Statistika

semestrální projekt

Zuzana Foltýnová
324542

U 96 náhodně vybraných studentů VŠE v Praze byly zjištěny následující údaje:
Pohlaví (0 – žena, 1 – muž)
Výška (tělesná výška v cm)
Hmotnost (tělesná hmotnost v kg)
Známka (známka z matematiky v 1. semestru)

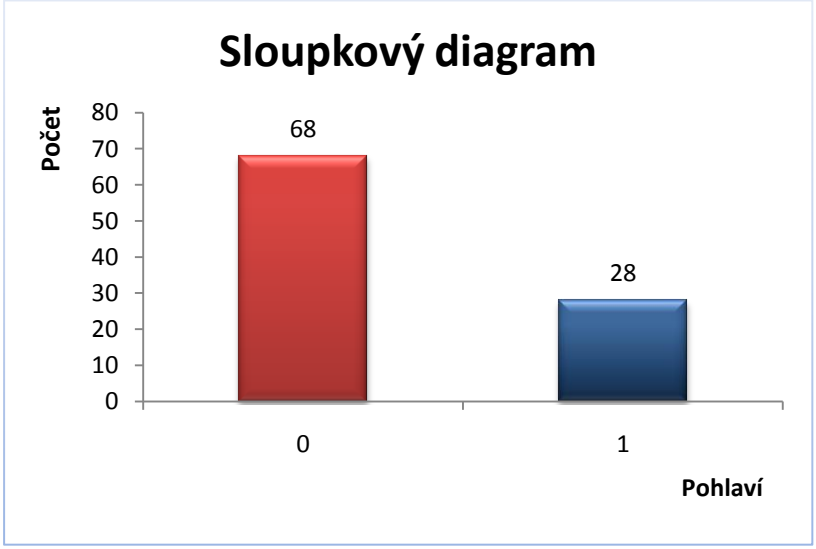
Úkol 1.
Zjistěte absolutní a relativní četnosti proměnných Pohlaví a Známka, přičemž pro proměnnou známka zjistěte též kumulativní absolutní a relativní četnosti. Pro proměnnou Pohlaví vytvořte sloupkový diagram, pro proměnnou Známka polygon četností.
Řešení:

I. Proměnná Pohlaví - absolutní a relativní četnost

Tabulka četností		
$x_{[j]}$	n_j	p_j
1	28	0,29
0	68	0,71

$x_{[j]}$ - proměnná pohlaví
n_j - absolutní četnost varianty $x_{[j]}$
p_j - relativní četnost varianty $x_{[j]}$

II. Proměnná Pohlaví - sloupkový diagram



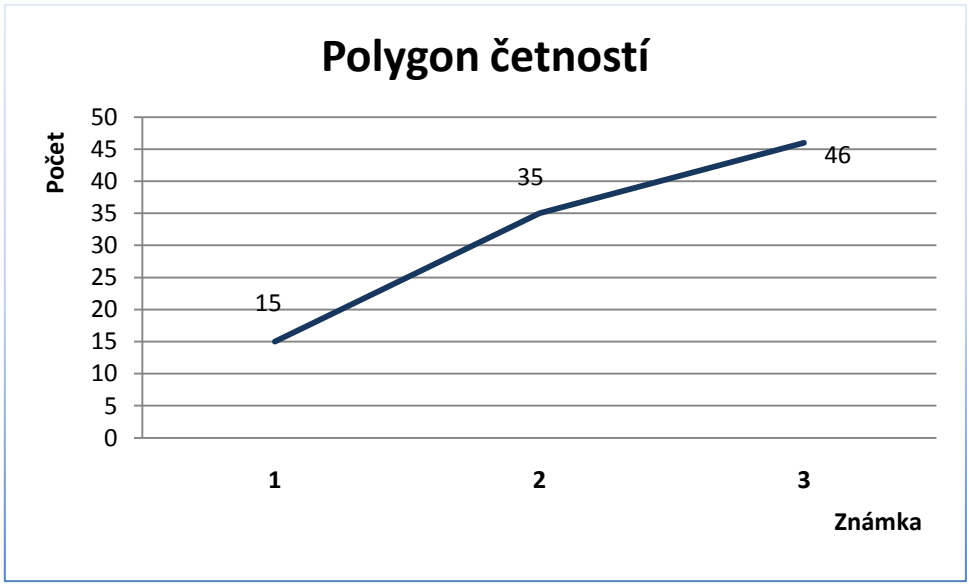
Komentář: Ve zkoumaném datovém souboru jsou ženy (0) zastoupeny v nadpoloviční většině oproti mužům (1).

