,15) = 1 - \Phi(-1,15)$$

$$= 1 - 0,125 = 0,875.$$

Ασκήσεις στο σύνολο της ύλης

5. Ένα αρχείο κειμένου περιέχει 1000 λέξεις. Κάθε λέξη, ανεξάρτητα από τις άλλες, είναι ανορθόγραφη με πιθανότητα p.
(α) Αν p = 0,015, υπολογίστε την πιθανότητα να περιέχει τουλάχιστον 20 ανορθόγραφες λέξεις.
(β) Αν p = 0,001, υπολογίστε την πιθανότητα να περιέχει τουλάχιστον 3 ανορθόγραφες λέξεις.

(α) $X = \{\text{πλήθος ανορθόγραφων λέξεων στις } 1.000\} \sim B(1.000, 0.015)$

$X \sim B(1.000, 0.015)$, $n \cdot p = 1000 \cdot 0.015 = 15 > 5$

$npq = 1.000 \cdot 0.015 \cdot 0.985 = 14.775 > 5$

$X \sim N(15, 14.775)$

$$P(X \geq 20) = P_N(X \geq 19.5) = P_N\left(Z \geq \frac{19.5 - 15}{\sqrt{14.775}}\right) = 1 - \Phi(1,171) = 0,121$$

(β) $n \cdot p = 1 < 5$, $X \sim B(1.000, 0.001)$

$P(X \geq 3) = 1 - P(X < 3) = 1 - P(X=0) - P(X=1) - P(X=2) = \cdots$

Ασκήσεις στο σύνολο της ύλης

6. Είναι γνωστό ότι το 5% των φορολογικών δηλώσεων έχουν κάποιο αριθμητικό λάθος. Να βρείτε την πιθανότητα μεταξύ 2.000 φορολογικών δηλώσεων να υπάρξουν περισσότερες από 12 δηλώσεις με αριθμητικό λάθος.

$$X = \{\text{δηλώσεις με λάθος στις } 2.000\} \sim B(2.000, 0.05) \approx N(100, 95)$$

$$P(X \geq 12) = P_N(X \geq 11.5) = P_N\left(Z \geq \frac{11,5 - 100}{\sqrt{95}}\right) = P_N(Z \geq -9.08)$$

$$= 1 - \phi(-9.08) = 1.$$

Ασκήσεις στο σύνολο της ύλης

7. Για την τυχαία μεταβλητή X, γνωρίζουμε ότι $f_X(x) = e^{-(x-\theta)}$, $\theta < x < +\infty$.

Να δειχθεί ότι η τ.μ. $Y = 2(X - \theta)$ ακολουθεί την X^2-κατανομή με δύο βαθμούς ελευθερίας.

Υπόδειξη: Αν $X \sim X^2(2)$, τότε $f_X(x) = \frac{1}{2}e^{-x/2}$, $x > 0$,

$Y = g(X)$, g μονοτονη, παρ. $\Rightarrow f_Y(y) = f_X(g^{-1}(y)) \cdot \left|\frac{d}{dy} g^{-1}(y)\right|$

$Y = 2(X-\theta) = g(X)$ με $g(x) = 2(x-\theta) = 2x - 2\theta$ ↗,

$x \in (\theta, +\infty)$, g ↗ $\Rightarrow g((\theta, +\infty)) = (g(\theta), \lim_{x \to +\infty} g(x)) = (0, +\infty)$: πεδίο ορισμού $g^{-1}(y)$

$g(x) = y \Leftrightarrow 2(x-\theta) = y \Leftrightarrow x = \frac{y+2\theta}{2} \Rightarrow g^{-1}(y) = \frac{y+2\theta}{2} \Rightarrow \frac{d}{dy} g^{-1}(y) = \frac{1}{2}$.

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y) =$$

$$= e^{-\left(\frac{y}{2}+\theta-\theta\right)} \cdot \frac{1}{2}$$

$$= \frac{1}{2} \cdot e^{-\frac{y}{2}}, \; y>0 = \chi^2(2) \sim Exp(1) \sim \Gamma\left(1, \frac{1}{2}\right)$$

Περιεχόμενα 7ου μαθήματος

- Κεντρικό Οριακό Θεώρημα.
- Εισαγωγή στην Στατιστική
- Δειγματοληψία
- Περιγραφική Στατιστική
- Αμερόληπτοι και συνεπείς εκτιμητές.
- Εκτιμητές μέγιστης πιθανοφάνειας.
- Διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού
- Διάστημα εμπιστοσύνης για την αναλογία ενός χαρακτηριστικού στον πληθυσμό.

Γνωστικοί στόχοι 7ου μαθήματος

Στο τέλος αυτού του μαθήματος, ο φοιτητής πρέπει να είναι σε θέση :

- Να χρησιμοποιεί το κεντρικό οριακό θεώρημα για να υπολογίζει πιθανότητες με προσέγγιση κανονικής κατανομής.
- Να γνωρίζει τη διαφορά μεταξύ πιθανοθεωρητικής και μη πιθανοθεωρητικής δειγματοληψίας.
- Να μπορεί να περιγράψει με απλά στατιστικά τα γεωμετρικά χαρακτηριστικά της κατανομής ενός δείγματος τιμών μίας μεταβλητής.
- Να μπορεί να σχεδιάσει και να εκμαιεύσει πληροφορίες από τα βασικά διαγράμματα που περιγράφουν την κατανομή τιμών μίας μεταβλητής (ιστόγραμμα, ραβδόγραμμα, κυκλικό διάγραμμα, θηκόγραμμα)
- Να μπορεί να αποδείξει ότι ο αριθμητικός μέσος είναι αμερόληπτος και συνεπής εκτιμητής για την αναμενόμενη τιμή του πληθυσμού.
- Να υπολογίζει τον εκτιμητή μέγιστης πιθανοφάνειας για μία άγνωστη παράμετρο.
- Να υπολογίζει διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού.

Προσέγγιση των B(n, p) και Poisson(λ) από την Κανονική

Αποδεικνύεται ότι:

Αν **$X \sim B(n, p)$**, $n > 30$ και $np > 5$, $nq > 5$ τότε **$X \approx N(np, npq)$**, $q = 1- p$.

Αν **$X \sim Poisson(\lambda)$** και $\lambda > 20$, τότε **$X \approx N(\lambda, \lambda)$** ($\mu = \sigma^2 = \lambda$).

Στους τελικούς υπολογισμούς, έχει καθιερωθεί η εξής προσαρμογή ως μία πρακτική που δίνει καλύτερα προσεγγιστικά αποτελέσματα (continuity correction factor):

- $P_B(X \le \alpha) = P_N(X < \alpha + ½)$, $P_P(X \le \alpha) = P_N(X < \alpha + ½)$,
- $P_B(X \ge \alpha) = P_N(X > \alpha - ½)$, $P_P(X \ge \alpha) = P_N(X > \alpha - ½)$,
- $P_B(X = \alpha) = P_N(\alpha - ½ < X < \alpha + ½)$, $P_P(X = \alpha) = P_N(\alpha - ½ < X < \alpha + ½)$

Γενική δυνατότητα προσέγγισης αθροίσματος τυχαίων μεταβλητών από την κανονική

Αναδεικνύεται με φυσικό τρόπο το ερώτημα:

Εκτός από τη διωνυμική και την Poisson, ποιες άλλες κατανομές μπορούν να προσεγγιστούν από την κανονική;

Η απάντηση είναι απλή: Όταν η τυχαία μεταβλητή που μας ενδιαφέρει ορίζεται ή μπορεί να θεωρηθεί ως ένα άθροισμα επιμέρους τυχαίων μεταβλητών που είναι ανεξάρτητες μεταξύ τους με ανάλογα γεωμετρικά χαρακτηριστικά στις κατανομές τους (αναμενόμενη τιμή και τυπική απόκλιση). Αυτό, είναι κάτι που συμβαίνει στις δύο κατανομές που είδαμε:

- Αν $X \sim B(n, p)$, τότε η X μπορεί να θεωρηθεί ως $X = X_1 + X_2 + \ldots + X_n$, όπου X_i είναι οι n επιμέρους Bernoulli δοκιμές με $EX_i = p$ και $VarX_i = p(1 – p)$, $i = 1, 2, \ldots, n$.
- Αν $X \sim Poisson(\lambda)$, τότε η X μπορεί να θεωρηθεί ως άθροισμα $X = X_1 + X_2 + \ldots + X_{[\lambda]}$, όπου $X_i \sim Poisson(1)$, με $EX_i = 1$ και $VarX_i = 1$, $i = 1, 2, \ldots, [\lambda]$, ([λ]:ακέραιο μέρος του λ).

Κεντρικό Οριακό Θεώρημα

Αν $X_1, X_2, \ldots, X_n$ είναι τ.μ. ανεξάρτητες μεταξύ τους με ίδια αναμενόμενη τιμή και τυπική απόκλιση, τότε το θεώρημα που εξασφαλίζει την γνώση της κατανομής του $X = X_1 + X_2 + \ldots + X_n$, είναι το περίφημο Κεντρικό Οριακό Θεώρημα (Central Limit Theorem – CLT). Αυτό έχει πολλές εκδοχές ως προς τις αναγκαίες προϋποθέσεις. Αυτή που ταιριάζει στους σκοπούς μας είναι η εξής:

Κεντρικό Οριακό Θεώρημα

Αν $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές, με ίσες αναμενόμενες τιμές μ, ίσες πεπερασμένες διακυμάνσεις σ^2 και

$$Y = X_1 + X_2 + \ldots + X_n,$$

τότε $\mathbf{(Y - n\cdot\mu) / (n^{1/2}\sigma) \sim N(0, 1)}$ ή ισοδύναμα $\mathbf{Y \sim N(n\cdot\mu, n\cdot\sigma^2)}$, καθώς $n \to \infty$.

