Χαρακτηριστικό, Τυχαίο Δείγμα και παρατηρήσεις

X: χαρακτηριστικό (δέντρα)

έχω δείγμα μεγέθους $n (=100)$

$X_1, X_2, X_3, \ldots, X_{100}$ τυχαίες μεταβλητές (κάθε φορά που κάνω δειγματοληψία έχουν άλλη τιμή)

Έχω πάρει ένα συγκεκριμένο δείγμα

$\underline{x_1 = 0}$ $x_2 = 5$ $x_3 = 1$ $x_4 = 0$ $\underline{x_5 = 0}$ ...... $x_{100} = 1$

(έχω διακριτό χαρακτηριστικό)

	0	1	2	3	4	5
συχνότητα	46	32	8	11	2	1

↓

από τις $x_1, \ldots, x_{100}$ οι 46 ισούται με 0

Επιλογή «μοντέλου» για την καλύτερη περιγραφή του χαρακτηριστικού

$X \sim \text{Poisson}(\theta)$ → παράμετρος: αν την προσδιορίσω "ξέρω" τα πάντα για το χαρακτηριστικό

ακολουθεί → μοντέλο

$P(X=x) = e^{-\theta}\theta^{x}/x!$ $\quad x=0,1,2\ldots$

χαρακτηριστικό → συγκεκριμένη τιμή

πιθανότητα

$x! = 1\cdot 2\cdot 3\cdots x$

$0! = 1.$

X = γεγονότα στον χώρο ή στον χρόνο

x παραγοντ

$$\sum_{x=0}^{\infty} P(X=x) = \sum_{x=0}^{\infty} e^{-\theta}\frac{\theta^{x}}{x!} = 1 \qquad \theta>0$$

Εκτιμητές των παραμέτρων- εκτίμηση από το συγκεκριμένο δείγμα

Το θ είναι η μέση τιμή: (πληθυσμού)

$$E(X) = \sum_{x=0}^{\infty} x P(X=x) \longrightarrow$$ από το θεωρητικό μοντέλο

expected value $\qquad E(X)=\theta$

$$\bar{X} = \sum_{i=1}^{n} X_i \frac{1}{n} \longrightarrow$$ από τις παρατηρήσεις του δείγματος

δειγματικός μέσος

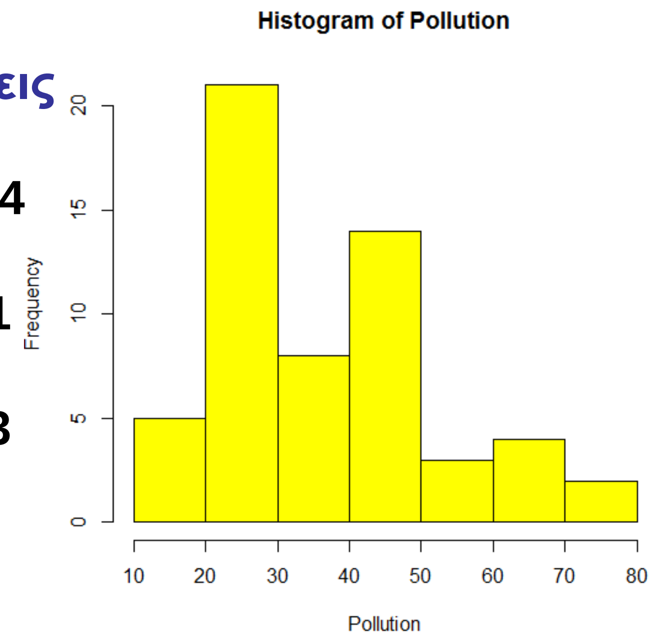
$$\hat{\theta} = \bar{X}$$ (εκτιμητής: συνάρτηση των παρατηρήσεων)

εκτιμώ την άγνωστη παράμετρο από την αντίστοιχη δειγματική της τιμή.