III. Proměnná Známka - tabulka četností (absolutní a relativní četnost, kumulativní absolutní a relativní četnost)

Tabulka četností				
$x_{[j]}$	n_j	p_j	N_j	F_j
1	15	0,16	15	0,16
2	35	0,36	50	0,52
3	46	0,48	96	1

$x_{[j]}$ - proměnná pohlaví
n_j - absolutní četnost varianty $x_{[j]}$
p_j - relativní četnost varianty $x_{[j]}$
N_j - kumulativní absolutní četnost prvních j variant
F_j - kumulativní relativní četnost prvních j variant

IV. Proměnná Známka - polygon četností



Komentář: Nejčastěji obdržená známka je 3, s 46 výskyty. Nejméně obdržená známka je 1, a to pouze s 15 výskyty.

Úkol 2.
Pro proměnné Pohlaví a Známka sestavte kontingenční tabulky absolutních a relativních četností, sloupcově a řádkově podmíněných relativních četností. Kolik procent žen má z matematiky jedničku? Kolik procent studentů, kteří mají jedničku, jsou muži?
Řešení:

I. Proměnná Pohlaví a Známka - kontingenční tabulka absolutních a relativních četností

Kontingenční tabulka absolutních četností					
-	**y**	**1**	**2**	**3**	$\mathbf{n_{j.}}$
x	$\mathbf{n_{jk}}$				
0		11	27	30	68
1		4	8	16	28
$\mathbf{n_{.k}}$		15	35	46	**n = 96**

x - proměnná pohlaví
y - proměnná známka
$\mathbf{n_{jk}}$ - absolutní četnost dvojice $(x_{[j]},y_{[k]})$
$\mathbf{n_{j.}}$ - absolutní četnost varianty $x_{[j]}$
$\mathbf{n_{.j}}$ - absolutní četnost varianty $y_{[k]}$

Kontingenční tabulka relativních četností					
-	**y**	**1**	**2**	**3**	$\mathbf{p_{j.}}$
x	$\mathbf{p_{jk}}$				
0		0,12	0,28	0,31	0,71
1		0,04	0,08	0,17	0,29
$\mathbf{p_{.k}}$		0,16	0,36	0,48	**1,00**

x - proměnná pohlaví
y - proměnná známka
$\mathbf{p_{jk}}$ - relativní četnost dvojice $(x_{[j]},y_{[k]})$
$\mathbf{p_{j.}}$ - relativní četnost varianty $x_{[j]}$
$\mathbf{p_{.j}}$ - relativní četnost varianty $y_{[k]}$

II. Proměnná Pohlaví a Známka - kontingenční tabulka sloupcově a řádkově podmíněné relativní četnosti

Kontingenční tabulka sloupcově podmíněných relativních četností				
-	**y**	**1**	**2**	**3**
x	$\mathbf{p_{j(k)}}$			
0		0,73	0,77	0,65
1		0,27	0,23	0,35
Σ		**1,00**	**1,00**	**1,00**

x - proměnná pohlaví
y - proměnná známka
$\mathbf{p_{j(k)}}$ - sloupcově podmíněná relativní četnost varianty $x_{[j]}$ za předpokladu $y_{[k]}$

Komentář: 27% studentů, kteří mají jedničku, jsou muži.

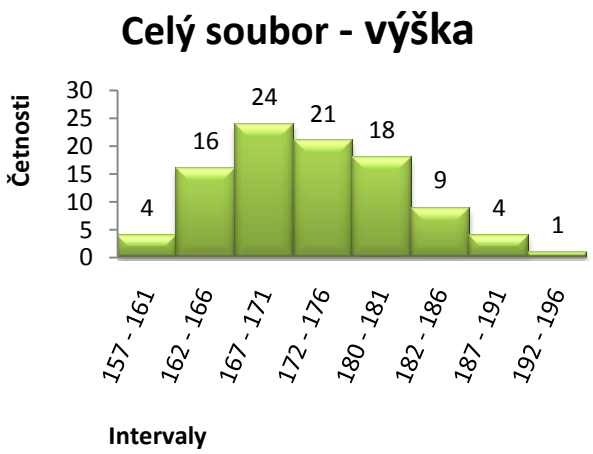
Kontingenční tabulka řádkově podmíněných relativních četností					
-	**y**	**1**	**2**	**3**	**Σ**
x	$\mathbf{n_{jk}}$				
0		0,16	0,40	0,44	**1,00**
1		0,14	0,29	0,57	**1,00**

