5/3/2020

Μέτρο διασποράς

ΕΥΡΟΣ $R = \max x_i - \min x_i$

$\bar{x} = \frac{1}{n} \Sigma x_i$

$2, 2, 4, 4, 15 \rightarrow s$

ΔΙΑΚΥΜΑΝΣΗ s^2

Απευθείας από $x_1, \dots, x_n$

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

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\underbrace{\Sigma x_i}_{n\bar{x}} + n\bar{x}^2 =$$

$EX = \mu$, $\quad Var X = \sigma^2$

$Var X = E(X-\mu)^2 = EX^2 - (EX)^2$

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

Η διακύμανση δείχνει πως κατανέμονται τα δεδομένα γύρω από τη μέση τιμή.

Συνηθέστερα, $s^2 = \frac{1}{n-1} \Sigma (x_i - \bar{x})^2$ και ο λόγος είναι γιατί τότε

η $ES^2 = \sigma^2$

αναμενόμενη τιμή του s^2

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

$\ell \rightarrow$ το πλήθος των διακεκριμένων μετρήσεων

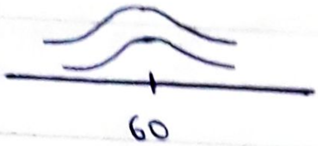
ΔΙΑΤΕΤΑΓΜΕΝΟΣ: $s^2 = \frac{1}{n-1} \sum_{j=1}^{\ell} (x_j - \bar{x})^2 f_j$

$k \rightarrow$ το πλήθος των κλάσεων

ΟΜΑΔΟΠΟΙΗΜΕΝΟΣ: $s^2 = \frac{1}{n-1} \sum_{j=1}^{k} (m_j - \bar{x})^2 f_j$

$x_1, \dots, x_{15}$
$\bar{x} = 60$
$S^2_x = 7$

$y_1, \dots, y_{18}$
$\bar{y} = 60$
$S^2_y = 3$



Παρατήρηση: Αν έχω 2 ή περισσότερα δείγματα, αυτό που έχει την μικρότερη διακύμανση είναι πιο ομοιογενές.

Συντελεστής μεταβλητότητας C.V. (coefficient of variation)

$C.V = \frac{S}{|\bar{x}|}$ → θετική ρίζα της διακύμανσης - τυπική απόκλιση $Var X = S^2$ ↓ διακύμανση

ή $\frac{S}{|\bar{x}|} \times 100\%$

Ο συντελεστής μεταβλητότητας χρησιμοποιείται για να συγκριθούν δύο δείγματα που έχουν διαφορετικές μέσες τιμές ως προς την μεταβλητότητά τους.

Ορισμός: Ένα δείγμα λέγεται ομοιογενές αν ο συντελεστής μεταβλητότητας είναι μικρότερος ή ίσος από το 10%.

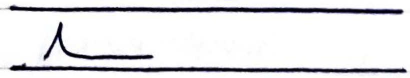
Μέτρα σχήματος και παρατηρήσεις για τη μορφή των δεδομένων

1η: $\bar{x}$, M, k

- Αν $\bar{x}$, M, k είναι "κοντινά" και επιπλέον έχουμε μονοκόρυφη κορυφή (k μοναδικό) τότε λέμε ότι έχουμε συμμετρική κατανομή (π.χ. κανονική)
- Αν $\bar{x} > M > k$ λέμε ότι έχουμε θετική ασυμμετρία ενώ αν έχουμε $\bar{x} < M < k$ λέμε ότι έχουμε αρνητική ασυμμετρία.

θετική

αρνητική



$$b_1 = \frac{\sum_{i=1}^{n} (x_i - x)^3}{n S^3}$$ ⇝ συντελεστής λοξότητας
- ≈ 0 συμμετρική
- < 0 θετική ασυμμετρία, η λοξότητα δεξιά
- > 0 λοξότητα αριστερά

$$b_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n s^4} \rightsquigarrow \# \text{ κύρτωσης}$$

ΛΕΠΤΟΚΥΡΤΕΣ >3

ΜΕΣΟΚΥΡΤΕΣ ≈ 3

ΠΛΑΤΥΚΥΡΤΕΣ

ΔΕΙΓΜΑΤΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

Έστω $x_1, x_2, \dots, x_n$ είναι τυχαίο δείγμα (τ.δ.) $\to x_1, \dots, x_n$ είναι ισόνομες τ.μ. (ανεξάρτητες μεταξύ τους) από έναν πληθυσμό με σ.π.π. ή σ.π. $f(x, \theta)$ όπου θ άγνωστη παράμετρος

$f(x, \lambda) = \lambda e^{-\lambda x}, \ x \geq 0$

$\lambda = 5$

Έστω $x_1, \dots x_n$ από θ άγνωστη παράμετρος. Ο στόχος είναι να εξάγουμε συμπεράσματα για την άγνωστη παράμετρο θ με την βοήθεια συναρτήσεων των $x_1, x_2, \dots, x_n$. Προφανώς αυτές οι συναρτήσεις δεν περιέχουν άγνωστες παραμέτρους. Λέγονται στατιστικές συναρτήσεις ή απλώς στατιστικά.

