Στατιστική

Ορισμός

Στατιστική είναι το σύνολο των μεθόδων και θεωριών που εφαρμόζονται σε αριθμητικά δεδομένα προκειμένου να ληφθεί κάποια απόφαση σε συνθήκες **αβεβαιότητας**.

Βασικές έννοιες

- Η μελέτη ενός ***πληθυσμού*** είναι συνήθως δύσκολη ή αδύνατη.
- Τις περισσότερες φορές αντί του πληθυσμού εξετάζεται ένα ***δείγμα***.
- Η επιλογή δείγματος με μια ορισμένη μέθοδο ονομάζεται ***δειγματοληψία***.

- Για την εξέταση ενός δείγματος ή πληθυσμού χρησιμοποιούνται ***αριθμητικά δεδομένα*** τα οποία ονομάζονται και ***μεταβλητές.***

- Οι μεταβλητές αυτές μπορεί να είναι:

  ποσοτικές ή ***ποιοτικές***

  συνεχείς ή ***διακριτές***

  μετρήσιμες ή ***προσεγγίσεις***.

- Όλες οι θεωρίες και μέθοδοι που χρησιμοποιούνται για διερευνήσουν ένα δείγμα ή ένα πληθυσμό αποτελούν την ***Περιγραφική Στατιστική***

- Όλες οι θεωρίες και μέθοδοι που χρησιμοποιούνται για να εξάγουν συμπεράσματα για τον πληθυσμό με βάση τα στοιχεία ενός δείγματος αποτελούν την ***Επαγωγική Στατιστική***

Επαγωγική Στατιστική:

Εκτίμηση παραμέτρων πληθυσμού

Έλεγχος υποθέσεων

- Η εξαγωγή συμπερασμάτων για τον πληθυσμό μπορεί να αφορά ***χαρακτηριστικά μεταβλητών*** ή ***αιτιώδεις σχέσεις***

- Χρήση συμπερασμάτων

  Προβλέψεις

  Λήψη αποφάσεων

Εφαρμογές στα οικονομικά

Η Στατιστική έχει κυρίως εφαρμογή σε εκείνη την κατηγορία φαινομένων που ***δεν ελέγχονται πλήρως*** από τον ερευνητή. Εκεί δηλαδή που υπάρχει κάποιος βαθμός ***αβεβαιότητας***.

Τα οικονομικά φαινόμενα ανήκουν στην κατηγορία αυτή.

Συνήθως το φαινόμενο που μελετάται έχει συμβεί και είναι αδύνατο να επαναληφθεί.

Τα στατιστικά στοιχεία που χρησιμοποιούνται συνήθως έχουν την μορφή

Χρονολογικών σειρών
Διαστρωματικών δεδομένων
Δυναμικών διαστρωματικών δεδομένων

Παραδείγματα:

- Μέτρηση οικονομικής δραστηριότητας (δείκτες)
- Ποσοτικός προσδιορισμός σχέσεων οικονομικών μεταβλητών (συναρτήσεις ζήτησης, προσφοράς, κόστους κ.λ.π.)
- Ανάλυση διαχρονικών διακυμάνσεων
- Έρευνα αγοράς

Τυχαίες μεταβλητές και κατανομές πιθανότητας

Τυχαία μεταβλητή
Η μεταβλητή της οποίας οι τιμές καθορίζονται τυχαία.

Σε κάθε τιμή της τυχαίας μεταβλητής αντιστοιχεί και μια ***πιθανότητα***, η πιθανότητα η μεταβλητή να πάρει την συγκεκριμένη τιμή.

Η σχέση μεταξύ των τιμών μιας τυχαίας μεταβλητής (X_i) και των αντίστοιχων πιθανοτήτων $f(X_i)$ μας δίνει την ***κατανομή πιθανότητας*** της τυχαίας μεταβλητής.

$$f(X_i) = P(X = X_i)$$ ***Βασικές ιδιότητες*** $$f(X_i) \geq 0$$ $$\sum_i f(X_i) = 1$$

Παράδειγμα 1 :

200 διαφορετικά εισοδήματα και συχνότητα εμφάνισης

X_i	f_i	$f(X_i)$	$F(X_i)$
180	10	0.050	0.050
190	15	0.075	0.125
200	20	0.100	0.225
210	25	0.125	0.350
220	25	0.125	0.475
230	30	0.150	0.625
240	35	0.175	0.800
250	25	0.125	0.925
260	10	0.050	0.975
270	5	0.025	1.00
	200	1.00	

Εμπειρική κατανομή

Πιθανότητα = Σχετική Συχνότητα

$$f(X_i) = \frac{f_i}{\sum f_i}$$

$$F(X_i) = P(X \le X_i)$$

Παράδειγμα 2 : Συνολικός αριθμός «Γραμμάτων» σε 2 ρίψεις ενός νομίσματος.

Δειγματικός χώρος S={(Γ,Γ), (Κ,Γ), (Γ,Κ), (Κ,Κ)}

Αποτελέσματα	**Συνολικός αριθμός Γραμμάτων** X_i	**Πιθανότητα** $f(X_i)=P(X=X_i)$
(Γ,Γ)	2	1/4
(Κ,Γ)	1	1/4
(Γ,Κ)	1	1/4
(Κ,Κ)	0	1/4

f(0)=1/4

f(1)=1/4+1/4=1/2

f(2)=1/4

Συνάρτηση πιθανότητας κ συνάρτηση αθροιστικής κατανομής

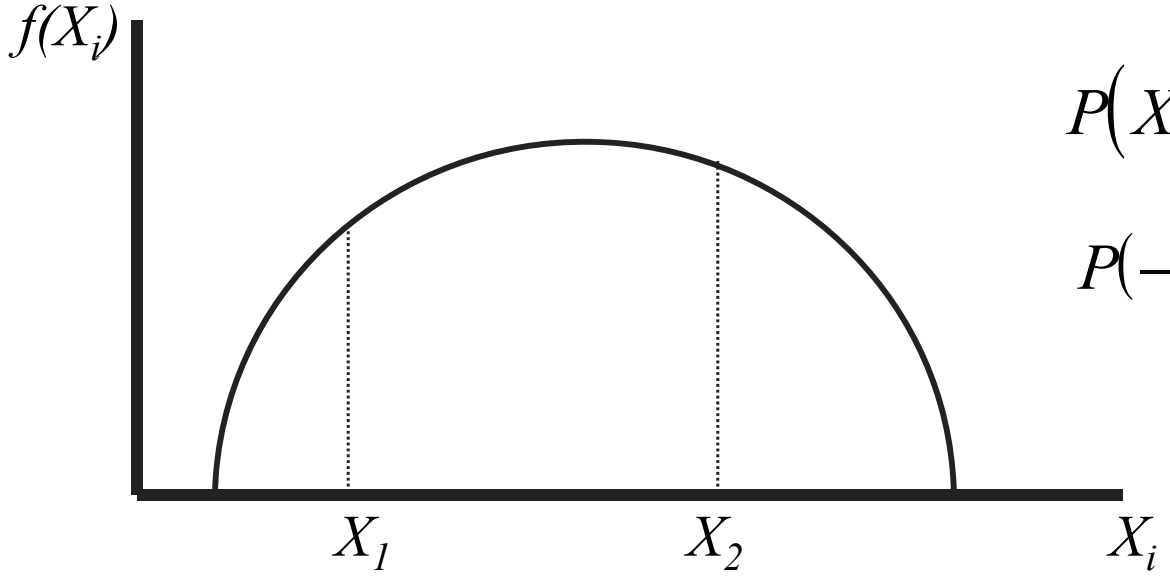
X	$f(X_i)$	$F(X)$
0	**1/4**	**1/4**
1	**1/2**	**3/4**
2	1/4	4/4

- Η $f(X_i)$ ονομάζεται ***συνάρτηση πιθανότητας*** ή ***συνάρτηση πυκνότητας.***

- Η $F(X_i)$ ονομάζεται ***συνάρτηση αθροιστικής κατανομής***

- Στην περίπτωση ***συνεχούς*** μεταβλητής η πιθανότητα μιας συγκεκριμένης τιμής είναι ίση με το ***μηδέν***.