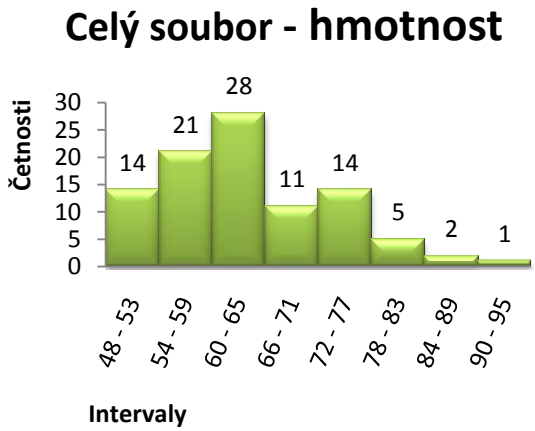
x - proměnná pohlaví
y - proměnná známka
$\mathbf{p_{(j)k}}$ - řádkově podmíněná relativní četnost varianty $y_{[k]}$ za předpokladu $x_{[j]}$

Komentář: 16 % žen má z matematiky jedničku

Úkol 3.
Podle Sturgersova pravidla stanovte optimální počet třídicích intervalů pro proměnné Výška a Hmotnost a nakreslete jejich histogramy, a to
a) pro celý soubor
b) pro ženy
c) pro muže.
Řešení:

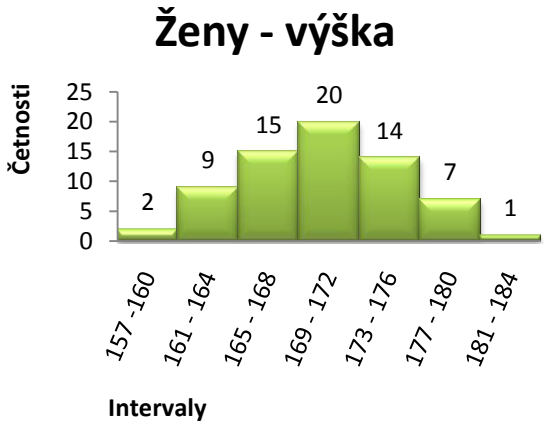
I. Histogramy pro celý soubor
Celkový rozsah souboru je 96, tedy podle Sturgersova pravidla je optimální počet třídicích intervalů r = 1 + 3,3*log (96) ≈ **8**.
Rozsah intervalů (výška) = (max - min) /8 = (193 - 160)/8 ≈ **5**
Rozsah intervalů (hmotnost) = (max - min)/8 = (91 - 48)/8 ≈ **6**

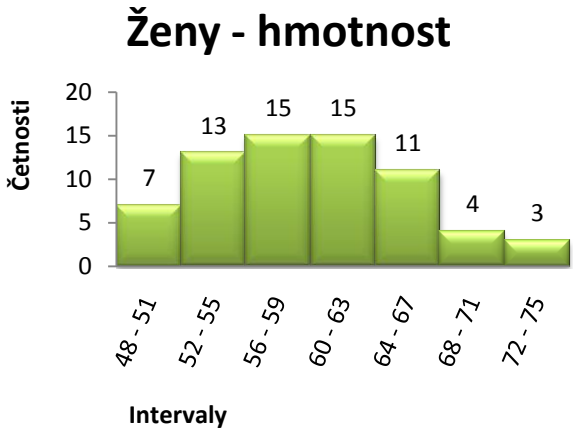




Komentář: Ve zkoumaném souboru převládají osoby s výškou v intervalu 167 - 171 cm či váhou mezi 60 - 65 kg. Naopak nejméně jsou zastoupeny osoby s výškou pod 161 cm a nad 192 cm a s váhou větší než 84 kg.

