OBSAH

2 OPIS JEDNOROZMERNÝCH ŠTATISTICKÝCH SÚBOROV

RIEŠENÉ PRÍKLADY

🕮 Príklad 2.1

Telekomunikačná spoločnosť sledovala dĺžku náhodne vybraných telefonických hovorov v minútach:

3, 2, 2, 4, 1, 2, 1, 0, 2, 4, 1, 1, 3, 2, 0, 3, 5, 2, 1, 2

Budeme riešiť nasledujúce úlohy:

a) Vypočítame priemernú dĺžku hovorov z netriedených hodnôt.
b) Vypočítame priemernú dĺžku hovorov z triedených hodnôt.
c) Určíme modus.
d) Určíme medián z netriedených hodnôt.
e) Určíme medián z triedených hodnôt.

☞ Riešenie

a) Vypočítame priemernú dĺžku hovorov z netriedených hodnôt.

Pri výpočte použijeme jednoduchý aritmetický priemer. Podľa vzťahu (2.1) dostaneme:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{3+2+2+4+1+2+1+\ldots}{20} = \frac{41}{20} = 2{,}05$$

b) Vypočítame priemernú dĺžku hovorov z triedených hodnôt.

Volajúcich roztriedime podľa dĺžky telefonického hovoru do tab. 2.1.

Výpočtová tabuľka Tabuľka 2.1

x_i	n_i	$x_i n_i$	N_i
0	2	0	2
1	5	5	7
2	7	14	14
3	3	9	17
4	2	8	19

pokračovanie tab. 2.1

x_i	n_i	$x_i n_i$	N_i
5	1	5	20
Spolu	**20**	**41**	×

Pre triedený štatistický znak pri výpočte použijeme vážený aritmetický priemer. Podľa vzťahu (2.2) získame:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{n_1 + n_2 + \ldots + n_k} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} = \frac{41}{20} = 2{,}05$$

⇨ *Interpretácia*

Podľa úlohy a) aj b) priemerná dĺžka telefonického hovoru bola 2,05 minúty.

c) Určíme modus.

Pri určovaní modusu budeme vychádzať z tab. 2.1. Najväčšiu absolútnu početnosť (n_i = 7) má hodnota znaku 2, čiže táto hodnota je modus ($\hat{x}$ = 2).

⇨ *Interpretácia*

Modus dĺžky telefonického hovoru je dve minúty, teda najviac telefonických hovorov malo dĺžku trvania dve minúty. Ide o typickú hodnotu trvania telefonických hovorov v empirickom súbore.

d) Určíme medián z netriedených hodnôt.

Netriedené údaje o dĺžke telefonických hovorov usporiadame podľa veľkosti:

0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5

Vzhľadom na to, že ide o párny počet pozorovaní, prostredná hodnota nie je jednoznačne určená a určí sa ako aritmetický priemer dvoch prostredných usporiadaných hodnôt znaku. V našom prípade 10. aj 11. jednotka nadobúdajú rovnakú hodnotu znaku (2), preto túto hodnotu môžeme považovať za medián ($\tilde{x}$ = 2).

e) Určíme medián z triedených hodnôt.

Z tab. 2.1 využijeme stĺpec kumulatívnych absolútnych početností (N_i). Polovica rozsahu súboru je 10, čiže medián zodpovedá tej obmene znaku (triede), pri ktorej je kumulatívna absolútna početnosť po prvýkrát vyššia ako polovica rozsahu súboru. V našom prípade ide o obmenu znaku 2, platí $\tilde{x}$ = 2.

⇨ *Interpretácia*

Podľa úlohy d) aj e) je medián trvania telefonického hovoru dve minúty, teda polovica volajúcich má dĺžku telefonického hovoru dve minúty alebo menej a polovica dve minúty alebo viac.

🕮 Príklad 2.2

K dispozícii máme údaje o vekovej štruktúre skupiny študentov nastupujúcich do 1. ročníka štúdia na Ekonomickej univerzite v Bratislave v študijnom programe *hospodárska informatika* za rok 2015/16. Údaje sú uvedené v tab. 2.2.

Veková štruktúra novoprijatých študentov Tabuľka 2.2

Vek v ukončených rokoch	Počet študentov
18	12
19	54
22	8
21	6
Spolu	**80**

Vypočítame priemerný vek študentov nastupujúcich na štúdium do 1. ročníka Ekonomickej univerzity v Bratislave v študijnom programe *hospodárska informatika* za akademický rok 2015/16. Pri výpočte ako váhy použijeme:

a) absolútne triedne početnosti,
b) relatívne triedne početnosti.

