Глава 5
Узорачка расподела

1. Расподела основног скупа. Узорачка расподела

Расподела основног скупа је расподела вероватноћа случајне променљиве X у основном скупу. Аритметичка средина μ ове променљиве и њена стандардна девијација σ су параметри основног скупа и они су увек њене константе.

Пример 1. Основни скуп чини 5 студената факултета Менаџмент у спорту који слушају предмет Статистика у спорту на другој години основних студија. Коначан њихов успех изражен у поенима, освојеним на предиспитним обавезама и самом испиту, је дат следећим низом:

$$70, \quad 78, \quad 80, \quad 80, \quad 95$$

Нека је X случајна променљива којом се описује овај успех уочене групе студената. Формирајмо одговарајућу табелу расподеле фреквенција и релативних фреквенција:

x	f	**релативна фреквенција** $= p(x)$
70	1	0,20
78	1	0,20
80	2	0,40
95	1	0,20

Није тешко видети да је у нашем случају $N = 5$, $\mu = 80{,}60$ и $\sigma = 8{,}09$. □

Посматрајмо различите узорке из X исте дужине n и израчунавајмо аритметичке средине ових узорака. У општем случају, очигледно, добијене вредности неће бити једнаке. Ако сада под основним скупом сматрамо све могуће узорке дужине n из задатог скупа X, онда са $\overline{X}$ означимо случајну променљиву која описује вредности аритметичких средина ових узорака; ову променљиву зовемо *узорачка аритметичка средина*. Ова променљива, или како кажемо још *статистика*, $\overline{X}$ има своју расподелу фреквенција и релативних фреквенција које се тада називају *узорачка расподела фреквенција* и *узорачка расподела вероватноћа*. Слично овоме могу се посматрати и друге статистике које "уочавају", на пример, медијану, модус, или стандардну девијацију узорка.

Пример 2. Посматрајмо претходни пример и означимо наше студенте са A, B, C, D и E. Нека је

$$X(A) = 70, \quad X(B) = 78, \quad X(C) = 80, \quad X(D) = 80, \quad X(E) = 95$$

Уочимо све узорке дужине 3 датог основног скупа и одговарајућим табелама прикажимо вредности статистике $\overline{X}$

узорак	**резултат у узорку**	$\overline{x}$
ABC	70 78 80	76,00
ABD	70 78 80	76,00
ABE	70 78 95	81,00
ACD	70 80 80	76,67
ACE	70 80 95	81,67
ADE	70 80 95	81,67
BCD	78 80 80	79,33
BCE	78 80 95	84,33
BDE	78 80 95	84,33
CDE	80 80 95	85,00

и њену расподелу фреквенција и расподелу вероватноћа (расподела релативних фреквенција)

$\overline{x}$	f	$p(x)$
76,00	2	0,20
76,67	1	0,10
79,33	1	0,10
81,00	1	0,10
81,67	2	0,20
84,33	2	0,20
85,00	1	0,10

Случајна (*узорачка*) грешка је разлика између вредности статистике узорка и вредности параметра посматраног основног скупа. За аритметичку средину је

$$\textbf{случајна грешка} = \overline{x} - \mu$$

Грешке које настају приликом прикупљања и бележења података и приликом њиховог уношења у табеле се називају *неслучајним* (*систематским*) грешкама.

Пример 3. У претходном примеру имамо да је

$$\mu = \frac{70 + 78 + 80 + 80 + 95}{5} = 80{,}60.$$

За узорак ACE је

$$\overline{x} = \frac{70 + 80 + 95}{3} = 81{,}67.$$

Случајна грешка је тада

$$\overline{x} - \mu = 81{,}67 - 80{,}60 = 1{,}07.$$

Ако је грешком у узорку ACE за студента E регистрован број поена 82, тада је

$$\overline{x} = \frac{70 + 82 + 95}{3} = 82{,}33,$$

па је

$$\overline{x} - \mu = 82{,}33 - 80{,}60 = 1{,}73.$$

Дакле, овде је

случајна грешка = 1,07
неслучајна грешка = 0,66

2. Аритметичка средина и стандардна девијација статистике $\overline{X}$

Стандардна девијација статистике $\overline{X}$ се зове и стандардном грешком статистике $\overline{X}$.

Аритметичка средина и стандардна грешка добијена на основу узорачке расподеле промењиве $\overline{X}$ зову аритметичка средина и стандардна грешка статистике $\overline{X}$ и означавају се са $\mu_{\overline{X}}$ и $\sigma_{\overline{X}}$. На пример, у последњем примеру је

$$\mu_{\overline{X}} = 80{,}60, \quad \sigma_{\overline{X}} = 3{,}30$$

Аритметичка средина статистике $\overline{X}$ је увек једнака аритметичкој средини основног скупа:

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

Аритметичка средина статистике $\overline{X}$ се зове и оцена аритметичке средине основног скупа μ.

Када је очекивана вредност (или средња вредност) статистике узорка једнака вредности одговарајућег параметра скупа, за статистику узорка се каже да представља непристрасну оцену. Како је $\mu_{\overline{X}} = \mu$, то је $\overline{X}$ непристрасна оцена за μ.