II. Histogramy pro ženy
Počet žen v souboru je 68. Podle Sturgesova pravidla je optimální počet třídicích intervalů
r = 1 + 3,3*log (68) ≈ **7**
Rozsah intervalů (výška) = (max - min)/7 = (182 - 160)/7 ≈ **4**
Rozsah intervalů (hmotnost) = (max - min)/7 = (76 - 48)/7 ≈ **4**

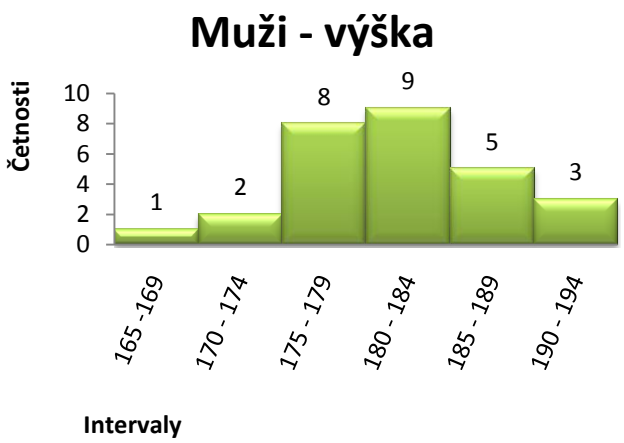


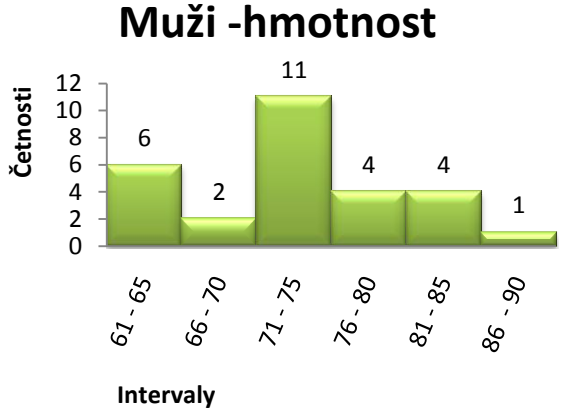


Komentář: Ve zkoumaném souboru převládají ženy s výškou v intervalu 169 - 172 cm či váhou mezi 59 - 63 kg. Naopak nejméně jsou zastoupeny ženy s výškou pod 160 cm a nad 181 cm a s váhou větší než 72 kg.

III. Histogramy pro muže

Počet mužů v souboru je 28. Podle Sturgesova pravidla je optimální počet třídících intervalů
r = 1 + 3,3*log (28) ≈ **6**
Rozsah intervalů (výška) = (max - min)/6 = (193 - 168)/6 ≈ **5**
Rozsah intervalů (hmotnost) = (max - min)/6 = (91 - 61)/6 ≈ **5**





Komentář: Co se týče mužů, nejvíce jsou zastoupeni ti s výškou v intervalu 180 - 184 cm či s váhou mezi 71 - 75 kg. Naopak nejméně jsou zastoupeni muži s výškou pod 169 cm či s váhou větší než 86 kg.

Úkol 4.
Vypočtěte minimum, maximum, medián, průměr, směrodatnou odchylku, šikmost a špičatost proměnných Výška a Hmotnost
a) pro celý soubor
b) pro ženy
c) pro muže.
Řešení:

I. Pro celý soubor

Hodnoty pro celý soubor								
-	**počet**	**minimum**	**maximum**	**medián**	**průměr**	**směrodatná odchylka**	**šikmost**	**špičatost**
výška	96	160	192	172,0	173,19	7,336	0,428	-0,299
hmotnost	96	48	90	62,5	63,38	9,327	0,588	-0,135

II. Pro ženy

Hodnoty pro ženy								
-	**počet**	**minimum**	**maximum**	**medián**	**průměr**	**směrodatná odchylka**	**šikmost**	**špičatost**
výška	68	160	181	170	169,94	5,063	0,017	-0,638
hmotnost	68	48	75	58	59,26	6,392	0,322	-0,446

III. Pro muže

Hodnoty pro muže								
-	**počet**	**minimum**	**maximum**	**medián**	**průměr**	**směrodatná odchylka**	**šikmost**	**špičatost**
výška	28	168	192	181	181,07	5,894	-0,249	0,089
hmotnost	28	61	90	73	73,36	7,670	0,172	-0,533

Komentář: Směrodatná odchylka a průměr mohou být silně ovlivněny extrémními hodnotami. Je-li šikmost kladná, rozložení dat má prodloužený pravý konec, mluvíme o kladně zešikmeném rozložení. V záporně zešikmeném rozložení má naopak rozložení prodloužený levý konec.
Je-li špičatost záporná, jedná se o ploché rozložení dat, je-li kladná, jde o strmé rozložení dat. V případě normálního rozložení je pak špičatost rovna 0.