  Για τον λόγο αυτό χρησιμοποιείται η πιθανότητα διαστήματος



$$P\left(X_1 < X < X_2\right) = \int_{X_1}^{X_2} f(X)dX$$

$$P\left(-\infty < X < \infty\right) = \int_{-\infty}^{\infty} f(X)dX = 1$$

$$P\left(X_1\right) = P\left(X_2\right) = 0$$

$$F\left(X_i\right) = P\left(-\infty < X < X_i\right) = \int_{-\infty}^{X_i} f\left(X\right)dX$$

$$P\left(X_1 < X < X_2\right) = F\left(X_2\right) - F\left(X_1\right)$$

$$F\left(-\infty\right) = 0$$

$$F\left(+\infty\right) = 1$$

$$F\left(X_1\right) < F\left(X_2\right) \text{ αν } \quad X_1 < X_2$$

Δεσμευμένη συνάρτηση πιθανότητας

Η $f(x \mid y)$ ονομάζεται δεσμευμένη συνάρτηση πιθανότητας της X και μας δίνει την πιθανότητα ότι η X θα πάρει την τιμή x δεδομένου ότι η Y παίρνει την τιμή y.

Η έννοια της δεσμευμένης συνάρτησης πιθανότητας μπορεί να επεκταθεί και σε περισσότερες από μια μεταβλητές:

Χαρακτηριστικά κατανομών πιθανότητας

Μαθηματική Ελπίδα τυχαίας μεταβλητής

$$E(X) = \sum X_i f\left(X_i\right)$$

ή $E\left(X\right) = \int_{-\infty}^{+\infty} X_i f\left(X_i\right) dX$ όταν η μεταβλητή είναι συνεχής

Ονομάζεται και ***προσδοκώμενη τιμή***

$$E(X) = \mu$$

Η μαθηματική ελπίδα δεν είναι παρά ο σταθμικός μέσος όταν ως συντελεστές στάθμισης χρησιμοποιούνται οι αντίστοιχες πιθανότητες.

Παραδείγματα:

Έστω ότι η μεταβλητή Χ είναι ο αριθμός που εμφανίζεται στην όψη ενός ζαριού. Η μαθηματική ελπίδα της Χ είναι:

$$E[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

Θεωρήματα για την μαθηματική ελπίδα:

- Θεώρημα 1:

Αν c είναι μια σταθερά τότε:

α) $\mathrm{E}c = c$

b) $\mathrm{E}\left[cg(X)\right] = c\mathrm{E}\left[g(X)\right]$

- Θεώρημα 2:

$$\mathrm{E}\left[\sum_{i=1}^{K} g_i(X)\right] = \sum_{i=1}^{K} \mathrm{E}g_i(X)$$

Δηλαδή, η προσδοκώμενη τιμή του αθροίσματος K συναρτήσεων είναι ίση με το άθροισμα των προσδοκώμενων τιμών των συναρτήσεων.

Διακύμανση τυχαίας μεταβλητής (συμβολίζεται και ως σ^2)

$$Var(X) = E(X - E(X))^2$$ ή αν μ είναι ο μέσος τότε:

$$\sigma^2 = E(X^2) - \mu^2$$

Απόδειξη:

$$\begin{aligned} V(X) &= E(X - E(X))^2 \\ &= E(X - \mu)^2 \\ &= E(X^2 + \mu^2 - 2\mu X) \\ &= E(X^2) + \mu^2 - 2\mu E(X) \\ &= E(X^2) + \mu^2 - 2\mu\mu \\ &= E(X^2) - \mu^2 \end{aligned}$$

Τυπική Απόκλιση

$$\sigma_X = \sqrt{Var(X)}$$

Συνέπειες των Θεωρημάτων 1 και 2

$$\mathrm{E}(aX+b)=\mathrm{E}(aX)+\mathrm{E}b=a\mathrm{E}(X)+b$$

$$\begin{aligned}
Var(aX+b) &= E\big((aX+b)-E(aX+b)\big)^2 \\
&= E\big(aX+b-aE(X)-b\big)^2 \\
&= E\big(aX-aE(X)\big)^2 = a^2E\big(X-E(X)\big)^2 \\
&= a^2Var(X)
\end{aligned}$$

Η έννοια της ***Συνδιακύμανσης***

Η συνδιακύμανση σ_{XY} ή Cov(X,Y) είναι μέτρο του βαθμού συσχετίσεως δύο μεταβλητών X και Y.

$$Cov(XY) = E(X - E(X))(Y - E(Y))$$

$$= E(XY - YE(X) - XE(Y) + E(X)E(Y))$$

$$= E(XY) - E(X)E(Y)$$

Αν X και Y ***ανεξάρτητες*** τυχαίες μεταβλητές

$$E(XY) = E(X)E(Y) \quad \Rightarrow \quad Cov(XY) = 0$$

Το αντίστροφο δεν ισχύει γιατί ***Cov(XY)*=0** σημαίνει ότι δεν υπάρχει ***γραμμική σχέση*** ενώ ***ανεξαρτησία*** σημαίνει ότι δεν υπάρχει ***καμία σχέση***.

$$Cov(XY) = \sigma_{XY}$$

Για την μέτρηση του βαθμού συσχετίσεως δύο μεταβλητών χρησιμοποιούμαι τον συντελεστή συσχέτισης ρ.

$$\rho = \frac{Cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Ο συντελεστής συσχέτισης είναι καθαρός αριθμός (ανεξάρτητος από μονάδες μέτρησης) και επομένως επιτρέπει συγκρίσεις του βαθμού συσχέτισης ανάμεσα σε διάφορες μεταβλητές.

