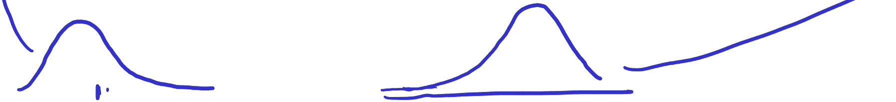
MEM-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

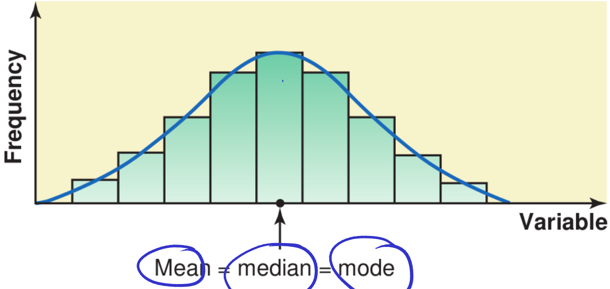
Κώστας Σμαραγδάκης (kesmarag@gmail.com)

3η εβδομάδα (διάλεξη θεωρίας)

- Δηλώνουν κατά πόσο οι τιμές μιας μεταβλητής κατανέμονται συμμετρικά ως προς ένα μέτρο κεντρικής τάσης.
- Όταν το πλήθος των τιμών μιας μεταβλητής είναι μεγαλύτερο για τιμές αριστερά του μέτρου κεντρικής τάσης λέμε ότι η μεταβλητή ακολουθεί κατανομή με **θετική ασυμμετρία**.
- Όταν το πλήθος των τιμών μιας μεταβλητής είναι μεγαλύτερο για τιμές δεξιά του μέτρου κεντρικής τάσης λέμε ότι η μεταβλητή ακολουθεί κατανομή με **αρνητική ασυμμετρία**.



$$\bar{x} = M = M_0$$



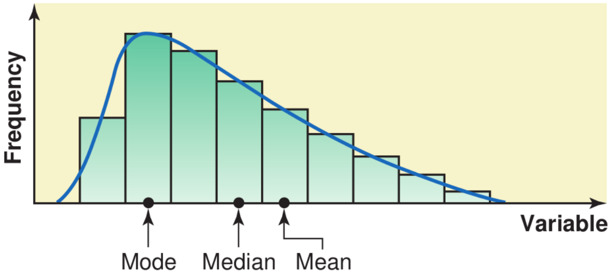
μέση τιμη.

διαμεσος

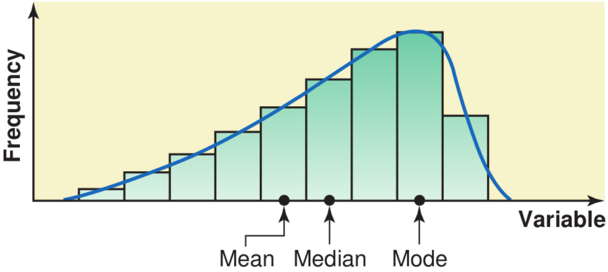
επικρατεστερη τιμη

Μέτρα Ασυμμετρίας - Θετική Ασυμμετρία

$$M_0 < M < \bar{x}$$



$$\bar{x} < M < M_0$$



Ο συντελεστής ασυμμετρίας του Pearson ποσοτικοποιεί την ασυμμετρία.

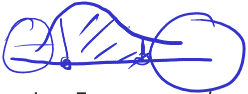
$$Sk_p = \frac{\bar{x} - M_0}{s}$$

Παρατηρούμε ότι ο συντελεστής είναι ανεξάρτητος της μονάδας μέτρησης της μεταβλητής.
Απουσία έντονης ασυμμετρίας η διάμεσος με τη επικρατέστερη τιμή συνδέονται από την ακόλουθη εμπειρική σχέση:

$$\bar{x} - M_0 \approx 3(\bar{x} - M)$$

Οπότε προκύπτει ο συντελεστής εκφρασμένος με τη βοήθεια της διαμέσου:

$$\tilde{Sk}_p = \frac{3(\bar{x} - M)}{s}$$



Ο συντελεστής ασυμμετρίας του Bowley δεν απαιτεί τον υπολογισμό της μέσης τιμής και δίνεται από τη σχέση:

$$Sk_b = \frac{(Q_3 - M) - (M - Q_1)}{Q_3 - Q_1}$$

- Είναι καταλληλότερος στη περίπτωση ύπαρξης ακραίων τιμών.
- Το βασικό του μειονέκτημα είναι ότι λαμβάνει υπόψη από το 50 % των παρατηρήσεων (κεντρικότερες).
- Εαν η διάμεσος είναι πλησιέστερα στο Q_1 σε σχέση με το Q_3 παρατηρείται θετική ασυμμετρία.
- Εαν η διάμεσος είναι πλησιέστερα στο Q_3 σε σχέση με το Q_1 παρατηρείται αρνητική ασυμμετρία.