$T = T(x_1, \dots, x_n) \to$ είναι τ.μ. $\to$ έχουν κατανομή $\to$ πρέπει να την προσδιορίσω

$N(\mu, \sigma^2)$ σ^2: γνωστή $\quad \sum \frac{x_i}{\sigma}$

Αυτές οι κατανομές συνήθως λέγονται δειγματικές κατανομές και μας βοηθούν στην εξαγωγή συμπερασμάτων από το δείγμα για τον πληθυσμό (Επαγωγική στατιστική).

Στη συνέχεια θα μελετήσουμε κάποιες κατανομές που προκύπτουν από την κανονική κατανομή και εμφανίζονται συχνά ως δειγματικές κατανομές.

X^2 - κατανομή με n βαθμούς ελευθερίας

τυπική κατανομή

Ορισμός: Έστω $x_1, x_2, \dots, x_n$ τ.δ. από την $N(0,1)$. Τότε η κατανομή της $Y = x_1^2 + x_2^2 + \dots + x_n^2$ λέμε ότι ακολουθεί X^2 κατανομή με n βαθμούς ελευθερίας και συμβολίζεται χ_n^2

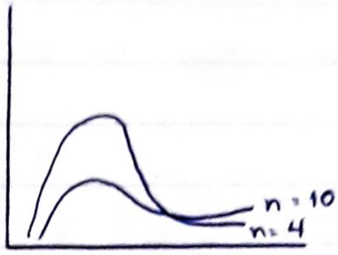
$X_1 \sim N(0,1)$ } ανεξάρτητα
$X_2 \sim N(0,1)$

$Z = X_1^2 + X_2^2$

$Z \sim \chi_2^2$

$\chi_n^2 = \Gamma\alpha\mu\mu\alpha\left(\frac{n}{2}, 2\right)$

ταυτίζεται



$E\chi_n^2 = \frac{n}{2} \cdot 2$, $Var\, X = 2^2 \cdot \frac{n}{2} = 2n$

Όταν $Z \sim N(0,1)$ συμβολίζουμε Z_a. το σημείο εκείνο τ.ω. $P(Z \geq Z_a) = a$

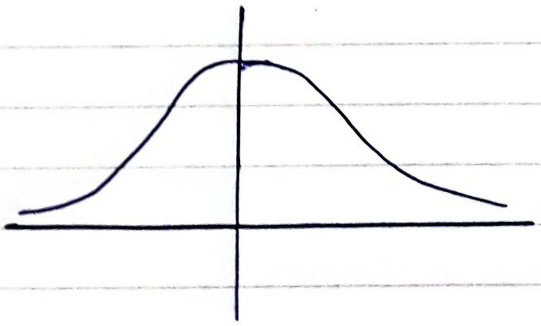
Όταν χ_n^2 συμβολίζω με $\chi^2_{n,a}$

$P(\chi_n^2 \geq \chi^2_{n,a}) = a$

t-κατανομή με n βαθμούς ελευθερίας

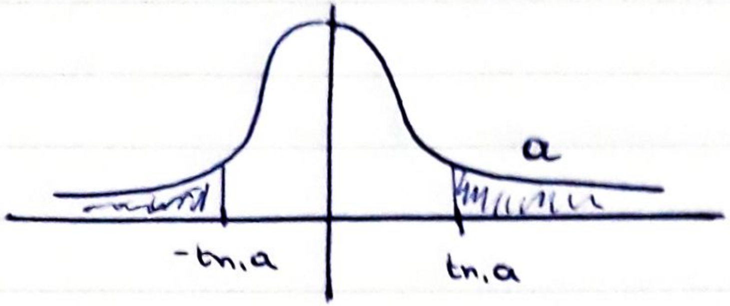
Έστω $X \sim N(0,1)$ και $Y \sim \chi_n^2$. Επιπλέον X, Y ανεξάρτητες. Τότε η κατανομή της $Z = \frac{X}{\sqrt{Y/n}} \sim t_n$. λέμε ότι ακολουθεί την t με n βαθμούς ελευθερίας.

Παρατήρηση: Η κατανομή αυτή είναι γνωστή ως κατανομή Student που είναι το ψευδώνυμο του Gosset.



συμμετρική γύρω από το 0.
(σελ. 75)

Όσο αυξάνονται οι βαθμοί ελευθερίας ($n \to \infty$) η t_n με n βαθμούς ελευθερίας τείνει στην $N(0,1)$. $t_{n,a}$ είναι εκείνο το σημείο $P(t_n \geq t_{n,a}) = a$



$-t_{n,a}$ είναι το σημείο εκείνο $P(t_n \geq \underbrace{-t_{n,a}}_{t_{n,1-a}}) = 1-a$

$-t_{n,a} = t_{n,1-a}$

$P(t_n \leq -t_{n,a}) = a$

Η μέση τιμή της t_n : $E(t_n) = 0$, $n > 1$

$$EX = \int_{\text{π.ο.}} x f_X(x)\,dx$$

$t_1 \to$ Cauchy δεν υπάρχει η $E(t_1)$

$$Var(t_n) = \frac{n}{n-2}, \quad n > 2$$