ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

5η εβδομάδα (διάλεξη θεωρίας)

0 <

6/7 των παρατηρήσεων

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

$\Phi_1 = \frac{x_1}{\sum x_j}$, $\Phi_2 = \frac{x_1 + x_2}{\sum x_j}$, ..., $\Phi_n = \frac{\sum x_j}{\sum x_j}$

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

$\Phi_0 = 0$

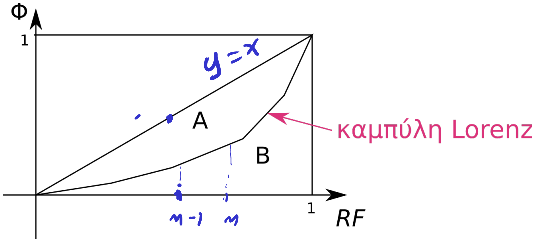
$$RF_n = n/N$$

$n = 0$ $RF_n = 0$

$n = N$ $RF_N = 1$

- Θεωρούμε την καμπύλη που ορίζεται από τα σημεία

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



1 [ ] m_1 f_1 $F_1 = f_1$ $RF_j = \frac{F_j}{N}$

2 [ ] m_2 f_2 $F_2 = f_1 + f_2$

$\vdots$

k [ ] m_k f_k $F_k = f_1 + \dots + f_k$

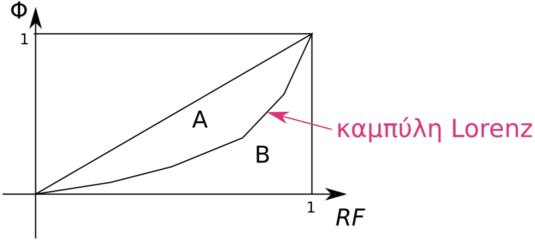
$$\phi_i = \frac{m_i f_i}{\sum_{j=1}^{K} m_j f_j}, \quad \Phi_i = \sum_{j=1}^{i} \phi_j$$

$\phi_K = 1$, $\phi_0 = 0$

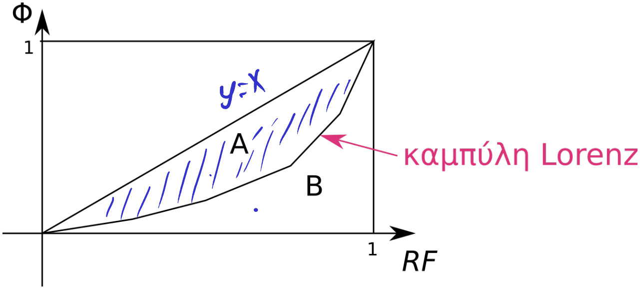
- Θεωρούμε την καμπύλη που ορίζεται από τα σημεία

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

RF_j



Καμπύλη Lorenz - Συντελεστής του Gini



$$\text{Gini} = \frac{\text{area}(A)}{\underbrace{\text{area}(A) + \text{area}(B)}_{"1/2}}, \quad 0 \leq \text{Gini} \leq 1$$

- 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.

$\Phi_0 = 0 \quad \Phi_1 = \frac{5000}{100000} = \frac{1}{20} \quad \Phi_2 = \frac{15000}{100000} = \frac{3}{20} \quad \Phi_3 = \frac{30000}{100000} = \frac{3}{10}$

$\Phi_4 = 0.5 \quad \Phi_5 = 1$

$RF_0 = 0 \quad RF_1 = 0{,}2,\ RF_2 = 0.4,\ RF_3 = 0.6,\ RF_4 = 0.8,\ RF_5 = 1.$

$\{(0,0), (0.2, 1/20), (0.4, 3/20), (0.6, 3/10), (0.8, 0.5), (1,1)\}$

$$\frac{1875000}{10425000}$$

Παράδειγμα

	m	f	mf	φ	Φ	RF	F
[0,5000)	2500	250	625000	0.06	0.06	0.25	250
[5000,10000)	7500	350	2625000	0.252	0.312	0.6	600
[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	
[25000,30000)	27500	55	1512500	0.145	1	1	
Total		1000	10425000	1			

$$\frac{600}{1000}$$

$$\{(0,0),(0.06,0.25),(0.312,0.6),(0.492,0.75),(0.693,0.87),(0.855,0,945),(1,1)\}$$