Άσκηση

Δίνονται οι ακόλουθες διατεταγμένες παρατηρήσεις μιας μεταβλητής:

$$3, 5, 5, 6, 8, 10, 14, 15, 16, 17, 17, 19, 21, 22, 23, 25, 30, 31, 31, 34$$

Υπολογίστε τους συντελεστές ασυμμετρίας $\tilde{Sk}_p, Sk_b$. Παρουσιάζουν οι παρατηρήσεις κάποια ασυμμετρία;

Sk_p

Ο συντελεστής Fisher-Pearson ορίζεται ως:

$$g_1 = \frac{\frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^3}{s^3}$$

Τροποποιημένος συντελεστής ασυμμετρίας Fisher-Pearson

$$G_1 = \frac{N^2}{(N-1)(N-2)} g_1$$

Ο συντελεστής G_1 χρησιμοποιείται από την βιβλιοθήκη pandas (python) για τον υπολογισμό της ασυμμετρίας (θα το δούμε στο 4ο εργαστήριο).

Άσκηση

Υπολογίστε τον τροποποιημένο συντελεστή ασυμμετρίας Fisher-Pearson για τις παρατηρήσεις: -2, -1, 0, 1, 2, 6

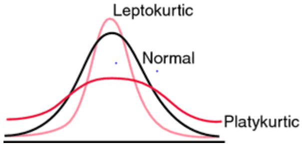
— Τέλος 6ης διάλεξης. —

Ως κυρτότητα ορίζεται ο βαθμός αιχμηρότητας της κορυφής που παρουσιάζει η καμπύλη σχετικών συχνοτήτων συγκρινόμενη με την αντίστοιχη καμπύλη της κανονικής κατανομής. Υπολογίζεται για μονόκορφες συμμετρικές ή σχεδόν συμμετρικές κατανομές.

$$\text{kurtosis} = \frac{\frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^4}{s^4}$$

Με βάση τη τιμή του kurtosis λαμβάνουμε τους χαρακτηρισμούς:

- kurtosis ≈ 3: Μεσόκυρτη (Κανονική)
- kurtosis < 3: Πλατύκυρτη
- kurtosis > 3: Λεπτόκυρτη

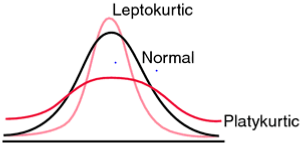


Η βιβλιοθήκη pandas (python) χρησιμοποιεί μια τροποποιημένη έκφραση για το συντελεστή κύρτωσης (θα το δούμε στο 4ο εργαστήριο).

$$\text{kurt} = \frac{\frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^4}{s^4} - 3$$

Με βάση τη τιμή του kurtosis λαμβάνουμε τους χαρακτηρισμούς:

- $\text{kurt} = 0$: Μεσόκυρτη (Κανονική)
- $\text{kurt} < 0$: Πλατύκυρτη
- $\text{kurt} > 0$: Λεπτόκυρτη



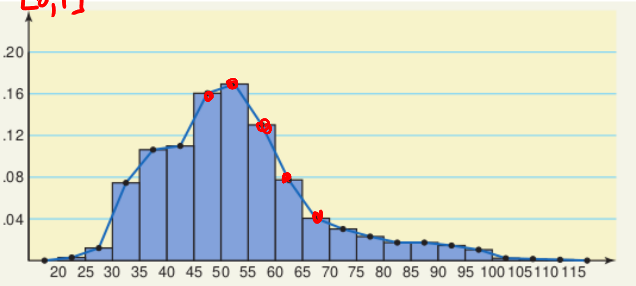
Περιγράφοντας Στατιστικές Κατανομές

1. Γραφική αναπαράσταση δεδομένων με χρήση ιστογράμματος
2. Αναγνώριση προτύπων και εντοπισμός πιθανών ακραίων τιμών
3. Υπολογισμός περιγραφικών μέτρων για τη συνοπτική περιγραφή των παρατηρήσεων

- Πολλές φορές η συνολική τάση των τιμών μιας μεταβλητής για μεγάλο αριθμό παρατηρήσεων είναι τέτοια που μπορεί να περιγραφεί από μια συνεχή συνάρτηση.

