

- Θεωρία Πιθανοτήτων: Γνωρίζουμε την κατανομή ενός τυχαίου πειράματος
- Στατιστική: Δεν γνωρίζουμε την κατανομή, όμως προσπαθούμε να την συμπεράνουμε από ένα δείγμα
 - Μη-παραμετρική: Δεν γνωρίζουμε τίποτα για την κατανομή

- Παραμετρική: Γνωρίζουμε ότι ανήκει σε μία οικογένεια κατανομών και προσπαθούμε να εκτιμήσουμε μια άγνωστη παράμετρο

↓ Με αυτή θα ασχοληθούμε

• Ορισμοί:

Θ. Πιθανοτήτων	Στατιστική	Σχόλια
Πληθυσμός (N)	Δείγμα (n)	Ένα δείγμα προκύπτει από τον πληθυσμό μέσω δειγματοληψίας
Παράμετρος (π.χ. μέση τιμή (μ), διασπορά (σ^2))	Εκτίμηση (π.χ. δειγματικός μέσος όρος ($\bar{x}$), δειγματική διασπορά (s^2))	Οι εκτιμήσεις <u>απεικονίζουν</u> τις αντίστοιχες παραμέτρους

- Αν έχουμε ένα δείγμα μεγέθους n , τότε θεωρούμε n ανεξάρτητες και ισόνομες τ.μ. $X_1, X_2, \dots, X_n$.

↓ Όλες ακολουθούν την ίδια κατανομή με τις ίδιες παραμέτρους



• Είδη στατιστικών στοιχείων

- Ποσοτικές (αριθμοί παιδιών, εισόδημα, διάρκεια ζωής, ύψος)
- Ποιοτικές (χρώμα ματιών, επάγγελμα)

• Μέθοδοι αναπαρίστασης

A) Γραφικές

• Μέθοδοι αναπαράστασης

A) Γραφικές

• Πίνακες συχνότητας:

Έστω τυχαίο δείγμα $X_1, \dots, X_n$ $\Rightarrow k \leq n$
 Έστω $y_1, \dots, y_k$ διαφορετικές τιμές

Συχνότητα v_i της τιμής y_i : το πλήθος των x_i που παίρνουν την τιμή y_i

$$v_1 + v_2 + \dots + v_k = n$$

Σχετική συχνότητα f_i : το αντίστοιχο ποσοστό

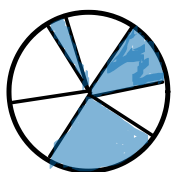
$$f_i = \frac{v_i}{n} \quad 0 \leq f_i \leq 1 \quad f_i = \frac{v_i}{n} \cdot 100\%$$

Αθροιστική συχνότητα N_i : πλήθος τιμών μικρότερες ή ίσες του y_i

Αθροιστική σχετική συχνότητα F_i : ποσοστό τιμών μικρότερες ή ίσες του y_i

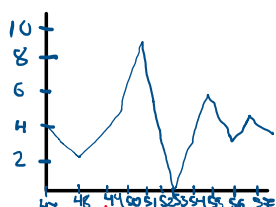
Πίνακας συχνότητας, συνοψίζει κάποια ή όλα από τα παραπάνω

Διάστημα $[a_i, b_i)$	Συχνότητα (v_i)	Σχετική συχνότητα (f_i)	Αθροιστική συχνότητα (N_i)	Αθροιστική σχετική συχνότητα (F_i)
$-24 \leq \text{HPR} < -14$	1	0.025	1	0.025
$-14 \leq \text{HPR} < -7$	5	0.125	6	0.15
$-7 \leq \text{HPR} < 3$	8	0.2	14	0.35
$3 \leq \text{HPR} < 13$	15	0.375	29	0.725
$13 \leq \text{HPR} < 23$	9	0.225	38	0.95
$23 \leq \text{HPR} < 33$	2	0.05	40	1



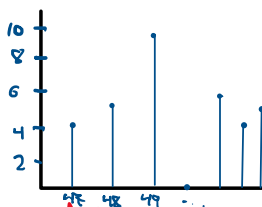
Κυκλικό Διάγραμμα
συχνότητας

$$\Gammaωγία \alpha_i = f_i \cdot 360^\circ$$

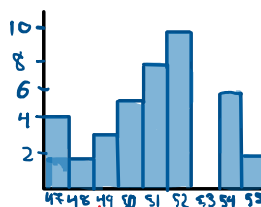


Πολύγωνο
συχνότητας

Οπτικοποίηση της
μεταβολής της
συχνότητας



Διαγράμμιση
συχνότητας



Ιστογράμμο
συχνότητας

- Υψος πολυγωνίου
c.m. εμβαδόν = f_i
- Συνολικό εμβαδόν = n
(ή 1 αν
σχετικές
συχνότητες)