☞ Riešenie

Vek je diskrétny kvantitatívny znak, ktorý v tomto prípade nadobúda malý počet obmien. K dispozícii máme triedené údaje, takže pri výpočte použijeme vážený aritmetický priemer. Pomocné výpočty obsahuje tab. 2.3.

Výpočtová tabuľka Tabuľka 2.3

x_i	n_i	$x_i n_i$	f_i	$x_i f_i$
18	12	216	0,150	2,700
19	54	1026	0,675	12,825
22	8	176	0,100	2,200
21	6	126	0,075	1,575
Spolu	**80**	**1544**	**1,000**	**19,300**

a) Pri výpočte použijeme absolútne triedne početnosti a podľa vzťahu (2.2) dostaneme:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{1544}{80} = 19{,}3$$

b) Pri výpočte použijeme relatívne triedne početnosti a podľa vzťahu (2.3) získame:

$$\overline{x} = \sum_{i=1}^{k} x_i f_i = 19{,}3$$

⇨ *Interpretácia*

Priemerný vek študentov nastupujúcich na denné štúdium do 1. ročníka štúdia Ekonomickej univerzitv Bratislave v študijnom programe hospodárska informatika za akademický rok 2015/2016 bol 19,3 roka.

🕮 Príklad 2.3

Pri čerpaní objemu nádrže sa používajú súčasne tri čerpadlá. Čerpadlo A vyčerpá 1 hl za 40 sekúnd, čerpadlo B za 10 sekúnd a čerpadlo C za 5 sekúnd.

a) Vypočítame priemerný čas potrebný na vyčerpanie 1 hl objemu nádrže na jedno čerpadlo, ak pracujú všetky tri súčasne.
b) Vypočítame priemerný čas potrebný na vyčerpanie 1 hl objemu nádrže za predpokladu, že pri čerpaní pracujú súčasne dve čerpadlá typu A, tri čerpadlá typu C a štyri čerpadlá typu C.

☞ Riešenie

a) Vypočítame priemerný čas potrebný na vyčerpanie 1 hl objemu nádrže na jedno čerpadlo, ak pracujú všetky tri súčasne.

Medzi objemom nádrže a časom potrebným na jej vyčerpanie je nepriamy vzťah. Čím dlhší je čas čerpania, tým menší objem nádrže vyčerpáme a naopak.

Čas, za ktorý vyčerpá 1 hl konkrétne čerpadlo, označíme ako x_i, pričom ide o pomerné číslo s meracou jednotkou (s/hl). Ak čerpajú všetky tri čerpadlá súčasne, ich čas čerpania môžeme považovať za konštantný, takže poznáme váhy z čitateľa, čiže $n = 3$ (váhy sú dané nepriamo). Priemerný čas potrebný na vyčerpanie 1 hl nádrže pri súčasnom pracovaní všetkých troch čerpadiel vypočítame následne pomocou jednoduchého harmonického priemeru. Podľa vzťahu (2.12) dostaneme:

$$\overline{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{3}{\frac{1}{40} + \frac{1}{10} + \frac{1}{5}} = 9,23$$

Výsledok overíme úvahou:

Priemerný čas potrebný na vyčerpanie 1hl objemu nádrže možno priamo vypočítať ako podiel celkového času potrebného na vyčerpanie nádrže a objemu nádrže. Za jednotku času, napr. za jednu sekundu je úhrnný objem vyčerpanej nádrže 0,325 hl (1/40+1/10+1/5), t. j. priemerný čas na vyčerpanie 1 hl, ak pracujú súčasne tri čerpadlá, je 9,23 sekundy (3/0,325).

⇨ *Interpretácia*

Priemerný čas potrebný na vyčerpanie 1 hl objemu nádrže bude 9,23 sekundy.

b) Vypočítame priemerný čas potrebný na vyčerpanie 1 hl objemu nádrže za predpokladu, že pri čerpaní pracujú súčasne dve čerpadlá typu A, tri čerpadlá typu C a štyri čerpadlá typu C.

V tomto prípade došlo pri čerpaní 1 hl nádrže k zvýšeniu počtu jednotlivých druhov čerpadiel, čiže je nutné vážiť čas čerpania počtom čerpadiel, takže využijeme vážený tvar harmonického priemeru podľa vzťahu (2.14):

$$\overline{x}_h = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{9}{\frac{2}{40} + \frac{3}{10} + \frac{4}{5}} = \frac{9}{\frac{46}{40}} = 7{,}83$$

⇨ *Interpretácia*

Priemerný čas potrebný na vyčerpanie 1 hl objemu nádrže bude 7,83 sekundy (priemerný čas je kratší vzhľadom na väčší počet výkonnejších čerpadiel).

