MEM-205 Περιγραφική Στατιστική

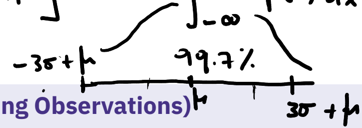
Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

20-02-2023

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

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



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:

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

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

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

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

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.

Γνωρίζουμε ότι X έχει $\mathrm{Var}\,X = \sigma^2$ $\qquad$ $\mathrm{Var}[aX] = a^2\,\mathrm{Var}\,X$

$$X = \mu + \sigma Z$$

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

Τυποποίηση Κανονικής Κατανομής

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

Θεωρούμε τον γραμμικό μετασχηματισμό:

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

Προκύπτει η νέα τυποποιημένη συνάρτηση πυκνότητας πιθανότητας

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

Standard Normal Probabilities

$z = -0.15$

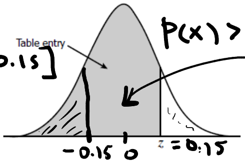
$P[Z \leq -0.15] = P[Z > 0.15]$

$1 - P[Z \leq 0.15]$

$-\infty$

$A \cap B = \emptyset$

$P(A \cup B) = P(A) + P(B)$



$P[Z \leq z] = P[Z < z] = \int_{-\infty}^{z} f(z)\,dz$

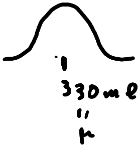
Table entry for z is the area under the standard normal curve to the left of z.

$P[Z \leq 0.15] = 0.5546$

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	$\boxed{.5596}$	.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

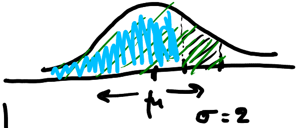
Άσκηση

Μια εταιρία παράγει ένα νέο αναψυκτικό. Το μηχάνημα που γεμίζει τα μπουκάλια έχει ρυθμιστεί να παρέχει 330 ml αναψυκτικού ανά μπουκάλι. Ωστόσο έχει παρατηρηθεί 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.



$$P[\overset{x_1}{331} \le X \le \overset{x_2}{332}] = ;$$

$$P[X \le 332] - P[X \le 331]$$



$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{1}{2}$$

$$z_2 = \frac{x_2 - \mu}{\sigma} = 1$$

$$= P[Z \le 1] - P[Z \le \tfrac{1}{2}]$$

$$= 0.8413 - 0.6915$$

$\leftarrow 1$
x_n

Έστω $x_1 \leq x_2 \leq \cdots \leq x_N$ παρατηρήσεις μιας μεταβλητής X.

$x_1 = 1000$
$x_2 = 2000$
$x_3 = 3000$

$\Phi_1 = \frac{1000}{6000} = \frac{1}{6}$

$\Phi_2 = \frac{3000}{6000} = \frac{1}{2}$

$$0 \leq \Phi_n = \frac{\sum_{j=1}^{n} x_j}{\sum_{j=1}^{N} x_j} \leq 1$$

$$RF_n = n/N$$

n

$x_1 = x_2 = x_3$

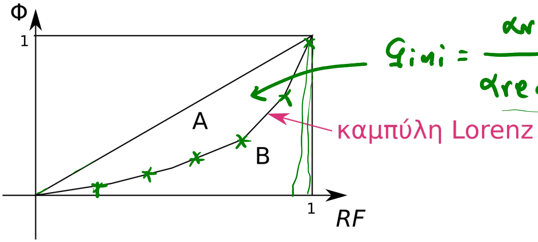
$\Phi_n = \Phi_{n-1} + \frac{1}{N}$

$\Phi_n = \frac{n\, x_1}{3 x_1} = \frac{n}{3} = \frac{n-1}{3} + \frac{1}{3}$

$\frac{n-1}{3} = \Phi_{n-1}$

- Θεωρούμε την καμπύλη που ορίζεται από τα σημεία

$$\{(0,0), (RF_1, \Phi_1), (RF_2, \Phi_2), \ldots, (RF_N = 1, \Phi_N = 1)\}$$

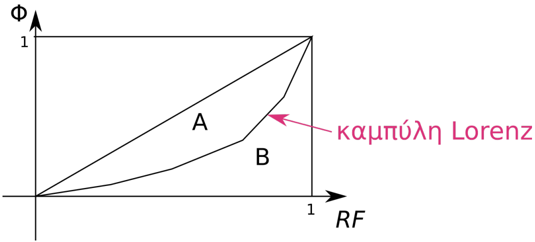


$$\text{Gini} = \frac{\text{area}(A)}{\text{area}(A) + \text{area}(B)} = \frac{1}{2}$$

$$\phi_j = \frac{m_j f_j}{\sum_{k=1}^{K} m_k f_k}, \quad \Phi_i = \sum_{j=1}^{i} \phi_j$$

