ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

15-02-2023

- Μέση τιμή δείγματος: $\bar{\mathrm{X}}$
- Μέση τιμή πληθυσμού: μ

$\{x_1, x_2, \ldots, x_N\}$ $\quad \bar{X} \xrightarrow[N\to\infty]{} \mu$

Έστω $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{N}$ παρατηρήσεις που αντιστοιχούν σε ένα τυχαίο δείγμα ενός πληθυσμού.
Έχουμε ορίσει ως μέση τιμή των παρατηρήσεων του δείγματος την ποσότητα:

$$\bar{\mathrm{X}} = 1/\mathrm{N} \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{x}_\mathrm{n}$$

Αυτή η μέση τιμή εκφράζει μόνο το δείγμα και όχι τον πληθυσμό, αν και για μεγάλο N προσεγγίζει την αντίστοιχη μέση τιμή μ του πληθυσμού.

Ανεξάρτητα των τιμών του δείγματος ισχύει η ανισότικη σχέση

$$\sum_{n=1}^{N}(x_n - \bar{X})^2 \leq \sum_{n=1}^{N}(x_n - \mu)^2$$

με ισότητα μόνο αν $\bar{X} = \mu$.

$f(x) = \sum_{n=1}^{N}(x_n - x)^2$

έχει ελάχιστη τιμή για $x = \overline{X}$

Παράδειγμα

Έστω το πείραμα της ρίψης ενός αμερόληπτου ζαριού.

$$\mu = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5$$

Ρίχνουμε το ζάρι 3 φορές και λαμβάνουμε τα αποτελέσματα: 3,2,6
Έχουμε $\bar{X} = 3.66$

$$\sum_{i=1}^{3} (x_i - \bar{X})^2 = 8.66 < 8.75 = \sum_{i=1}^{3} (x_i - \mu)^2$$

Διασπορά πληθυσμού

Ορίζεται ως η μέση τιμή του συνόλου τιμών

$$\{(\mathrm{x}-\mu)^2\}$$

Διασπορά

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2$$

για κάθε παρατήρηση x του πληθυσμού. Η διασπορά του πληθυσμού συμβολίζεται με σ^2.

Διασπορά στατιστικού δείγματος

$$\mathrm{s}^2 = \frac{1}{\mathrm{N}-1}\sum_{\mathrm{n}=1}^{\mathrm{N}}(\mathrm{x_n}-\bar{\mathrm{X}})^2$$

$N \to \infty$

$\bar{X} \to \mu$

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

$s^2 \le \sigma^2$

Μπορούμε να γράψουμε ισοδύναμα:

$$\mathrm{s}^2 = \frac{\sum_{\mathrm{n}=1}^{\mathrm{N}}\mathrm{x_n^2} - \frac{(\sum_{\mathrm{n}=1}^{\mathrm{N}}\mathrm{x_n})^2}{\mathrm{N}}}{\mathrm{N}-1}$$

Όσο το N αυξάνεται έχουμε $\mathrm{s}^2 \to \sigma^2$.

Διασπορά στατιστικού δείγματος

$$s^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{X})^2$$

Γιατί διαιρούμε με $N-1$ και όχι απλά με N;

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \Rightarrow N\bar{x} = \sum_{n=1}^{N-1} x_n + x_N \Rightarrow$$

$$\Rightarrow x_N = N\bar{x} - \sum_{m=1}^{N-1} x_m$$

Εαν $\bar{x}$ και $x_1, \ldots, x_{N-1}$ γνωστα τοτε x_N υπολογ.