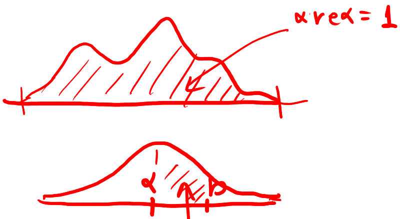
$[0,1]$

f_i



$\sum f_j = 1$

Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function)



Μια συνάρτηση πυκνότητας πιθανότητας $p(x)$:

- Είναι μη αρνητική

$$p(x) \geq 0, \forall x$$

- Το εμβαδόν της επιφάνειας μεταξύ της καμπύλης που ορίζεται από την $p(x)$ και του οριζόντιου άξονα είναι μονάδα.

$$\int_{-\infty}^{+\infty} p(x) dx = 1$$

Μια τέτοια συνάρτηση περιγράφει το συνολική τάση των τιμών μιας κατανομής. Το εμβαδόν κάτω από την καμπύλη $y = p(x)$, για ένα εύρος τιμών του x, εκφράζει την πιθανότητα (σχετική συχνότητα) εμφάνισης παρατηρήσεων στο συγκεκριμένο εύρος τιμών.

Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function)

Συνεχείς μεταβλητές.

Πιθανότητα

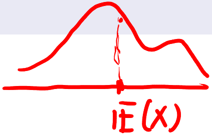
Πυκνότητα.

$$P(X \in [a,b]) = P([a,b]) = P(a \leq X \leq b) = \int_a^b p(x)dx$$

Μέση τιμή - Αναμενόμενη τιμή

$\sum \frac{f_i}{n} \cdot m_i$

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} xp(x)dx$$



Διασπορά

Var.

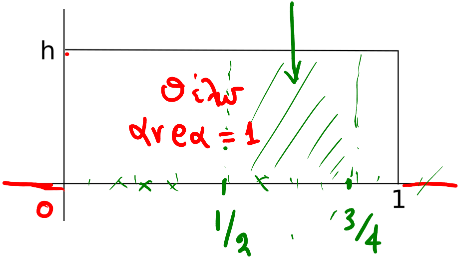
$$\mathbb{V}(X) = \int_{-\infty}^{+\infty} (x - \mathbb{E}(X))^2 p(x)dx$$

Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function)

Για ποιο h είναι πυκνότητα πιθανότητας;

$h = 1$

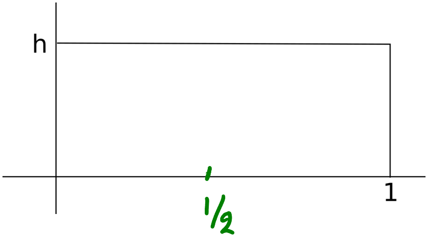
$$f(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & x \notin [0,1] \end{cases}$$



Έστω X έχει την f ως πυκνότητα.

$$P\{X \in [1/2, 3/4]\} = 1 \cdot \left(3/4 - 1/2\right) = \frac{1}{4}.$$

$$E\{X\} = \int_{-\infty}^{+\infty} x f(x)\,dx = \int_0^1 x\,dx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}$$

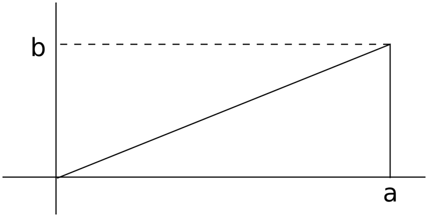


$$u = x - 1/2 \qquad du = dx$$

$$V\{X\} = \int_{-\infty}^{+\infty} \left(x - \frac{1}{2}\right)^2 f(x)\,dx = \int_0^1 (x - 1/2)^2\,dx =$$

$$= \int_{-1/2}^{1/2} u^2\,du = \left[\frac{u^3}{3}\right]_{-1/2}^{1/2} = 2 \cdot \frac{(1/2)^3}{3} = \frac{1/4}{3} = 1/12$$

Συνάρτηση Πυκνότητας Πιθανότητας (Probability Density Function)