- Πολλές φορές είναι απαραίτητη η ομαδοποίηση των δεδομένων σε διαστήματα, αν τα δεδομένα είναι ποσοτικά
- Όχι πολλές ομάδες, όχι λίγες

B) Αριθμητικά περιγραφικά μέτρα

Έστω $x_1, \dots, x_n$ οι τιμές του δείγματος και $y_1, \dots, y_k$ οι διαφορετικές τιμές
 και v_i, f_i οι αντίστοιχες συχνότητες

- Μέτρα θέσης (ή κεντρικής τάσης)

και v_i, g_i οι αντίστοιχες συχνότητες

- Μέτρα θέσης (ή κεντρικής τάσης)

• Μέση τιμή: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \sum_{i=1}^n \frac{v_i}{n} \cdot x_i = \sum_{i=1}^n p_i \cdot x_i$

$\swarrow \quad \downarrow \quad \searrow$
ισοδύναμοι

• Χαρμφή (mode): Η τιμή με την μεγαλύτερη συχνότητα:
 $M_o = \max(x_i \mid \forall i: v_i \leq v_i)$

• Διάμεσος: Η τιμή που χωρίζει το δείγμα σε δύο ίσα μέρη
Αν διατάξουμε το δείγμα (αυξανόμενα) $x_{(1)}, x_{(2)}, \dots, x_{(n)}$
$$\bar{d} = \begin{cases} x_{(n)} & , n \text{ περιττός} \\ \frac{x_{(n)} + x_{(n+1)}}{2} & , n \text{ άρτιος} \end{cases}$$

• Ποσοστημόριο: το α -ποσοστημόριο είναι η τιμή για την οποία
το $\alpha \cdot 100\%$ των τιμών είναι μικρότερες ή ίσες
από αυτή, ενώ $(1-\alpha)100\%$ μεγαλύτερες:

- $P_{0.25} = Q_1$: πρώτο τεταρτημόριο
- $P_{0.5} = Q_2 = \bar{d}$: δεύτερο "
- $P_{0.75} = Q_3$: τρίτο "

- Μέτρα διασποράς ή μεταβλητότητας

- Εύρος R: Διαφορά μικρότερης και μεγαλύτερης τιμής

- Ενδοτεταρτημοριακή απόκλιση: $Q_3 - Q_1$

• Περιέχει το 50% των τιμών

• $Q_3 - Q_1 \downarrow \Rightarrow$ υψηλή συκέντρωση των τιμών $\Leftrightarrow$ μικρή διασπορά

- Μέση απόκλιση: $MD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$

• Μέση απόκλιση $\downarrow \Rightarrow$ Συκέντρωση γύρω από μέση τιμή $\uparrow$

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

- (Δειγματική) τυπική απόκλιση: $s = \sqrt{s^2}$

- Μέση Διαφορά κατά Gini: $d = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|$

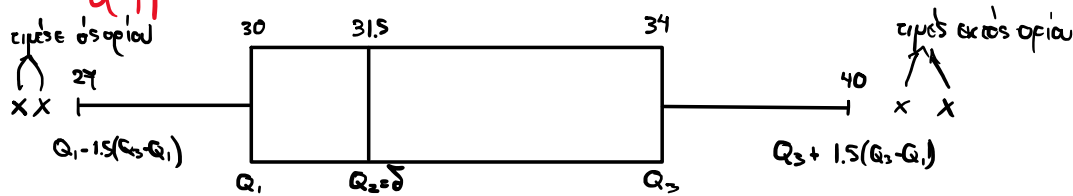
• Μέση απόλυτη διαφορά από 2 (ένδειξη ομοιομορφίας)

Συντελεστής μεταβλητότητας

$$CV = \frac{s}{\bar{x}}$$

Γενικά, αν $CV \leq 0.1 \Rightarrow$ ομοιομορφία

• Ομοεπαγματο



Μήκος Ομοεπαγματος: R

Μήκος Κουτιού: $Q_3 - Q_1$: ενδοεπατημοιοτική απόσταση

Γ) Ζεύγη στατιστικών δεδομένων

Έχουμε δείγμα με ζεύγη τιμών (x_i, y_i) και ψάχνουμε να βρούμε κατά πόσο συσχετίζονται οι x_i με τις y_i τιμές