Παίρνει τιμές: $-1 < \rho < 1$

ρ=0, οι μεταβλητές δε συσχετίζονται

ρ=1, τέλεια θετική συσχέτιση

ρ= -1, τέλεια αρνητική συσχέτιση

Η κανονική κατανομή

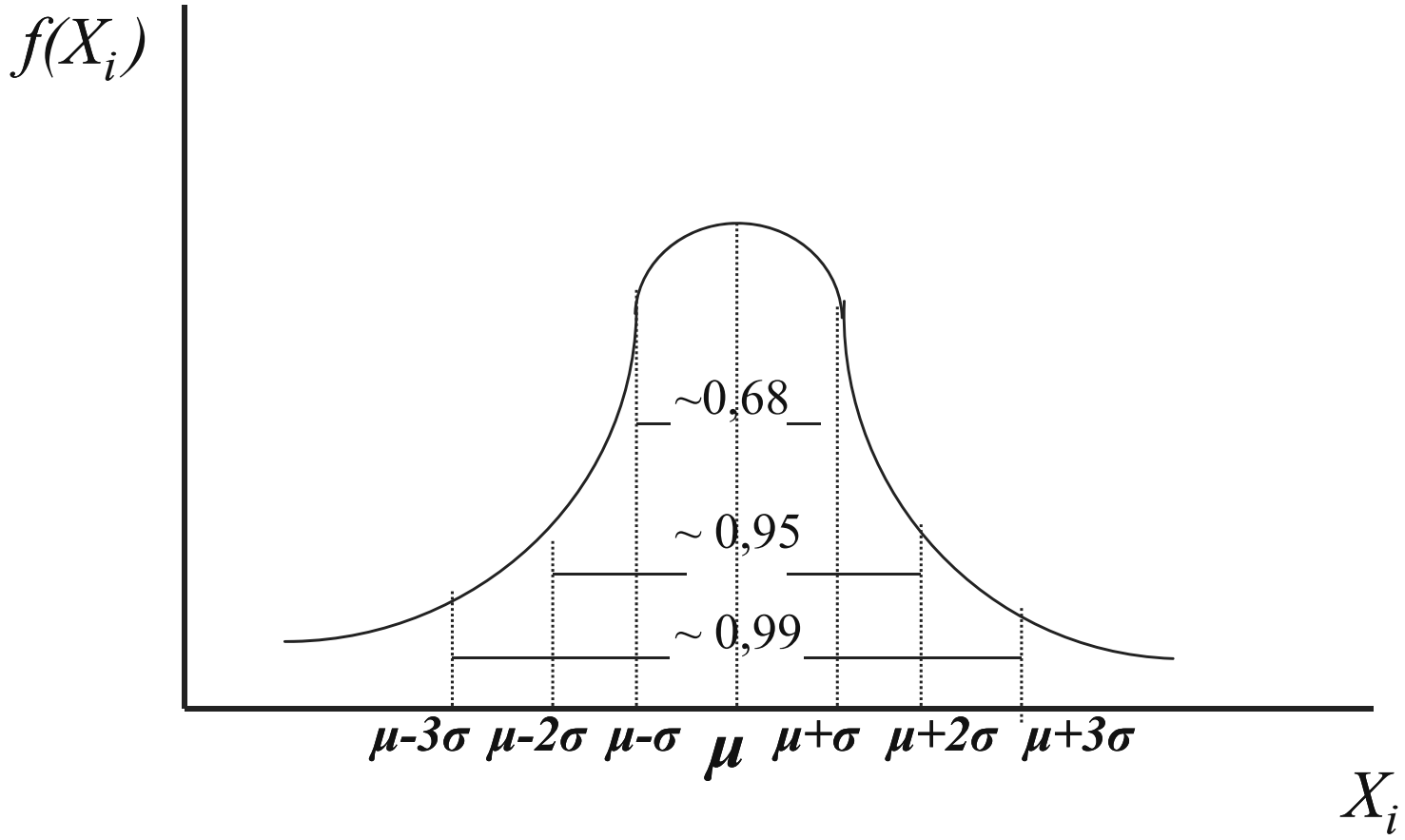
Στις περισσότερες περιπτώσεις η κατανομή πιθανότητας δεν είναι γνωστή.

Για τον λόγο αυτό καταφεύγουμε σε υποθέσεις.

Μια από τις περισσότερο δημοφιλείς υποθέσεις είναι ότι η τυχαία μεταβλητή που εξετάζεται ακολουθεί ***την Κανονική Κατανομή***.

Αν μια τυχαία μεταβλητή με μέσο μ και διακύμανση σ^2 ακολουθεί την κανονική κατανομή, δηλαδή $X \sim N(\mu, \sigma^2)$, τότε

$$f(X_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} \qquad -\infty < X < \infty$$


f(X_i)
~0,68
~ 0,95
~ 0,99
μ-3σ μ-2σ μ-σ μ μ+σ μ+2σ μ+3σ
X_i

Αν $X \sim N(\mu,\sigma^2)$, και $\boxed{Z_i = \frac{X_i - \mu}{\sigma}}$

τότε $Z \sim N(0,1)$ ***Τυποποιημένη Κανονική Κατανομή***

$$\mathrm{E}(Z) = \mathrm{E}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}\mathrm{E}(X) - \frac{\mu}{\sigma} = 0$$

$$\mathrm{V}(Z) = \mathrm{E}\left(Z - \mathrm{E}(Z)\right)^2 = \mathrm{E}(Z-0)^2 = \mathrm{E}\left(\frac{X-\mu}{\sigma}\right)^2 =$$

$$= \frac{1}{\sigma^2}\mathrm{E}(X-\mu)^2 = \frac{1}{\sigma^2}\mathrm{V}(X) = \frac{1}{\sigma^2}\sigma^2 = 1$$

Συνεπάγεται ότι

$$P(X_i \le a) = P\left(\frac{X_i - \mu}{\sigma} \le \frac{a-\mu}{\sigma}\right) = P\left(Z_i \le \frac{a-\mu}{\sigma}\right)$$