Стандардна грешка $\sigma_{\overline{X}}$ статистике $\overline{X}$ није једнака стандардној девијацији σ основног скупа (осим у случају када је $n = 1$). За израчунавање стандардне девијације статистике $\sigma_{\overline{X}}$ користимо формулу

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

али у случају или ако се из коначног скупа бирају узорци са понављањем или ако се из бесконачног (веома великог по броју елемената) основног

скупа узимају узорци без понављања, тј. ако је величина узорка мала у односу на основни скуп. Узорак је мали ако важи да је

$$\frac{n}{N} \leq 0{,}05.$$

Ако узорак није мали, онда се користи формула

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Вредност $\sqrt{\frac{N-n}{N-1}}$ се зове поправни фактор за коначне основне скупове.

Приметимо да је увек $\sigma_{\overline{X}} < \sigma$, ако је $n < 1$. Такође, приметимо да $\sigma_{\overline{X}}$ када дужина узорка n расте.

Стандардна грешка $\sigma_{\overline{X}}$ представља просек квадрата одступања аритметичких средина узорка од аритметичке средине основног скупа μ. Аритметичка средина статистике узорка $\overline{X}$ конзистентна оцена аритметичке средине основног скупа μ.

Пример 4. Просечна зарада по сату у предузећу од 5.000 запослених износи 27,50 долара са стандардном девијацијом од 3,70 долара. Нека је $\overline{X}$ просечна зарада по сату случајног узорка запослених изабраних из тог предузећа. Израчунајте аритметичку средину и стандардну грешку статистике $\overline{X}$ за узорак величине: (а) 30; (б) 75; (в) 200.

(а) Очигледно, да је

$$\mu_{\overline{X}} = \mu = 27{,}50$$

Како је

$$n = 30, \quad N = 5000, \quad \frac{n}{N} = \frac{30}{5000} = 0{,}006 \leq 0{,}05,$$

то је

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3{,}70}{\sqrt{30}} = 0{,}676.$$

(б) Очигледно, да је

$$\mu_{\overline{X}} = \mu = 27{,}50$$

Како је

$$n = 75, \quad N = 5000, \quad \frac{n}{N} = \frac{75}{5000} = 0{,}015 \leq 0{,}05,$$

то је

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3{,}70}{\sqrt{75}} = 0{,}427.$$

(в) Очигледно, да је

$$\mu_{\overline{X}} = \mu = 27{,}50$$

Како је

$$n = 200, \quad N = 5000, \quad \frac{n}{N} = \frac{200}{5000} = 0{,}04 \leq 0{,}05,$$

то је

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3{,}70}{\sqrt{200}} = 0{,}262.$$

3. Облик узорачке расподеле статистике $\overline{X}$

Можемо, очигледно, уочити два случаја, када

1) основни скуп има нормалну расподелу;

2) Основни скуп нема нормалну расподелу

Размотримо ова два случаја.

Претпоставимо, прво, да су узорци из основног скупа са нормалном расподелом. Долази се до следећих закључака:

1) $\mu_{\overline{X}} = \mu$.

2) $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$, ако је $\frac{n}{N} \leq 0{,}05$.

3) Облик узорачке расподеле статистике $\overline{X}$ је нормалан, без обзира на величину узорка

% слика

Претпоставимо, сада, да су узорци из основног скупа који нема нормалну расподелу. Из централне граничне теореме тада следи да аритметичке средине $\overline{X}$ великих узорака имају приближно нормалну расподелу са аритметичком средином $\mu_{\overline{X}} = \mu$ и стандардном девијацијом $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$, ако је при томе $\frac{n}{N} \leq 0{,}05$. Иначе узорци су велики ако је $n \geq 30$.

% слика

4. Примена узорачке расподеле $\overline{X}$

Из централне граничне теореме за велике узорке имамо следеће правило:

Ако изаберемео све велике узорке исте величине из једног основног скупа и израчунамо аритметичке средине ових узорака закључујемо да се приближно

68,26 % ових средина налази у интервалу $[\mu - 1 \cdot \sigma_{\overline{X}}, \mu + 1 \cdot \sigma_{\overline{X}}]$
95,44 % ових средина налази у интервалу $[\mu - 2 \cdot \sigma_{\overline{X}}, \mu + 2 \cdot \sigma_{\overline{X}}]$
99,74 % ових средина налази у интервалу $[\mu - 3 \cdot \sigma_{\overline{X}}, \mu + 3 \cdot \sigma_{\overline{X}}]$

Пример 5. Претпостављамо да тежина свих паковања једне врсте кекса има нормалну расподелу са аритметичком средином 32 унце и стандардном девијацијом 3 унце. Одреди вероватноћу да просечна тежина $\overline{X}$ у случајном узорку од 20 паковања буде у интервалу између 31,8 и 31,9 унци.

