Jurečková Mária

Molnárová Iveta

Štatistika

s Excelom

AOS 2005

Za odbornú a jazykovú stránku zodpovedajú autori.

Jednotlivé kapitoly spracovali:
doc. RNDr. Mária Jurečková, CSc., kap. 1, 4, 5, 7
RNDr. Iveta Molnárová, kap. 2, 3, 6, 8, 9

Recenzenti:
doc. RNDr. Ľudmila Lysá, PhD.
doc. RNDr. František Kôpka, CSc.
RNDr. Eva Drobná, PhD.

Vydala Akadémia ozbrojených síl gen. M.R. Štefánika v Liptovskom Mikuláši
Prvé vydanie, 2005

ISBN 80-8040-257-4

OBSAH

PREDSLOV

Vzrastajúce množstvo a dôležitosť informácii sú výrazné črty dnešnej spoločnosti. Ukazuje sa preto nevyhnutné vedieť informácie spracovať, sprehľadniť a správne interpretovať, aby mohli byť dobrým podkladom pre dôležité rozhodnutia. Matematická štatistika má v tomto procese nezastupiteľné miesto a je významným nástrojom na prípravu pôdy pre kvalifikované rozhodnutia.

Predkladaná učebnica ponúka prehľad základných štatistických metód používaných v rozhodovacom procese v etape štatistického spracovávania, vyhodnocovania a analýzy údajov. Teoretický výklad jednotlivých metód je doplnený návodom na ich realizáciu s využitím EXCELu, ktorý je dostupný každému užívateľovi Microsoft Office. Od čitateľa sa vyžadujú iba základné zručnosti pri práci s EXCELom. Každá kapitola končí aplikáciou vo forme riešeného príkladu a príkladmi, ktoré sú určené čitateľovi na vlastné precvičenie. Príklady sú riešené v českej verzii EXCELu 2000. Na konci učebnice sú výsledky neriešených príkladov a súbory dát, ktoré sa využívajú pri ich riešení.

Učebnica je určená predovšetkým študentom Akadémie ozbrojených síl, ale i ostatných vysokých škôl pre základný kurz matematickej štatistiky. Môže byť ale i vhodnou pomôckou všetkým tým, ktorí majú záujem o skvalitnenie svojho rozhodovania.

Autorky aj touto cestou ďakujú recenzentom doc. RNDr. Ľudmile Lysej, PhD. z Katolíckej univerzity v Ružomberku, doc. RNDr. Františkovi Kôpkovi, CSc. zo Žilinskej univerzity a RNDr. Eve Drobnej, PhD. z Akadémie ozbrojených síl gen. M. R. Štefánika za pripomienky, ktoré pomohli k skvalitneniu obsahu i formy predkladanej učebnice.

1. ZÁKLADNÉ POJMY

1.1 Úvod do štatistiky

Matematická štatistika je vedná disciplína, ktorej úlohou je vyhodnotiť hromadné javy v prírode a spoločnosti na základe poznatkov z teórie pravdepodobnosti. Umožňuje preniknúť k podstate neprehľadného množstva získaných údajov a tým lepšie spoznať zákony objektívnej reality. Na odhalenie zákonitostí skúmaných hromadných javov je potrebné v etape zisťovania informácií zhromaždiť veľké množstvo údajov a nájsť vhodnú formu ich vyjadrenia. Po spracovaní príslušných hodnôt potom nastupuje ich analýza pre praktické využitie. Matematická štatistika je nástroj na zvýšenie kvality rozhodovania v ľubovoľnej sfére ľudskej činnosti.

Dlhé obdobie bola štatistika spojená so zbieraním faktov najmä v demografii a v ekonomike, ale v súčasnej dobe sa štatistické metódy používajú na skúmanie hromadných javov akéhokoľvek druhu. Prispela k tomu i široká prístupnosť počítačov a množstvo softvérových produktov, ktoré majú zabudované i štatistické funkcie. Existujú však i špeciálne softvérové systémy na štatistickú analýzu (STATGRAPHICS, SAS, atď.). Jednotlivé programové produkty sa líšia obsahom procedúr, ale i spôsobom ovládania. Ich výber závisí i od špecifík oblastí, v ktorých chceme matematickú štatistiku aplikovať.