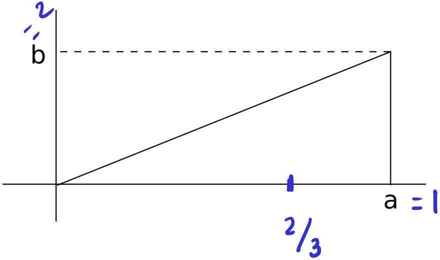
$$\frac{1}{2}\alpha b = 1 \Rightarrow \boxed{\alpha b = 2} \quad (*)$$

$$f(x) = \begin{cases} \frac{b}{\alpha} x & , x \in [0, \alpha] \\ 0 & , \text{διαφορετικά} \end{cases}$$

$$\mathbb{E}\{X\} = \int_{-\infty}^{+\infty} x f(x)\, dx =$$

$$= \int_0^\alpha x \frac{b}{\alpha} x \, dx =$$

$$= \frac{b}{\alpha} \int_0^\alpha x^2 \, dx =$$



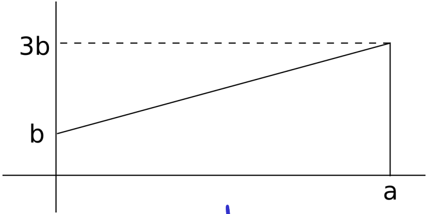
$$= \frac{b}{\alpha}\left[\frac{x^3}{3}\right]_0^{\alpha} = \frac{\alpha^2 b}{3}$$

για $\alpha = 1$ $b = 2$

$$E\{X\} = \frac{1 \cdot 2}{3} = 2/3$$

$$\frac{1}{2}(b+3b)\cdot\alpha = 1$$

$$4b\cdot\alpha = 2 \Leftrightarrow \boxed{\alpha b = 1/2}$$



$$f(x) = \begin{cases} \frac{2b}{\alpha}x + b, & x\in[0,\alpha] \\ 0, & \text{διαφορετικά} \end{cases}$$

$$E\{X\} = \int_{-\infty}^{+\infty} x f(x)\,dx = \int_0^{\alpha} \frac{2b}{\alpha}x^2 + bx\,dx =$$

$$= \frac{2b}{\alpha}\left[\frac{x^3}{3}\right]_0^{\alpha} + b\left[\frac{x^2}{2}\right]_0^{\alpha} = \frac{2b\alpha^2}{3} + \frac{b\alpha^2}{2}, \quad \alpha b = 1/2$$

Καλείται η κατανομή με συνάρτηση πυκνότητας πιθανότητας που δίνεται στη μορφή

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

$p(x) \geq 0 \quad \forall x \in \mathbb{R}$

Προσδιορίζεται από δύο παραμέτρους $(\mu,\ \sigma^2)$. Συμβολίζεται ως $\mathcal{N}(\mu, \sigma^2)$

$$\mathbb{E}(X) = \mu, \quad \mathbb{V}(X) = \sigma^2$$

ακολουθεί

μέση τιμή

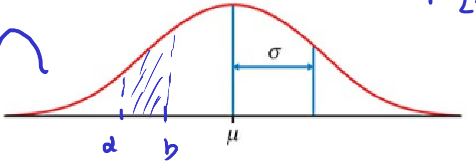
$X \sim \mathcal{N}(\mu, \sigma^2)$

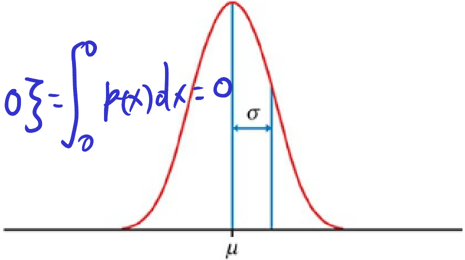
διασπορα.

Κανονική κατανομή

$0 \leq P\{X \in [\alpha, \beta]\} = \int_\alpha^\beta p(x)\,dx \leq 1$

$P\{X = 0\} = \int_0^0 p(x)\,dx = 0$





Κανόνας 68-95-99.7

Εάν η μεταβλητή X ακολουθεί κανονική κατανομή με μέση τιμή $\mathcal{N}(\mu, \sigma^2)$ τότε:

- Περίπου το 68% των παρατηρήσεων της ανήκουν στο διάστημα $[\mu - \sigma, \mu + \sigma]$

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$$

- Περίπου το 95% των παρατηρήσεων της ανήκουν στο διάστημα $[\mu - 2\sigma, \mu + 2\sigma]$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$

- Περίπου το 99.7% των παρατηρήσεων της ανήκουν στο διάστημα $[\mu - 3\sigma, \mu + 3\sigma]$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997$$

Παράδειγμα: $X \sim \mathcal{N}(175, 4)$ $\quad P\{X \in [175 - \underbrace{2}_{\sqrt{4}}, 175 + 2]\} \approx 0.68$

$P\{X \in [175 - 3\cdot 2, 175 + 3 \cdot 2]\} \approx 0.997$

169 , 181

Τυποποίηση Παρατηρήσεων (Standardizing Observations)

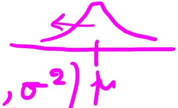
Εάν x μια παρατήρηση της X η οποία ακολουθεί την κανονικής κατανομής $\mathcal{N}(\mu, \sigma^2)$, η τυποποιημένη τιμή του x ορίζεται ως:

z-score.