κάθε φορά $\bar{x}$ είναι διαφορετικό (έχει άλλη τιμή για κάθε δείγμα)

και έτσι παίρνω μια εκτίμηση για το θ. Μπορώ να βάλω την συγκεκριμένη τιμή στο μοντέλο και να το προσδιορίσω πλήρως

Δείγμα από τη συγκέντρωση ενός συγκεκριμένου ρύπου (σε mgr/cm³) σε δείγματα αέρος από 57 πόλεις



68 63 42 27 30 36 28 32 79 27 22 23 24 25 24

65 43 25 74 51 36 42 28 31 28 25 45 12 57 51

12 32 49 38 42 27 31 50 38 21 16 24 69 47 23

22 43 27 49 48 23 12 19 46 30 49 49

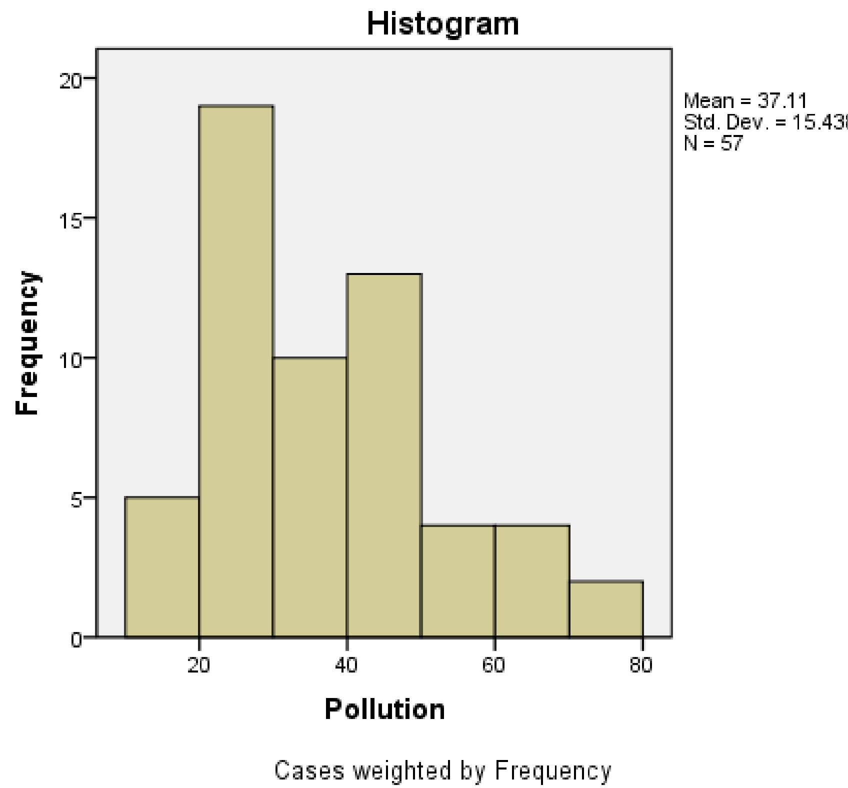
$$\overline{\mathrm{X}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X_i}}{\mathrm{n}} = \frac{68 + 63 + 42 + \cdots + 30 + 49 + 49}{57} = \frac{2099}{57} = 36.82\ \mathrm{mgr/cm^3}$$

$$S^2 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{X_i} - \overline{\mathrm{X}})^2}{\mathrm{n} - 1} = \frac{(68 - 36.82)^2 + (63 - 36.82)^2 + \cdots + (49 - 36.82)^2}{56}$$

$$= \frac{971.90 + 685.15 + \cdots + 148.24}{56} = \frac{14366.25}{56} = 256.54 = (16.02)^2$$

$$S = 16.02\ \mathrm{mgr/cm^3}$$

Όρια κλάσης L_i -U_i	Κεντρική τιμή x_i	Συχνότητα f_i	Αθροιστική Συχνότητα F_i
[10-20)	**15**	**5**	**5**
[20-30)	**25**	**19**	**24**
[30-40)	**35**	**10**	**34**
[40-50)	**45**	**13**	**47**
[50-60)	**55**	**4**	**51**
[60-70)	**65**	**4**	**55**
[70-80]	**75**	**2**	**57**
		57	