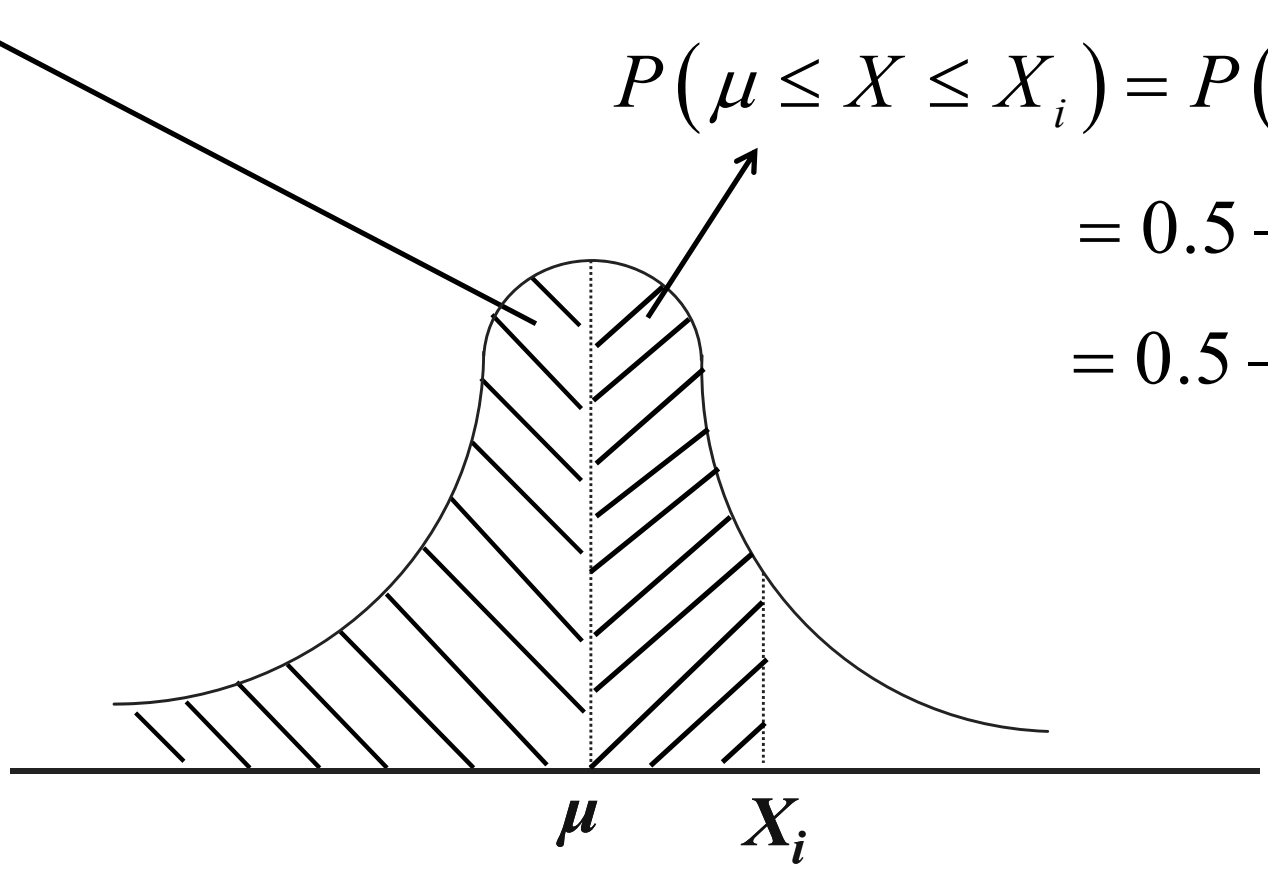
Υπολογίζεται από τους πίνακες

$$P\left(X_i \le \mu\right) = P\left(X_i \ge \mu\right) = 0.5$$

$$P\left(\mu \le X \le X_i\right) = P\left(-X_i \le X \le \mu\right)$$

$$= 0.5 - P\left(X \ge X_i\right)$$

$$= 0.5 - P\left(X \le -X_i\right)$$



Παράδειγμα 1:

$$\text{αν} \quad X \sim N\left(\mu = 8, \sigma^2 = 16\right)$$

- Να υπολογιστεί $P\left(X \le 5\right)$

$$P\left(X \le 5\right) = P\left(\frac{X-\mu}{\sigma} \le \frac{5-\mu}{\sigma}\right) = P\left(Z \le \frac{5-8}{4}\right)$$

$$= P\left(Z \le -0.75\right) = P\left(Z \ge 0.75\right) = 0.5 - P\left(0 \le Z \le 0.75\right) =$$

- Να υπολογιστεί $P\left(8 \le X \le 14\right)$

$$P\left(8 \le X \le 14\right) = P\left(\frac{8-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{14-\mu}{\sigma}\right) = P\left(\frac{8-8}{4} \le Z \le \frac{14-8}{4}\right)$$

$$= P\left(0 \le Z \le 1{,}5\right) = 43{,}32\%$$

Παράδειγμα 2:

αν $X \sim N\left(\mu = 25, \sigma^2 = 16\right)$

- Να υπολογιστεί $P\left(X \le 30\right)$

$$P\left(X \le 30\right) = P\left(\frac{X-\mu}{\sigma} \le \frac{30-\mu}{\sigma}\right) = P\left(Z \le \frac{30-25}{4} = 1,25\right)$$

- Να υπολογιστεί $P\left(X \le 20\right)$

$$P\left(X \le 20\right) = P\left(\frac{X-\mu}{\sigma} \le \frac{20-\mu}{\sigma}\right) = P\left(Z \le \frac{20-25}{4} = -1,25\right) = 0,5 - 0,3944 = 10,56\%$$

- Να υπολογιστεί $P\left(20 \le X \le 30\right)$

$P\left(20 \le X \le 30\right)$

Η κατανομή χ^2_n

Η κατανομή $F_{m,n}$

Η κατανομή t_n (Student)

Η ιστορία πίσω από το όνομα Student:

Ο William Sealy Gosset εργαζόταν στην εταιρεία ζυθοποιείας του Arthur **Guinness** & Son στο Δουβλίνο.

Ένας άλλος ερευνητής στην ζυθοποιεία είχε δημοσιεύσει ένα άρθρο το οποίο περιείχε μυστικά της εταιρίας.

Για την αποτροπή της κοινοποίησης περαιτέρω πληροφοριών η εταιρία είχε απαγορεύσει την δημοσίευση οποιουδήποτε άρθρου από υπαλλήλους της.

Έτσι ο Gosset έπρεπε να εφεύρει ένα ψευδώνυμο (*Student)* για να μην γίνει αντιληπτός.

Το επιστημονικό του επίτευγμα έμεινε γνωστό ως “Student's t-distribution”.

Αλλιώς μπορεί να το λέγαμε “Gosset's t-distribution”.

Κατανομή δειγματοληψίας

Αν X τυχαία μεταβλητή με $E(X)=\mu$ και $Var(x)=\sigma^2$, τόσο το μ όσο και το σ^2 (παράμετροι του πληθυσμού) είναι ***σταθεροί*** αριθμοί.

Αντίθετα $$\overline{X} = \frac{\sum X_i}{n}$$

$$S^2 = \frac{\sum \left(X_i - \overline{X}\right)^2}{n-1}$$

Τυχαίες μεταβλητές

Γιατί σε διαφορετικά δείγματα τα $\overline{X}$ S^2

είναι διαφορετικά

Για την εξαγωγή συμπερασμάτων σχετικά με τον πληθυσμό πρέπει να γνωρίζουμε τα χαρακτηριστικά της ***κατανομής δειγματοληψιας.***

Τι κατανομή ακολουθεί; – Ποιος είναι ο μέσος; – Ποια είναι η διακύμανση;

Δείγμα από κανονικό πληθυσμό

Αν $X_1, X_2, \ldots X_n,$ τυχαίο δείγμα από κανονικό πληθυσμό με μέσο μ και διακύμανση σ^2.

τότε $$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \overline{X} \sim N\left(\mu, \sigma^2_{\overline{X}}\right)$$

επίσης $Z \sim N(0, 1)$ όπου $$Z_i = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Σύμφωνα με το ***Κεντρικό Οριακό Θεώρημα*** τα παραπάνω ισχύουν ακόμη και όταν ο πληθυσμός δεν είναι κανονικός αρκεί το δείγμα είναι μεγάλο.

Παράδειγμα 1:

Έστω πληθυσμός $X \sim N\left(\mu = 20, \sigma^2 = 144\right)$

- Όταν $n = 36$ $\quad \bar{X} \sim N\left(20, \frac{144}{36}\right) \quad \Rightarrow \quad \bar{X} \sim N(20,4)$
- Όταν $n = 64$ $\quad \bar{X} \sim N\left(20, \frac{144}{64}\right) \quad \Rightarrow \quad \bar{X} \sim N(20,\ 2,25)$

Όταν το μέγεθος του δείγματος αυξάνει η διακύμανση της κατανομής του μέσου τείνει στο μηδέν.