1.2 Štatistický súbor a štatistické znaky

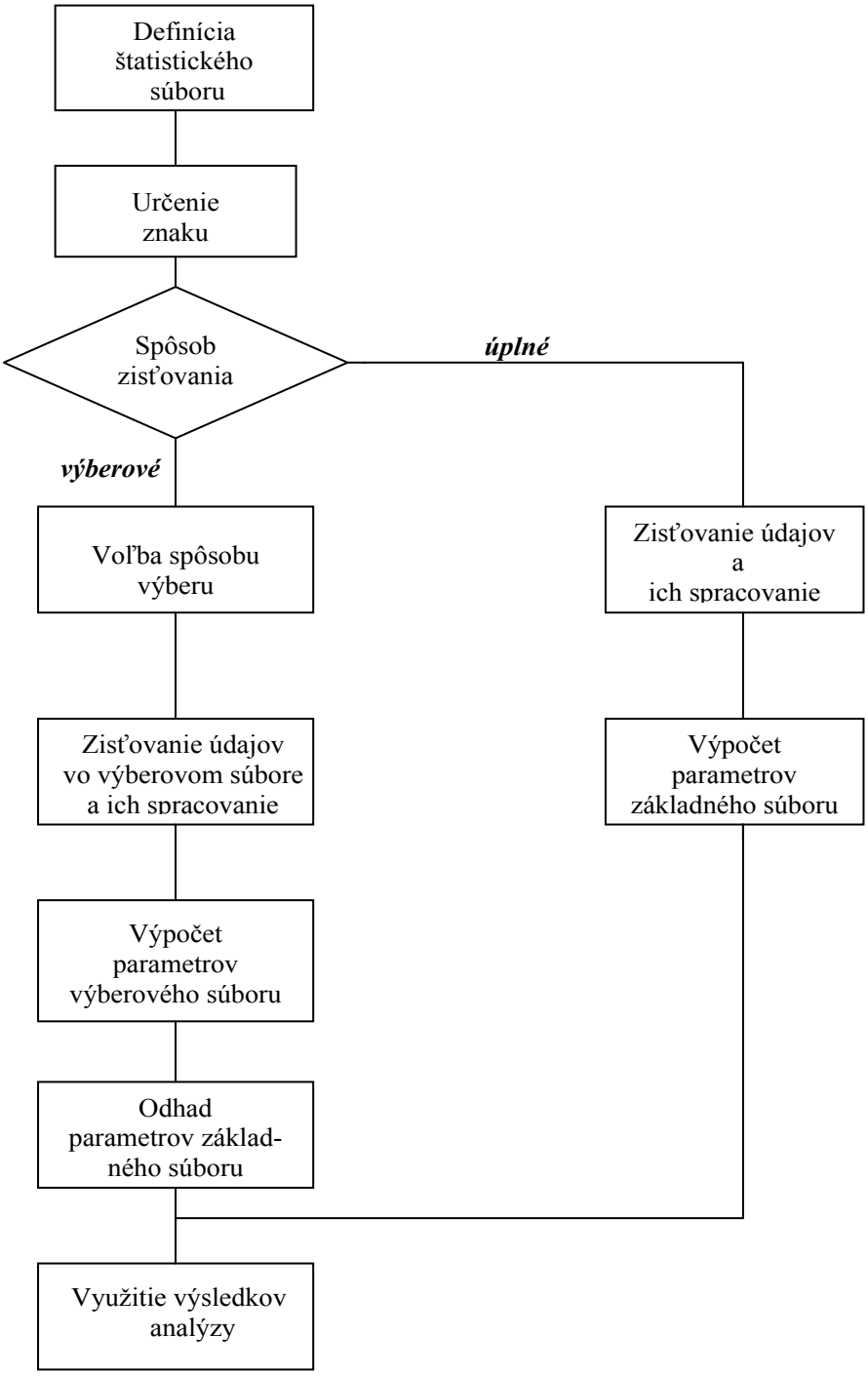
Štatistika sa zaoberá javmi, ktoré nazývame **hromadné javy**. Budeme rozlišovať dva typy hromadných javov. Jeden z nich je založený na veľkom počte opakovaných pozorovaní (meranie, váženie) istej vlastnosti jedného prvku. Konečným cieľom je čo najlepšie priblíženie sa k skutočnej hodnote sledovanej vlastnosti. Svoju pozornosť však sústredíme na druhý typ hromadných javov. V tomto prípade sledujeme na množine, ktorá pozostáva z veľkého počtu prvkov vlastnosť, ktorú má v istej miere každý prvok tejto množiny. Skúmaním hromadného javu cez dostatočný počet individuálnych javov (udalostí) poznávame jeho podstatu, vnútorné vzťahy a súvislosti. Napríklad úroveň absolventov danej školy nemôžeme hodnotiť na

základe jedného absolventa. Hypotézu, že na Slovensku stúpa počet rodín, ktoré majú najviac jedno dieťa, nemožno formulovať bez pozorovania väčšieho počtu rodín. Množinu prvkov, na ktorých sledujeme istý hromadný jav (udalosť), nazývame **štatistický súbor.** Štatistickým súborom môže byť množina osôb, zvierat, inštitúcií a pod., ktorá je definovaná vecne, lokálne a časovo. Prvok štatistického súboru nazývame **štatistická jednotka**. Počet jednotiek v súbore nazývame **rozsah súboru. Štatistický znak** je vlastnosť, ktorú sledujeme na štatistických jednotkách. Skúmané vlastnosti štatistických jednotiek môžu mať rôzny charakter. Delíme ich na:

- **kvantitatívne (kardinálne) znaky**, hodnoty ktorých sú reálne čísla. Kvantitatívne znaky delíme na
  - **diskrétne** (známka z matematiky, počet detí v rodine),
  - **spojité** (telesná výška, príjem).
- **kvalitatívne (kategoriálne) znaky,** ktoré slovne vyjadrujú vlastnosť štatistickej jednotky. Rozlišujeme kvalitatívny znak
  - **nominálny**, ktorého hodnoty nie je možné usporiadať tak, aby hodnota znaku s vyšším poradím vyjadrovala vyšší stupeň vlastnosti, ako hodnota s nižším poradím (farba očí, štátna príslušnosť),
  - **ordinálny**, ktorého hodnoty môžeme prirodzene usporiadať (hodnosť v armáde).

Pri štatistickom spracovávaní údajov často krát nahrádzame kvalitatívny znak kvantitatívnym. Napríklad, kvalitatívne hodnotenie študenta (výborný, veľmi dobrý, dobrý, nevyhovel) môžeme vyjadriť číselným hodnotením (1, 2, 3, 4).

Základný štatistický súbor rozsahu N predstavuje množinu všetkých štatistických jednotiek. V prípade, že nie je možné skúmať základný súbor (z časových, finančných alebo iných dôvodov), vytvárame z neho **výberový súbor** rozsahu $n < N$ a z hodnôt x_i sledovaného znaku X z výberového súboru odhadujeme vlastnosti (parametre) základného súboru. Obr. 1.1. ukazuje všeobecný postup pri riešení štatistických úloh.



Obr. 1.1 Všeobecný postup pri riešení štatistických úloh

1.3 Popisné metódy

Pri spracovaní štatistického súboru je dôležitou úlohou **triedenie súboru**. Pri triedení vytvárame skupiny (triedy) štatistických jednotiek, ktoré majú rovnaké hodnoty skúmanej veličiny. V prípade, že skúmaná veličina nadobúda diskrétne hodnoty, je každá trieda zvyčajne reprezentovaná jednou hodnotou tejto veličiny. Ak skúmaný znak je spojitá veličina, triedu reprezentuje interval jej možných hodnôt. Pri triedení je dôležité dodržať nasledujúce zásady:

- **zásada úplnosti** - každá štatistická jednotka musí byť zaradená do niektorej triedy,
- **zásada jednoznačnosti -** každá štatistická jednotka je zaradená práve do jednej triedy.