$$\overline{X} = \frac{\sum_{i=1}^{\kappa} x_i f_i}{\sum_{i=1}^{\kappa} f_i} = \frac{15 \times 5 + 25 \times 19 + \cdots + 65 \times 4 + 75 \times 2}{5 + 19 + \cdots + 4 + 2} = \frac{2115}{57} = 37.11 \text{ mgr/cm}^3$$

$$S^2 = \frac{\sum_{i=1}^{\kappa} (X_i - \overline{X})^2 f_i}{n-1} = \frac{(15-37.11)^2 \times 5 + (25-37.11)^2 \times 19 + \cdots + (75-37.11)^2 \times 2}{56}$$

$$= \frac{2443.21 + 2784.21 + \cdots + 2872.02}{56} = \frac{13347.37}{56} = 238.34 = (15.43)^2$$

$$S = 15.43 \text{ mgr/cm}^3$$

Statistics

Pollution

N	Valid	57
	Missing	0
Mean		37.11
Mode		25
Std. Deviation		15.438
Variance		238.346
Percentiles	25	24.79[a]
	50	34.66
	75	47.65

a. Percentiles are calculated from grouped data.

Pollution

Όρια κλάσης L_i-U_i		Frequency	Percent	Αθροιστική Συχνότητα F_i	Cumulative Percent
[10-20)	15	5	8.8	**5**	8.8
[20-30)	25	19	33.3	**24**	42.1
[30-40)	35	10	17.5	**34**	59.6
[40-50)	45	13	22.8	**47**	82.5
[50-60)	55	4	7.0	**51**	89.5
[60-70)	65	4	7.0	**55**	96.5
[70-80]	75	2	3.5	**57**	100.0
	Total	57	100.0		

$$\delta = Q_2 = L_3 + \frac{\frac{n}{2} - F_2}{f_3} c = 30 + \frac{28.5 - 24}{10} 10$$

$$= 30 + 4.5 = 34.5 \text{ mgr/cm}^3$$

Διάμεσος (Median) δ

- ενός πεπερασμένου συνόλου τιμών μιας μεταβλητής ή παρατηρήσεων είναι εκείνη η τιμή που χωρίζει το σύνολο σε δύο ίσα μέρη
  Δηλ. 50% παρατηρήσεων $\geq \delta$
  —"— $\leq \delta$
- διατάσσουμε τις παρατηρήσεις κατά αύξουσα σειρά μεγέθους. Εάν το πλήθος των παρατηρήσεων n είναι περιττό τότε η διάμεσος είναι η μεσαία τιμή. Εάν το n είναι άρτιο δ = ημιάθροισμα των δύο μεσαίων παρατηρήσεων
- Σε ομαδοποιημένες παρατηρήσεις η διάμεσος δίνεται από τον τύπο

$$\delta = L_i + \frac{\frac{n}{2} - F_{i-1}}{f_i} \cdot c$$

L_i: το κάτω αληθινό όριο της κλάσεως που βρίσκεται η διάμεσος

n: πλήθος παρατηρήσεων

$F_{i-1} = \sum_{j=1}^{i-1} f_j$ το άθροισμα των συχνοτήτων μέχρι την προηγούμενη κλάση

c: εύρος της κλάσεως

με i τέτοιο ώστε $F_{i-1} < n/2 \leq F_i$

$$\frac{n}{2} = 28.5 \qquad F_2 = 24 < 28.5 \leq 34 = F_3$$

Όρια κλάσης L_i - U_i	Κεντρική τιμή x_i	Συχνότητα f_i	Αθροιστική Συχνότητα F_i
[10-20)	15	5	5
[20-30)	25	19	24
[30-40)	35	10	34
[40-50)	45	13	47
[50-60)	55	4	51
[60-70)	65	4	55
[70-80]	75	2	57
		57	