• Δειγματικός συντελεστής συσχέτισης:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

απόλυτη αρνητική συσχέτιση

• $-1 \leq r \leq 1$ → απόλυτη θετική συσχέτιση

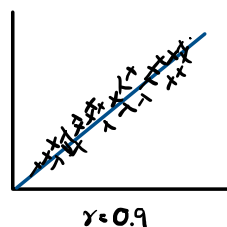
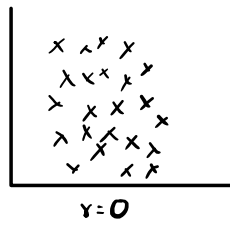
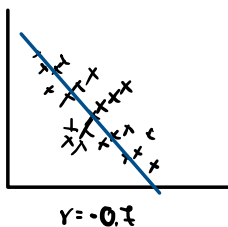
• Αν $b > 0$, α σταθερά και $y_i = bx_i + a \Rightarrow r = 1$

• Αν $b < 0$, α " και $y_i = bx_i + a \Rightarrow r = -1$

• Αν $r = 0$ ο σ.σ. των x_i και y_i , τότε ο σ.σ. των

$a + bx_i$ και $c + dy_i$ είναι επίσης r

αν b και d έχουν ίδιο πρόσημο



• Ιδιότητα:

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

Άσκηση 1

Wednesday, December 11, 2024 2:42 PM

Συμπληρώστε τους παρακάτω πίνακες:

α)

x	v_i	f_i	N_i	F_i
1	10	0.2	10	0.2
2	20	0.4	30	0.6
4	17	0.34	47	0.94
6	3	0.06	50	1
Σύνολο	50	1	-	-

β)

x_i	v_i	f_i	N_i	F_i
1	12	0.2	12	0.2
2	21	0.35	33	0.55
3	24	0.4	57	0.95
5	3	0.05	60	1
Σύνολο	60	1	-	-

$$v = \sum_{i=1} v_i = 10 + 20 + 17 + 3 = 50$$

$$f_i = \frac{v_i}{v} \Rightarrow$$

$$v = \frac{v_i}{f_i} = \frac{24}{0.4} = 60$$

$$f_i = \frac{v_i}{v} \Rightarrow v_i = f_i \cdot v \Rightarrow$$

$$v_5 = 0.05 \cdot 60 = 3$$

$$v_2 = 60 - 12 - 24 - 3 = 21$$

Άσκηση 2

Wednesday, December 13, 2023 11:06 AM

Τιμή (y _i)	Συχνότητα (v _i)
1	9
2	8
3	5
4	5
5	6
6	7

Ο παρακάτω πίνακας συχνοτήτων δίνει τις τιμές που ήρθαν μετά από 40 ρίψεις ενός ζαριού. Να βρεθούν:

α) ο δειγματικός μέσος,		στ) η ημιενδοτεταρτημοριακή απόκλιση,
β) η διάμεσος,		ζ) το εύρος,
γ) η κορυφή (ή επικρατούσα τιμή)		η) η μέση απόκλιση,
δ) το P _{0.20} ποσοστημόριο,		θ) η διασπορά,
ε) τα τεταρτημόρια,		ι) η τυπική απόκλιση

$$a) \bar{x} = \sum_{i=1}^n f_i \cdot y_i = \frac{\sum_{i=1}^n v_i \cdot y_i}{n} = \frac{1 \cdot 9 + 2 \cdot 8 + 3 \cdot 5 + 4 \cdot 5 + 5 \cdot 6 + 6 \cdot 7}{40} = 3.05$$

$$b) \delta = \begin{cases} x_{(r)} & , \quad n \text{ περιττός} \\ \frac{x_{(r)} + x_{(r+1)}}{2} & , \quad n \text{ άρτιος} \end{cases}$$

$$\delta = \frac{3+3}{2} = 3$$

$$c) M_0 = 1$$

δ) P_{0.20} : είναι η τιμή όπου το 20% βρίσκεται πριν από αυτή και το 80% γανω από αυτή