V prípade spojitých hodnôt znaku *X*, ktoré reprezentuje číselná os, zásadu jednoznačnosti zabezpečíme vytváraním triednych intervalov typu $\langle a_i, b_i)$, resp. $(a_i, b_i\rangle$, kde a_i je dolná a b_i horná hranica príslušnej triedy. Počet intervalov volíme najčastejšie od 8 – 20, pričom prihliadame na povahu skúmaného znaku a na rozsah štatistického súboru. V prípade, že pracujeme so štatistickým súborom rozsahu *N*, môžeme počet intervalov k určiť podľa predpisu $k = 1 + 3{,}3 \log N$. V niektorých prípadoch zvolíme stredy príslušných intervalov za reprezentantov hodnôt znaku v jednotlivých intervaloch.

Početnosť, ktorá je priradená jednotlivým triedam, môže byť:

- **absolútna** n_i - udáva počet jednotiek súboru, u ktorých sa vyskytuje hodnota x_i skúmaného znaku *X*,
- **relatívna** $\frac{n_i}{N}$ - udáva podiel absolútnej početnosti a rozsahu súboru,
- **kumulatívna absolútna** F_i - je súčtom absolútnych početností všetkých *j*- tých tried , kde $j \le i$,

- **kumulatívna relatívna** $\frac{F_i}{N}$ - udáva podiel kumulatívnej absolútnej početnosti a rozsahu súboru.

Triedenie pomocou zvoleného softvéru na počítači vyžaduje vloženie údajov do pamäte počítača a voľbu vhodnej štatistickej procedúry. Dobrou názornou predstavou o rozdelení početností je grafická prezentácia výsledkov. Graficky môžeme získané údaje znázorniť histogramom. Jednotlivé početnosti sú zobrazené obdĺžnikmi, ktorých základne sú rovné šírke intervalu a výšky zodpovedajú príslušným absolútnym, resp. relatívnym početnostiam jednotlivých intervalov. Pre grafické znázornenie štruktúry štatistického súboru je však možné použiť i rôzne iné diagramy, resp. grafy, ktoré poskytuje zvolený softvér. V prípade, že sledujeme výskyt hodnôt nominálneho znaku, je vhodná voľba číselného kódu pre jednotlivé hodnoty kvalitatívneho znaku.

Pri vyhodnocovaní súborov, na ktorých sledujeme rôzne kvalitatívne ale i kvantitatívne znaky s veľkým počtom pozorovaní, je vhodné použiť **kontingenčnú tabuľku**, ktorá nám umožní analyzovať súbor ako celok, ale i jeho jednotlivé časti. V ponuke štatistických softvérov je procedúra, ktorá nám umožní túto analýzu.

EXCEL

Na triedenie štatistického súboru v EXCELi volíme príkaz ***Nástroje/Analýza údajov***. Z dialógového panelu si zo zoznamu vyberieme ***Histogram***. Táto voľba je vhodná na triedenie dát a tvorbu histogramových grafov. Uvedená procedúra používa vstupnú oblasť (zadané hodnoty skúmaného kvantitatívneho znaku) a oblasť tried (zadané číselné hranice tried). Ak nezadáme hranice tried, EXCEL sám zvolí počet tried a vytvorí triedy o rovnakej šírke. Pri voľbe ***Vytvoriť graf*** sa v samostatnom okne objaví graf, ktorého umiestnenie zvolíme v možnostiach výstupu. Pri voľbe ***kumulatívneho percentuálneho podielu*** sa v grafe zobrazí

i percentuálna kumulatívna relatívna početnosť. Iný spôsob získania histogramu je cez voľbu vytvorenia stĺpcového grafu.

Pre tvorbu kontingenčnej tabuľky si zvolíme ***Údaje(Data)/Zostava kontingenčnej tabuľky a kontingenčného grafu.*** Po uvedenej voľbe sa objaví okno *Sprievodca kontingenčnou tabuľkou.* V ňom špecifikujeme oblasť, v ktorej sa nachádza súbor so vstupnými údajmi, ktoré chceme analyzovať. V ďalšom kroku pristúpime k tvorbe vlastnej tabuľky. Vo všeobecnosti má tabuľka štyri rozmery. Sú pod názvami *STRANA (PAGE), RIADOK (ROW), STĹPEC (COLUMN)* a *ÚDAJE (DATA).* Okienko *STRANA* nie je nevyhnutné vyplniť. V prípade nevyplnenia, tabuľka sa ráta zo súboru ako celku, v prípade vyplnenia, ráta z hodnôt premennej, ktorú sme do okienka preniesli. Ak okienko *RIADOK* vyplníme konkrétnou premennou, bude výsledná tabuľka obsahovať toľko riadkov, koľko je rôznych hodnôt zvolenej premennej a navyše riadok pre súbor spolu. V prípade nevyplnenia tohto okienka, bude tabuľka obsahovať len jeden riadok s hodnotami za súbor ako celok. Podobne je to i v prípade zaplnenia okienka *STĹPEC*. Zaplnenie okienka *ÚDAJE* je povinné. Vyplnením tohto okienka špecifikujeme, podľa hodnôt ktorej premennej sa početnosti počítajú, konkrétny tvar týchto početností, prípadne ďalšie údaje, ako napr. súčet, priemer, maximum, minimum a ďalšie. Je možné zvoliť si i rôzne formy prezentácie výstupu. Ak v sprievodcovi kontingenčnou tabuľkou a kontingenčným grafom zvolíme voľbu kontingenčný graf, objaví sa spolu s kontingenčnou tabuľkou i graf.