Овде је $n < 30$, али је основни скуп са нормалном расподелом, па је

$$\mu_{\overline{X}} = \mu = 32 \text{ унце}, \quad \sigma_{\overline{X}} = \frac{0{,}3}{\sqrt{20}} = 0{,}06708204$$

Треба одредити вредност $P(31{,}8 < \overline{X} < 31{,}9)$. Уочимо стандардизовану променљиву

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$

и њене вредности за

$$\overline{x} = 31{,}8: \qquad Z = \frac{31{,}8 - 32}{0{,}06708204} = -2{,}98$$
$$\overline{x} = 31{,}9: \qquad Z = \frac{31{,}9 - 32}{0{,}06708204} = -1{,}49$$

Дакле, имамо да је

$$P(31{,}8 < \overline{X} < 31{,}9) = P(-2{,}98 < Z < -1{,}49) =$$
$$= P(Z < -1{,}49) - P(Z < -2{,}98) = 0{,}681 - 0{,}0014 = 0{,}0667. \square$$

5. Пропорција основног скупа и узорка

Пропорција основног скупа и пропорција узорка обележавају се са p и са $\hat{p}$:

$$p = \frac{N_1}{N}, \qquad \hat{p} = \frac{X}{n},$$

где је

N — укупан број елемената у основном скупу;
n — укупан број елемената у узорку;
N_1 — број елемената у основном скупу са одређеном карактеристиком;
X — број елемената у узорку са одређеном карактеристиком

Пример 6. У граду живи 789.654 породице, 563.282 има кућу у свом власништву. Изабран је случајни узорак од 240 породица, од којих 158 поседује кућу у свом власништву. Одредити пропорције породица које поседују кућу у основном скупу и у случајном узорку.

Имамо да је

$$p = \frac{N_1}{N} = \frac{563.282}{789.654} = 0{,}71$$

и

$$\hat{p} = \frac{X}{n} = \frac{158}{240} = 0{,}66$$

Дакле, случајна грешка је

$$\hat{p} - p = -0{,}05. \square$$

Расподела вероватноћа пропорције узорка $\hat{p}$ се зове узорачком расподелом пропорције и представља скуп парова вредности које може узети статистика $\hat{p}$ и одговорајућих вероватноћа.

Аритметичка средина пропорције узорка се обележава са $\mu_{\hat{p}}$ и једнака је пропорцији основног скупа, p, односно

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

Пропорција узорка $\hat{p}$, назива се оцена пропорције основног скупа, p. Из горе наведеног је јасно да је статистика $\hat{p}$ непристрасна оцена.

Стандардна девијација (стандардна грешка) пропорције узорка се означава са $\sigma_{\hat{p}}$ и израчунава се или по формули

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}},$$

ако је $n/N \leq 0{,}05$, где је p — пропорција основног скупа, а $q = 1 - p$, или по формули

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \sqrt{\frac{N-n}{N-1}},$$

ако је $n/N \leq 0{,}05$, где је $\sqrt{\frac{N-n}{N-1}}$ поправни фактор за коначне скупове.

Из централне граничне теореме следи да је узорачка расподела статистике $\hat{p}$ приближно нормална, када се ради о великим узорцима. У пракси узорак је велик ако је испуњен услов

$$np > 5 \quad \text{и} \quad nq > 5.$$

Пример 7. Према једном истраживању 50% американаца је задовољно својим послом. Претпоставимо да се овај резултат односи и на основни скуп Американаца. Нека је $\hat{p}$ пропорција Американаца који су задовољни својим послом у случајном узорку од 1000 Американаца. Израчунајте аритметичку средину и стандардну грешку статистике $\hat{p}$ и описати облик узорачке расподеле

Добијамо да је

$$\mu_{\hat{p}} = p = 0{,}50, \quad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = 0{,}0158. \square$$

Следећим примером је илустрована примена узорачке расподеле статистике $\hat{p}$.

Пример 8. У једној анкети 38% Американаца је изјавило да је веома задовољно својим животом. Нека је $\hat{p}$ пропорција Американаца који су веома задовољни својим животом у случајном узорку од 1000 Американаца. Израчунајте вероватноћу да се вредност пропорције $\hat{p}$ нађе у интервалу између 0,40 и 0,42.

Како је

$$n = 1000, \quad p = 0{,}38, \quad q = 1 - p = 0{,}62,$$

то је

$$\mu_{\hat{p}} = p = 0{,}38, \quad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = 0{,}01534927.$$

Како је

$$np = 1000 \cdot 0{,}38 = 380 > 5 \quad \text{и} \quad nq = 620 > 5,$$

то $\hat{p}$ има нормалну расподелу. Уочимо стандардну нормалну расподелу

$$Z = \frac{\hat{P} - p}{\sigma_{\hat{p}}}.$$

Одавде добијамо

$$\hat{p} = 0{,}40, \quad z = \frac{0{,}40 - 0{,}38}{0{,}01534927} = 1{,}30$$
$$\hat{p} = 0{,}42, \quad z = \frac{0{,}42 - 0{,}38}{0{,}01534927} = 2{,}61$$

па је

$$\begin{aligned} P(0{,}40 < \hat{p} < 0{,}42) &= P(1{,}30 < z < 2{,}61) = P(z < 2{,}61) - P(z < 1{,}30) = \\ &= 0{,}9955 - 0{,}9032 = 0{,}0923. \square \end{aligned}$$