Διασπορά ομαδοποιημένων δεδομένων

$$s^2 = \frac{1}{N-1} \sum_{j=1}^{K} f_j (m_j - \bar{X})^2$$

Μπορούμε να γράψουμε ισοδύναμα:

$$s^2 = \frac{\sum_{j=1}^{K} m_j^2 f_j - \frac{(\sum_{j=1}^{K} m_j f_j)^2}{N}}{N-1}$$

$$s^2 = \frac{\sum_{j=1}^{K} m_j^2 f_j - \frac{(\sum_{j=1}^{K} m_j f_j)^2}{N}}{N-1}$$

Άσκηση - Διασπορά ομαδοποιημένων δεδομένων

	f	m	$m \cdot f$	m^2	$m^2 f$
[0,2)	3	1	3	1	3
[2,4)	4	3	12	9	9·4
[4,6)	5	5	25	25	25·5
[6,8)	2	7	14	49	49·2
[8,10)	4	9	36	81	81·4
[10,12)	2	11	22	121	121·2
N Total	20		$\sum m_j f_j$		$\sum m_j^2 f_j$

Αποτελεί το πιο συχνά χρησιμοποιούμενο μέτρο μεταβλητότητας.
Ορίζεται ως η τετραγωνική ρίζα της διασποράς.

- Τυπική απόκλιση πληθυσμού:

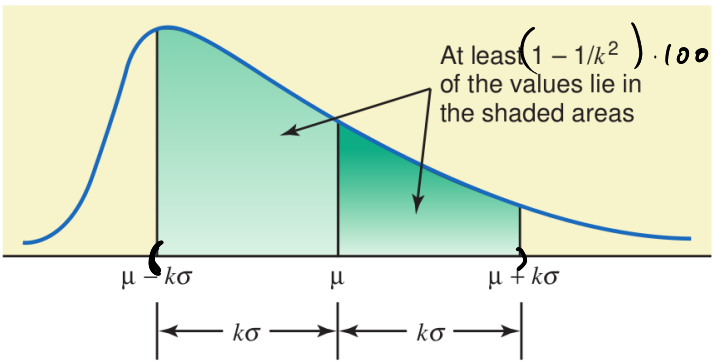
$$\sigma = \sqrt{\sigma^2}$$

- Τυπική απόκλιση δείγματος:

$$\mathrm{s} = \sqrt{\mathrm{s}^2}$$

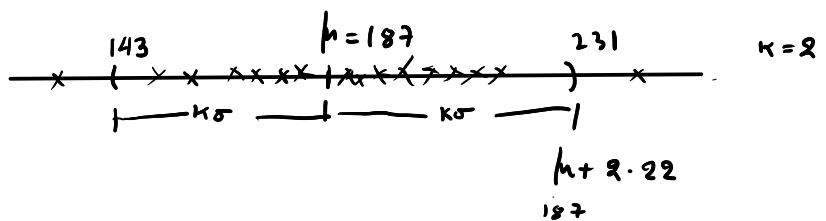
Η τυπική απόκλιση εκφράζεται στην ίδια μονάδα μέτρησης με τη μεταβλητή που αναφέρεται.

Για κάθε $\mathrm{k} > 1$, τουλάχιστον $(1 - 1/\mathrm{k}^2)$ των παρατηρήσεων ανοίκουν στο διάστημα $[\mu - \mathrm{k}\sigma, \mu + \mathrm{k}\sigma]$



Άσκηση

Η μέση συστολική αρτηριακή πίεση 4000 γυναικών που υποβλήθηκαν σε εξέταση για υψηλή πίεση αίματος βρέθηκε να είναι 187 mm Hg με τυπική απόκλιση 22. Χρησιμοποιώντας το Θεώρημα του Chebyshev βρείτε το ελάχιστο ποσοστό των γυναικών αυτής της ομάδας με συστολική αρτηριακή πίεση μεταξύ 143 και 231 mm Hg.