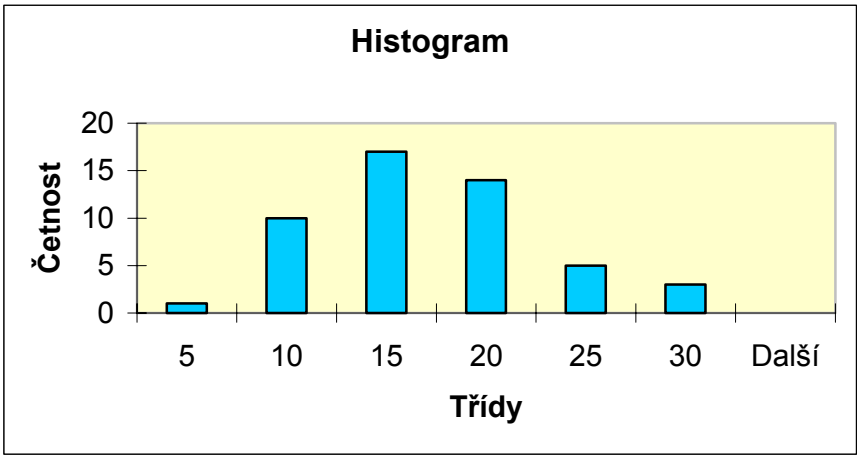
Príklad 1.1

V tabulkovej prílohe sa nachádza súbor PRIJÍMACIE SKÚŠKY. Obsahuje výsledky študentov na prijímacích skúškach na vysokú školu z matematiky, fyziky a psychotestov. Okrem týchto údajov obsahuje aj informácie o type absolvovanej strednej školy, známku z matematiky po 2. semestri (desatinné číslo zohľadňuje počet opravných termínov) a priemernú známku zo všetkých predmetov v 2. semestri. Všetky údaje sú roztriedené do dvoch skupín podľa príslušnosti študentov k jednej z dvoch fakúlt.

Ukážeme si, ako urobíme triedenie súboru a príslušný histogram použitím EXCELu. Našou úlohou je urobiť triedenie bodov, ktoré študenti 2. fakulty v danom roku získali na prijímacích skúškach z matematiky. Použijeme súbor PRIJÍMACIE SKÚŠKY a zadáme horné hranice triednych intervalov, ktoré sú 5, 10, 15, 20, 25, 30. Volíme príkazy ***Nástroje/Analýza údajov/Histogram***. Špecifikujeme oblasť, kde sa nachádzajú hodnoty, ktoré chceme analyzovať a oblasť, v ktorej sa nachádzajú hranice triednych intervalov. Potvrdením okienka *Vytvoriť graf* získame nasledujúci výstup.

Tab. 1.1 Tabuľka rozdelenia absolútnych početností

Třídy	*Četnost*
5	1
10	10
15	17
20	14
25	5
30	3
Další	0



Obr. 1.1 Histogram

Príklad 1.2

Zaujíma nás, aký mali študenti 1. fakulty priemerný počet bodov z matematiky a fyziky na prijímacích skúškach v sledovaném roku, pričom chceme urobiť triedenie v závislosti od typu strednej školy a hodnotenia z psychotestov.

Ukážeme si, ako urobíme kontingenčnú tabuľku, ktorá bude prehľadným triedením sledovaného znaku. Volíme príkazy ***Údaje(Data)/Zostava kontingenčnej tabuľky a kontingenčného grafu.***

V okne *Sprievodca kontingenčnou tabuľkou* špecifikujeme oblasť, v ktorej sa nachádza súbor so vstupnými údajmi, ktoré chceme analyzovať. V našom prípade to budú stĺpce stred. škola, mat., fyzika a psych. zo súboru PRIJÍMACIE SKÚŠKY/1.fakulta. Teraz pristúpime k špecifikácii vlastnej tabuľky. Do okienka *RIADOK* presunieme z ponuky na pravej strane obrazovky znak stred. škola a do okienka *STĹPEC* znak psych. Do okienka *DATA* presunieme znak mat. a fyzika. Dvakrát klikneme myšou na prenesené znaky v okienku *DATA* a objaví sa okno *Pole kontingenčnej tabuľky*, kde si v okienku *Súhrn* vyberieme pre naše potreby *priemer*. Ak nás navyše zaujíma i percentuálne triedenie súboru, t.j. aké percento študentov z celku je z daného typu strednej školy a má dané hodnotenie z psychotestu, vložíme do okienka *DATA* napríklad znak mat. a z okienka súhrn vyberieme *počet hodnôt*. Voľbou *Možnosti* si z ponuky *Zobraziť data ako* vyberieme *% z celku*. Potom už len zvolíme miesto na umiestnenie výslednej tabuľky a potvrdíme *dokončiť*. Na zvolenom mieste sa objaví výsledná tabuľka (Tab. 1.2).

Z nasledujúcej tabuľky Tab. 1.2 vidíme, že najvyšší priemerný počet bodov z matematiky (21,17) získali gymnazisti, ktorí z psychotestov získali hodnotenie 1. Najvyšší priemerný počet bodov z fyziky (17,67) dosiahli opäť gymnazisti, ale s hodnotením 3 zo psychotestov. Podľa očakávania, výrazne najslabšie priemerné hodnotenie z matematiky (4,09) mali absolventi učilíšť a z fyziky (4,67) absolventi obchodných akadémii. Z posledného riadku sa napríklad dozvieme, že až 68,42% študentov malo na psychotestoch hodnotenie 3.