1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, ..., 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6

$$\text{Άρα } P_{0.20} = 1$$

$$e) Q_1 = P_{0.25} = 2$$

$$Q_2 = P_{0.5} = \delta = 3$$

$$Q_3 = P_{0.75} = 5$$

$$\sigma c) \quad Q = Q_3 - Q_1 = 5 - 2 = 3$$

$$3) \quad R = 6 - 1 = 5$$

$$n) \quad MD = \frac{1}{v} \cdot \sum_{i=1}^6 v_i |y_i - \bar{y}| = \frac{1}{40} \cdot \sum_{i=1}^6 v_i |y_i - 3.05| \approx 0.4$$

$$\theta) \quad s^2 = \frac{1}{v-1} \sum_{i=1}^v |x_i - \bar{x}|^2 = \frac{1}{v-1} \left(\left(\sum_{i=1}^v x_i \right)^2 - v \cdot \bar{x} \right) \approx 5.023$$

$$\therefore s = \sqrt{s^2} = \sqrt{5.023} \approx 2.24$$

Έστω ότι οι θερμοκρασίες στις 14.00 για το πρώτο δεκαήμερο του Ιουνίου είναι οι εξής:

32.5, 32.1, 32.3, 31.8, 31.7, 32.4, 32, 31.7, 32.2, 31.9

Υπολογίστε τη δειγματική διασπορά.

$$s^2 = ?$$

1ος τρόπος

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

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{10} \sum_{i=1}^n x_i = 32.06$$

2ος τρόπος

x_i	32.5	32.1	32.3	31.8	31.7	32.4	32	31.7	32.2	31.9
-32	.5	.1	.3	-.2	-.3	.4	0	-.3	.2	-.1
$y_i = x_i - 32$	5	1	3	-2	-3	4	0	-3	2	-1

$$y_i = 10(x_i - 32)$$

$$\bar{y} = \frac{\sum_{i=1}^{10} y_i}{10} = \frac{\sum_{i=1}^{10} (x_i - 32) \cdot 10}{10} = \sum_{i=1}^{10} (x_i - 32) = (\bar{x} - 32) \cdot 10$$

$$\bar{y} = \frac{5+1+3-2-3+4+0-3+2-1}{10} = 0.6$$

$$(\bar{x} - 32) \cdot 10 = \bar{y} \Rightarrow \bar{x} - 32 = \frac{\bar{y}}{10} \Rightarrow \bar{x} = \frac{\bar{y}}{10} + 32 = \frac{0.6}{10} + 32 \Rightarrow$$

$$\boxed{\bar{x} = 32.06}$$

$$\boxed{x = 2.06}$$

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2 \right) \Rightarrow$$

$$\Rightarrow s_y^2 = \frac{1}{10-1} \left[5^2 + 1^2 + 3^2 + 2^2 + 3^2 + 4^2 + 0^2 + 3^2 + 2^2 + 1^2 - 10 \cdot 0.6^2 \right] =$$

$$= \frac{1}{9} (48 - 3.6) \approx 8.26$$

$$\boxed{y_i = ax_i + b \Rightarrow s_y^2 = a^2 s_x^2}$$

$$\text{Άρα: } y_i = 10x_i - 320 \Rightarrow s_y^2 = 10^2 s_x^2 \Rightarrow s_x^2 = \frac{s_y^2}{100} = \frac{8.26}{100} \Rightarrow \boxed{s_x^2 = 0.0826}$$

Έστω ότι το 2018 στις εξετάσεις του μαθήματος Πιθανότητες οι βαθμολογίες είχαν μέση τιμή 7 και τυπική απόκλιση 1.8 ενώ το 2019 η μέση τιμή των βαθμών ήταν 6 και η τυπική απόκλιση ήταν 0.5. Πότε υπήρχε μεγαλύτερη ομοιογένεια;

Συντελεστής μεταβλητότητας: $CV = \frac{s}{\bar{x}}$

$$\bar{x}_1 = 7 \quad \bar{x}_2 = 6$$

$$s_1 = 1.8 \quad s_2 = 0.5$$

$$CV_1 = \frac{s_1}{\bar{x}_1} = \frac{1.8}{7} \approx 0.26 \quad \text{ή} \quad CV_1 = 26\%$$

$$CV_2 = \frac{s_2}{\bar{x}_2} = \frac{0.5}{6} \approx 0.083 \quad \text{ή} \quad CV_2 = 8.3\%$$

Πιο ομοιογενές είναι το δείγμα που έχει μικρότερο CV.

Άρα, $CV_2 < CV_1 \Rightarrow$ Το 2019 υπήρχε μεγαλύτερη ομοιογένεια.

(Μάλιστα, $CV_2 < 10\% \Rightarrow$ Θεωρείται ομοιογενές)