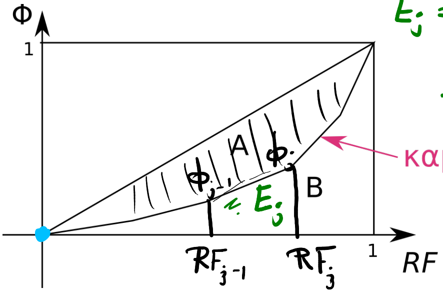
- Θεωρούμε την καμπύλη που ορίζεται από τα σημεία

$$\{(0,0), (RF_1, \Phi_1), (RF_2, \Phi_2), \ldots, (RF_K = 1, \Phi_K = 1)\}$$



$$E_j = \frac{1}{2}(\phi_{j-1}+\phi_j)\overbrace{(RF_j - RF_{j-1})}^{1/N} =$$

$$= \frac{1}{2N}(\phi_{j-1}+\phi_j) = \frac{1}{2N}\sum \Phi_j$$



$$\text{area}(B) = \sum E_j$$

$$\text{area}(A) = \tfrac{1}{2} - \sum E_j$$

$$\text{Gini} = \frac{\text{area(A)}}{\text{area(A)} + \text{area(B)}}, \quad 0 \leq \text{Gini} \leq 1$$

$$\text{Gini} = 1 - \frac{1}{2}\underbrace{\sum_{j=1}^{N} E_j}_{\text{area}(B)}$$

- Αποτελεί μέτρο ανισοκατανομής, δηλαδή ελέγχει κατά πόσο ανισοκατανέμεται η συνολική τιμή μιας μεταβλητής.
- Βρίσκει εφαρμογή σε οικονομικές μελέτες, για παράδειγμα μελέτη για την ανισοκατανομή των μισθών των εργαζομένων μιας επιχείρησης.

Παράδειγμα

Έστω οι ετησιοι μισθοί των 5 εργαζομένων μιας εταιρείας.

$$x_1 = 5000,\ x_2 = 10000,\ x_3 = 15000,\ x_4 = 20000,\ x_5 = 50000$$

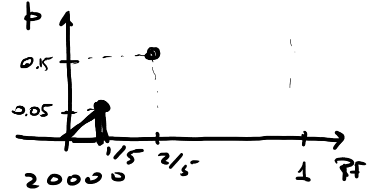
Σχεδιάστε τη καμπύλη Lorenz και υπολογίστε τον συντελεστή του Gini.

		Φ	RF	ΣΦ
#1	5000	0.05	1/5	0.05
#2	10000	0.15	2/5	0.2
#3	15000	0.3	3/5	0.45
#4	20000	0.5	4/5	0.8
#5	50000	1	1	1.5
	100000			

5000/100000

$$\frac{5000 + 10000}{100000}$$

$$\frac{5000 + \dots + 20000}{100000}$$



$$\frac{0.05 + 0.2 + 0.45 + 0.8 + 1.5}{2.5} = \text{area}(B)$$

$$\text{Gini} = 1 - 0.5 \cdot \text{area}(B)$$

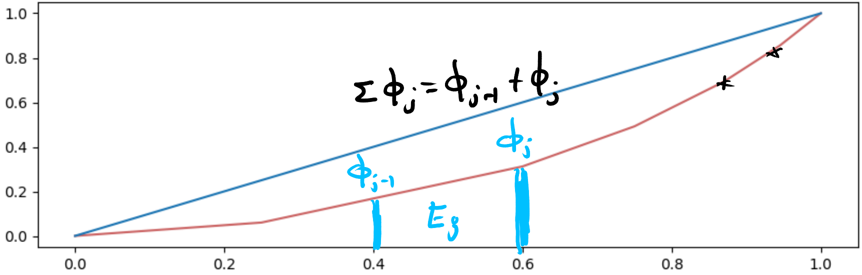
Καμπύλη Lorenz - Συντελεστής του Gini

1687500/10425000   625000/10.425.000

Παράδειγμα

	m	f	mf	φ	ΣΦ	RF
[0,5000)	2500	250	625000	0.06	0.06	0.25 = 250/1000
[5000,10000)	7500	350	2625000	0.252	0.312	0.6 = 600/1000
[10000,15000)	12500	150	1875000	0.18	0.492	0.75
[15000,20000)	17500	120	2100000	0.201	0.693	0.87
[20000, 25000)	22500	75	1687500	0.162	0.855	0.945 = 945/1000
[25000,30000)	27500	55	1512500	0.145	1	1
Total		1000	10425000	1		