Τουλάχιστο $(1 - 1/4) \cdot 100\% = 75\%$

- Είναι το πηλίκο της τυπικής απόκλισης δια της μέσης τιμής. Συμβολίζεται ως CV:

$$\text{CV} = \frac{\text{S}}{\bar{\text{X}}}$$

$x_n > 0 \quad \forall n \quad \bar{X} > 0$

- Είναι χρήσιμος για τη σύγκριση της ομοιογένειας δυο συσχετισμένων μεταβλητών με διαφορετικές μονάδες μέτρησης ή στο να συγκρίνουμε την ομοιογένεια μεταβλητών με ίδιες μονάδες μέτρησης αλλά με διαφορετικές μέσες τιμές.
- Επίσης χρησιμοποιείται για το χαρακτηρισμό ένος δείγματος ως ανομοιογενές (CV ≥ 0.1) ή ομοιογενές (CV < 0.1) .

Παράδειγμα

Έστω δείγματα με τις ημερήσιες μετρήσεις θερμοκρασίας 2 πολέων στη διάρκεια ενός έτους. Για την πόλη Α η μέση θερμοκρασία ήταν 20 βαθμούς ^{o}C και η τυπική απόκλιση 2, ενώ για την Β η μέση θερμοκρασία ήταν 15 βαθμούς ^{o}C και η τυπική απόκλιση 1.8

Παράδειγμα

Σε δυο γραπτές δοκιμασίες οι μαθητές μιας τάξης είχαν επιδόσεις που περιγράφονται παρακάτω:
δοκιμασία Α (κλίμακα 0-20): μέση τιμή 14, τυπική απόκλιση 1.4
δοκιμασία Β (κλίμακα 0-100): μέση τιμή 70, τυπική απόκλιση 3.5

$\bar{y} = a\bar{x} + b$

$y = ax + b$

$\bar{x}$

s_x^2

cv_x

N_x

M_{ox}

$Y = 3x + 5$

3 5 8 10

$Y = -3x + 5$

$s_y^2 = a^2 s_x^2$

$cv_y = \frac{s_y}{\bar{y}} = \frac{a\, s_x}{a\bar{x} + b}$

$M_y = aM_x + b$

$Q_{1y} =$

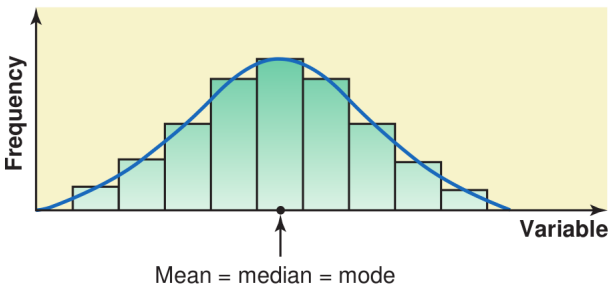
$a < 0 \quad a\,Q_{3x} + b$

$a > 0 \quad a\,Q_{1x} + b$

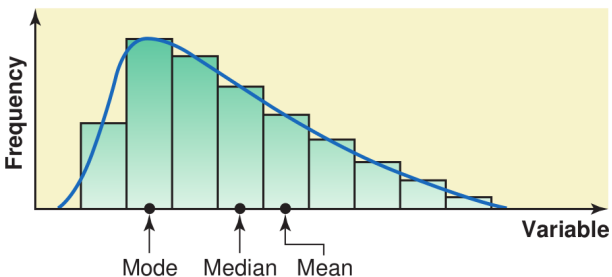
$M_{oy} = aM_{ox} + b$

- 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**.

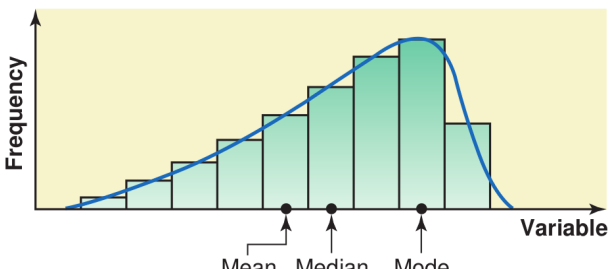
$$\bar{X} = M = M_0$$



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



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



Ο συντελεστής ασυμμετρίας του Pearson ποσοτικοποιεί την ασυμμετρία.