🕮 Príklad 2.4

O priemernej mesačnej mzde a vyplatenom mzdovom fonde v troch rôznych prevádzkach máme údaje v tab. 2.4.

Prevádzky podľa mzdového fondu a priemernej mzdy Tabuľka 2.4

Prevádzka	Priemerná mzda (v € na pracovníka)	Mesačný mzdový fond v €
1	550	9 900
2	600	12 000
3	680	10 200
Spolu	×	**32 100**

Vypočítajme priemernú mesačnú mzdu pracovníka v celom podniku.

☞ Riešenie

Priemerná mzda jedného pracovníka v prevádzke je pomerné číslo. Priemer z pomerných čísel sa určí ako vážený harmonický priemer, keď poznáme iba nepriame váhy – čitateľa pomerných čísel. Ako váhy v našom prípade použijeme celkový mesačný mzdový fond (podiel mzdového fondu a priemernej mzdy vyjadruje počet pracovníkov). Na základe vzťahu (2.14) získame:

$$\overline{x}_h = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{9\,900 + 12\,000 + 10\,200}{\frac{9\,900}{550} + \frac{12\,000}{600} + \frac{10\,200}{680}} = \frac{32\,100}{53} = 605{,}66$$

5 ŠTATISTICKÉ SKÚMANIE ZÁVISLOSTI

5.1 Regresná a korelačná analýza

RIEŠENÉ PRÍKLADY

🕮 Príklad 5.1

Predajca jazdených áut chce preskúmať vzťah medzi počtom najazdených kilometrov a ponukovou cenou jazdených automobilov Škoda Octavia s objemom motora 1,6 litra. Výberové údaje sú uvedené v tabuľke.

Automobil	1	2	3	4	5	6	7	8	9	10	11
Ponuková cena (v tis. €)	4,4	3,9	3,5	5,6	6,0	6,7	7,6	8,1	8,3	9,4	10,2
Najazdené kilometre (v tis. km)	196	182	207	169	136	151	113	89	158	98	74

a) Odhadneme regresnú priamku charakterizujúcu závislosť ponukovej ceny automobilov Škoda Octavia od počtu najazdených kilometrov.
b) Na hladine významnosti 0,05 overíme štatistickú významnosť regresného modelu.
c) Na hladine významnosti 0,05 overíme štatistickú významnosť regresného koeficienta.
d) So spoľahlivosťou 0,95 odhadneme priemernú zmenu ponukovej ceny automobilov Škoda Octavia spôsobenú nárastom počtu najazdených kilometrov o 1 000.
e) So spoľahlivosťou 0,95 odhadneme ponukovú cenu automobilov Škoda Octavia, ktoré majú najazdených 100 000 km.
f) Tesnosť skúmanej závislosti budeme kvantifikovať korelačnými charakteristikami.
g) Na hladine významnosti 0,05 overíme štatistickú významnosť koeficienta korelácie.
h) Na hladine významnosti 0,05 overíme, či miera intenzity lineárnej závislosti medzi ponukovou cenou automobilov Škoda Octavia a počtom najazdených kilometrov má hodnotu $-0{,}8$.
i) So spoľahlivosťou 0,95 odhadneme intenzitu lineárnej závislosti medzi ponukovou cenou automobilov Škoda Octavia a počtom najazdených kilometrov.
j) Pokúsime sa nájsť linearizovateľný regresný model, ktorý vyrovnáva empirické údaje lepšie ako lineárny regresný model. Urobíme odhad parametrov takéhoto modelu.

☞ Riešenie

a) Odhadneme regresnú priamku charakterizujúcu závislosť ponukovej ceny automobilov Škoda Octavia od počtu najazdených kilometrov.

Odhadujeme lokujúcu konštantu β_0 a regresný koeficient β_1 lineárneho regresného modelu (5.4) daného vzťahom $y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$. Regresný koeficient odhadneme na základe vzťahu (5.14), podľa ktorého $b_1 = \frac{\operatorname{cov} xy}{s_x^2}$. Vysvetľovanou premennou Y je v našom príklade ponuková cena automobilov Škoda Octavia a vysvetľujúcou premennou je počet najazdených kilometrov. Vypočítame priemerné hodnoty oboch premenných, kovarianciu medzi analyzovanými premennými a rozptyl premennej X. Čiastkové výpočty pre $\overline{x}$, $\overline{y}$, $\operatorname{cov} xy$ a s_x^2 sú realizované v tab. 5.1.