0.492+0.201

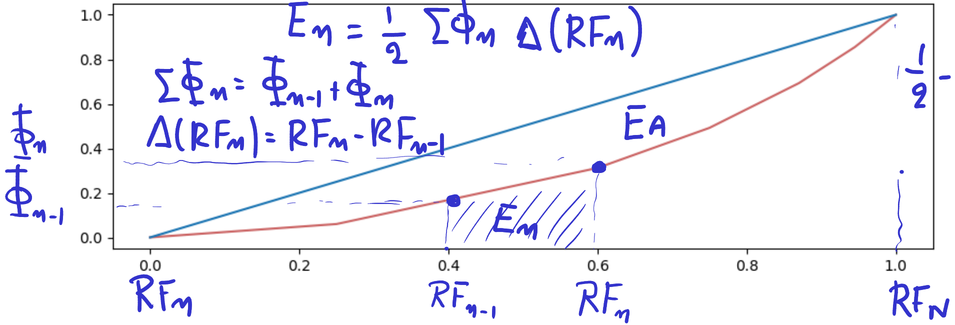
$$E_n = \frac{1}{2}\left(\Phi_{n-1} + \Phi_n\right)\left(RF_n - RF_{n-1}\right) \qquad E_A = \frac{1}{2} - \sum_{n=1}^{N} E_n$$

$$E_n = \frac{1}{2} \Sigma\Phi_n \, \Delta(RF_n)$$

$$\Sigma\Phi_n = \Phi_{n-1} + \Phi_n$$

$$\Delta(RF_n) = RF_n - RF_{n-1}$$

$$= \frac{1}{2} - \frac{1}{2}\sum_{n=1}^{N} \Sigma\Phi_n \Delta(RF_n)$$



$$Gini = 1 - \sum \Sigma\Phi_n \Delta(RF_n)$$

$\Sigma\Phi_n = \Phi_n + \Phi_{n-1}$ $\Delta(RF_n) = RF_n - RF_{n-1}$

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 = 0.804*0.15
[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

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

Καμπύλη Lorenz - Συντελεστής του Gini

$0.5(N-1)$

$0.5 \cdot 6 = 3$

8

$0.5 \cdot 7 = 3.5$

$3 \to X_4 + 0.5 \cdot (X_5 - X_4) = 0.5 \cdot (X_4 + X_5) = \frac{X_4 + X_5}{2}$

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

$X \sim N(\mu, \sigma^2)$ $\quad f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$ $\quad P(X \in [a,b]) = \int_a^b f(x)\,dx$

x

Δειγματική κατανομή της $\bar{X}$

Η στατιστική κατανομή της $\bar{X}$ καλείται δειγματική κατανομή της $\bar{X}$.

$x_1, x_2, \ldots, x_N \leftarrow 1^η$ πραγματοποίηση $\quad \bar{X}$

$x_1', x_2', \ldots, x_N' \leftarrow 2^η$ " $\quad \bar{X}'$

$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$

Γενικά η στατιστική κατανομή οποιοδήποτε στατιστικού του δείγματος καλείτε δειγματική κατανομή του συγκεκριμένου στατιστικού.

Δειγματικό Σφάλμα

Είναι η διαφορά μεταξύ της τιμής ενός στατιστικού ενός δείγματος και της αντίστοιχης τιμής του στατιστικού που αφορά τον πληθυσμό. Στη περίπτωση της μέσης τιμής έχουμε:

$N=100$

$1, 5, 6, 2, \ldots, 3$ $\quad \bar{x} = 3.6$

$3, 6, 6, 1, \ldots, 1$ $\quad \bar{x}' = 3.3$

$$\text{Δειγματικό σφάλμα} = \bar{X} - \mu$$

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

$\bar{X} \leftarrow$ Τ.μ της Δειγματικής μέσης τιμής

Παράδειγμα

5

Έστω ότι σε ένα μάθημα υπηρξάν μόνο 5 εγγεγραμένοι φοιτητές και οι τελική τους αξιολόγηση ήταν: 5, 3, 7, 10, 6
Βρείτε τη μέση τιμή όλων των δειγμάτων με τρία στοιχεία. Στη συνέχεια υπολογίστε τη δειγματική κατανομή της $\bar{X}$ των δειγμάτων με τρία στοιχεία.