• Για ομαδοποιημένες παρατηρήσεις

$$P_k = L_i + \left(\frac{\frac{kn}{100} - F_{i-1}}{f_i}\right) c$$

με i τέτοιο ώστε $F_{i-1} < k*n/100 \le F_i$

• 1ο τεταρτημόριο $Q_1 = L_i + \left(\frac{\frac{n}{4} - F_{i-1}}{f_i}\right) c$

• 2ο τεταρτημόριο (διάμεσος) $\delta = L_i + \left(\frac{\frac{n}{2} - F_{i-1}}{f_i}\right) c$

• 3ο τεταρτημόριο $Q_3 = L_i + \left(\frac{\frac{3n}{4} - F_{i-1}}{f_i}\right) c$

$$\frac{n}{4} = 14.25$$

$$Q_1 = L_2 + \frac{\frac{n}{4} - F_1}{f_2} c = 20 + \frac{14.25 - 5}{19} 10 = 20 + 4.49 = 24.49 \text{ mgr/cm}^3$$

$$F_1 = 5 < 14.25 \le 24 = F_2$$

$$\frac{3n}{4} = 42.75$$

$$Q_3 = L_4 + \frac{\frac{3n}{4} - F_3}{f_4} c = 40 + \frac{42.75 - 34}{13} 10 = 40 + 6.73 = 46.73 \text{ mgr/cm}^3$$

$$F_3 = 34 < 42.75 \le 47 = F_4$$

• σε ομαδοποιημένες παρατηρήσεις η κορυφή δίνεται από τον τύπο $M_o = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot c$

$\Delta_1 = f_i - f_{i-1}$

$\Delta_2 = f_i - f_{i+1}$

με i τέτοιο ώστε $f_i > f_j$ για κάθε $j \neq i$

$$max f_i = f_2 = 19$$

$\Delta_1 = f_2 - f_1$=19-5=14 $\quad$ $\Delta_2 = f_2 - f_3$ =19-10=9

$$M = L_2 + \frac{\Delta_1}{\Delta_1 + \Delta_2} c = 20 + \frac{14}{14 + 9} 10 = 20 + 6.09 = 26.09 \text{ mgr/cm}^3$$

Εναλλακτικός τύπος υπολογισμού της διασποράς

κεντρική τιμή ομαδοποιημένης παρατηρήσεις

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

διασπορά

k: πλήθος κλάσεων

x_i: αντιπρόσωπος για τις f_i το πλήθος παρατηρήσεις της i κλάσης

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

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{k}x_i^2 f_i - 2\bar{x}\sum_{i=1}^{k}x_i f_i + \bar{x}^2\sum_{i=1}^{k}f_i\right\}$$

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

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{k}x_i^2 f_i - 2n\bar{x}^2 + n\bar{x}^2\right\}$$

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

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

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{k}x_i^2 f_i - \frac{\left(\sum^{k} x_i f_i\right)^2}{n}\right\}$$

Γραμμικός Μετασχηματισμός των παρατηρήσεων

Έστω $y_i = \alpha x_i + \beta$ $\quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

β → αλλαγή θέσης

α → κλίμακας

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n}(\alpha x_i + \beta)}{n} = \frac{\sum_{i=1}^{n} \alpha x_i}{n} + \frac{\sum_{i=1}^{n} \beta}{n} = \alpha \frac{\sum_{i=1}^{n} x_i}{n} + \frac{n\beta}{n}$$

$$= \alpha \bar{x} + \beta$$

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\alpha x_i + \beta - (\alpha \bar{x} + \beta)\right)^2$$

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

$$= \alpha^2 \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \Rightarrow S_y^2 = \alpha^2 \cdot S_x^2 \Rightarrow S_y = |\alpha| S_x$$

Τα ποσοστιαία σημεία και η κορυφή μετασχηματίζονται «όπως ακριβώς» και οι παρατηρήσεις.

προσοχή για α<0