Tab. 1.2 Kontingenčná tabuľka

		psych.			
stred.škola	Data	1	2	3	Celkový součet
G	průměr z mat.	21,17	18	6,33	14,81
	průměr z fyz.	15,33	14,5	17,67	16
	Počet z mat.	7,89%	5,26%	7,89%	21,05%
OA	průměr z mat.			9,17	9,17
	průměr z fyz.			4,67	4,67
	Počet z mat.	0,00%	0,00%	7,89%	7,89%
P	průměr z mat.	17,25	14	7,38	9,58
	průměr z fyz.	15,5	13,33	11,19	12,03
	Počet z mat.	5,26%	7,89%	34,21%	47,37%
U	průměr z mat.	7,3	1,5	4,00	4,09
	průměr z fyz.	5	1	5,64	5,06
	Počet z mat.	2,63%	2,63%	18,42%	23,68%
Celkem průměr z mat.		17,55	13,25	6,56	9,35
Celkem průměr z fyzika		13,67	11,67	9,69	10,63
Celkem Počet z mat.		15,79%	15,79%	68,42%	100,00%

1.4 Parametre základného súboru

Súbor hodnôt, ktoré sme získali pri sledovaní znaku X, charakterizuje komplexne skúmaný jav. Pri ich analýze nás však zväčša zaujíma iba niekoľko kľúčových údajov, z ktorých získame postačujúce informácie o skúmanom súbore. Tieto údaje nazývame **parametre súboru**, resp. **popisné štatistiky,** prípadne **číselné charakteristiky.** Medzi základné parametre súboru patria aritmetický priemer (stredná hodnota), modus a medián, ktoré charakterizujú polohu (úroveň) hodnôt znaku, variačné rozpätie, rozptyl, štandardná odchýlka, ktoré charakterizujú variabilitu hodnôt analyzovaného znaku. Informáciu o rozložení hodnôt znaku X nám poskytne i koeficient špicatosti a koeficient šikmosti. Bližšie si ich popíšeme.

Aritmetický priemer $\bar{x}$ vyjadruje, aký objem hodnôt znaku X pripadá v priemere na jednu jednotku súboru. Je definovaný vzťahom

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.1}$$

N - rozsah súboru,

x_i - hodnota znaku X u i-tej jednotky.

Vážený aritmetický priemer $\bar{x}$ používame, ak pracujeme s triedeným súborom hodnôt znaku *X*. Na jeho výpočet použijeme vzťah

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{m} x_j n_j \tag{1.2}$$

N - rozsah súboru,

m - počet tried v súbore,

n_j - absolútna početnosť j-tej triedy (j = 1, 2, ...,m),

x_j - hodnota znaku *X*, ktorá reprezentuje j-tu triedu.

Spomenieme aspoň niektoré dôležité vlastnosti aritmetického priemeru:

a) $\sum_{i=1}^{N}(x_i - \bar{x}) = 0$, t.j. súčet odchýlok hodnôt znaku od jeho aritmetického priemeru je nulový.

b) $\frac{1}{N}\sum_{i=1}^{N}(x_i + a) = \bar{x} + a$, t.j. ak pripočítame ku všetkým hodnotám ľubovoľnú konštantu a, potom aritmetický priemer sa líši od pôvodného priemeru o túto konštantu.

c) $\frac{1}{N}\sum_{i=1}^{N} ax_i = a\bar{x}$, t.j. ak všetky hodnoty znaku násobíme nenulovou konštantou, potom i aritmetický priemer z týchto hodnôt získame násobením pôvodného priemeru touto konštantou.

Modus M_O je najčastejšie sa vyskytujúca hodnota znaku X, resp., v prípade triedeného súboru hodnota reprezentanta triedy s najväčšou absolútnou početnosťou.

Medián M_e je hodnota, ktorá súbor zistených hodnôt delí na 2 rovnako početné skupiny, t.j. skupiny, z ktorých prvá obsahuje 50% štatistických jednotiek, ktoré majú hodnotu znaku X menšiu ako medián, druhá obsahuje 50% zvyšných štatistických jednotiek, ktoré majú hodnotu väčšiu ako medián. Ak zoradíme všetky hodnoty znaku podľa veľkosti do neklesajúcej (resp. nerastúcej) postupnosti, tak mediánom bude hodnota, ktorá je v strede uvažovanej postupnosti.

$$M_e = x_{k+1}, \text{ ak } N = 2k + 1, \tag{1.3}$$

$$M_e = \frac{x_k + x_{k+1}}{2}, \text{ ak } N = 2k. \tag{1.4}$$

V prípade triedeného súboru

$$M_e = a + \frac{\frac{N+1}{2} - n_1}{n_2} \tag{1.5}$$

a - horná hranica triedy, ktorá predchádza mediálny interval,

N - rozsah súboru,

n_1 - počet všetkých prvkov pod mediálnym intervalom,

n_2 - počet prvkov mediálneho intervalu,

h - šírka triedy.

Skutočnosť, že poznáme priemer, modus, resp. medián daného súboru nie je postačujúcou informáciou pre popis daného súboru. Pre štatistický súbor je charakteristická i variabilita (menlivosť) zistených hodnôt. Na vyjadrenie variability súboru použijeme nasledujúce parametre.

Variačné rozpätie v_r je iba približnou charakteristikou variability hodnôt sledovaného znaku. Je definovaný ako rozdiel najväčšej a najmenšej hodnoty kvantitatívneho znaku, t.j.

$$v_r = x_{max} - x_{min} . \tag{1.6}$$

Rozptyl σ^2 predstavuje aritmetický priemer druhých mocnín (štvorcov) odchýlok od priemeru $\bar{x}$. Je definovaný predpisom

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 . \tag{1.7}$$

V prípade triedeného súboru ho vypočítame podľa vzťahu

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{m}(x_j - \bar{x})^2 n_j \tag{1.8}$$

N - rozsah súboru,

m - počet tried v súbore,

n_j - absolútna početnosť j-tej triedy (j = 1, 2, ...,m),

x_j - hodnota znaku X, ktorá reprezentuje j-tu triedu.

K dôležitým vlastnostiam rozptylu patrí:

a) Rozptyl konštanty je rovný nule.

b) Ak pripočítame ku všetkým hodnotám znaku konštantu, rozptyl sa nezmení.

c) Ak všetky hodnoty znaku vynásobíme konštantou a, potom rozptyl takto vzniknutých hodnôt je rovný súčinu rozptylu pôvodného súboru a druhej mocniny konštanty a.

Štandardná (smerodajná) odchýlka σ je definovaná ako $\sigma = \sqrt{\sigma^2}$ a udáva, ako sa v priemere v danom súbore odchyľujú hodnoty znaku od aritmetického priemeru.