Παράδειγμα 2:

Έστω πληθυσμός $X \sim N(20,144)$

Από τον οποίο επιλέγεται τυχαίο δείγμα $n = 36$

Ζητείται $P(18 \le \bar{X} \le 24)$

$$P(18 \le \bar{X} \le 24) = P\left(\frac{18-\mu}{\sigma_{\bar{X}}} \le \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \le \frac{24-\mu}{\sigma_{\bar{X}}}\right)$$

$$= P\left(\frac{18-\mu}{\sigma/\sqrt{n}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{24-\mu}{\sigma/\sqrt{n}}\right) \quad = P\left(\frac{18-20}{12/6} \le \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \le \frac{24-20}{12/6}\right)$$

$$= P(-1 \le Z \le 2) = P(-1 \le Z \le 0) + P(0 \le Z \le 2)$$

ΑΣΚΗΣΕΙΣ

Άσκηση 1. Έστω ο πληθυσμός {2,4,6,8}

α) να βρεθούν ο μέσος και η διακύμανση του πληθυσμού

β) να γραφούν όλα τα δυνατά τυχαία δείγματα μεγέθους ν=3 χωρίς επαναφορά

γ) να βρεθεί η κατανομή δειγματοληψίας του μέσου $\bar{X}$ και της διακύμανσης S^2

δ) να επαληθευθούν οι σχέσεις $E(\bar{X}) = \mu$ και $V(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$

Λύση:

α)

β)

ΑΣΚΗΣΕΙΣ

γ) Μέσοι

Διακυμάνσεις

$\bar{X}$	$f(\bar{X})$	S^2	$f(S^2)$
4,0	¼	2,66	½
4,66	¼	6,222	½
5,333	¼		
6	¼		

ΑΣΚΗΣΕΙΣ

δ) να επαληθευθούν οι σχέσεις $\mathrm{E}\left(\bar{\mathrm{X}}\right) = \mu$ και $V(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$

Δηλαδή, και

ΑΣΚΗΣΕΙΣ

Άσκηση 2. Η τυχαία μεταβλητή Χ κατανέμεται κανονικά με μέσο μ=3 και διακύμανση $\sigma^2=9$. Ποια η πιθανότητα $P(-2\leq X\leq 3,5)$;

Λύση:

Στατιστική επαγωγή: Εκτίμηση

Ας θεωρήσουμε μια τυχαία μεταβλητή Χ, η οποία χαρακτηρίζεται από μια παράμετρο θ (π.χ. μέσος, διακύμανση κτλ). Για να εκτιμήσουμε την θ πρέπει να διαλέξουμε μια συνάρτηση των παρατηρήσεων του δείγματος.

Εκτιμητής Ο τύπος (συνάρτηση) με βάση τον οποίο υπολογίζεται η εκτίμηση της παραμέτρου του πληθυσμού:

Εκτίμηση Η αριθμητική τιμή $\hat{\theta}$ που παίρνει ο εκτιμητής για δεδομένο δείγμα

Οι εκτιμητές είναι τυχαίες μεταβλητές αφού είναι συναρτήσεις των παρατηρήσεων του δείγματος που είναι τυχαίες μεταβλητές.

Οι εκτιμήσεις διαφέρουν από δείγμα σε δείγμα.

Στατιστική επαγωγή: Εκτίμηση

Ένας εκτιμητής $\hat{\theta}$ δεν μπορεί να δίνει πάντοτε την αληθινή τιμή της παραμέτρου του πληθυσμού.

Τα βασικά χαρακτηριστικά ενός εκτιμητή, $\hat{\theta}$, μιας παραμέτρου του πληθυσμού, θ, είναι:

Σφάλμα δειγματοληψίας

Σφάλμα μεροληψίας

Διακύμανση

Μέσος του τετραγώνου του σφάλματος

Ισχύει ότι:

(Μέσος τετραγώνου σφάλματος) = (Διακύμανση) + (Σφάλμα μεροληψίας)^2

Επιθυμητές ιδιότητες εκτιμητών

Οι εκτιμητές που μπορούν να χρησιμοποιηθούν για μια άγνωστη παράμετρο του πληθυσμού είναι άπειροι.

Χρειάζονται λοιπόν κριτήρια επιλογής.

Τα κριτήρια επιλογής βασίζονται στις επιθυμητές ιδιότητες των εκτιμητών.

Οι επιθυμητές ιδιότητες των εκτιμητών αποτελούν τα κριτήρια με βάση τα οποία επιλέγουμε εκτιμητές που υπολογίζονται με διαφορετικές μεθόδους.

Ορισμένες από αυτές τις ιδιότητες έχουν εφαρμογή σε μικρά δείγματα ενώ άλλες σε μεγάλα δείγματα.

Αμεροληψία

Το $\hat{\theta}$ αποτελεί αμερόληπτη εκτίμηση του θ αν

$$E\left(\hat{\theta}\right) = \theta$$

Η αμεροληψία από μόνη της δεν είναι αρκετή γιατί:

α) μπορεί να υπάρχουν περισσότεροι από ένας αμερόληπτοι εκτιμητές και

β) δεν μας λέει τίποτα για την διασπορά της κατανομής του εκτιμητή.

Αποτελεσματικότητα

Μεταξύ δύο αμερόληπτων εκτιμήσεων μιας παραμέτρου του πληθυσμού αποτελεσματικότερη είναι αυτή που έχει την μικρότερη διακύμανση

Η επιλογή του αποτελεσματικού εκτιμητή, που είναι γνωστός ως *άριστος αμερόληπτος εκτιμητής*, είναι δύσκολη, γιατί προϋποθέτει την σύγκριση μεταξύ πολλών (άπειρων) εκτιμητών.

Γι' αυτό περιορίζουμε την σύγκριση μεταξύ του συνόλου των αμερόληπτων εκτιμητών που είναι *γραμμικές* εκτιμήσεις των παρατηρήσεων του δείγματος.

Αυτός ο εκτιμητής ονομάζεται *άριστος γραμμικός αμερόληπτος εκτιμητής (BLUE)*

Ένας εκτιμητής $\hat{\theta}$ είναι *άριστος γραμμικός αμερόληπτος εκτιμητής* αν:

i. $\hat{\theta}$ είναι γραμμική συνάρτηση* των παρατηρήσεων του δείγματος

ii. $\hat{\theta}$ είναι αμερόληπτος

iii. $V(\hat{\theta}) < V(\theta^*)$, όπου θ^* οποιοσδήποτε άλλος γραμμικός αμερόληπτος εκτιμητής

* Σημείωση: Με τον όρο γραμμική συνάρτηση εννοούμε μια συνάρτηση της μορφής $\sum \alpha_i X_i$.

Συνέπεια

Μια εκτιμήτρια είναι συνεπής αν η κατανομή της τείνει να συγκεντρωθεί στην πραγματική τιμή του πληθυσμού καθώς το μέγεθος του δείγματος τείνει στο άπειρο.

$$P \lim \hat{\theta} = \theta$$

Για να διαπιστώσουμε αν ένας εκτιμητής είναι συνεπής εξετάζουμε την διακύμανση και το σφάλμα μεροληψίας, όταν το μέγεθος του n τείνει στο άπειρο.

Αν και τα δύο μεγέθη μικραίνουν και τελικά μηδενίζονται τότε ο εκτιμητής είναι συνεπής.