$$\mathrm{Sk_p} = \frac{\bar{\mathrm{X}} - \mathrm{M_0}}{\mathrm{s}}$$

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:

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

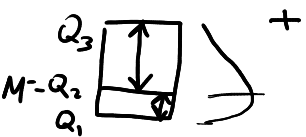
Οπότε προκύπτει ο συντελεστής εκφρασμένος με τη βοήθεια της διαμέσου:

$$\tilde{\mathrm{Sk}}_\mathrm{p} = \frac{3(\bar{\mathrm{X}} - \mathrm{M})}{\mathrm{s}}$$

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

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

$Q_3 + Q_1 - 2M$



- 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 $\mathrm{Q_1}$ σε σχέση με το $\mathrm{Q_3}$ παρατηρείται θετική ασυμμετρία.
- Εαν η διάμεσος είναι πλησιέστερα στο $\mathrm{Q_3}$ σε σχέση με το $\mathrm{Q_1}$ παρατηρείται αρνητική ασυμμετρία.

Άσκηση

Δίνονται οι ακόλουθες διατεταγμένες παρατηρήσεις μιας μεταβλητής:

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

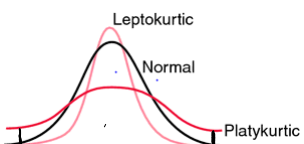
Υπολογίστε τους συντελεστές ασυμμετρίας $\tilde{\mathrm{Sk}}_{\mathrm{p}}, \mathrm{Sk}_{\mathrm{b}}$. Παρουσιάζουν οι παρατηρήσεις κάποια ασυμμετρία;

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.

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

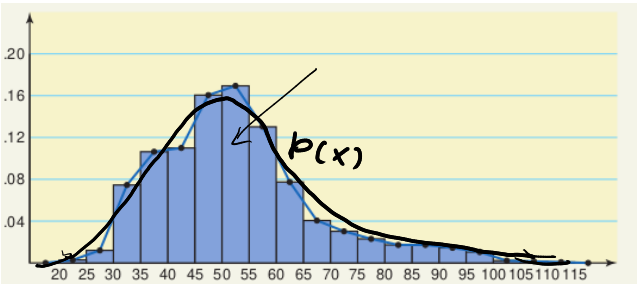
Με βάση τη τιμή του kurtosis λαμβάνουμε τους χαρακτηρισμούς:

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



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.



Μια συνάρτηση πυκνότητας πιθανότητας $\mathrm{p(x)}$:

- Είναι μη αρνητική

$$\mathrm{p(x)} \geq 0, \forall \mathrm{x}$$

$$\int_a^b p(x)\,dx = \mathbb{P}\left(X \in (a,b)\right)$$

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

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

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

Πιθανότητα

$$P(x = \alpha) = 0$$

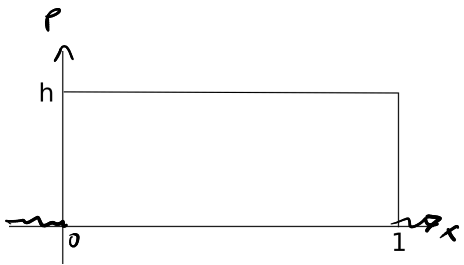
$$\mathrm{P}(\mathrm{X} \in [\mathrm{a}, \mathrm{b}]) = \mathrm{P}([\mathrm{a}, \mathrm{b}]) = \mathrm{P}(\mathrm{a} \leq \mathrm{X} \leq \mathrm{b}) = \int_{\mathrm{a}}^{\mathrm{b}} \mathrm{p}(\mathrm{x})\mathrm{dx}$$