Štandardná chyba (chyba strednej hodnoty) je definovaná ako $\frac{\sigma}{\sqrt{N}}$.

Nasledujúce dva parametre informujú o rozložení hodnôt analyzovaného znaku v zmysle väčšieho zoskupenia týchto hodnôt v určitej časti variačného rozpätia v porovnaní s ostatnými časťami.

Koeficient šikmosti S charakterizuje symetriu rozdelenia. Jeho hodnotu vypočítame podľa vzťahu

$$S = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\sigma^3}. \tag{1.9}$$

V prípade triedeného súboru

$$S = \frac{\frac{1}{N}\sum_{j=1}^{m}(x_j - \bar{x})^3 n_j}{\sigma^3}. \tag{1.10}$$

Ak $S = 0$, rozdelenie je symetrické, ak $S > 0$, hovoríme o kladnej asymetrii (zošikmení) rozdelenia, ak $S < 0$ o zápornej asymetrii rozdelenia. Znamienko koeficienta šikmosti signalizuje smer zošikmenia a jeho absolútna hodnota silu zošikmenia.

Koeficient špicatosti K charakterizuje koncentráciu hodnôt sledovaného znaku okolo jeho priemeru. Na výpočet použijeme vzorec

$$K = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^4}{\sigma^4} - 3, \tag{1.11}$$

resp.

$$K = \frac{\frac{1}{N}\sum_{j=1}^{m}\left(x_j - \bar{x}\right)^4 n_j}{\sigma^4} - 3, \tag{1.12}$$

ak pracujeme s triedeným súborom.

Meranie koncentrácie v tomto prípade, je teda meraním špicatosti (excesu) rozdelenia početnosti. Ak $K > 0$, rozdelenie je „špicatejšie“, čo znamená, že vrchol rozdelenia veľmi výrazne vyniká. $K < 0$ sa prejavuje v „plochosti" tvaru rozdelenia početnosti. Na vyjadrenie stupňa koncentrácie sa používa porovnanie s normálnym rozdelením, s ktorým sa bližšie zoznámime v kapitole 3. V prípade normálneho rozdelenia je koeficient špicatosti i šikmosti rovný 0.

EXCEL

Na výpočet parametrov základného štatistického súboru použitím EXCELu využijeme jeho ponuku štatistických funkcií. Niektoré základné z nich sú uvedené v nasledujúcej tabuľke.

Tab. 1.2 Štatistické funkcie v EXCELi

AVERAGE	aritmetický priemer
MODE	modus
MEDIAN	medián
MAX	maximálna hodnota
MIN	minimálna hodnota
VARP	rozptyl
STDEVP	štandardná (smerodajná) odchýlka
SKEW	koeficient šikmosti
KURT	koeficient špicatosti

Na získanie uceleného obrazu o parametroch skúmaného súboru je vhodné z ponuky ***Analytických nástrojov*** vybrať procedúru ***Popisná štatistika***. Pomocou tohto nástroja získame tabuľku, ktorá prehľadne udáva nielen hodnoty základných parametrov, ale i ďalšie údaje, ako napr. minimálna a maximálna hodnota, počet prvkov, súčet hodnôt, a pod. Treba poznamenať, že v tabuľke popisnej štatistiky pod názvom rozptyl výberu a smerodajná odchýlka nenájdeme údaje, ktorý získame použitím spomínaných štatistických funkcií VARP a STDEVP (Tab. 1.2), ale výberový rozptyl (medzi štatistickými funkciami je pod názvom VAR) a výberová štandardná (smerodajná) odchýlka (medzi štatistickými funkciami je pod názvom STDEV), o ktorých sa zmienime v kapitole 4.

Príklad 1.4

Vypočítame základné parametre súboru z Príkladu 1.1, t.j. súboru, ktorý tvoria body z matematiky, ktoré študenti 2. fakulty v danom roku získali na prijímacích skúškach. Volíme príkazy ***Nástroje/Analýza údajov/Popisná štatistika***. V okienku Vstupná oblasť špecifikujeme oblasť, v ktorej sa nachádzajú analyzované údaje, ponecháme špecifikáciu, že údaje sú usporiadané v stĺpcoch. V prípade, že v prvom riadku vybraných údajov je názov znaku, v našom prípade je to **mat.**, potvrdíme okienko *Popisky*. V ďalšom okienku špecifikujeme výstupnú oblasť a potom vyznačíme, ktoré parametre chceme na výstupe. Pre naše potreby stačí vybrať *Celkový prehľad.* Po stlačení *OK* dostaneme na výstupe nasledujúcu tabuľku.

Tab. 1.3 Popisná štatistika

mat.	
Stř. hodnota	15,12
Chyba stř. hodnoty	0,83
Medián	14,50
Modus	9,00
Směr. odchylka	5,88
Rozptyl výběru	34,56
Špičatost	-0,02
Šikmost	0,44
Rozdíl max-min	26,50
Minimum	3,50
Maximum	30,00
Součet	756,00
Počet	50,00

Z tabuľky vidíme, že napr. priemerný počet bodov z matematiky dosiahnutý študentmi na 2. fakulte v danom roku je 15,12.

Pozor, v okienku *Rozptyl výberu* sa nachádza tzv. výberový rozptyl, o ktorom sa zmienime v kapitole 4. Z neho je potom vypočítaná štandardná odchýlka (*Smer. odchýlka*) a chyba strednej hodnoty. Na výpočet týchto parametrov pre náš základný súbor použijeme príslušné štatistické funkcie. Konkrétne, použitím funkcie VARP (STDEVP) dostaneme, že rozptyl (štandardná odchýlka) analyzovaného súboru je 33,87 (5,82). Podobne i koeficienty šikmosti a špicatosti sú v tejto tabuľke počítané ako tzv. štandardizované koeficienty šikmosti a špicatosti pre výberové súbory. Rozdiely medzi hodnotami koeficientov šikmosti a špicatosti v tabuľke popisnej štatistiky a medzi hodnotami, ktoré získame použitím funkcií SKEW a KURT sú však zanedbateľné.