$\mathcal{N}(0, 1^2)$

$$z = \frac{x - \mu}{\sigma}$$

$X \sim \mathcal{N}(\mu, \sigma^2)$



$Y = X - \mu \sim \mathcal{N}(0, \sigma^2)$

$Z = \frac{Y}{\sigma} \sim \mathcal{N}(0,1)$

Η τυποποιημένη τιμή συχνά καλείται ως **z-score** της παρατήρησης.

- Το z-score εκφράζει τον αριθμό των τυπικών αποκλίσεων που χωρίζουν την αρχική παρατήρηση x από τη μέση τιμή μ.

Chebychef: Για X τ.μ. $\sim$ οποιαδήποτε κατανομή

$P\{X \in [\mu - \sigma, \mu + \sigma]\} \geq 1 - 1/1 = 0$

- Την κανονική κατανομή $\mathcal{N}(0,1)$ με μέση τιμή μηδέν και τυπική απόκλιση μονάδα την καλούμε τυπική κανονική κατανομή.

Τυποποίηση Κανονικής Κατανομής

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mathcal{N}(\mu, \sigma^2) \to \mathcal{N}(0,1)$$

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

Θεωρούμε τον γραμμικό μετασχηματισμό:

$$X = \mu + \sigma Z \leftarrow Z = \frac{X-\mu}{\sigma}$$

$$Z \sim \mathcal{N}(0,1) \Leftrightarrow X \sim \mathcal{N}(\mu, \sigma^2)$$

Προκύπτει η νέα τυποποιημένη συνάρτηση πυκνότητας πιθανότητας

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

$\alpha > 0$

$P(Z \leq -\alpha) = 1 - P(Z \leq \alpha)$

$X \sim N(1,1) \quad P(X \leq 0.5) = P(Z \leq z(0.5)) = P(Z \leq -0.5) = 1 - P(Z \leq 0.5)$

$= 1 - 0.6915$

$Z \sim N(0,1)$

$z(0.5) = \frac{0.5-1}{1} = -0.5$

Standard Normal Probabilities

$z = \boxed{0.5}5$

$P(Z \leq 0.55) = 0.7088$

$z = ;$ τω $P(Z \leq z) = 0.593 \sim z \approx 0.24$

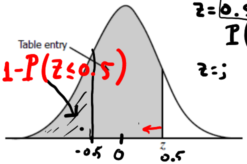


Table entry for z is the area under the standard normal curve to the left of z.

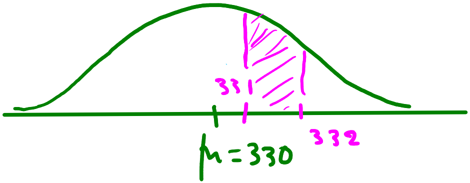
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	$\boxed{.7088}$	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

Άσκηση

Η math.uoc παράγει ένα νέο αναψυκτικό την Stat Cola. Το μηχάνημα που γεμίζει τα μπουκάλια έχει ρυθμιστεί να παρέχει 330 *ml* αναψυκτικού ανά μπουκάλι. Ωστόσο έχει παρατηρήθει ότι η πραγματική ποσότητα δεν είναι σταθερή αλλά περιγράφεται από την κανονική κατανομή με μέση τιμή 330 *ml* και τυπική απόκλιση 2 *ml*. Τι ποσοστό μπουκαλιών περιέχει από 331 εώς 332 *ml* αναψυκτικού.

330 ml $\quad \sigma = 2$ ml $\qquad P(X \in [331, 332]) = ;$

$X \sim N(330, 2^2)$



$$z_1 = \frac{331-330}{2} = \frac{1}{2}$$

$$z_2 = \frac{332-330}{2} = 1$$

$$P(X \in [331, 332]) = \underbrace{P(Z \leq z_2)}_{0.8413} - \underbrace{P(Z \leq z_1)}_{0.6915}$$