Έχουμε συνολικά 10 δείγματα. Γιατί; 3 10

$$(5,3,7) \to \bar{x} = 5,\ (5,3,10) \to \bar{x} = 6,\ (5,3,6) \to \bar{x} = 4.67,\ (5,7,10) \to \bar{x} = 7.33,\ (5,7,6) \to \bar{x} = 6$$
$$(5,10,6) \to \bar{x} = 7,\ (3,7,10) \to \bar{x} = 6.67,\ (3,7,6) \to \bar{x} = 5.33,\ (3,10,6) \to \bar{x} = 6.33,\ (7,10,6) \to \bar{x} = 7.67$$

$\bar{X}$ $\{5, 6, 4.67, \ldots, 7.67\}$ ← πραγματοποιήσεις για την $\bar{X}$

$$\mu = \frac{5+3+7+10+6}{5}$$

$N=10$

Μέση Τιμή και Τυπική Απόκλιση της $\bar{X}$

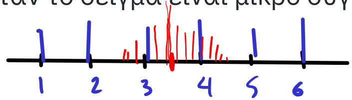
- Η μέση τιμή της δειγματικής κατανομής της $\bar{X}$ συμβολίζεται ως $\mu_{\bar{X}}$
- Η ~~μέση τιμή~~ Τυπική απόκλιση. της δειγματικής κατανομής της $\bar{X}$ συμβολίζεται ως $\sigma_{\bar{X}}$

$$\mu_{\bar{X}} = \mu$$

αριθμός στοιχείων του δείγματος.

Όταν το δείγμα είναι μικρό συγκριτικά με το πληθυσμό $(N/N_p \leq 0.05)$

αριθμός στοιχείων του πληθυσμού.



$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

Όταν η παραπάνω συνθήκη δεν ικανοποιήται χρησιμοποιούμε την έκφραση:

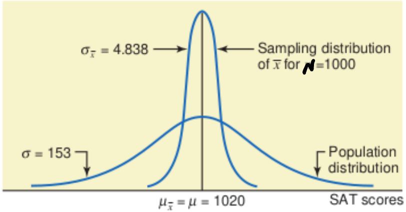
$N/N_p > 0.05$

$$\sigma_{\bar{X}} = \sqrt{\frac{N_p - N}{N_p - 1}}\frac{\sigma}{\sqrt{N}}$$

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

$\mu = \mu_{\bar{X}}$ $\qquad$ $\frac{\sigma^2}{N} = \sigma^2_{\bar{X}}$ $\qquad$ $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$

Εάν X ακολουθεί την $\mathcal{N}(\mu, \sigma^2)$ τότε η $\bar{X}$ ακολουθεί την $\mathcal{N}(\mu_{\bar{X}}, \sigma^2_{\bar{X}})$

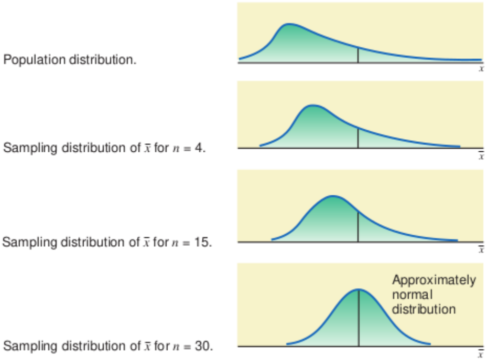


$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{N})$ $\qquad$ $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{25})$

Δειγματοληψία από Πληθυσμό που δεν ακολουθεί Κανονική κατανομή

Σύμφωνα με το **κεντρικό οριακό θεώρημα**, για μεγάλο μέγεθος του δείγματος, η δειγματική κατανομή της $\bar{X}$ προσεγγίζει τη κανονική κατανομή $(\mu_{\bar{X}}, \sigma_{\bar{X}})$ ανεξάρτητα της κατανομής που ακολουθεί η X.

Σε αυτή τη περίπτωση θεωρούμε ένα δείγμα επαρκώς μεγάλο όταν $N \geq 30$.