Z ostatných údajov si všimnime napríklad koeficient šikmosti, ktorý je 0,44, čo signalizuje pozitívnu asymetriu. Z ďalších údajov v tabuľke je praktická informá-

cia o minimálnej a maximálnej hodnote získaných bodov ktoré sú na 2. fakulte 3,5 a 30. Z posledného riadku sa dozvieme, že rozsah sledovaného súboru je 50.

Príklady na precvičenie

1.5 Urobte triedenie a histogram súboru PRIJÍMACIE SKÚŠKY /1.fakulta/mat. Akceptujte hranice tried ponúkané EXCELom.

1.6 Urobte triedenie a histogram súboru PRIJÍMACIE SKÚŠKY/1.fakulta/ fyzika. Najskôr akceptujte hranice tried ponúkané EXCELom a potom urobte triedenie pre hranice 5, 10, 15, 20, 25, 30. Coronate oba histogramy.

1.7 Urobte triedenie a histogram súboru PRIJÍMACIE SKÚŠKY/1.fakulta/psych. do 3 tried.

1.8 V tvare kontingenčnej tabuľky urobte percentuálne triedenie súboru PRIJÍMACIE SKÚŠKY/2. fakulta/ fyzika v závislosti od typu strednej školy a hodnotenia z psychotestov. Urobte analýzu kontingenčnej tabuľky.

1.9 V tvare kontingenčnej tabuľky urobte triedenie súboru PRIJÍMACIE SKÚŠKY/2. fakulta/ matematika v závislosti od typu strednej školy a hodnotenia z psychotestov. Urobte analýzu kontingenčnej tabuľky.

1.10 Urobte kontingenčnú tabuľku pre určenie priemerného počtu bodov získaných na prijímacích skúškach z matematiky v závislosti od typu strednej školy a hodnotenia z psychotestov. Použite súbory:

a) PRIJÍMACIE SKÚŠKY/1.fakulta
b) PRIJÍMACIE SKÚŠKY/2.fakulta

Urobte analýzu kontingenčnej tabuľky.

1.11 Urobte kontingenčnú tabuľku pre určenie priemerného počtu bodov získaných na prijímacích z fyziky v závislosti od typu strednej školy a hodnotenia z psychotestov. Použite súbory:

a) PRIJÍMACIE SKÚŠKY/1.fakulta
b) PRIJÍMACIE SKÚŠKY/2.fakulta

Urobte analýzu kontingenčnej tabuľky.

1.12 Určite priemer, rozptyl, štandardnú odchýlku, modus, medián, koeficient šikmosti a špicatosti súborov:

a) PRIJÍMACIE SKÚŠKY/1.fakulta/mat.
b) PRIJÍMACIE SKÚŠKY/1.fakulta/fyzika
c) PRIJÍMACIE SKÚŠKY/2. fakulta/ fyzika

Porovnajte údaje, ktoré získate použitím štatistických funkcií ponúkaných EXCELom s údajmi, ktoré získate použitím procedúry *Popisná štatistika*.

2. PRAVDEPODOBNOSŤ

Matematický základ štatistických metód je teória pravdepodobnosti, ktorá skúma zákonitosti v takých javoch a procesoch, kde sa vyskytujú prvky náhodnosti. Táto kapitola obsahuje tie pojmy teórie pravdepodobnosti, ktoré využijeme pri popise štatistických metód. Vhodný matematický aparát na axiomatické vybudovanie teórie náhodných javov je teória množín.

2.1 Náhodný pokus a náhodný jav (udalosť)

Okrem javov, ktoré sú určené deterministicky (fyzikálne a matematické zákony, vzťahy medzi ekonomickými veličinami) sa často stretávame s javmi, ktoré sa nedajú opísať týmto spôsobom. Na výsledok činnosti, pri zachovaní rovnakých podmienok, môžu vplývať aj druhotné faktory, ktorých vplyv nevieme predvídať, preto dopredu nie je známy ani výsledok tejto činnosti. Preto takúto činnosť - fyzický experiment voláme **náhodný pokus** (hod kockou alebo viacerými kockami, losovanie prvku z nejakej množiny, rozdanie kariet, rôzne náhodné hry, narodenie dieťaťa istého pohlavia). Spoločnou vlastnosťou náhodných pokusov je ich opakovateľnosť.

Náhodný jav (udalosť) je výsledok náhodného pokusu, ktorý pod vplyvom náhody niekedy nastane, inokedy nie. Náhodné javy sa označujú *A, B, ...* .

Príklady :

- Pri hode kockou padne 5 bodov.
- Narodí sa dievča.
- Vyrobí sa zlý výrobok.
- V rade stojí najviac 5 ľudí.
- Zo všetkých obyvateľov mesta sa vylosuje dôchodca.

Je dôležité poznať množinu všetkých možných výsledkov náhodného pokusu. V ďalšom ku každému takémuto výsledku priradíme číslo, ktorým vyjadríme šancu, že daný pokus skončí s týmto výsledkom.

2.2 Vzťahy medzi náhodnými javmi a operácie s nimi

Osobitne dôležité postavenie majú elementárne náhodné javy (označujeme ich E_1, E_2, ...), ktoré zodpovedajú rôznym základným výsledkom náhodného pokusu. Pre každý pokus je dôležité poznať všetky elementárne náhodné javy, z nich vytvoriť množinu všetkých možných výsledkov pokusu

$$\Omega = \{E_1, E_2, ..., E_n, ...\}.$$

Presnejšie: Nech množina $\Omega = \{E_1, E_2, ..., E_n, ...\}$ je neprázdna, aspoň dvojprvková. Jej prvky sú **elementárne javy**, **náhodný jav** je ľubovoľná podmnožina množiny Ω.