$$E_j = \frac{1}{2}\left(\Sigma\Phi_j\right) * \Delta RF_j$$



$\leftarrow \Delta RF_j \rightarrow$

$$\Delta RF_j = RF_j - RF_{j-1}$$

Παράδειγμα

	Φ	RF	ΣΦ	Δ(RF)	ΣΦ x Δ(RF)
[0,5000)	0.06	0.25	0.06	0.25	0.015
[5000,10000)	0.312	0.6	0.372	0.35	0.130
[10000,15000)	0.492	0.75	0.804	0.15	0.121
[15000,20000)	0.693	0.87	1.185	0.12	0.142
[20000,25000)	0.855	0.945	1.548	0.075	0.116
[25000,30000)	1	1	1.855	0.055	0.102
Total					0.626

0.25 - 0

0.75 - 0.6

0.945 - 0.87

$\cdot \frac{1}{2}$ = εμβαδόν

$$\text{Gini} = 1 - 0.626 = 0.374$$

Καμπύλη Lorenz - Συντελεστής του Gini

Gini * 100%

	Member state	2011	2012	2013	2014	2015	2016	2017	2018
1	**Bulgaria**	35.0	33.6	35.4	35.4	37.0	37.7	40.2	39.6
2	**Lithuania**	33.0	32.0	34.6	35.0	37.9	37.0	37.6	36.9
3	**Latvia**	35.1	35.7	35.2	35.5	35.4	34.5	34.5	35.6
4	***Serbia***[n 1]	—	—	*38.0*	*38.6*	*38.2*	*38.6*	*37.8*	*35.6*
5	**Romania**	33.5	34.0	34.6	35.0	37.4	34.7	33.1	35.1
6	**Italy**	32.5	32.4	32.8	32.4	32.4	33.1	32.7	33.4
7	**Luxembourg**	27.2	28.0	30.4	28.7	28.5	31.0	30.9	33.2
8	**Spain**	34.0	34.2	33.7	34.7	34.6	34.5	34.1	33.2
9	**Greece**	33.5	34.3	34.4	34.5	34.2	34.3	33.4	32.3
10	**Portugal**	34.2	34.5	34.2	34.5	34.0	33.9	33.5	32.1
11	**Germany**	29.0	28.3	29.7	30.7	30.1	29.5	29.1	31.1
12	**Estonia**	31.9	32.5	32.9	35.6	34.8	32.7	31.6	30.6
13	**Croatia**	31.2	30.9	30.9	30.2	30.4	29.8	29.9	29.7
14	**Cyprus**	29.2	31.0	32.4	34.8	33.6	32.1	30.8	29.1
15	**Ireland**	29.8	30.5	30.7	31.1	29.8	29.5	30.6	28.9

	Member state	2011	2012	2013	2014	2015	2016	2017	2018
16	**Hungary**	26.9	27.2	28.3	28.6	28.2	28.2	28.1	28.7
17	**Malta**	27.2	27.1	27.9	27.7	28.1	28.5	28.3	28.7
18	**France**	30.8	30.5	30.1	29.2	29.2	29.3	29.3	28.5
19	**Denmark**	26.6	26.5	26.8	27.7	27.4	27.7	27.6	27.9
20	**Poland**	31.1	30.9	30.7	30.8	30.6	29.8	29.2	27.8
21	**Netherlands**	25.8	25.4	25.1	26.2	26.7	26.9	27.1	27.0
22	**Sweden**	26.0	26.0	26.0	26.9	26.7	27.6	28.0	27.0
23	**Austria**	27.4	27.6	27.0	27.6	27.2	27.2	27.9	26.8
24	**Finland**	25.8	25.9	25.4	25.6	25.2	25.4	25.3	25.9
25	**Belgium**	26.3	26.5	25.9	25.9	26.2	26.3	26.0	25.6
26	**Czech Republic**	25.2	24.9	24.6	25.1	25.0	25.1	24.5	24.0
27	**Slovenia**	23.8	23.7	24.4	25.0	24.5	24.4	23.7	23.4
28	**Slovakia**	25.7	25.3	24.2	26.1	23.7	24.3	23.2	20.9
29	***Montenegro***[n 2][11]	—	—	*38.5*	*36.5*	*36.5*	*36.5*	*36.7*	
	European Union	30.5	30.4	30.6	30.9	30.8	30.6	30.3	30.4