ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

06-03-2023

Αριθμός συνδυασμών

Ο αριθμός των δυνατών συνδυασμών της επιλογής K διακεκριμένων στοιχείων από N συνολικά στοιχεία συμβολίζεται ως ${}_{\mathrm{N}}\mathrm{C}_{\mathrm{K}}$ και δίνεταί από τη σχέση:

$$ {}_{\mathrm{N}}\mathrm{C}_{\mathrm{K}} = \binom{\mathrm{N}}{\mathrm{K}} = \frac{\mathrm{N}!}{(\mathrm{N}-\mathrm{K})!\mathrm{K}!} $$

Παράδειγμα

Μια τράπεζα θέλει να προσλάβει 3 νέους ταμίες. Υπάρχουν για τις θέσεις 10 ισάξιοι υποψήφιοι οπότε επιλέγονται 3 στην τύχη. Πόσοι διαφορετικοί συνδιασμοί υπάρχουν;

$$ {}_{10}\mathrm{C}_{3} = \binom{10}{3} = \frac{10!}{(10-3)!3!} = 120 $$

Αριθμός διατάξεων

Ο αριθμός των πιθανών διατάξεων για την επιλογή K διακεκριμένων στοιχείων από N συνολικά στοιχεία συμβολίζεται ως ${}_{\mathrm{N}}\mathrm{P}_{\mathrm{K}}$ και δίνεται απο τη σχέση:

$$ {}_{\mathrm{N}}\mathrm{P}_{\mathrm{K}} = \frac{\mathrm{N}!}{(\mathrm{N}-\mathrm{K})!} = (N-K+1)\cdots N $$

Παράδειγμα

Ένα κουτί περιέχει 10 αριθμημένες μπάλες (0-9). Σχηματίζουμε τριψήφιο αριθμό επιλέγοντας τυχαία 3 μπάλες (χωρίς επανατοποθέτηση). Η πρώτη μπάλα θα αντιστοιχεί στις εκατοντάδες, η δεύτερη στις δεκάδες και η τελευταία στις μονάδες. Πόσοι αριθμοί μπορούν να σχηματισθούν;

$$ {}_{10}\mathrm{P}_{3} = \frac{10!}{(10-3)!} = 10 * 9 * 8 = 720 $$

$\bar{X}$ ως τυχαία μεταβλητή

Δειγματική κατανομή της $\bar{\mathrm{X}}$

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

Γενικά η στατιστική κατανομή οποιοδήποτε στατιστικού του δείγματος καλείτε ως 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:

μέση τιμή του πληθυσμού

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

δειγματική

Παράδειγμα

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

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 $\bar{\mathrm{X}}$ των δειγμάτων με τρία στοιχεία.

$\bar{X}_3$ $\quad \binom{5}{3} = \frac{5!}{3!\,2!} = 4\cdot 5/2 = 10$

Έχουμε συνολικά 10 δείγματα. Γιατί;

$$(5,3,7) \to \bar{\mathrm{X}} = 5,\ (5,3,10) \to \bar{\mathrm{X}} = 6,\ (5,3,6) \to \bar{\mathrm{X}} = 4.67,\ (5,7,10) \to \bar{\mathrm{X}} = 7.33,\ (5,7,6) \to \bar{\mathrm{X}} = 6$$

$$(5,10,6) \to \bar{\mathrm{X}} = 7,\ (3,7,10) \to \bar{\mathrm{X}} = 6.67,\ (3,7,6) \to \bar{\mathrm{X}} = 5.33,\ (3,10,6) \to \bar{\mathrm{X}} = 6.33,\ (7,10,6) \to \bar{\mathrm{X}} = 7.67$$

$\{5, 6, 4.67, \ldots, 7.67\}$

$\mathbb{E}[\bar{X}_3] = \frac{5+6+\cdots+7.67}{10} = \mu_{\bar{X}_3} =$

$\bar{X}_3$

$Var[\bar{X}_3] = \mathbb{E}[\bar{X}_3^2] - (\mathbb{E}[X_3])^2 = \sigma^2_{\bar{X}_3}$

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

.

$$\left(\mu_{\bar{X}} = \mu\right)$$

Εάν ο πληθυσμός είναι πεπερασμένος.

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

δείγματος.

Πλήθος των στοιχείων του πληθυσμού.

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

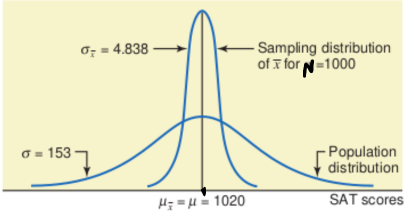
Όταν η παραπάνω συνθήκη δεν ικανοποιείται χρησιμοποιούμε την έκφραση:

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

$X \sim N(\mu, \sigma^2)$ $\bar{X}_N \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$

$X \sim \mathcal{N}(0, 1)$ $\bar{X}_3 \sim \mathcal{N}(0, 1/3)$

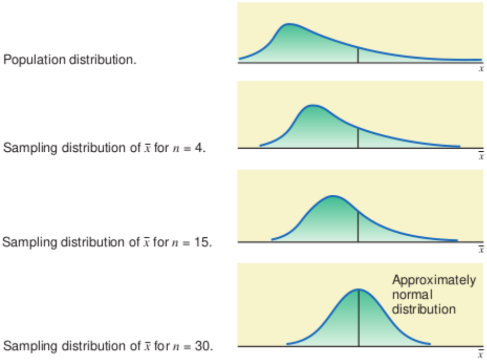
Εάν X ακολουθεί την $\mathcal{N}(\mu, \sigma^2)$ τότε η $\bar{\mathrm{X}}$ ακολουθεί την $\mathcal{N}(\mu_{\bar{\mathrm{X}}}, \sigma^2_{\bar{\mathrm{X}}})$



Δειγματοληψία από Πληθυσμό που δεν ακολουθεί Κανονική κατανομή

Σύμφωνα με το **κεντρικό οριακό θεώρημα**, για μεγάλο μέγεθος του δείγματος, η δειγματική κατανομή της $\bar{\mathrm{X}}$ προσεγγίζει τη κανονική κατανομή $(\mu_{\bar{\mathrm{X}}}, \sigma^2_{\bar{\mathrm{X}}})$ ανεξάρτητα της κατανομής που ακολουθεί η X.

