ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

20-02-2020

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 $N = 6$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{6}(-2-1+0+1+2+6) = 1$$

$$x - \bar{x} \rightarrow -3 \quad -2 \quad -1 \quad 0 \quad 1 \quad 5 \xrightarrow{(\,)^3} (-3)^3 + (-2)^3 + (-1)^3 + 1^3 + 5^3 = \dots$$

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{5}\left[(-3)^2 + (-2)^2 + \dots + 5^2\right] \rightarrow s \rightarrow s^3$$

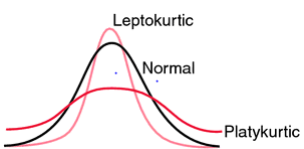
$$\frac{N^2}{(N-1)(N-2)} \cdot \frac{\quad}{\quad}$$

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

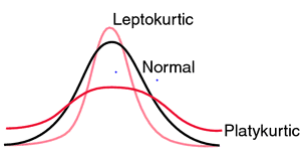


Η βιβλιοθήκη pandas (python) χρησιμοποιεί μια τροποποιημένη έκφραση για το συντελεστή κύρτωσης (θα το δούμε στο 4ο εργαστήριο).

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

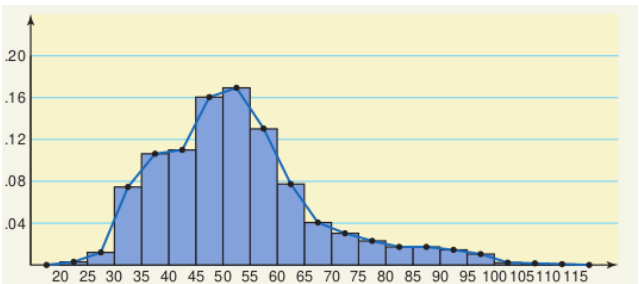
Με βάση τη τιμή του kurtosis λαμβάνουμε τους χαρακτηρισμούς:

- kurt $= 0$: Μεσόκυρτη (Κανονική)
- kurt < 0: Πλατύκυρτη
- kurt > 0: 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 $p(x)$:

- 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.

Πιθανότητα

$P([a,b]) = P((a,b]) = P([a,b)) = P((a,b))$.

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

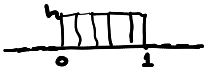
Ιδιότητα $P(X = k) = 0 \quad \forall k \qquad P(x=k) = \int_k^k p(x)dx = 0$

Μέση τιμή - Αναμενόμενη τιμή

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

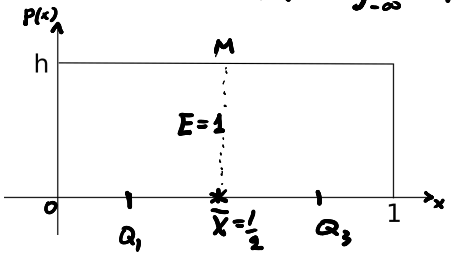
Διασπορά

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



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

$$E(x) = \int_{-\infty}^{+\infty} x\, p(x)\, dx = \int_0^1 x\, dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}$$



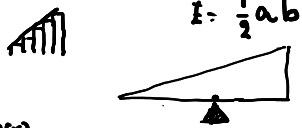
$$M = \frac{1}{2}$$

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

$$= \int_0^1 \Big(\overbrace{x - \tfrac{1}{2}}^{y}\Big)^2 dx \qquad dy = dx$$

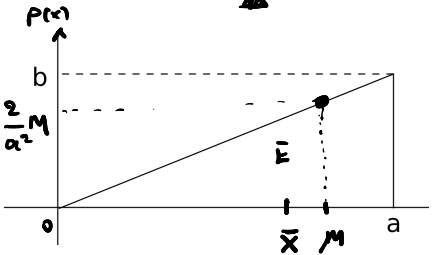
$$= \int_{-1/2}^{1/2} y^2\, dx = \left[\frac{y^3}{3}\right]_{-1/2}^{1/2} =$$

$$= 2 \cdot \frac{\left(\frac{1}{2}\right)^3}{3} = \frac{1}{12}$$



$E = \frac{1}{2}ab = 1 \qquad b = \frac{2}{a}$

$$P(x) = \begin{cases} \frac{b}{a}x = \frac{2}{a^2}x \\ 0, \quad x \notin [0,a] \end{cases}$$