Výpočtová tabuľka Tabuľka 5.1

i	y_i	x_i	$x_i y_i$	x_i^2	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$	$(x_i - \overline{x})^2$
1	4,4	196	862,4	38 416	53,0	–2,3	–121,9	2 809
2	3,9	182	709,8	33 124	39,0	–2,8	–109,2	1 521
3	3,5	207	724,5	42 849	64,0	–3,2	–204,8	4 096
4	5,6	169	946,4	28 561	26,0	–1,1	–28,6	676
5	6,0	136	816,0	18 496	–7,0	–0,7	4,9	49
6	6,7	151	1 011,7	22 801	8,0	0,0	0,0	64
7	7,6	113	858,8	12 769	–30,0	0,9	–27,0	900
8	8,1	89	720,9	7 921	–54,0	1,4	–75,6	2 916
9	8,3	158	1 311,4	24 964	15,0	1,6	24,0	225
10	9,4	98	921,2	9 604	–45,0	2,7	–121,5	2 025
11	10,2	74	754,8	5 476	–69,0	3,5	–241,5	4 761
Σ	**73,7**	**1 573**	**9 637,9**	**244 981**	**0,0**	**0,0**	**–901,2**	**20 042**

Na základe čiastkových výpočtov z tab. 5.1 vypočítame $\overline{x} = 143$ a $\overline{y} = 6{,}7$. Analyzované automobily majú teda v priemere najazdených 143 000 km a ich priemerná ponuková cena je 6 700 eur. Okrem toho vyčíslime priemerné hodnoty:

$$\overline{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} = \frac{9\,637{,}9}{11} = 876{,}173 \qquad \overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n} = \frac{244\,981}{11} = 22\,271$$

Kovarianciu medzi premennými X a Y vypočítame podľa vzťahu (5.15):

$$\operatorname{cov} xy = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \overline{x}) \cdot (y_i - \overline{y})\right] = \frac{1}{11} \cdot (-901{,}20) = -81{,}927$$

alebo podľa vzorca (5.16):

$$\operatorname{cov} xy = \overline{xy} - \overline{x} \cdot \overline{y} = 876{,}173 - 143 \cdot 6{,}7 = -81{,}927$$

Rozptyl premennej *X* vypočítame podľa vzťahu (2.27):

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \frac{1}{11} \cdot 20\,042 = 1\,822$$

alebo podľa vzorca (2.35):

$$s_x^2 = \overline{x^2} - (\overline{x})^2 = 22\,271 - 143^2 = 1\,822$$

Po dosadení do vzťahu (5.14) získame odhad regresného koeficienta:

$$b_1 = \frac{\operatorname{cov} xy}{s_x^2} = \frac{-81{,}927}{1\,822} = -0{,}045$$

Odhad lokujúcej konštanty vypočítame podľa vzťahu (5.17):

$$b_0 = \overline{y} - b_1 \cdot \overline{x} = 6{,}7 - (-0.045) \cdot 143 = 13{,}130$$

Lokujúca konštanta $b_0 = 13{,}130$ je odhadom lokujúcej konštanty β_0 a regresný koeficient $b_1 = -0{,}045$ je odhadom regresného koeficienta β_1. Neznáme parametre β_0 a β_1 sú parametre regresného modelu $y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$, ktorý na základe zadania predpokladáme v základnom súbore. Odhadom uvedeného regresného modelu je vyrovnávajúca priamka $\hat{y}_i = 13{,}130 - 0{,}045 \cdot x_i$.

⇨ *Interpretácia*

Lokujúca konštanta $\boldsymbol{b_0} = 13{,}130$ *: priemerná ponuková cena automobilov Škoda Octavia s nulovým počtom najazdených kilometrov je 13 130 eur.*

Regresný koeficient $\boldsymbol{b_1} = -0{,}045$ *je smernicou vyrovnávajúcej priamky. Z jeho hodnoty sme získali dôležitú informáciu, že ak sa zvýši počet najazdených kilometrov o 1 000 (meracia jednotka premennej X), klesne ponuková cena automobilu Škoda Octavia v priemere o 45 eur (*b_1 *násobíme jednotkou premennej Y, čo je tisíc eur).*