Αν η κατανομή ενός εκτιμητή συγκεντρώνεται σε ένα σημείο, έστω θ*, όταν n τείνει στο άπειρο, το σημείο θ* λέγεται όριο πιθανότητας και γράφεται:

$$P \lim \hat{\theta} = \theta^*$$

Είδαμε προηγουμένως ότι:
(Μέσος τετραγώνου σφάλματος) = (Διακύμανση) + (Σφάλμα μεροληψίας)^2

Επομένως, ο εκτιμητής είναι συνεπής αν:

Έλεγχος υποθέσεων

Στατιστική Επαγωγή

Εκτιμητική

Έλεγχος υποθέσεων

Διαδικασία ελέγχου

1. Διατύπωση της *μηδενικής* και της *εναλλακτικής* υπόθεσης

H_0: Μηδενική υπόθεση
H_1: Εναλλακτική υπόθεση

Παραδείγματα:

$$H_0: \mu = 0 \qquad H_0: \mu = 3$$
$$H_1: \mu \neq 0 \qquad H_1: \mu > 3$$

2. Υπολογισμός της κατάλληλης στατιστικής

Ο κανόνας με βάση τον οποίο αποφασίζεται αν θα αποδεχθούμε ή θα απορρίψουμε την μηδενική υπόθεση βασίζεται στον υπολογισμό της κατάλληλης στατιστικής (ποια κατανομή θα χρησιμοποιήσουμε).

Η στατιστική αυτή υπολογίζεται με βάση τις διαθέσιμες πληροφορίες του δείγματος.

3. Κριτήρια αποδοχής και απόρριψης

Περιοχή αποδοχής: Το απαραίτητο εύρος τιμών της στατιστικής για να γίνει αποδεκτή η H_0

Περιοχή απόρριψης: Το απαραίτητο εύρος τιμών της στατιστικής για να απορριφθεί η H_0

Κριτήριο ελέγχου: Έννοια

$H_0 : \theta = \theta_0$

$H_1 : \theta < \theta_0$

Ένα κριτήριο θα μπορούσε να είναι:

Εάν $\hat{\theta} \geq \kappa$, η H_0 γίνεται δεκτή

Εάν $\hat{\theta} < \kappa$, η H_0 απορρίπτεται υπέρ της H_1 όπου κ μια σταθερά

Περιοχή απόρριψης	Περιοχή αποδοχής
Η H_0 απορρίπτεται	Η H_0 γίνεται δεκτή

$\hat{\theta} < \kappa$ κ $\hat{\theta} \geq \kappa$

Δειγματικός χώρος της $\hat{\theta}$

Κριτήριο ελέγχου: Έννοια

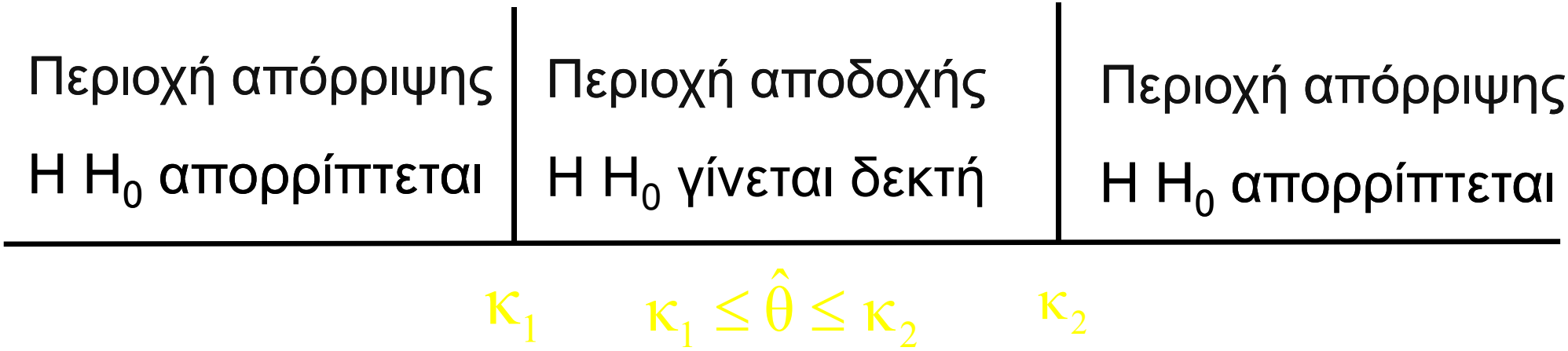
$H_0 : \theta = \theta_0$

$H_1 : \theta \quad \theta_0$

Ανάλογα με το προηγούμενο παράδειγμα, ένα κριτήριο θα μπορούσε να είναι:

Εάν $\kappa_1 \le \hat{\theta} \le \kappa_2$, η H_0 γίνεται δεκτή

Εάν $\hat{\theta} < \kappa_1 \quad \hat{\theta} > \kappa_2$, η H_0 απορρίπτεται υπέρ της H_1



Αποτελέσματα Ελέγχου

Το κριτήριο καθώς και οι πληροφορίες από το δείγμα δεν αποτελούν εγγύηση ότι θα καταλήξουμε στο σωστό συμπέρασμα. Κάθε έλεγχος μπορεί να οδηγήσει σε ένα από τα παρακάτω αποτελέσματα:

1. Σωστή απόφαση	Αποδοχή της ορθής H_0 Απόρριψη λανθασμένης H_0
2. Λανθασμένη απόφαση	Απόρριψη ορθής H_0 Σφάλμα Τύπου I *π.χ. Ενοχοποίηση αθώου*
3. Λανθασμένη απόφαση	Αποδοχή λανθασμένης H_0 Σφάλμα Τύπου II *π.χ. Αθώωση ενόχου*

Αποτελέσματα Ελέγχου

	H_0 σωστή	H_1 σωστή
Η H_0 γίνεται δεκτή	Σωστή απόφαση	Σφάλμα τύπου II
Η H_0 απορρίπτεται	Σφάλμα τύπου I	Σωστή απόφαση

- Κάθε προσπάθεια να μειώσουμε την πιθανότητα του Σφάλματος Ι συνεπάγεται αύξηση της πιθανότητας του Σφάλματος ΙΙ και αντίστροφα.

  Αν π.χ. απορρίπτουμε πάντα την H_0 για να αποφύγουμε το Σφάλμα ΙΙ τότε η πιθανότητα του Σφάλματος Ι γίνεται πολύ μεγάλη

- Συνήθως ο ερευνητής καθορίζει την πιθανότητα σφάλματος Ι την οποία θεωρεί αποδεκτή (π.χ. 5% ,η 1%) Η πιθανότητα αυτή ονομάζεται *επίπεδο σημαντικότητας.* Στη συνέχεια επιλέγεται μια διαδικασία απόφασης που ελαχιστοποιεί την πιθανότητα Σφάλματος ΙΙ

Έλεγχος του μέσου κανονικής κατανομής

Υποθέτουμε $X \sim N\left(\mu, \sigma^2\right)$

Μονόπλευρος έλεγχος *(one-sided test)*

$$H_0: \mu = \mu_0$$
$$H_1: \mu > \mu_0$$

Στην περίπτωση του μονόπλευρου ελέγχου μας ενδιαφέρει απλώς η μια πλευρά της κατανομής *π.χ. ο μέσος μηνιαίος μισθός σε ένα κλάδο είναι 1000 € ή μεγαλύτερος*

Δίπλευρος έλεγχος
(two-sided test)

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Στην περίπτωση του δίπλευρου ελέγχου η περιοχή απόρριψης περιλαμβάνει τις τιμές που βρίσκονται μακρυά από τον αριθμητικό μέσο, είτε αυτές είναι μεγαλύτερες είτε μικρότερες