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



$X \sim N(\mu, \sigma^2)$ $\quad$ $P[\bar{X}_N \in [\alpha, b]]$

$X \not\sim N$ $\quad$ $P[\bar{X}_N \in [\alpha, b]]$, $N \geq 30$

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

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

$$Z = \frac{\bar{\mathrm{X}} - \mu}{\sigma_{\bar{\mathrm{X}}}} \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

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

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

$$P[\bar{X}_{16} \in [\overset{z_1}{8}, \overset{z_2}{9}]] = ;$$

$$\mu = 8.4 \qquad \sigma = 1.8$$

$$\bar{X}_{16} \sim N\left(8.4, \frac{1.8^2}{16}\right)$$

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \sqrt{16}\,\frac{\bar{X} - 8.4}{1.8} = 4\,\frac{\bar{X} - 8.4}{1.8}$$

$$z_1 = 4\,\frac{8 - 8.4}{1.8} = 4 \cdot \frac{-0.4}{1.8} = -\frac{8}{9}$$

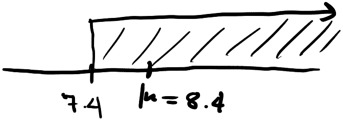
$$z_2 = 4\,\frac{9 - 8.4}{1.8} = 4\,\frac{0.6}{1.8} = \frac{4}{3}$$

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

St.norm.cdf(z1)

2.



$$P(\bar{X} \geq 7.4) = 1 - P(Z \leq z_3)$$

$$z_3 = \frac{7.4 - 8.4}{1.8} \cdot 4 = \frac{-4}{1.8}$$

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 $\hat{\mathrm{p}}$.

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

$N \to N_p$

$\hat{p} \to p$

Όπου:

- $\mathrm{N_p}$ το μέγεθος του πληθυσμού.
- $\mathrm{M_p}$ 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 $\hat{\mathrm{p}}$

- Η μέση τιμή της δειγματικής κατανομής της $\hat{\mathrm{p}}$ συμβολίζεται ως $\mu_{\hat{\mathrm{p}}}$
- Τυπική απόκλιση. Η ~~μέση τιμή~~ της δειγματικής κατανομής της $\hat{\mathrm{p}}$ συμβολίζεται ως $\sigma_{\hat{\mathrm{p}}}$

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

Όταν το δείγμα είναι μικρό συγκριτικά με το πληθυσμό ($\mathrm{N}/\mathrm{N}_\mathrm{p} \leq 0.05$)

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

Όταν η παραπάνω συνθήκη δεν ικανοποιείται χρησιμοποιούμε την έκφραση:

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

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

$$P(\hat{p} \in [\alpha, b]) = ;$$

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

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

$$\mathrm{z} = \frac{\hat{\mathrm{p}} - \mathrm{p}^{\,\text{“}\mu_{\hat{p}}}}{\sigma_{\hat{\mathrm{p}}}}$$

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

$Np > 5$ και $N(1-p) > 5$

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

$$\min(Np, N(1-p)) > 5$$

Παράδειγμα

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

$p = 0.53 \qquad N = 400 \qquad \hat{p} = 0.49$

$$\boxed{\begin{matrix} Np \\ N(1-p) \end{matrix} > 5}$$

$$z_1 = \frac{\hat{p} - p}{\sigma_{\hat{p}}} \qquad \sigma_{\hat{p}} = \sqrt{\frac{0.53 \cdot 0.47}{400}}$$

$$P[Z \leq z_1] = 0.054$$

$$z_1 = \frac{0.49 - 0.53}{\sqrt{\frac{0.53 \cdot 0.47}{400}}} = -\frac{0.04}{0.02495} = -1.602$$

Διαστήματα εμπιστοσύνης για αναλογίες στο πληθυσμό

- 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

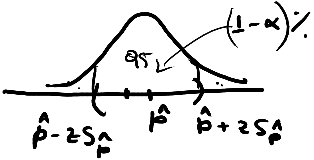
$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{N}}$$

Διάστημα εμπιστοσύνης της p

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

95 % $\alpha = 0.05$

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



όπου z το z-score για το οποίο $P(Z < z) = 1 - a/2$.

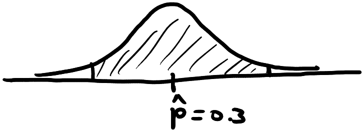
Τότε

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

Παράδειγμα

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.

$N=1000$ $\qquad \hat{p}=0.3$



$\alpha=0.01$

$P(Z\leq z)=0.995$ ← από τον πίνακα πρέπει να βρω το z.

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

$p \in \left[0.3 - z s_{\hat{p}}, 0.3 + z s_{\hat{p}}\right]$ με πιθανότητα 0.99