Μία απόδειξη είναι διαθέσιμη εδώ: https://en.wikipedia.org/wiki/Central_limit_theorem#Proof_of_classical_CLT

Video: CLT

$\Delta = (-2,467)^2 - 4\cdot(-25)\cdot 1.000 \approx 100.006$

Ασκήσεις στο Κεντρικό Οριακό Θεώρημα

1. Φορτηγό μεταφέρει κομμάτια συμπιεσμένου χαρτιού. Το βάρος κάθε κομματιού είναι τ.μ. με αναμενόμενη τιμή μ = 25 kgr και τυπική απόκλιση σ = 1,5 kgr. Πόσα το πολύ κομμάτια μπορεί να μεταφέρει το φορτηγό ώστε με πιθανότητα 95% το ολικό φορτίο να είναι μικρότερο από 1tn;

Λύση

X_i : βάρος i κομματιού , $\mu = 25$, $\sigma = 1,5$

ολικό φορτίο $Y = X_1 + X_2 + \dots + X_n \overset{\text{κ.ο.θ.}}{\sim} N(25n, 1,5^2 n)$

$n = ;$ ώστε $P(Y < 1.000) = 0,95 \iff P\left(Z < \frac{1.000 - 25n}{1,5\sqrt{n}}\right) = 0,95$

$\iff \Phi\left(\frac{1.000-25n}{1,5\sqrt{n}}\right) = 0,95 \iff \frac{1.000-25n}{1,5\sqrt{n}} = 1,645 \iff$

$\iff -25n - 2,467\sqrt{n} + 1.000 = 0$, $\sqrt{n}_{1,2} = \frac{2,467 \pm 316,2}{-50} \Rightarrow \sqrt{n} = 6,28 \Rightarrow n = 39,4 \approx \underline{39}$

Ασκήσεις στο Κεντρικό Οριακό Θεώρημα

2. Σε ένα ορυχείο υπάρχουν 100 μηχανήματα εξόρυξης μεταλλευμάτων τα οποία λειτουργούν ανεξάρτητα το ένα από το άλλο. Η ποσότητα (σε τόνους) που εξορύσσεται από κάθε ένα μηχάνημα περιγράφεται από μία τ.μ. με συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \frac{1}{6}x^3 e^{-x},\ x > 0.$$

(α) Να υπολογιστεί η πιθανότητα ότι τα 100 μηχανήματα θα εξορύξουν τουλάχιστον 420 τόνους μετάλλευμα.

(β) Πόσα παραπάνω μηχανήματα πρέπει να αγοραστούν, ώστε να είμαστε σίγουροι με 97,5% πιθανότητα πως θα εξορύσσονται τουλάχιστον 420 τόνοι του υλικού;

Υπόδειξη: Αξιοποιήστε τον ορισμό της συνάρτηση Γάμμα $\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-z} dz,\ \mathrm{Re}\, z > 0$, και την ιδιότητά της Γ(n) = (n – 1)!.

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx = \int_0^{+\infty} \frac{1}{6} x^4 e^{-x} dx = \frac{1}{6}\Gamma(5) = \frac{4!}{6} = \frac{24}{6} = 4.$$

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx = \int_0^{+\infty} \frac{1}{6} x^5 e^{-x} dx = \frac{1}{6}\Gamma(6) = \frac{5!}{6} = \frac{120}{6} = 20,\ \sigma^2 = E(X^2) - [E(X)]^2 = 4$$

$Y = X_1 + X_2 + \cdots + X_{100} \overset{\text{κ.ο.θ.}}{\sim} N(400, 400)$ (α) $P(Y > 420) = 1 - \Phi\left(\frac{420-400}{\sqrt{400}}\right) = 1 - \Phi(1) = 1 - 0{,}841 = 0{,}159$

Ασκήσεις στο Κεντρικό Οριακό Θεώρημα

2. Λύση

Στατιστική

Βασικές έννοιες

Στατιστικός πληθυσμός ή απλά **πληθυσμός** ονομάζεται κάθε σύνολο, τα στοιχεία του οποίου εξετάζουμε ως προς ένα ή περισσότερα χαρακτηριστικά τους. Τα στοιχεία του πληθυσμού ονομάζονται **μονάδες** ή **άτομα**. Ο πληθυσμός μπορεί να είναι **θεωρητικός** ή **πραγματικός.**

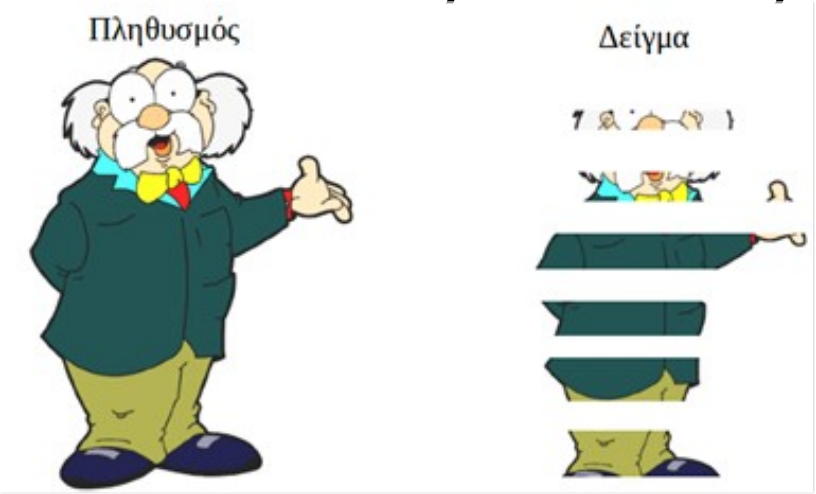
Δείγμα ονομάζεται το υποσύνολο του πληθυσμού το οποίο μπορούμε να καταγράψουμε υπό τους περιορισμούς (υλικούς και χρονικούς) της έρευνάς μας.

Οι κυριότερες μέθοδοι συλλογής στατιστικών δεδομένων είναι η **απογραφή** (census) και η **δειγματοληψία** (sampling).

Η **απογραφή** και η **δειγματοληψία**, μαζί με το **τυχαιοποιημένο πειραματικό σχέδιο** (randomized controlled trial) συνιστούν τις τρεις επιστημονικές μεθόδους με τις οποίες συλλέγονται στοιχεία και επεξεργάζονται με στατιστικές μεθόδους.

Όταν είναι εφικτή η απογραφή τότε αρκεί η **περιγραφική στατιστική** (descriptive statistics). Όταν η απογραφή είναι δύσκολη, οικονομικά και χρονικά ασύμφορη ή απλά αδύνατη, τότε είναι αναγκαία η επιλογή μιας μικρής ομάδας του πληθυσμού, δηλαδή ενός δείγματος. Συλλέγουμε τις παρατηρήσεις από το δείγμα και στη συνέχεια γενικεύουμε τα συμπεράσματα για ολόκληρο τον πληθυσμό με **επαγωγική** ή **συμπερασματική στατιστική** (inferential statistics).

Βασικές έννοιες



Το σφάλμα μίας δειγματοληψίας διαχωρίζεται σε **τυχαίο** και **συστηματικό**.

Τυχαίο σφάλμα δειγματοληψίας ονομάζεται η διαφορά μεταξύ των μετρήσεων του δείγματος και των πραγματικών μετρήσεων το οποίο θα υπάρχει στην έρευνά μας και δεν μπορούμε να το υπολογίζουμε επακριβώς εκτός αν καταφέρουμε να κάνουμε μία τέλεια εκτελεσμένη απογραφή!

Συστηματικό σφάλμα δειγματοληψίας ονομάζεται το σφάλμα που εμφανίζεται λόγω των σφαλμάτων που υπάρχουν στη σχεδίαση ή την υλοποίηση της δειγματοληψίας.

Στάδια δειγματοληψίας

Πλαίσιο δειγματοληψίας: ο φυσικός περιορισμός που ορίζεται στον πληθυσμό από το χρόνο και τόπο που διεξάγεται η δειγματοληψία.

Μέγεθος του δείγματος: ορίζεται είτε από το διαθέσιμο χρόνο και κόστος στην περίπτωση της μη πιθανοθεωρητικής δειγματοληψίας είτε με κατάλληλο υπολογισμό βάσει του επιθυμητού δειγματικού σφάλματος αν η δειγματοληψία πραγματοποιείται με κάποια πιθανοθεωρητική μέθοδο.

Πιθανοθεωρητική (propability sampling): η δειγματοληψία στην οποία κάθε μέλος του πληθυσμού έχει γνωστή πιθανότητα επιλογής πριν την υλοποίηση της δειγματοληψίας, δηλαδή είναι δυνατή η χρήση της Θεωρίας Πιθανοτήτων για τον υπολογισμό του τυχαίου σφάλματος της δειγματοληψίας.

Μη πιθανοθεωρητική (nonprobability sampling) ονομάζεται η δειγματοληψία στην οποία η πιθανότητα επιλογής των μελών του πληθυσμού είναι άγνωστη και δεν είναι δυνατή η εκ των προτέρων πιθανότητα επιλογής.

Είδη Πιθανοθεωρητικής Δειγματοληψίας

Απλή τυχαία δειγματοληψία (Simple Random Sampling): Κάθε μέλος του πληθυσμού έχει ίση πιθανότητα επιλογής στο δείγμα. Στην πράξη η απλή τυχαία δειγματοληψία συμβαίνει όταν υπάρχει η δυνατότητα να τοποθετηθεί ο πληθυσμός στη σειρά και μετά να επιλεχθεί το 10%- 15% με γεννήτρια τυχαίων αριθμών.

Συστηματική δειγματοληψία (Systematic Sampling): Η συστηματική δειγματοληψία συμβαίνει όταν θέλουμε να επιλέξουμε ένα τυχαίο δείγμα και είναι περισσότερο εύκολο να πάρουμε περιοδικό δείγμα αντί για τυχαίο, όπως για παράδειγμα σε δειγματοληψία μάρκετινγκ στην αγορά. Επιλέγεται μία αρχή με τυχαίο τρόπο και μετά επιλέγεται κάθε ν-οστό μέλος του καταλόγου. Στην πράξη: Τοποθετείται ο πληθυσμός στη σειρά 1, 2, ..., μετά επιλέγεται με τυχαίο τρόπο η πρώτη θέση (π.χ. 10), επιλέγεται το βήμα ανάλογα με το συνολικό μέγεθος του πληθυσμού (π.χ. 5) και μετά επιλέγεται το δείγμα από το $10^{ο}$, $15^{ο}$, $20^{ο}$ ... μέλος της σειράς.

Στρωματοποιημένη δειγματοληψία (Stratified Sampling): Ο ερευνητής ορίζει κάποια χαρακτηριστικά του πληθυσμού για τα οποία επιθυμεί οπωσδήποτε αναλογική εκπροσώπηση στο δείγμα του και επιλέγει απλό τυχαίο δείγμα αναλογικά από κάθε κατηγορία του πληθυσμού.

Είδη Μη Πιθανοθεωρητικής Δειγματοληψίας

Δειγματοληψία ευκολίας (convenience sampling): Το δείγμα αποτελείται από τις μονάδες του πληθυσμού που είναι διαθέσιμες εκείνη τη χρονική στιγμή.