$X \sim N(\mu, \sigma^2) \qquad Z = \frac{X - \mu}{\sigma}$

1. Για X που ακολουθεί κανονική κατανομή, υπολογισμός της πιθανότητας η $\bar{X}$ να ανήκει σε συγκεκριμένο διάστημα.
2. Για X που δεν ακολουθεί κανονική κατανομή, υπολογισμός της πιθανότητας η $\bar{X}$ να ανήκει σε συγκεκριμένο διάστημα όταν $\underline{N \geq 30}$.

Σε κάθε περίπτωση μπορούμε να υπολογίσουμε το **z-score** για την $\bar{X}$

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \sqrt{N}\,\frac{\bar{X} - \mu}{\sigma}$$

$N \geq 30 \qquad \bar{X} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$

$X \sim N(8.4,\ 1.8^2)$ $\quad \mu = 8.4 \quad \sigma = 1.8$

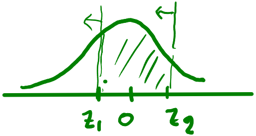
Ο χρόνος παράδοσης παραγγελιών σε ένα fast food στις ώρες αιχμής ακολουθεί κανονική κατανομή με μέση τιμή 8.4 λεπτά και τυπική απόκλιση 1.8 λεπτά. Για ένα τυχαίο δείγμα 16 παραγγελιών υπολογίστε την πιθανότητα η μέση τιμή του δείγματος να είναι:

1. Μεταξύ 8 και 9 λεπτών. $\mu = 8.4$

2. Τουλάχιστον 1 λεπτό λιγότερο από τη μέσο χρόνο παράδοσης που αντιστοιχεί σε όλο τον πληθυσμό.

$N = 16 \qquad \bar{X} \sim N\left(8.4, \frac{1.8^2}{16}\right) \qquad \mu_{\bar{x}} = 8.4 \qquad \sigma_{\bar{x}} = \frac{1.8}{4} = 0.45$

1. $P(\bar{X} \in [8,9])$



$Z = \sqrt{N}\,\frac{\bar{x} - \mu}{\sigma} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} =$

$z_1 = z(\bar{x} = 8) = \frac{8 - 8.4}{0.45} = \frac{-0.4}{0.45}$

$z_2 = z(\bar{x} = 9) = \frac{9 - 8.4}{0.45} = \frac{0.6}{0.45}$

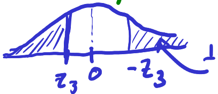
$P(\bar{x} \in [8,9]) = P(Z \in [z_1, z_2]) = P(Z \le z_2) - P(Z \le z_1)$

$1 - P(Z \le -z_3)$

$\parallel$

$P(Z \le z_3)$

2.



$1 - P(Z \le -z_3)$

$Z_3 = \frac{\mu - 1 - \mu}{\sigma_{\bar{x}}} = \frac{-1}{0.45}$

Μια αναλογία στο πληθυσμό προκύπτει ως το λόγο του αριθμού των στοιχείων του πληθυσμού που παρουσιάζουν μια χαρακτηριστική ιδιότητα με το μέγεθος του πληθυσού. Συμβολίζεται με p. Η αντίστοιχη αναλογία για ένα δείγμα συμβολίζεται με $\hat{p}$.

$$p = \frac{M_p}{N_p}, \quad \hat{p} = \frac{M}{N}$$

$N \to N_p$

$\hat{p} \to p$

Όπου:

- N_p το μέγεθος του πληθυσμού.
- M_p αριθμός στοιχείων του πληθυσμού που παρουσιάζουν την ιδιότητα που μελετάμε.
- N το μέγεθος του δείγματος.
- M αριθμός στοιχείων του δείγματος που παρουσιάζουν την ιδιότητα που μελετάμε.

Μέση Τιμή και Τυπική Απόκλιση του $\hat{p}$

- Η μέση τιμή της δειγματικής κατανομής της $\hat{p}$ συμβολίζεται ως $\mu_{\bar{p}}$
- Η μέση τιμή της δειγματικής κατανομής της $\hat{p}$ συμβολίζεται ως $\sigma_{\bar{p}}$

$\hat{p}$

$$\mu_{\hat{p}} = p$$

Όταν το δείγμα είναι μικρό συγκριτικά με το πληθυσμό ($N/N_p \leq 0.05$)