Úkol 5.
Vypočtěte a interpretujte Pearsonův koeficient korelace proměnných Výška a Hmotnost
a) pro celý soubor
b) pro ženy
c) pro muže.

Řešení:

I. Pearsonův koeficient korelace proměnných Výška a Hmotnost pro celý soubor
r_{12} = **0,739702**

Komentář: Existuje silná kladná korelace mezi proměnnými Výška a Hmotnost pro celý soubor. Můžeme tedy říci, že čím vyšší jsou hodnoty jedné proměnné, tím vyšší jsou hodnoty druhé a naopak.

II. Pearsonův koeficient korelace proměnných Výška a Hmotnost pro ženy
r_{12} = **0,5082377**

Komentář: Existuje středně silná kladná korelace mezi proměnnými Výška a Hmotnost pro ženy. Můžeme tedy říci, že čím vyšší jsou hodnoty jedné proměnné, tím vyšší jsou hodnoty druhé a naopak.

III. Pearsonův koeficient korelace proměnných Výška a Hmotnost pro muže
r_{12} = **0,487723**

Komentář: Existuje středně silná kladná korelace mezi proměnnými Výška a Hmotnost pro muže. Můžeme tedy říci, že čím vyšší jsou hodnoty jedné proměnné, tím vyšší jsou hodnoty druhé a naopak.

Úkol 6.
Najděte rovnici regresní přímky vyjadřující závislost proměnné Hmotnost na proměnné Výška. Jaký je index determinace a co vyjadřuje? Jaká je predikovaná hodnota hmotnosti pro výšku 175 cm? Nalezenou regresní přímku zakreslete do dvourozměrného tečkového diagramu.
Řešení:

I. Rovnice regresní přímky
Rovnice: **hmotnost = -99,502 + 0,9405*výška**

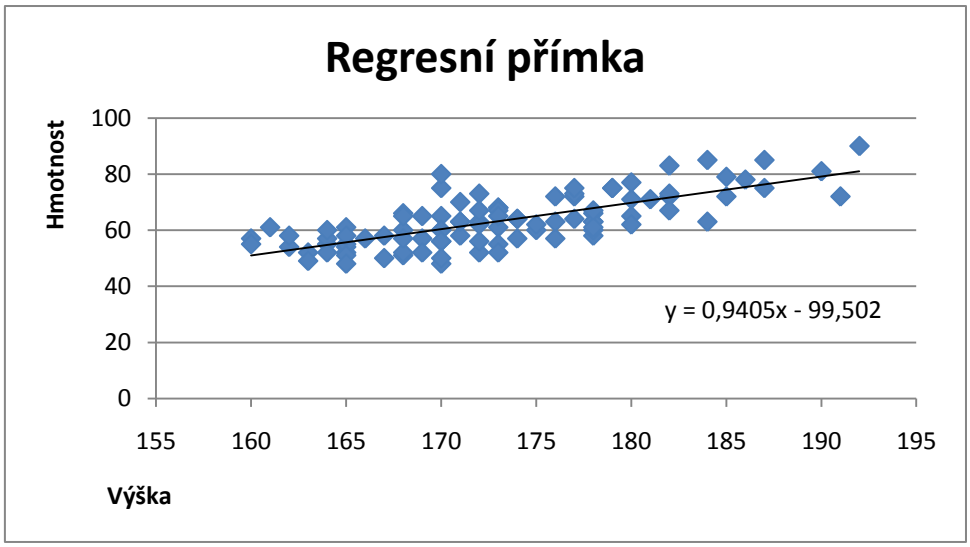
Komentář: Zvětší-li se výška o 1 bod, hmotnost se zvětší v průměru o 0,9405 bodu.

II. Index determinace
Index determinace je **0,5472**

Komentář: Model regresní přímky vysvětluje variabilitu proměnné hmotnost z 54%.

III. Predikovaná hodnota hmotnosti pro výšku 175
Vypočteme, že predikovaná hodnota hmotnosti pro výšku 175 je:
hmotnost = -99,502 + 0,9405*175 = **65,0855**

IV. Dvourozměrný tečkový diagram



Komentář: Z grafu je patrné, že regresní přímka je vhodná na modelování dané závislosti, jelikož body jsou přibližně rozmístěny kolem regresní přímky.