Δειγματοληψία σκοπιμότητας (purposive sampling): Ένας εκπαιδευμένος δειγματολήπτης επιλέγει τις μονάδες του πληθυσμού που θεωρεί πως ανταποκρίνονται σε προκαθορισμένο προφίλ

Δειγματοληψία αναλογίας (quota sampling): Επιλογή του δείγματος έτσι ώστε να αντανακλάται σε αυτό η δημογραφική δομή του πληθυσμού ως προς ένα ή περισσότερα χαρακτηριστικά.

Δειγματοληψία χιονοστιβάδας (Snowball Sampling): Αρχική επιλογή ενός δείγματος με πιθανοθεωρητική μέθοδο και σε δεύτερο στάδιο συνέχιση της δειγματοληψίας από φίλο σε φίλο, από γείτονα σε γείτονα, κλπ. Συνιστάται στις περιπτώσεις που είναι επιθυμία του ερευνητή, το δείγμα να έχει κάποια συγκεκριμένα κοινωνικά ή πολιτικά χαρακτηριστικά.

Περιγραφική Στατιστική
(Descriptive Statistics)

Περιγραφή Δεδομένων

Το πρώτο μέλημα ενός ερευνητή είναι να περιγράψει με όσο το δυνατόν περισσότερη ακρίβεια, σαφήνεια και καθαρότητα τα δεδομένα τα οποία συνέλεξε. Ο τρόπος και οι μέθοδοι που θα χρησιμοποιηθούν για την περιγραφή αυτή εξαρτάται από το είδος των μεταβλητών. Συνοπτικά, στους παρακάτω πίνακες παρουσιάζονται τα βασικά μέτρα και γραφήματα που μπορούν να χρησιμοποιηθούν για την παρουσίαση των τιμών μίας μεταβλητής.

<table>
<tr><th>Είδος Μεταβλητής</th><th colspan="2">Προτεινόμενα Υπολογιστικά Μέτρα</th><th>Προτεινόμενα Γραφήματα</th></tr>
<tr><td rowspan="2">Ποιοτική
(όπως χρώμα ματιών, φύλο κ.α.)</td><td colspan="2" rowspan="2">Πίνακας Συχνοτήτων</td><td>Ραβδόγραμμα</td></tr>
<tr><td>Κυκλικό Διάγραμμα</td></tr>
<tr><td rowspan="7">Ποσοτική
(όπως ύψος, βάρος κ.α.)</td><td rowspan="3">Μέτρα θέσης</td><td>Επικρατούσα Τιμή</td><td rowspan="7">Ιστόγραμμα και Πολύγωνο Συχνοτήτων
(Για διακριτές ποσοτικές με «λίγες» τιμές είναι αποδεκτό επίσης το ραβδόγραμμα και το κυκλικό διάγραμμα)</td></tr>
<tr><td>Μέση Τιμή</td></tr>
<tr><td>Διάμεση Τιμή</td></tr>
<tr><td rowspan="4">Μέτρα διασποράς</td><td>Εύρος</td></tr>
<tr><td>Διακύμανση</td></tr>
<tr><td>Τυπική Απόκλιση</td></tr>
<tr><td>Απόλυτη Απόκλιση</td></tr>
</table>

$f_i = \frac{\nu_i}{N}$

Πίνακας Συχνοτήτων

Ρωτήθηκαν 20 γυναίκες για το πλήθος των παιδιών που έχουν και έδωσαν τις παρακάτω απόκρισεις : 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 2, 0, 4, 2, 3, 1, 0. (α) Να συμπληρωθεί ο πίνακας συχνοτήτων των παραπάνω παρατηρήσεων. (β) Να γίνει το ραβδόγραμμα συχνοτήτων και το κυκλικό διάγραμμα συχνοτήτων των τιμών.

Πλήθος (x_i)	Συχνότητα (ν_i)	Σχετική Συχνότητα (f_i)	Αθροιστική Συχνότητα (N_i)	Αθροιστική Σχετική Συχνότητα (F_i)
0	6	0,3	6	0,3
1	7	0,35	13	0,65
2	5	0,25	18	0,9
3	1	0,05	19	0,95
4	1	0,05	20	1

Μέτρα Θέσης

Επικρατούσα τιμή

Τα κυριότερα μέτρα θέσης ή κεντρικής τάσης είναι η μέση τιμή, η διάμεση τιμή και η επικρατούσα τιμή.

Επικρατούσα τιμή (Mode)
Η επικρατούσα τιμή σε ένα σύνολο δεδομένων είναι απλά η τιμή με τη μεγαλύτερη συχνότητα εμφάνισης. Όταν δύο οι περισσότερες τιμές συμπίπτουν στη συχνότητα τότε ονομάζονται όλες επικρατούσες τιμές.

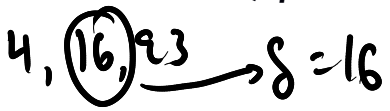
Παράδειγμα
Η επικρατούσα τιμή του δείγματος 1, 3, 3, 4, 5, 5, 6, 2, 3, 4, 3, 1, 5 είναι η τιμή 3 με συχνότητα 4.

Διάμεση τιμή

Διάμεση τιμή (Median)
Η διάμεση τιμή (ή διάμεσος) σε ένα σύνολο δεδομένων είναι απλά η μεσαία παρατήρηση αν το πλήθος των στοιχείων είναι περιττό ενώ είναι το ημιάθροισμα των δύο μεσαίων παρατηρήσεων αν το πλήθος των στοιχείων είναι άρτιο. Για να βρούμε τη διάμεσο κάνουμε τα εξής βήματα:

α) Ταξινομούμε τις παρατηρήσεις από τη μικρότερη στη μεγαλύτερη.

β) Η μεσαία παρατήρηση βρίσκεται στη θέση (n + 1) / 2.
Αν το (n + 1) / 2 είναι ακέραιος τότε η διάμεσος είναι η παρατήρηση που βρίσκεται στη θέση αυτή , ενώ αν είναι δεκαδικός τότε παίρνουμε το ημιάθροισμα των δύο παρατηρήσεων που βρίσκονται στις γειτονικές θέσεις.



Παράδειγμα
Η διάμεσος των 23, 4, 16 είναι το 16 ενώ η διάμεσος των παρατηρήσεων 23, 10, 13, 7 είναι το (10 + 13) / 2 = 11,5.

Μέση τιμή

Το νόημα της μέσης τιμής

Ως μέση τιμή ονομάζεται κάθε στατιστικό με το οποίο περιγράφεται το κέντρο των παρατηρήσεων. Ωστόσο, το «κέντρο» ενός συνόλου παρατηρήσεων που αποτελείται από πολλούς αριθμούς δεν ορίζεται μονοσήμαντα.

Ένας πρακτικός τρόπος ερμηνείας και κατανόησης της μέσης τιμής είναι η αναγνώρισή της ως το μέγεθος που θα μπορούσε να αντικαταστήσει το σύνολο των παρατηρήσεων ώστε να προκύπτει το ίδιο συνολικό αποτέλεσμα στο φυσικό πλαίσιο που ορίζεται η μεταβλητή.

Ο αριθμητικός μέσος αν και είναι η περισσότερο συχνή και απλή επιλογή δεν είναι πάντα η περισσότερο σωστή στην πράξη. Ανάλογα με το είδος των μονάδων μέτρησης της μεταβλητής, για τον υπολογισμό της μέσης τιμής μπορεί να επιλεχθεί:

- Ο αριθμητικός μέσος όταν η μεταβλητή εκφράζει καθαρές μονάδες (π.χ. ύψος ή βάρος)
- Ο αρμονικός μέσος όταν η μεταβλητή εκφράζει ρυθμό μεταβολής (π.χ. ταχύτητα)
- Ο γεωμετρικός μέσος όταν η μεταβλητή εκφράζει ποσοστιαίες μεταβολές (π.χ. επιτόκιο)

Αριθμητικός Μέσος

Αριθμητικός Μέσος (Mean ή Average)

Αν $x_1, x_2, \ldots, x_n$ είναι οι παρατηρήσεις του δείγματός μας τότε ο αριθμητικός μέσος ορίζεται να είναι:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

όπου n το μέγεθος του δείγματος ενώ αν τα στοιχεία $x_1, x_2, \ldots, x_n$ είναι όλος ο πληθυσμός για το οποίο γίνεται η έρευνα τότε γράφουμε:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Παράδειγμα

Αν 10 μαθητές έχουν βαθμολογία στη Στατιστική 12, 15, 10, 18, 17, 19, 15, 20, 13, 15 τότε η μέση βαθμολογία των μαθητών είναι:

$$\bar{x} = \frac{12 + 15 + 10 + 18 + 17 + 19 + 15 + 20 + 13 + 15}{10} = 15{,}4$$

Αριθμητικός Μέσος

Παρατήρηση
Ο αριθμητικός μέσος επηρεάζεται δυσανάλογα από τις πολύ μεγάλες ή τις πολύ μικρές παρατηρήσεις.

Πράγματι, η πρόσθεση ενός πολύ μεγάλου αριθμού στον αριθμητή του κλάσματος που ορίζει τη μέση τιμή θα τον αυξήσει δυσανάλογα σε σχέση με την αύξηση στον παρονομαστή η οποία θα είναι μόνο μία μονάδα.

Στην πράξη, αν υπάρχουν ιδιάζουσες τιμές στο δείγμα μας (πολύ μεγάλες ή πολύ μικρές παρατηρήσεις) τότε η μέση τιμή δεν αποτελεί αντιπροσωπευτικό στατιστικό του «κέντρου» των παρατηρήσεων και υπάρχουν οι εξής εναλλακτικές για την εκτίμηση του «κέντρου» της κατανομής:

(α) Ο υπολογισμός του αποκομμένου μέσου
(αφαιρείται το 5% έως 10% των πιο ακραίων παρατηρήσεων)

(β) η χρήση της διαμέσου (median)

Αρμονικός Μέσος

Αρμονικός Μέσος (Harmonic Mean)

Ο αρμονικός μέσος των παρατηρήσεων $x_1, x_2, \ldots, x_\nu$ ορίζεται να είναι η ποσότητα

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Ο αρμονικός μέσος είναι το κατάλληλο στατιστικό για τον υπολογισμό της μέσης τιμής όταν οι παρατηρήσεις εκφράζουν ρυθμούς μεταβολής.