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{N}}$$

Όταν η παραπάνω συνθήκη δεν ικανοποιείται χρησιμοποιούμε την έκφραση:

$$\sigma_{\hat{p}} = \sqrt{\frac{N_p - N}{N_p - 1}}\sqrt{\frac{p(1-p)}{N}}$$

Από το κεντρικό οριακό θεώρημα όταν Np και $N(1-p)$ αρκετά μεγάλοι αριθμοί η $\hat{p}$ ακολουθεί την κατανομή $\mathcal{N}(\mu_{\hat{p}}, \sigma^2_{\hat{p}})$. Σε αυτη τη περίπτωσή θεωρόυμε ότι επαρκεί $Np > 5$ και $N(1-p) > 5$

$Np > 5$ $\quad \hat{p} \sim N\left(p, \frac{p(1-p)}{N}\right)$

1. Υπολογισμός της πιθανότητας το $\hat{p}$ να είναι μικρότερο από μια συγκεκριμένη τιμή.
2. Υπολογισμός της πιθανότητας το $\hat{p}$ να ανοίκει σε ένα διάστημα.

Το **z-score** για τη δειγματική κατανομή της $\hat{p}$ δίνεται ως:

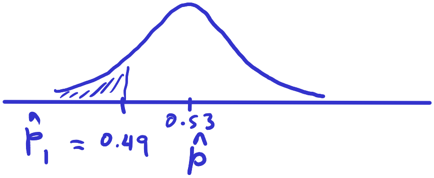
$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$$

Παράδειγμα

p

Ένας υποψήφιος δήμαρχος μιας μεγάλης πόλης ισχυρίζεται ότι έχει τη στήριξη του 53 % των ψηφοφόρων. Εάν δεχτούμε τον ισχυρισμό του ως αλήθηνο ποιά είναι η πιθανότητα σε ένα τυχαίο δείγμα 400 ψηφοφόρων λιγότεροι από 49 % να στηρίζουν τον υποψήφιο;

$p = 0.53 \qquad N = 400$



$$\hat{p} \sim N\left(0.53, \frac{0.53 \cdot 0.47}{400}\right)$$

$$Z = \frac{\hat{p}_1 - p}{\sigma_{\hat{p}}} = \frac{0.49 - 0.53}{\sqrt{\frac{0.53 \cdot 0.47}{400}}} = \frac{-0.04}{0.02495} = -1.602$$

$$P(Z \leq -1.602) = 1 - P(Z \leq 1.602) = 1 - 0.85 = 0.15$$

- Όταν δεν γνωρίζουμε τη τιμή του p δεν μπορούμε να υπολογίσουμε το $\sigma_{\hat{p}}$

Εκτιμήτρια της τυπικής απόκλισης της $\hat{p}$ για μεγάλο δείγμα

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{N}}$$

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

$$\hat{p} = 0.49$$

Διάστημα εμπιστοσύνης της p

Το $(1 - \overset{0.8}{a}) * 100\%$ διάστημα εμπιστοσύνης για την αναλογία p στο πληθυσμό είναι:

$\alpha = 0.2$

$$[\hat{p} - zs_{\hat{p}}, \hat{p} + zs_{\hat{p}}],$$

όπου z το z-score για το οποίο $P(Z \leq z) = 1 - a/2. = 1 - 0.1 = 0.9 \rightarrow z = 1.28$

$$[\hat{p} - 1.28\, s_{\hat{p}}, \hat{p} + 1.28\, s_{\hat{p}}]$$

Τότε

$$P(p \in [\hat{p} - zs_{\hat{p}}, \hat{p} + zs_{\hat{p}}]) = 1 - a$$

Παράδειγμα $\hat{p} = 0.3$

Σε δείγμα 1000 ατομών μιας χώρας το 30% μετρήθηκε να έχει ηλικία μικρότερη από 25 έτη. Βρείτε το 99% διάστημα εμπιστοσύνης για το ποσοστό του πληθυσμού της χώρας με ηλικία μικρότερη από 25 έτη.

$(1-\alpha)\cdot 100\% = 99\%$ $\quad \alpha = 0.01$

$P(Z \leq z) = 1 - \frac{0.01}{2} = 1 - 0.005 = 0.995$

$S_{\hat{p}} = \sqrt{\frac{0.3 \cdot 0.7}{1000}}$

$[0.3 - z \cdot S_{\hat{p}},\ 0.3 + z S_{\hat{p}}]$