π.χ. ο μέσος μηνιαίος μισθός σε ένα κλάδο είναι 1000 €

Η στατιστική που υπολογίζεται για τον έλεγχο του αριθμητικού μέσου είναι η

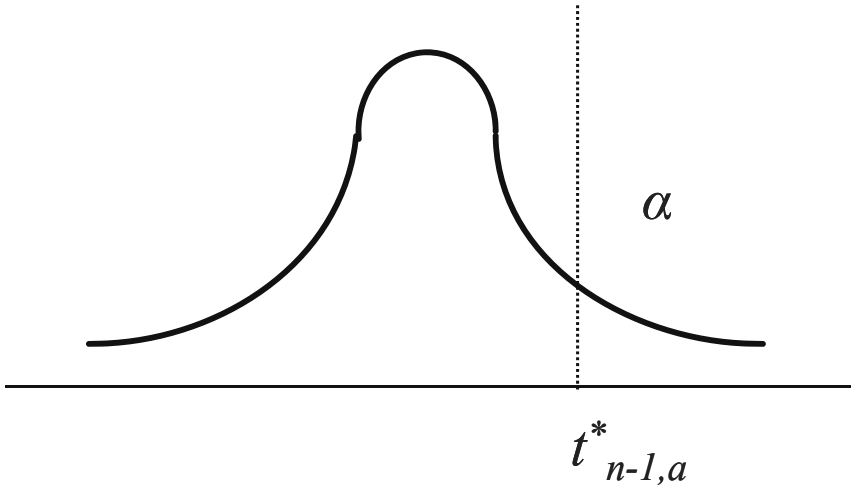
$$t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

Η στατιστική αυτή ακολουθεί την κατανομή t με $n\text{-}1$ βαθμούς ελευθερίας

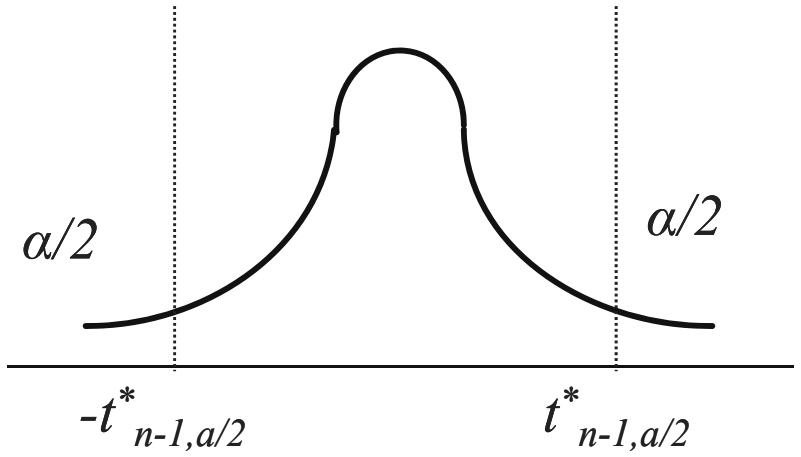
Επιλέγουμε το επίπεδο σημαντικότητας (την πιθανότητα σφάλματος Ι).
Συμβολίζεται με α. Δηλαδή $P(\Sigma\varphi\acute{\alpha}\lambda\mu\alpha\ I)=\alpha$ $\pi.\chi\ 0,05\ (5\%)$
Αυτό καθορίζει και την περιοχή απόρριψης.

Η περιοχή απόρριψης (κρίσιμη περιοχή) εξαρτάται από το είδος του ελέγχου (μονόπλευρος ή δίπλευρος)

μονόπλευρος



δίπλευρος



Το όριο ή τα όρια της περιοχής απόρριψης, δηλαδή η τιμή $t^*_{n-1,a}$ ή $t^*_{n-1,a/2}$ δίνονται στους πίνακες της κατανομής student-t.

Η τιμή της στατιστικής t που υπολογίζεται με βάση τα στοιχεία του δείγματος συγκρίνεται με τα όρια της περιοχής απόρριψης.

Δίπλευρος έλεγχος

Αν $|t| < t^*_{n-1,a/2}$ η μηδενική υπόθεση δεν απορρίπτεται

Μονόπλευρος έλεγχος

Αν $|t| < t^*_{n-1,a}$ η μηδενική υπόθεση δεν απορρίπτεται

Παράδειγμα 1:

Θέλουμε να ελέγξουμε την υπόθεση αν ο μέσος βαθμός σε ένα μάθημα είναι 6,0.

Επειδή δεν είναι δυνατό να ρωτηθούν όλοι οι φοιτητές επιλέγεται ένα δείγμα από το οποίο προέκυψαν τα ακόλουθα στοιχεία: 5 6 6 2 8 4 7 6 3 4

Με βάση τα στοιχεία αυτά υπολογίζονται

$$\bar{X} = 5,1 \qquad S^2 = \frac{\sum\left(X_i - \bar{X}\right)^2}{n-1} = 3,43 \qquad S = 1,85$$

$H_0 : \mu = 6$
$H_1 : \mu \neq 6$

Είναι δυνατό το δείγμα με μέσο 5.1 να προέρχεται από πληθυσμό με μέσο 6;
Είναι δυνατό η διαφορά να οφείλεται μόνο στην δειγματοληψία;

Υπολογίζουμε την στατιστική $t = \dfrac{\bar{X} - \mu}{S/\sqrt{n}} = \dfrac{5,1 - 6,0}{1,85/\sqrt{10}} = -1,53$

Από τους πίνακες $t_{9,(0,025)} = 2,262 > 1,53$ Η H_0 δεν απορρίπτεται

Άρα, ΝΑΙ είναι δυνατό το δείγμα με μέσο 5,1 να προέρχεται από πληθυσμό με μέσο 6 και η διαφορά οφείλεται μόνο στην δειγματοληψία.

Παράδειγμα 2:

Θέλουμε να ελέγξουμε την υπόθεση αν ο μέσος βαθμός είναι 7 ή μικρότερος του 7.

$H_0 : \mu = 7$

$H_1 : \mu < 7$

$t = \dfrac{5,1 - 7,0}{1,85/\sqrt{10}} = -3,24$

$t_{9,(0,05)} = 1,833 < 3,24$

Η H_0 απορρίπτεται

Διαστήματα εμπιστοσύνης

Πολλές φορές αντί να αποβλέπουμε σε εκτίμηση σημείου ενδιαφερόμαστε στην εκτίμηση διαστήματος.

Τα όρια του διαστήματος είναι και αυτά τυχαίες μεταβλητές.

Μπορούμε να υπολογίσουμε την πιθανότητα με την οποία το διάστημα αυτό μπορεί να περιλαμβάνει την αληθινή τιμή μιας παραμέτρου.