$1\,Km \sim \frac{1}{60}\,h$

Παράδειγμα

Ένα όχημα ταξιδεύει μία συγκεκριμένη διαδρομή με ταχύτητα 60 km/h (σε χρόνο t_1) και μετά επαναλαμβάνει την ίδια διαδρομή με 40 km/h (σε χρόνο t_2). Η μέση του ταχύτητα είναι

$1\,Km \sim \frac{1}{40}\,h$

$$\bar{x}_h = \frac{2}{\frac{1}{60} + \frac{1}{40}} = 48 \text{ km/h}$$

δηλαδή αν ταξιδέψει με 48km/h θα καλύψει την απόσταση των δύο διαδρομών στον ίδιο χρόνο ($t_1 + t_2$). Είναι αξιοσημείωτο πως η τιμή αυτή διαφέρει από τον αριθμητικό μέσο (50 km/h).

Γεωμετρικός Μέσος

Γεωμετρικός Μέσος (Geometric Mean)

Ο γεωμετρικός μέσος των παρατηρήσεων $x_1, x_2, \ldots, x_n$ ορίζεται να είναι η ποσότητα

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

Ο γεωμετρικός μέσος χρησιμοποιείται για τον υπολογισμό της μέσης τιμής ποσοστιαίων μεταβολών.

Παράδειγμα

Ο γεωμετρικός μέσος των αριθμών 3 και 5 είναι: $\bar{x}_g = \sqrt{3 \cdot 5} \approx 3{,}9$

ενώ ο γεωμετρικός μέσος των αριθμών 3, 5 και 8 είναι $\bar{x}_g = \sqrt[3]{3 \cdot 5 \cdot 8} \approx 4{,}9$

Άσκηση

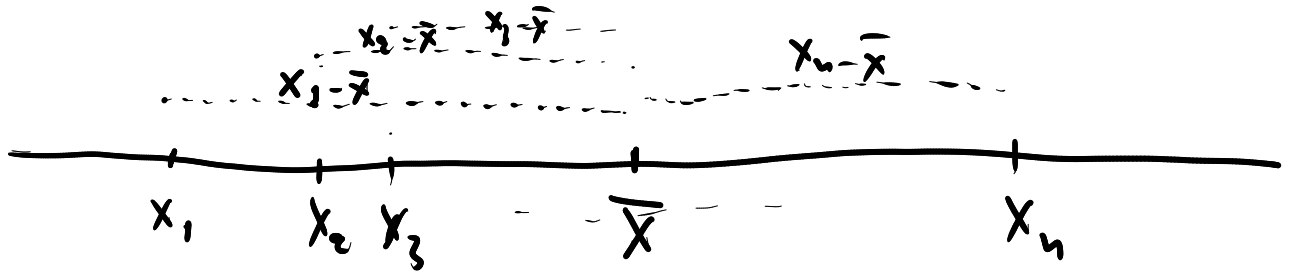
Μία μετοχή που αξίζει 100 ευρώ κερδίζει τον πρώτο χρόνο 10%, το δεύτερο χρόνο 15% και τον τρίτο χρόνο 20%. Να βρεθεί η μέση ποσοστιαία αύξηση της μετοχής.

Διάμεση Τιμή και Αριθμητικός Μέσος

Η διαφοροποίηση της διάμεσου με τη μέση τιμή ως ένδειξη ασυμμετρίας
Σε μία πλήρως συμμετρική κατανομή η μέση τιμή και η διάμεση τιμή πρέπει να ταυτίζονται.

Το αντίθετο δεν ισχύει.
Μπορεί μία κατανομή να έχει μέση τιμή ίση με τη διάμεσο αλλά να είναι ασύμμετρη. Π.χ. αυτό συμβαίνει με τις παρατηρήσεις -2, -1, 0, 0, 3

Στην πράξη, ωστόσο δεν συμβαίνει συχνά να είναι ίσες η μέση τιμή με τη διάμεσο. Αν η διάμεση τιμή είναι μικρότερη από τη μέση τιμή τότε αυτό μπορεί να ερμηνευθεί ως ένδειξη θετικής συμμετρίας (ουρά προς τα δεξιά της κατανομής) ενώ αν η διάμεση τιμή είναι μεγαλύτερη από τη μέση τιμή αυτό αποτελεί ένδειξη αρνητικής συμμετρίας (ουρά προς τα αριστερά της κατανομής).



$$\frac{1}{n}\sum_{i=1}^{n}|x_i-\bar{x}| = M.A.D.$$

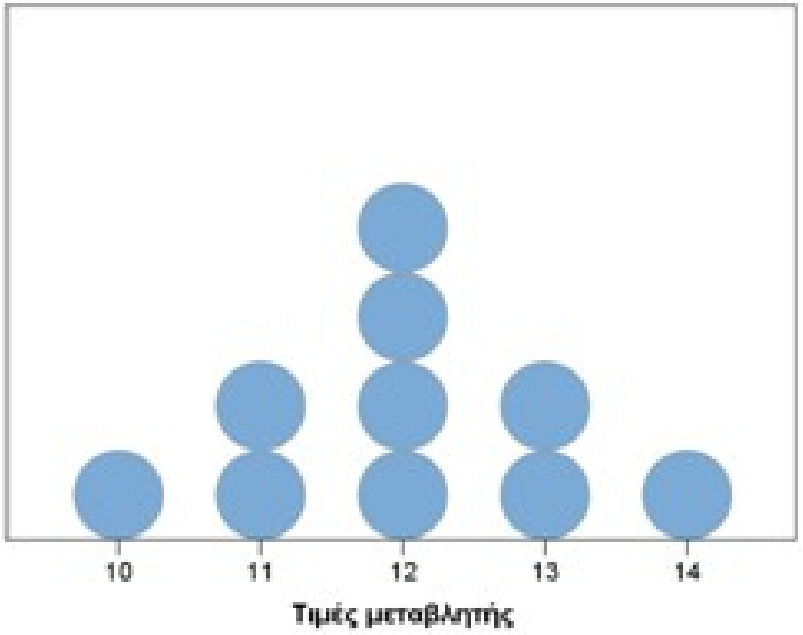
$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = s^2$$

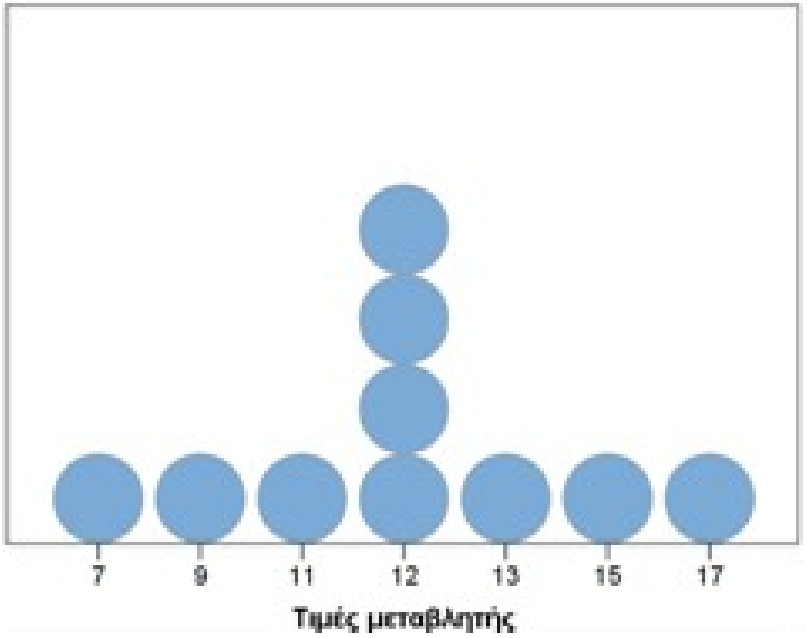
Μέτρα Διασποράς

Διακύμανση ή Variance

Μέτρα Διασποράς

Η περιγραφή της διασποράς μίας ομάδας παρατηρήσεων είναι απαραίτητη καθώς η μέση τιμή δεν δίνει πλήρη εικόνα για τη φύση της κατανομής. Χαρακτηριστικά, στην επόμενη εικόνα παρουσιάζονται δύο δείγματα με 10 τιμές που έχουν το ίδιο κέντρο (12) αλλά διαφορετική διασπορά.





Μέτρα Διασποράς

Με τη γενική ονομασία “Μέτρα Διασποράς” περιγράφουμε όλα τα στατιστικά που αποσκοπούν στην περιγραφή της διασποράς των παρατηρήσεων. Τα κυριότερα είναι τα εξής:

- Εύρος R (Range) $= Max - Min$
- Ενδοτεταρτημοριακό Εύρος IR (Interquartile Range)
- Μέση απόκλιση MAD (Mean Absolute Deviation)
- Διακύμανση Var (Variance)
- Τυπική απόκλιση StDev (Standard Deviation)

Εύρος

Το εύρος ενός δείγματος είναι απλά η διαφορά της μέγιστης από την ελάχιστη τιμή του.

R = Max – Min.

Το εύρος συνήθως συμβολίζεται με R από την αγγλική λέξη Range.

Παράδειγμα

Το εύρος των παρατηρήσεων -2, 0, 5, 4, 9, 11 είναι 11 – (-2) = 13.

Μέση απόκλιση, διακύμανση και τυπική απόκλιση

Μέση (απόλυτη) απόκλιση (mean absolute deviation) των παρατηρήσεων ονομάζεται η ποσότητα

$$\text{MAD} = \frac{1}{n}\sum_{k=1}^{n} |x_k - \bar{x}|$$

Διακύμανση ή διασπορά των παρατηρήσεων ονομάζεται η ποσότητα,

$$\sigma^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \bar{x})^2$$

Τυπική απόκλιση ονομάζεται η τετραγωνική ρίζα της διακύμανσης ή

$$\sigma = \sqrt{\sigma^2}$$

Κοινός στόχος και των τριών αυτών στατιστικών είναι να περιγράψουν την απόσταση των παρατηρήσεων από τη μέση τους τιμή.

Διακύμανση πληθυσμού vs Διακύμανση δείγματος

Άσκηση 2

Να βρεθεί η διακύμανση και η τυπική απόκλιση των 20 παρατηρήσεων

1, 0, 1, 1, 1, 5, 3, 4, 1, 2, 2, 3, 3, 2, 1, 0, 1, 4, 2, 3.