$$E(X) = \int_0^a x \frac{2}{a^2}x\,dx = \frac{2}{a^2}\int_0^a x^2\,dx =$$

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

$$= \frac{2}{3}a$$

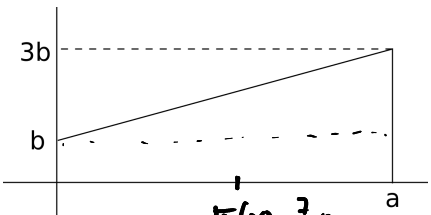
$$E_1 = \frac{1}{2}M \cdot \frac{2}{a^2}M = \frac{1}{2}$$

$$\left[\frac{M}{E(X)}\right]^2 = \left[\frac{\frac{\sqrt{2}}{2}a}{\frac{2}{3}a}\right]^2 = \qquad M^2 = \frac{a^2}{2} \Rightarrow M = \frac{\sqrt{2}}{2}a$$

$$= \left(\frac{\sqrt{2}\cdot 3}{4}\right)^2 = \frac{2\cdot 9}{16} = \frac{18}{16} > 1$$

$$E = \frac{1}{2}(b+3b)\cdot a = 2ab = 1 \iff b = \frac{1}{2a}$$

$$P(x) = \begin{cases} \frac{2b}{a}x + b = \frac{2}{2a^2}x + \frac{1}{2a} = \frac{1}{a^2}x + \frac{1}{2a} \\ 0\ ,\ x \in [0,a] \end{cases}$$



$$E(X) = \int_0^a x \cdot \left(\frac{1}{a^2}x + \frac{1}{2a}\right) dx =$$

$$= \frac{1}{a^2}\int_0^a x^2\,dx + \frac{1}{2a}\int_0^a x\,dx =$$

$$= \frac{1}{a^2}\left[\frac{x^3}{3}\right]_0^a + \frac{1}{2a}\left[\frac{x^2}{2}\right]_0^a =$$

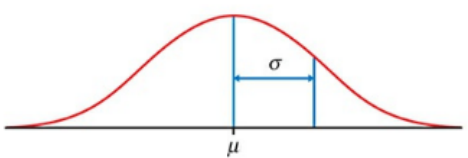
$$= \frac{a^3}{3a^2} + \frac{a^2}{4a} = \frac{a}{3} + \frac{a}{4} = \frac{7}{12}a$$

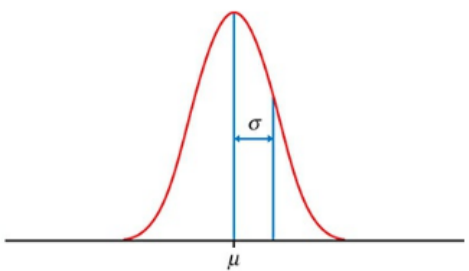
Καλείται η κατανομή με συνάρτηση πυκνότητας πιθανότητας που δίνεται στη μορφή

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

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

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





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 $[\mu - \sigma, \mu + \sigma]$

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

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

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

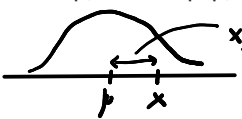
- Περίπου το 99.7% των παρατηρήσεων της ανήκουν στο διάστημα $[\mu - 3\sigma, \mu + 3\sigma]$

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

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:

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



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:

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

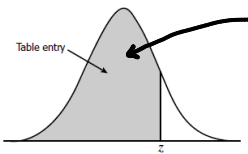
Προκύπτει η νέα τυποποιημένη συνάρτηση πυκνότητας πιθανότητας

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

$$P(z \le z_1) = \int_{-\infty}^{z_1} p(z)\,dz$$

Τυπική Κανονική Κατανομή (Standard Normal Distribution)

Standard Normal Probabilities



$P(Z \leq z)$

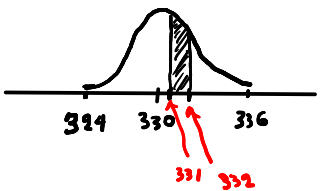
$P(Z \leq 0.65) = 0.7422$

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

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.



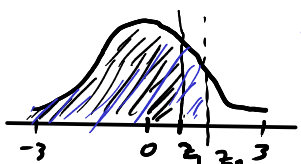
$$z = \frac{x - \mu}{\sigma}$$

$x_1 = 331 \longrightarrow z_1 = 0.5$

$x_2 = 332 \longrightarrow z_2$

$$z_1 = \frac{331 - 330}{2} = \frac{1}{2} = 0.5$$

$$z_2 = \frac{332 - 330}{2} = \frac{2}{2} = 1.0$$



$$P(z_1 \le Z \le z_2) = P(z \le z_2) - P(z \le z_1)$$

$$= 0.8413 - 0.6915 = 0.1498$$

14.98 %.

Τυπική Κανονική Κατανομή (Standard Normal Distribution)