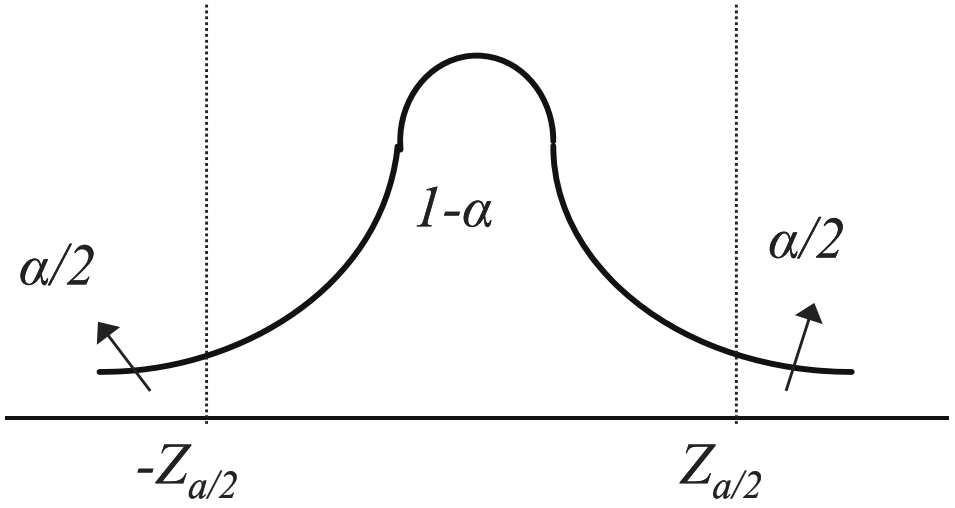
Το διάστημα αυτό ονομάζεται διάστημα εμπιστοσύνης.

Διαστήματα εμπιστοσύνης

Ξέρουμε ότι για τυχαία δείγματα από κανονικούς πληθυσμούς με μέσο μ και διακύμανση σ^2, η στατιστική $\left(\bar{X}-\mu\right)\Big/\frac{\sigma}{\sqrt{n}}$ ακολουθεί την τυποποιημένη κανονική κατανομή.

Έστω $Z_{\alpha/2}$ η τιμή του Z, για την οποία η πιθανότητα να υπερβληθεί είναι α/2 και (λόγω συμμετρίας) $-Z_{\alpha/2}$ η τιμή του Z, για την οποία η πιθανότητα να είναι μικρότερη ή ίση είναι α/2.

Άρα μπορούμε να γράψουμε: $P\left(-Z_{a/2} \le Z \le Z_{a/2}\right) = 1-a$



Διαστήματα εμπιστοσύνης

$$P\left(-Z_{a/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le Z_{a/2}\right) = 1-a$$

$$P\left(\bar{X} - Z_{a/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{a/2}\frac{\sigma}{\sqrt{n}}\right) = 1-a$$

Η πιθανότητα δηλαδή ότι η ανισότητα μέσα στην παρένθεση περιλαμβάνει τον μέσο είναι 1-α.

Παράδειγμα: Εάν σ=8, n=16, $\bar{X} = 25$, και α=5% (οπότε $Z_{\alpha/2}$=1,96)

$$25 - 1{,}96\frac{8}{\sqrt{16}} \le \mu \le 25 + 1{,}96\frac{8}{\sqrt{16}}$$

$$21{,}08 \le \mu \le 28{,}92$$

Με πιθανότητα 95% οι τυχαίες μεταβλητές παίρνουν τιμές (21,08 , 28,92) που περικλείουν το μέσο όρο.

Το διάστημα αυτό αποκαλείται διάστημα εμπιστοσύνης και η πιθανότητα 1-α συντελεστής ή επίπεδο εμπιστοσύνης.

Διαστήματα εμπιστοσύνης

Αν η διακύμανση του πληθυσμού είναι άγνωστη το διάστημα εμπιστοσύνης βασίζεται στην κατανομή t.

$$\bar{X} - t^*_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t^*_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

Π.χ. για $\bar{X}=43$, n=9, S=5 και α=5%:

$$43 - t^*_{8,(0,025)}\frac{5}{\sqrt{9}} \le \mu \le 43 + t^*_{8,(0,025)}\frac{5}{\sqrt{9}}$$

$$43 - 2,306\frac{5}{\sqrt{9}} \le \mu \le 43 + 2,306\frac{5}{\sqrt{9}}$$

$$39,06 \le \mu \le 46,94$$

Δηλαδή, με συντελεστή εμπιστοσύνης 95% ο μέσος κυμαίνεται (39,06 , 46,94).

Είναι προφανές ότι κάθε υπόθεση για το μ μέσα στο διάστημα

$$\bar{X} - t^*_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le \mu_0 \le \bar{X} + t^*_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

δεν απορρίπτεται

<u>Παράδειγμα:</u>

Ζητείται το διάστημα μέσα στο οποίο βρίσκεται ο πραγματικός μέσος βαθμός του πληθυσμού του μαθήματος με πιθανότητα 95%

$$\bar{X} = 5{,}1 \ , \ S = 1{,}85 \ , \ \text{n=10}$$

$$3,776 \leq \mu \leq 6,423$$

Αν επιλέξουμε α=0.01 $3,19 \leq \mu \leq 7,0$

ΑΛΓΕΒΡΑ ΜΗΤΡΩΝ

Βασικοί ορισμοί

Μήτρα διαστάσεων MxN είναι μια ορθογώνια κατάταξη αριθμών σε M γραμμές και N στήλες

Μήτρες διαστάσεων Mx1 ή 1xM ονομάζονται διάνυσμα στήλης και γραμμής αντίστοιχα.

Αν M=N η μήτρα ονομάζεται *τετραγωνική*. Μια τετραγωνική μήτρα με όλα τα στοιχεία ίσα με μηδέν εκτός από την κύρια διαγώνιο ονομάζεται *διαγώνια* μήτρα.

Βασικοί ορισμοί

Μια διαγώνια μήτρα με όλα τα στοιχεία στην κύρια διαγώνιο ίσα με την μονάδα ονομάζεται *μοναδιαία*

Βασικοί ορισμοί

Γενικά αν όμως η Α λέγεται *ταυτοδύναμη*.

Ανάστροφη μήτρα είναι η μήτρα που προκύπτει όταν οι γραμμές πάρουν την θέση των στηλών και οι στήλες την θέση των γραμμών.

Αν η Α έχει διαστάσεις MxN, τότε η Α′ έχει διαστάσεις NxM.

Ισχύουν τα παρακάτω:

Αν Α′=Α η μήτρα λέγεται *συμμετρική*.

Βασικοί ορισμοί

Ο *βαθμός* μια μήτρας **A** ορίζεται ως η τάξη (ή οι διαστάσεις) της μεγαλύτερης ορίζουσας που μπορεί να προκύψει από την **A** που είναι διαφορετική από το μηδέν.

Εάν η ορίζουσα μιας τετραγωνικής μήτρας **A** είναι ίση με το μηδέν λέγεται *ιδιάζουσα* αλλιώς λέγεται *μη-ιδιάζουσα*.

Η αντίστροφη μια τετραγωνικής μήτρας **A** συμβολίζεται με $\mathbf{A}^{-1}$ και είναι αντίστροφη αν και μόνον αν:

Ισχύουν οι εξής ιδιότητες:

Λύση συστημάτων γραμμικών εξισώσεων

Έστω το γραμμικό σύστημα:

Το οποίο μπορεί να με την μορφή μητρών να ξαναγραφεί:

ή **AX=C**

Λύση συστημάτων γραμμικών εξισώσεων

Μέθοδος Cramer:

Η λύση της άγνωστης X_j δίνεται από την σχέση:

Όπου $\Delta = |\mathbf{A}|$

Και Δ_j η ορίζουσα που προκύπτει αν αντικαταστήσουμε την *j* στήλη του πίνακα **A** με το διάνυσμα στήλη **C**.