Κώδικας R

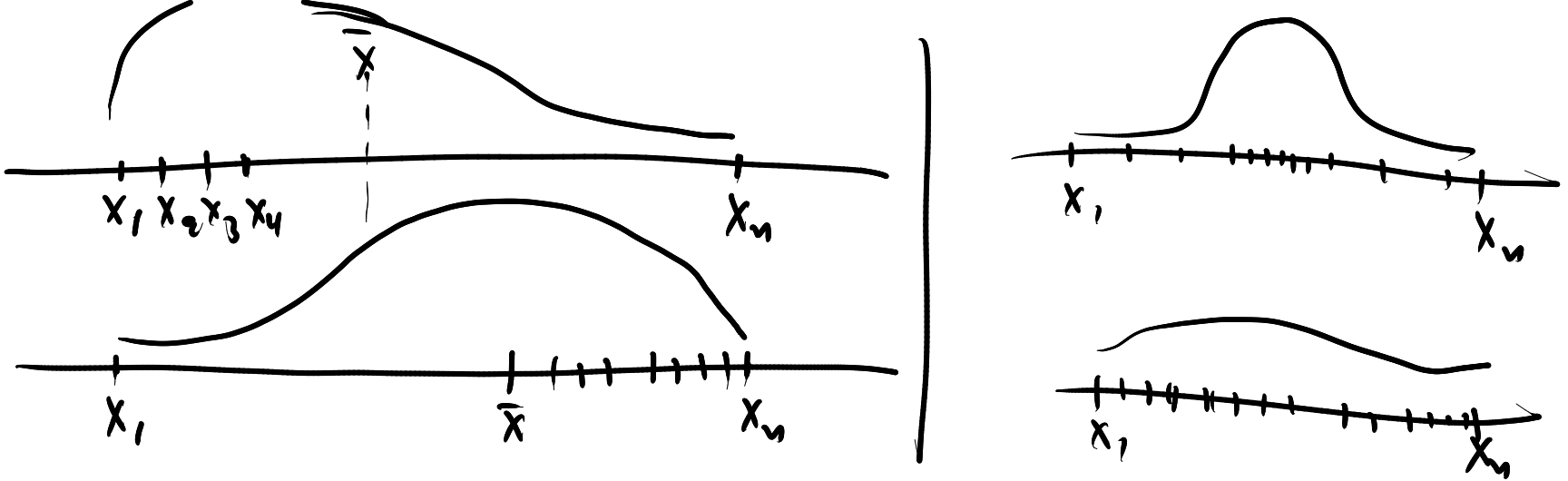
```
x = c(1, 0, 1, 1, 1, 5, 3, 4, 1, 2, 2, 3, 3, 2, 1, 0, 1, 4, 2, 3)
summary(x)
sum((x - mean(x))^2)/length(x)
```

$$\bar{x} = \frac{1+0+1+1+\dots+2+3}{20} = \frac{40}{20} = 2$$

$$s^2 = \frac{1}{20}\left[2\cdot(0-2)^2 + 7\cdot(1-2)^2 + 4\cdot(2-2)^2 + 4\cdot(3-2)^2 + 2\cdot(4-2)^2 + (5-2)^2\right]$$

$$= \frac{1}{20}\cdot(8+7+0+4+8+9) = \frac{1}{20}\cdot 36 = 1,8.$$

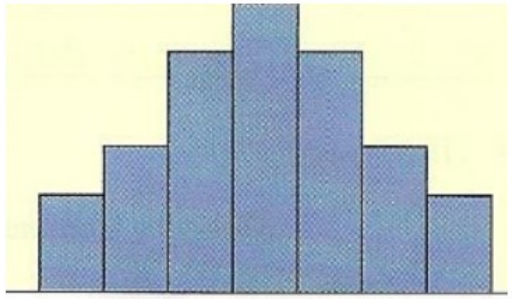
$$s = \sqrt{s^2} = \sqrt{1,8} = 1,34$$

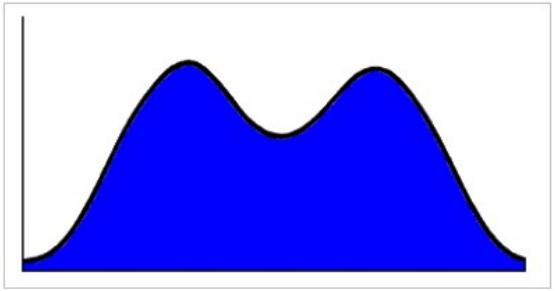


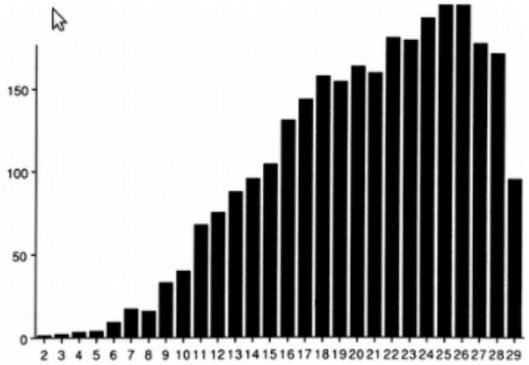
Ασυμμετρία και κυρτότητα κατανομής

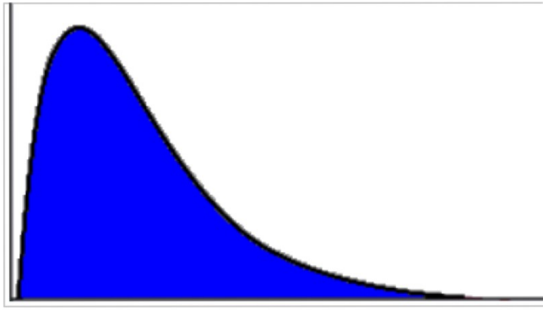
Ασυμμετρία μίας κατανομής (Skewness)

Μία κατανομή συχνοτήτων (ή σχετικών συχνοτήτων) ονομάζεται συμμετρική όταν είναι φανερό πως υπάρχει ένας κατακόρυφος άξονας ο οποίος λειτουργεί ως καθρέπτης της μισής κατανομής στην άλλη μισή. Χαρακτηριστικό παράδειγμα είναι η κανονική κατανομή χωρίς αυτό να σημαίνει πως δεν μπορεί να είναι συμμετρική μία περισσότερο “ανώμαλη” κατανομή.









Ασυμμετρία και κυρτότητα μίας κατανομής

Συντελεστής ασυμμετρίας (skew)

$$\gamma = \frac{\mu_3}{\sigma^3} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^3$$

Συντελεστής κυρτότητας (kurtosis)

$$\alpha = \frac{\mu_4}{\sigma^4} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^4$$

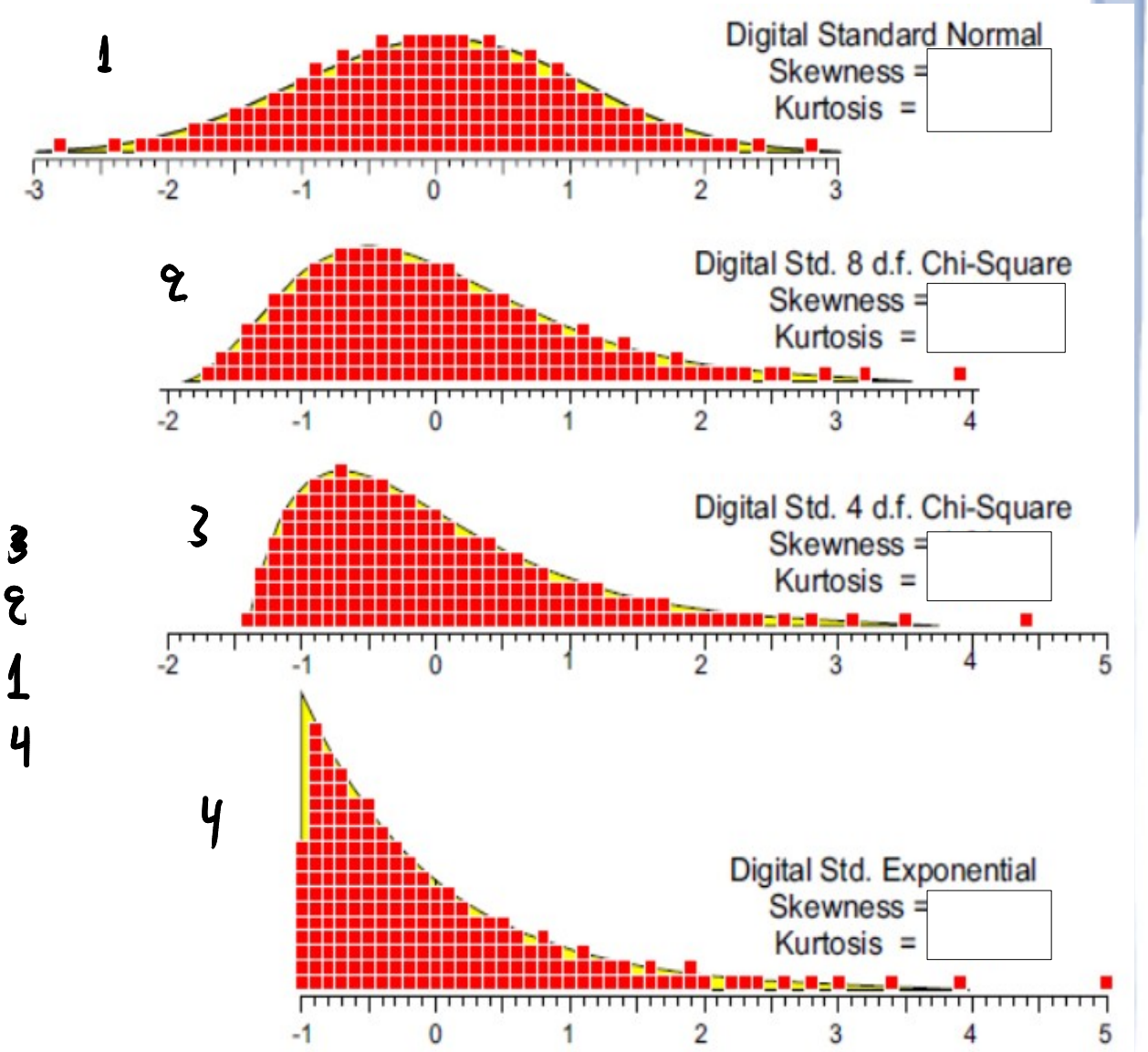
...όπου $\mu_k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^k$, $k = 1, 2, \ldots$ οι κεντρικές ροπές k τάξης.