Μέση τιμή - Αναμενόμενη τιμή

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

Διασπορά

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



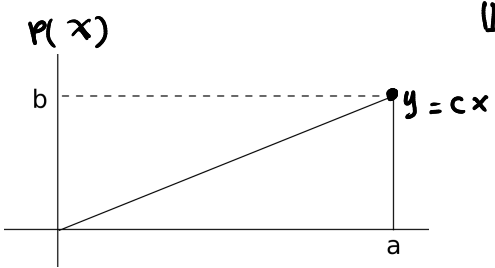
$h=1$

$$P(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & \text{allow} \end{cases}$$

$$E[X] = \int_{-\infty}^{+\infty} x\, P(x)\, dx =$$

$$= \int_0^1 x\, dx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}$$

$\frac{1}{2} a b = 1 \Rightarrow ab = 2$



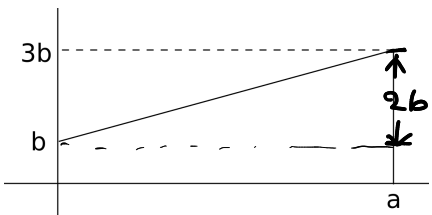
$$\mathbb{E}[X] = \int_0^a x \frac{b}{a} x \, dx =$$

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

$$p(x) = \frac{b}{a} x$$

$\frac{4b}{2} \cdot a = 1$

$b \cdot a = \frac{1}{2}$



$p(x) = \frac{2b}{a} x + b$

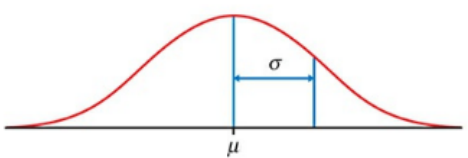
$E[X] = \int_0^a \left(\frac{2b}{a} x + b\right) x \, dx =$

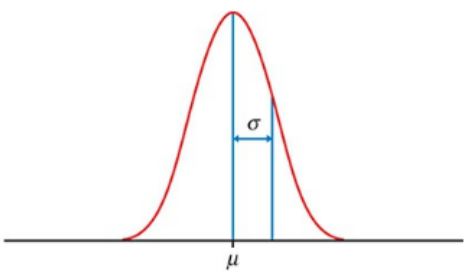
Καλείται η κατανομή με συνάρτηση πυκνότητας πιθανότητας που δίνεται στη μορφή

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

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

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





Κανόνας 68-95-99.7

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

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

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

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

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

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

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

Τυποποίηση Παρατηρήσεων (Standardizing Observations)

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

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

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.

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

Τυποποίηση Κανονικής Κατανομής

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

Θεωρούμε τον γραμμικό μετασχηματισμό:

$$\mathrm{Z} = \frac{\mathrm{X}-\mu}{\sigma}$$

Προκύπτει η νέα τυποποιημένη συνάρτηση πυκνότητας πιθανότητας

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

Τυπική Κανονική Κατανομή (Standard Normal Distribution)

Standard Normal Probabilities

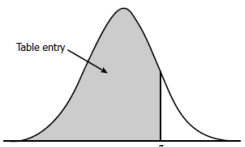


Table entry for z is the area under the standard normal curve to the left of z.

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	.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

Μια εταιρία παράγει ένα νέο αναψυκτικό. Το μηχάνημα που γεμίζει τα μπουκάλια έχει ρυθμιστεί να παρέχει $330\ \mathrm{ml}$ αναψυκτικού ανά μπουκάλι. Ωστόσο έχει παρατηρήθει ότι η πραγματική ποσότητα δεν είναι σταθερή αλλά περιγράφεται από την κανονική κατανομή με μέση τιμή $330\ \mathrm{ml}$ και τυπική απόκλιση $2\ \mathrm{ml}$. Τι ποσοστό μπουκαλιών περιέχει από 331 εώς $332\ \mathrm{ml}$ αναψυκτικού.

Τυπική Κανονική Κατανομή (Standard Normal Distribution)