Keďže ***regresný koeficient*** *je záporný, regresná priamka je klesajúca, a teda medzi ponukovou cenou automobilov a počtom najazdených kilometrov je nepriama lineárna závislosť. O nepriamej lineárnej závislosti medzi analyzovanými premennými nás informuje aj záporná hodnota kovariancie (*$\operatorname{cov} xy = -81{,}927$ *).*

Riešenie v SAS Enterprise Guide

Postupujeme v krokoch:

Tasks → *Regression* → *Linear Regression*

Otvorí sa okno, v ktorom môžeme otvárať viaceré záložky. V prvej z nich, teda v záložke *Data*, zadáme do položiek *Dependent variable* a *Explanatory variables* vysvetľovanú a vysvetľujúcu premennú. Po stlačení tlačidla *Run* dostaneme výstup na obr. 5.1. Ten okrem iného obsahuje bodové odhady parametrov regresnej priamky (tabuľka *Parameter Estimates*, stĺpec *Parameter Estimate*). Sú to odhady získané metódou najmenších štvorcov, ktorá bola použitá na odvodenie vzťahov, ktoré sme aplikovali pri výpočte bodového odhadu regresného koeficienta a bodového odhadu lokujúcej konštanty (pozri vzťahy (5.10) až (5.17)).

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	40.52297	40.52297	38.73	0.0002
Error	9	9.41703	1.04634		
Corrected Total	10	49.94000			

Root MSE	1.02291	**R-Square**	0.8114
Dependent Mean	6.70000	**Adj R-Sq**	0.7905
Coeff Var	15.26725		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	13.13008	1.07829	12.18	<.0001
Pocet_km	1	–0.04497	0.00723	–6.22	0.0002

Obr. 5.1 **Základná analýza lineárneho regresného modelu vystihujúceho závislosť ponukovej ceny automobilov Škoda Octavia od počtu najazdených kilometrov (SAS Enterprise Guide)**

Riešenie v Statgraphics Centurion (Statgraphics Plus)

Postupujeme v krokoch:

Relate → *One Factor* → *Simple Regression*

(Relate → *Simple Regression)*

Otvorí sa okno, v ktorom do položiek *Y* a *X* zadáme vysvetľovanú a vysvetľujúcu premennú. Po stlačení tlačidla *OK* získame výstup zobrazený na obr. 5.2.

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Intercept	13,1301	1,07829	12,1768	0,0000
Slope	–0,0449656	0,00722545	–6,22322	0,0002

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	40,523	1	40,523	38,73	0,0002
Residual	9,41703	9	1,04634		
Total (Corr.)	49,94	10			

Correlation Coefficient = –0,900796
R-squared = 81,1433 percent
R-squared (adjusted for d.f.) = 79,0481 percent
Standard Error of Est. = 1,02291
Mean absolute error = 0,701893
Durbin-Watson statistic = 2,07553 (P=0,4471)
Lag 1 residual autocorrelation = –0,0465187

Obr. 5.2 **Základná analýza lineárneho regresného modelu vystihujúceho závislosť ponukovej ceny automobilov Škoda Octavia od počtu najazdených kilometrov (Statgraphics Centurion)**

Tabuľka *Coefficients* zodpovedá tabuľke *Parameter Estimates* zobrazenej na obr. 5.1. Riadok *Intercept* sa vzťahuje na lokujúcu konštantu a riadok *Slope* sa vzťahuje na regresný koeficient. V stĺpci *Least Squares Estimates* sú teda v týchto dvoch riadkoch uvedené: bodový odhad lokujúcej konštanty a bodový odhad regresného koeficienta.

b) Na hladine významnosti 0,05 overíme štatistickú významnosť regresného modelu.

Testujeme pravdivosť nulovej hypotézy H_0: *regresný model nie je štatisticky významný* oproti alternatívnej hypotéze H_1: *regresný model je štatisticky významný.* Na testovanie týchto hypotéz potrebujeme urobiť rozklad variability vysvetľovanej premennej. Celková variabilita ponukovej ceny automobilov Škoda Octavia, ktorá je daná vzťahom $SST = \sum_{i=1}^{11} (y_i - \overline{y})^2$, je vypočítaná v súčtovom riadku piateho stĺpca výpočtovej tab. 5.2. Variabilitu premennej *Y*, ktorú nevieme vysvetliť regresným modelom, vypočítame podľa vzťahu $SSE = \sum_{i=1}^{11} (y_i - \hat{y}_i)^2$. Podľa vyrovnávajúcej regresnej