Στη βιβλιογραφία συναντώνται και οι εκδοχές με αποκλειστική χρήση των κεντρικών ροπών: $\gamma = \frac{\mu_3}{(\mu_2)^{3/2}}$, $\alpha = \frac{\mu_4}{\mu_2^2}$

Άσκηση 1

Αντιστοιχίστε τις παρακάτω τιμές ασυμμετρίας και κυρτότητας με τις 4 δειγματικές κατανομές του σχήματος:

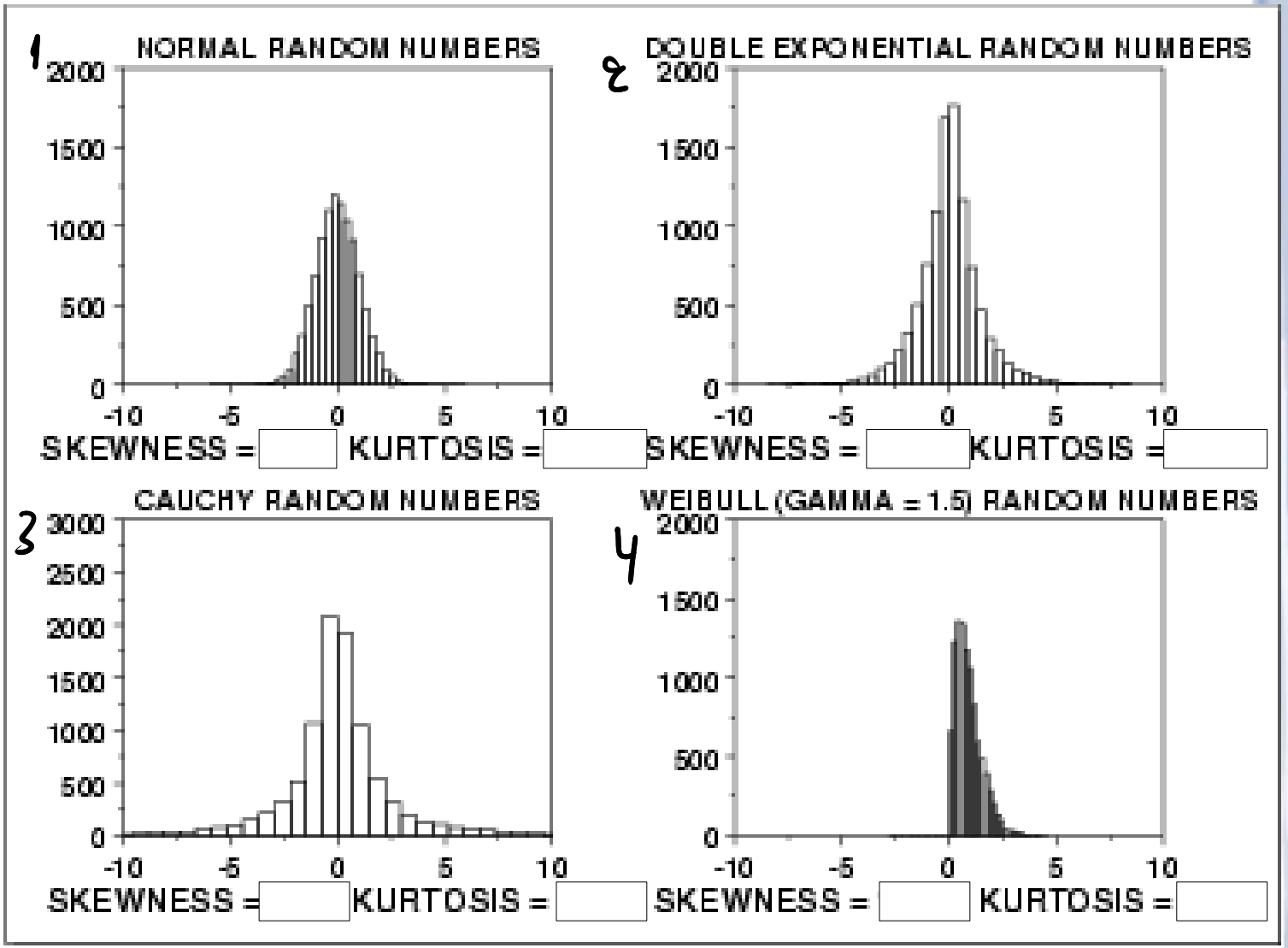
Ασυμμετρία	Κυρτότητα	
1,31	5,18	3
0,93	4,03	2
0,00	2,88	1
1,84	7,34	4



Άσκηση 2

Αντιστοιχίστε τις παρακάτω τιμές ασυμμετρίας και κυρτότητας με τις 4 δειγματικές κατανομές του σχήματος:

	Ασυμμετρία	Κυρτότητα
4	69,9	6.693
2	0,06	5,90
1	0,03	2,96
3	1,08	4,46



Επαγωγική Στατιστική
(Inferential Statistics)

Από το δείγμα στον πληθυσμό

Όταν η πρόσβαση στο σύνολο του πληθυσμού είναι μία εφικτή επιλογή τότε η περιγραφική στατιστική είναι αρκετή για έναν ερευνητή: με τη χρήση κατάλληλων διαγραμμάτων και πινάκων ο ερευνητής αποκτά πλήρη και συγκεκριμένη εικόνα για τα στοιχεία του πληθυσμού που έχει συλλέξει και μπορεί να πάρει τις επιχειρηματικές (ή άλλες) αποφάσεις του.

Ωστόσο, στις περισσότερες πραγματικές περιπτώσεις συλλογής δεδομένων δεν είναι δυνατή η κάλυψη όλου του πληθυσμού λόγω χρονικών, τεχνικών και οικονομικών περιορισμών και επιλέγεται η κάλυψη ενός υποσυνόλου του πληθυσμού, δηλαδή η μέτρηση ενός δείγματος.

Ανακύπτει με φυσικό τρόπο η ανάγκη της γενίκευσης των παρατηρήσεων από το δείγμα σε όλον τον πληθυσμό με τρόπο ώστε να προσδιορίζεται το σφάλμα που φανερά αντιστοιχεί σε αυτήν τη γενίκευση.

Στο σημείο αυτό, εμφανίζεται η **επαγωγική στατιστική** και οι μέθοδοι της.

Από το δείγμα στον πληθυσμό

Με τον όρο **Επαγωγική Στατιστική** (Inferential Statistics) περιγράφονται όλες οι στατιστικές διαδικασίες που έχουν ως σκοπό την εκμαίευση χρήσιμων δεδομένων για τον πληθυσμό από τα στοιχεία ενός αντιπροσωπευτικού δείγματός του.

Χρησιμοποιούμε Επαγωγική Στατιστική όταν δεν έχουμε πρόσβαση στον πληθυσμό.

Απαραίτητη προϋπόθεση είναι η επιλογή ενός δείγματος που να αντιπροσωπεύει όσο το δυνατόν καλύτερα τον πληθυσμό που μας ενδιαφέρει.

Αν ο ερευνητής δεν είναι βέβαιος πως το δείγμα του αντιπροσωπεύει τον πληθυσμό ως προς τα χαρακτηριστικά που τον ενδιαφέρουν τότε το αποτέλεσμα θα είναι μακριά από την αλήθεια και θα οδηγήσει σε λανθασμένες αποφάσεις.

Εκτιμητές

Εκτιμητές

Όταν γνωρίζουμε την κατανομή που ακολουθεί μία τυχαία μεταβλητή και την αντίστοιχη συνάρτηση πιθανότητας f(x), τότε έχουμε δυνατότητα υπολογισμού όλων των γεωμετρικών χαρακτηριστικών της, όπως είναι η αναμενόμενη τιμή, η διακύμανση κλπ.

Στην πράξη όμως, τις περισσότερες φορές, αν και γνωρίζουμε το είδος της κατανομής, δεν γνωρίζουμε επακριβώς τη συνάρτηση πιθανότητας, αλλά ούτε και τις παραμέτρους του πληθυσμού.

Στην περίπτωση αυτή, συλλέγουμε ένα δείγμα τιμών από τον πληθυσμό και προσπαθούμε να εκτιμήσουμε από αυτό τις άγνωστες παραμέτρους που μας ενδιαφέρουν.

Οι μέθοδοι που έχουν τον παραπάνω στόχο ανήκουν στο γνωστικό αντικείμενο της **Επαγωγικής Στατιστικής.**

Εκτιμητές

Στην επαγωγική στατιστική, αξιοποιούμε ένα δείγμα του πληθυσμού $x = (x_1, \ldots, x_n)$, για να ορίσουμε ένα δειγματικό στατιστικό $\theta_n = \theta_n(x)$ και να εκτιμήσουμε μία άγνωστη παράμετρο θ του πληθυσμού.

Παραδείγματα

Για την παράμετρο $\theta = \mu$ χρησιμοποιούμε τον $T_n = (x_1 + x_2 + \ldots + x_n) / n$.

Για την παράμετρο $\theta = \sigma^2$ χρησιμοποιούμε τον $\sigma_n^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \mu)^2$

Αν $X \sim U(\alpha, \beta)$, τότε για το $\theta = \beta$, χρησιμοποιούμε το $T_n = \max\{x_1, x_2, \ldots, x_n\}$

Αν $X \sim Exp(\lambda)$, τότε για το $\theta = \lambda$, χρησιμοποιούμε το $\lambda_n = \frac{n}{x_1 + x_2 + \ldots + x_n}$

Ο εκτιμητής ως μία Τυχαία Μεταβλητή

Στην πράξη, ένα δείγμα του πληθυσμού $x = (x_1, \ldots, x_n)$, είναι ένα αριθμητικό διάνυσμα. Ωστόσο, πρέπει να γίνει κατανοητό πως στη θέση κάθε μίας παρατήρησης θα μπορούσε να είναι οποιαδήποτε άλλη από τον θεωρητικό πληθυσμό. Δηλαδή, το τυχαίο σύνολο τιμών x μπορεί να γίνει αντιληπτό ως ένα διάνυσμα τυχαίων μεταβλητών:

$$X = (X_1, \ldots, X_n).$$

Στο πλαίσιο αυτό, αναγνωρίζουμε και τον εκτιμητή θ_n ως μία τυχαία μεταβλητή η οποία αποδίδει εκτιμήσεις για την άγνωστη παράμετρο θ από το τυχαίο δείγμα που του παρέχεται.

Αμερόληπτοι και συνεπείς εκτιμητές

Πόσο καλές είναι οι εκτιμήσεις των παραμέτρων που προκύπτουν από τους τύπους που χρησιμοποιούμε;

Στην Στατιστική είναι επιθυμητό ο εκτιμητής θ_n να είναι

- **αμερόληπτος** (unbiased)

$$\text{Bias}(\theta_n, \theta) = E_{x|\theta}(\theta_n - \theta) = 0.$$

(ανεξάρτητα από το μέγεθος του δείγματος, ο εκτιμητής δίνει τη σωστή τιμή της παραμέτρου ως αναμενόμενη τιμή)

- **συνεπής** (consistent)

$$\text{plim}_{n\to\infty}\theta_n = \lim_{n\to\infty} P(|\theta_n - \theta| > \varepsilon) = 0, \text{ για κάθε } \varepsilon > 0.$$

(η αύξηση του μεγέθους του δείγματος οδηγεί σε καλύτερη πρόβλεψη της παραμέτρου)

Και οι δύο έννοιες αποκτούν νόημα με την προϋπόθεση πως γνωρίζουμε το είδος της κατανομής που ακολουθεί ο πληθυσμός από τον οποίο πήραμε το δείγμα. Αυτό που δεν γνωρίζουμε είναι μία παράμετρος αυτής της κατανομής.

Ο αρ. μέσος είναι αμερόληπτος εκτιμητής του μ

Αν $\theta = \mu$ είναι η άγνωστη μέση τιμή όλου του πληθυσμού και $X = (X_1, \ldots, X_n)$, τότε ο αριθμητικός μέσος T_n είναι ένας αμερόληπτος εκτιμητής για τη μ. Πράγματι,

$$\begin{aligned}
\text{Bias}(T_n, \mu) &= E_{x|\mu}(T_n - \mu) \\
&= E_{x|\mu}(T_n) - \mu \\
&= E_{x|\mu}[(X_1 + \ldots + X_n)/n] - \mu \\
&= [E(X_1) + \ldots + E(X_n)]/n - \mu \\
&= (\mu + \ldots + \mu)/n - \mu \\
&= \mu - \mu \\
&= 0.
\end{aligned}$$

Ο αρ. μέσος είναι ένας συνεπής εκτιμητής του μ

Ο αριθμητικός μέσος T_n είναι ένας συνεπής εκτιμητής για τη μέση τιμή του πληθυσμού. Η απόδειξη γίνεται με τη βοήθεια του Κεντρικού Οριακού Θεωρήματος από το οποίο παίρνουμε $T_n \sim N(\mu, \sigma^2/n)$ ή $T_n - \mu \sim N(0, \sigma^2/n)$ και για κάθε $\varepsilon > 0$,

$$P(|T_n - \mu| > \varepsilon) = P(T_n - \mu > \varepsilon) + P(T_n - \mu < -\varepsilon)$$

$$= 2P(T_n - \mu > \varepsilon)$$

$$= 2[1 - P(T_n - \mu \leq \varepsilon)]$$

$$= 2[1 - P(Z < n^{1/2}\varepsilon/\sigma)]$$

$$= 2(1 - \Phi(n^{1/2}\varepsilon/\sigma)).$$

$T_n = \frac{X_1 + X_2 + \dots + X_n}{n} \rightarrow N(n\cdot\mu, n\sigma^2)$

$T_n \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow T_n - \mu \sim N(0, \frac{\sigma^2}{n})$

Από την τελευταία σχέση παίρνουμε: $\lim_{n\to\infty} P(|T_n - \mu| > \varepsilon) = \lim_{n\to\infty} 2(1 - \Phi(n^{1/2}\varepsilon/\sigma)) = 0$

Αμερόληπτος εκτιμητής για τη διακύμανση σ^2

Χρήσιμες ιδιότητες της διακύμανσης

Μία χρήσιμη ιδιότητα της διακύμανσης είναι η παρακάτω:

Θεώρημα
Αν X, Y ανεξάρτητες τότε Var(X + Y) = Var(X) + Var(Y)

Απόδειξη

$$
\begin{aligned}
\operatorname{Var}(X+Y) &= \mathrm{E}\left[(X+Y)^2\right] - (\mathrm{E}[X+Y])^2 \\
&= \mathrm{E}\left[X^2 + 2XY + Y^2\right] - (\mathrm{E}[X] + \mathrm{E}[Y])^2 \\
&= \mathrm{E}\left[X^2\right] + 2\,\mathrm{E}[XY] + \mathrm{E}\left[Y^2\right] - \left(\mathrm{E}[X]^2 + 2\,\mathrm{E}[X]\,\mathrm{E}[Y] + \mathrm{E}[Y]^2\right) \\
&= \mathrm{E}\left[X^2\right] + \mathrm{E}\left[Y^2\right] - \mathrm{E}[X]^2 - \mathrm{E}[Y]^2 \\
&= \operatorname{Var}(X) + \operatorname{Var}(Y).
\end{aligned}
$$

Χρήσιμες ιδιότητες της διακύμανσης

Πόρισμα 1
Αν $X_1, X_2, \ldots, X_n$ ανεξάρτητες τ.μ. με ίδια διακύμανση σ^2, τότε $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

Απόδειξη

$$\mathrm{Var}\left(\overline{X}\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Πόρισμα 2
Αν $X_1, X_2, \ldots, X_n$ ανεξάρτητες τ.μ. με ίδια διακύμανση σ^2, τότε $\mathrm{E}\left[(\overline{X}-\mu)^2\right] = \frac{1}{n}\sigma^2$.

Απόδειξη

Προκύπτει άμεσα από την παρατήρηση πως $\mu_{\bar{X}} = \mu_X = \mu$ και το Πόρισμα 1.

Πηγή: https://en.wikipedia.org/wiki/Variance#Sum_of_uncorrelated_variables_(Bienaym%C3%A9_formula)

Ένας αμερόληπτος εκτιμητής για το σ^2

Παράδειγμα 2

Αν η αναμενόμενη τιμή μ της ΤΜ X είναι γνωστή και σ^2 είναι η άγνωστη διακύμανση της, τότε η παράσταση

$$\sigma_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2$$

είναι αμερόληπτος εκτιμητής για το σ^2.

Απόδειξη

$$E(\sigma_n^2 - \sigma^2) = E(\sigma_n^2) - \sigma^2 = \frac{1}{n}\sum_{k=1}^{n} E(X_k - \mu)^2 - \sigma^2 = \frac{1}{n}\sum_{k=1}^{n} \sigma^2 - \sigma^2 = \sigma^2 - \sigma^2 = 0.$$

Ένας **μη** αμερόληπτος εκτιμητής για το σ^2

Παράδειγμα 2

Αν η αναμενόμενη τιμή μ της ΤΜ Χ ΔΕΝ είναι γνωστή, τότε ο εκτιμητής $S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^2$

ΔΕΝ είναι αμερόληπτος εκτιμητής για το σ^2.

Διαισθητικά, το γεγονός αυτό οφείλεται στο γεγονός πως το άθροισμα λαμβάνει τη μικρότερη τιμή του όταν η απόσταση των παρατηρήσεων μετριέται από τη μέση τους τιμή.

(Εφαρμογή 2, σελ. 98, σχολικό βιβλίο μαθηματικών γενικής παιδείας Γ Λυκείου.)

2. **Να αποδειχτεί ότι η συνάρτηση**

$$f(\lambda) = \sum_{i=1}^{\nu}(x_i - \lambda)^2 = (x_1 - \lambda)^2 + (x_2 - \lambda)^2 + \ldots + (x_\nu - \lambda)^2$$

γίνεται ελάχιστη, όταν $\lambda = \bar{x}$.

Το ερώτημα που ανακύπτει είναι:

Πόσο μεγαλύτερη πρέπει να γίνει η ποσότητα S^2_n, ώστε να προσεγγίζει με ικανοποιητικό τρόπο την άγνωστη διακύμανση σ^2 του πληθυσμού;

Ένας **μη** αμερόληπτος εκτιμητής για το σ^2

Παράδειγμα 2

Αν η αναμενόμενη τιμή μ της ΤΜ Χ ΔΕΝ είναι γνωστή, τότε ο εκτιμητής $S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^2$ ΔΕΝ είναι αμερόληπτος εκτιμητής για το σ^2.

Απόδειξη

Ένας (δειγματικός) αμερόληπτος εκτιμητής για το σ^2

Παράδειγμα 3

Αν $S^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2$ βρήκαμε ότι $E(S^2) = (n-1)\frac{\sigma^2}{n}$, από όπου συνάγεται πως το στατιστικό

$$s_n^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2$$

είναι αμερόληπτος εκτιμητής για το σ^2. Πράγματι:

$$E(s_n^2 - \sigma^2) = \frac{n}{n-1}(n-1)\frac{\sigma^2}{n} - \sigma^2 = \sigma^2 - \sigma^2 = 0.$$

Ο $s_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2$ είναι ο βασικός τύπος υπολογισμού της διακύμανσης στη γλώσσα R (συνάρτηση var), στο SPSS και στα λογιστικά φύλλα (συνάρτηση VAR)

Δύο Χρήσιμες Ανισότητες

Ανισότητα Markov

Έστω ότι X είναι μία ΤΜ με θετικές τιμές. Τότε για κάθε α > 0, ισχύει

$$P(X \geq \alpha) \leq \frac{E(X)}{\alpha}.$$

Ανισότητα Chebyshev

Έστω ότι X είναι μία ΤΜ τέτοια ώστε $\sigma^2 < +\infty$, $\sigma^2 \neq 0$. Τότε για κάθε k > 0, ισχύει

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Σημείωση: Ισοδύναμα μπορούμε να γράψουμε $P(|X - \mu| \geq k) \leq \frac{Var(X)}{k^2}$.

Ανισότητα Markov

Έστω ότι X είναι μία ΤΜ με θετικές τιμές. Τότε για κάθε α > 0, ισχύει

$$P(X \geq \alpha) \leq \frac{E(X)}{\alpha}.$$

Απόδειξη (για X συνεχής ΤΜ)

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} x f(x)\,dx = \int_{0}^{\infty} x f(x)\,dx \\
&= \int_{0}^{a} x f(x)\,dx + \int_{a}^{\infty} x f(x)\,dx \\
&\geq \int_{a}^{\infty} x f(x)\,dx \geq \int_{a}^{\infty} a f(x)\,dx = a\int_{a}^{\infty} f(x)\,dx = a\,\Pr(X \geq a)
\end{aligned}$$

$$\Pr(X \geq a) \leq \mathrm{E}(X)/a$$

Πηγή: https://en.wikipedia.org/wiki/Markov%27s_inequality

Ανισότητα Markov

Έστω ότι X είναι μία ΤΜ με θετικές τιμές. Τότε για κάθε α > 0, ισχύει

$$P(X \geq \alpha) \leq \frac{E(X)}{\alpha}.$$

Απόδειξη (για X διακριτή ΤΜ)

$$\begin{aligned} E(X) &= \sum_{x \geq a} xP(X = x) + \sum_{x < a} xP(X = x) \\ &\geq \sum_{x \geq a} aP(X = x) + 0 \\ &= a \sum_{x \geq a} P(X = x) \\ &= aP(X \geq a) \end{aligned}$$

Πηγή: https://math.dartmouth.edu/~m20x18/markov

Ανισότητα Markov

Άσκηση: Έστω η τμ $X \sim Exp(\lambda)$.
(α) Να εφαρμοστεί η ανισότητα Markov για να βρεθεί άνω φράγμα για την $P(X \geq \alpha)$, $\alpha > 0$.
(β) Να συγκριθεί η εκτίμηση με την πραγματική τιμή της $P(X \geq \alpha)$.

Λύση

$$P(X \geq a) \leq \frac{1}{a} \cdot E(X)$$

$$X \sim Exp(\lambda), \quad E(X) = \frac{1}{\lambda}.$$

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0$$

(α) $$P(X \geq a) \leq \frac{1}{a \cdot \lambda}.$$

(β) $$P(X \geq a) = e^{-\lambda \cdot a}$$ Είναι $$e^{-\lambda a} \leq \frac{1}{a \cdot \lambda}.$$

Ανισότητα Markov

Άσκηση: Έστω ότι ένα πείραμα έχει πιθανότητα επιτυχίας p και επαναλαμβάνεται μέχρι να συμβεί η πρώτη επιτυχία και η τμ X μετράει το πλήθος δοκιμών μέχρι την πρώτη επιτυχία.
(α) Να εφαρμοστεί η ανισότητα Markov για να βρεθεί άνω φράγμα για την $P(X \geq \alpha)$, $\alpha > 0$.
(β) Να συγκριθεί η εκτίμηση με την πραγματική τιμή της $P(X \geq \alpha)$.

Λύση
Υπόδειξη: X ~ Geom(p) και EX = 1/p.

$$X \sim G_T(p) \quad P(X=k) = (1-p)^{k-1} \cdot p, \quad E(X) = \frac{1}{p}$$

(α) $$P(X \geq a) \leq \frac{1}{a} \cdot E(X) = \frac{1}{a \cdot p}$$

(β) $$P(X \geq a) = \sum_{\substack{n=a \\ n-a=0}}^{+\infty} (1-p)^{n-1} \cdot p = p \sum_{k=0}^{+\infty} (1-p)^{k+a-1} = p \cdot (1-p)^{a-1} \cdot \frac{1}{1-(1-p)} =$$
$$= (1-p)^{a-1}.$$

Ανισότητα Chebyshev

Έστω ότι X είναι μία ΤΜ τέτοια ώστε $\sigma^2 < +\infty$, $\sigma^2 \neq 0$. Τότε για κάθε $k > 0$, ισχύει

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \text{ ή ισοδύναμα } P(|X - \mu| \geq k) \leq \frac{\mathrm{Var}(X)}{k^2}.$$

Απόδειξη

Η απόδειξη της ανισότητας Chebyshev είναι μία κατάλληλη εφαρμογή της ανισότητας Markov:

$$P(|X - \mu| \geq k\sigma) = P((X - \mu)^2 \geq k^2\sigma^2)$$

$$\leq \frac{E(X - \mu)^2}{k^2\sigma^2}$$

$$= \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

Ανισότητα Chebyshev

Έστω ότι X είναι μία ΤΜ τέτοια ώστε $\sigma^2 < +\infty$, $\sigma^2 \neq 0$. Τότε για κάθε $k > 0$, ισχύει

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Παραδείγματα:

- Αν $k = 1$, τότε η ανισότητα που προκύπτει είναι προφανής: $P(|X - \mu| \geq \sigma) \leq 1$.
- Αν $k = \sqrt{2}$, τότε $P(|X - \mu| \geq \sqrt{2}\sigma) \leq \frac{1}{2}$, δηλαδή η πιθανότητα μία τιμή της X να βρίσκεται έξω από το διάστημα $(\mu - \sqrt{2}\sigma, \mu + \sqrt{2}\sigma)$, είναι μικρότερη από 0,5 = 50%.
- Αν $k = 5$, τότε $P(|X - \mu| \geq 5\sigma) \leq \frac{1}{25}$, δηλαδή η πιθανότητα μία τιμή της X να βρίσκεται έξω από το διάστημα $(\mu - 5\sigma, \mu + 5\sigma)$, είναι μικρότερη από 0,04 = 4%.

Ανισότητα Chebyshev

Άσκηση: Έστω η τμ $X \sim Exp(\lambda)$. Να εφαρμοστεί η ανισότητα Chebyshev για να βρεθεί άνω φράγμα για την $P(|X - 1/\lambda| \geq \beta)$, $\beta > 0$.

Λύση

Ανισότητα Chebyshev

Άσκηση: Το πλήθος των πελατών σε ένα κατάστημα έχει μέση τιμή 100 και διακύμανση 225. Να βρεθεί ένα άνω όριο για την πιθανότητα να υπάρχουν περισσότεροι από 120 ή λιγότεροι από 80 πελάτες στο κατάστημα.

Λύση

Ανισότητα Chebyshev

Άσκηση (Ασθενής Νόμος των Μεγάλων Αριθμών - Weak Law of Large Numbers)

Έστω ότι $X_1, X_2, \ldots, X_n$ είναι ΤΜ με ίδια κατανομή πιθανότητας. τέτοια ώστε $\sigma^2 < +\infty$, $\sigma^2 \neq 0$. Τότε, να δείξετε ότι $\text{plim}_{n \to +\infty} \bar{X} = \mu$ ή ισοδύναμα ότι για κάθε $\varepsilon > 0$, είναι:

$$P\left(|\bar{X} - \mu| \geq \varepsilon\right) \to 0, \text{ καθώς } n \to \infty.$$

Ένας συνεπής εκτιμητής για το σ^2

Παράδειγμα 4

Αν $X \sim N(\mu, \sigma^2)$, και σ^2 άγνωστο τότε η ποσότητα $s_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2$ αποδεικνύεται ότι είναι συνεπής εκτιμητής του σ^2.

Απόδειξη: Γνωρίζουμε ότι $E(s_n^2) = \sigma^2$. Αρκεί να δείξω ότι $\text{plim}_{n\to\infty} s_n^2 = \sigma^2$.

Έστω $\varepsilon > 0$. Είναι:

$$P\left(\left|s_n^2 - \sigma^2\right| \geq \varepsilon\right) = P\left(\left|s_n^2 - \mu_{s_n^2}\right| \geq \varepsilon\right) \leq \frac{\text{Var}(s_n)}{\varepsilon^2} = \frac{1}{\varepsilon^2(n-1)^2}\text{Var}\left[\sum_{k=1}^{n}(X_i - \bar{X})^2\right]$$

$$= \frac{\sigma^4}{(n-1)^2}\text{Var}\left[\sum_{k=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2\right] = \frac{\sigma^4}{(n-1)^2}\text{Var}(Z_n) = \frac{\sigma^4}{\varepsilon^2(n-1)^2}2(n-1) = \frac{2\sigma^4}{\varepsilon^2(n-1)} \to 0,$$

καθώς $n \to \infty$.

Για την απόδειξη χρησιμοποιήθηκαν:
(α) η ανισότητα του Chebyshev: Αν X Τ.Μ. με $\sigma^2 < \infty$, τότε $P(|X - \mu| > k) \leq \sigma^2/k^2$,
(β) η ιδιότητα $\text{Var}(\lambda X) = \lambda^2\text{Var}(x)$,
(γ) Το γεγονός ότι $Z_n \sim \chi^2(n-1)$ και $\text{Var}(Z_n) = 2(n-1)$.

Κριτήρια

Θεώρημα (Κριτήριο συνέπειας)

Έστω ένας εκτιμητής θ_n, ο οποίος βασίζεται σ' ένα δείγμα μεγέθους n. Αν

$$\lim_{n \to +\infty} E(\theta_n) = \mu$$

και

$$\lim_{n \to +\infty} Var(\theta_n) = 0$$

τότε ο θ_n είναι ένας συνεπής εκτιμητής της θ.

Μία απόδειξη είναι διαθέσιμη εδώ: https://stats.stackexchange.com/questions/17706/how-to-show-that-an-estimator-is-consistent

Θεώρημα (Κριτήριο αμεροληψίας)

Έστω ένας εκτιμητής θ_n, ο οποίος βασίζεται σ' ένα δείγμα μεγέθους n. Αν ο εκτιμητής είναι συνεπής και έχει πεπερασμένη διακύμανση ($Var(\theta_n) < +\infty$) τότε ο θ_n είναι αμερόληπτος.

Μία απόδειξη είναι διαθέσιμη εδώ: https://math.stackexchange.com/questions/239146/consistency-and-asymptotically-unbiasedness

Άσκηση

Σε ένα πληθυσμό υπάρχει μία ιδιότητα στα αντικείμενά του με (άγνωστη) πιθανότητα p. Παρατηρούμε n αντικείμενα και καταγράφουμε το πλήθος X μεταξύ αυτών που έχουν την ιδιότητα. Να δείξετε ότι η ποσότητα $p_n = X/n$ είναι αμερόληπτος και συνεπής εκτιμητής της πιθανότητας p.

Τελικά σχόλια

Ένας εκτιμητής μπορεί να είναι αμερόληπτος αλλά όχι συνεπής

π.χ. Αν επιλέξω n παρατηρήσεις από τον πληθυσμό και ορίσω ως εκτιμητή της αναμενόμενης τιμής τον αριθμητικό μέσο T_2 των 2 μεγαλύτερων τιμών τότε αυτός είναι ένας αμερόληπτος ($E(T_2) = \mu$) αλλά όχι συνεπής εκτιμητής ($\text{plim}_{n\to\infty} T_2 \neq \mu$).

Ένας εκτιμητής μπορεί να είναι συνεπής αλλά όχι αμερόληπτος

π.χ. Αν επιλέξω n παρατηρήσεις και ορίσω $T_n = (x_1 + x_2 + \ldots + x_n)/n + 1/n$, τότε ο T_n είναι μεροληπτικός ($\text{Bias}(T_n) = E(T_n - \mu) = 1/n$), αλλά είναι συνεπής (απόδειξη με χρήση του κριτηρίου συνέπειας και του Κ.Ο.Θ.).

Ο εντοπισμός αμερόληπτων και συνεπών εκτιμητών για γενικές κατανομές είναι συνήθως δύσκολη υπόθεση.

π.χ. *Estimation Of Parameters of a Lognormal distribution.* Retrieved April 17, 2023, from https://www.jstor.org/stable/43834395