Základné vzťahy medzi javmi a operácie s nimi

a) Množina Ω predstavuje jav, ktorý nastane vždy po vykonaní pokusu, volá sa **istý jav** a označuje sa *I*. t.j. $I = \Omega$ (Obr. 2.1 *a*) .

b) Prázdna podmnožina množiny Ω reprezentuje **nemožný jav**. Je to jav, ktorý po vykonaní pokusu nikdy nenastane, označuje sa ϕ (Obr. 2.1 *b*).

c) **Jav *A* je časťou javu *B***, ak s nastaním javu *A* nastane vždy aj jav *B*, čo označujeme $A \subset B$ (Obr. 2.1 *c*). Pre reláciu $\subset$ platia vzťahy:

$$A \subset A\ ; \quad A \subset I\ ; \quad \phi \subset A\ ; \quad (A \subset B) \wedge (B \subset C) \Rightarrow A \subset C\ .$$

d) **Javy *A, B* sa rovnajú (sú rovnocenné)**, ak platí $A \subset B$ a súčasne $B \subset A$, čo označujeme $A = B$ (Obr. 2.1 *d*).

e) **Jav $\bar{A}$ je opačný jav k javu *A***, ak jav $\bar{A}$ nastane práve vtedy, keď nenastane jav *A* (Obr. 2.1 *e*).

f) **Jav *C***, pri ktorom súčasne nastane jav *A* aj *B* nazývame **prienik (súčin) javov *A, B*** a označujeme $C = A \cap B$ (Obr. 2.1 *f*).

g) **Jav *C*,** pri ktorom nastane aspoň jeden z javov *A* alebo *B,* **nazývame zjednotenie (súčet) javov *A, B*** a označujeme $C = A \cup B$ (Obr. 2.1*g*).

h) **Jav *C* je rozdiel javov *A* a *B*** (v tomto poradí) a nastane práve vtedy, keď nastane jav *A* a zároveň nenastane *B*, čo označujeme $C = A - B$ (Obr. 2.1 *h*).

i) **Javy *A, B* sú disjunktné (nezlučiteľné),** ak ich súčasné nastanie nie je možné, t.j. $A \cap B = \phi$ (Obr. 2.1 *i*).

j) Skupina javov $A_1, A_2, ..., A_n$ tvorí **úplný systém javov**, ak $A_i \cap A_j = \phi$, pre $i \neq j$, $i, j = 1,2,...,n$ a ak platí $A_1 \cup A_2 \cup ... \cup A_n = \Omega$.

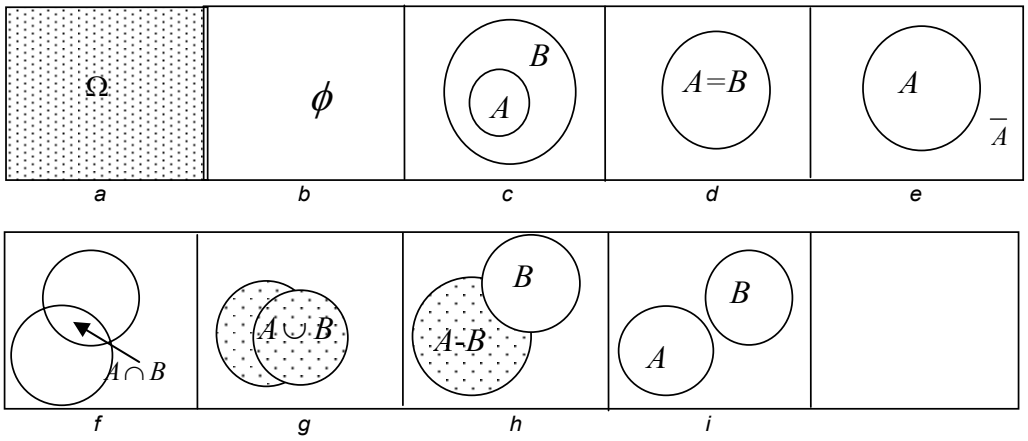
k) Jav *A* je **zložený jav**, ak sa dá vyjadriť ako zjednotenie aspoň dvoch javov, rôznych od nemožného javu a samotného javu *A*.

l) **Elementárny jav** $E \neq \phi$ má vlastnosti:

- Neexistuje taký jav $B \neq \phi$, pre ktorý by platilo $B \subset E$.
- Rôzne elementárne javy sú vždy disjunktné.
- Každý zložený jav sa dá vyjadriť ako zjednotenie elementárnych javov.
- Ku každému zloženému javu *B* existuje elementárny jav *E* tak, že $E \subset B$.

Poznámka 2.1

Vidíme, že s javmi pracujeme ako s množinami, môžeme ich znázorňovať Vennovými diagramami ako na Obr. 2.1



Obr. 2.1 Znázornenie náhodných javov pomocou Vennových diagramov

Príklad 2.1

Hádžeme hracou kockou. Nech jav *A* je padnutie párneho počtu bodov, jav *B* padnutie 2 alebo 4 alebo 6 bodov, jav *C* padnutie 2 bodov, jav *D* padnutie najviac 3 bodov, jav *F* padnutie viac ako 6 bodov. Potom vzťahy medzi javmi môžeme vyjadriť takto:

- $C \subset D$, $A = B$
- $\overline{D}$ - jav, že padnú aspoň 4 body
- F - nemožný jav
- $A \cap D = C$
- $B - C$ - jav, že padnú 4 body alebo padne 6 bodov
- $\overline{(A \cup D)}$ - jav, že padne 5 bodov.

2.3 Pravdepodobnosť a jej vlastnosti

Zatiaľ sme sa zaoberali len vzťahmi medzi náhodnými javmi. Náhodné javy majú aj inú objektívnu vlastnosť. Pri mnohonásobnom opakovaní náhodného pokusu môžeme pozorovať isté pravidelnosti, zákonitosti. Napríklad, niektoré javy nastávajú častejšie ako iné. Skúmaním zákonitostí pri pôsobení náhody sa zaoberá teória **pravdepodobnosti**. Každý náhodný jav môžeme kvantitatívne (číselne) ohodnotiť, priradiť mu číslo, ktoré bude vyjadrovať mieru možnosti, že tento jav nastane. Toto číslo sa nazýva **pravdepodobnosť javu *A***, označujeme ho *P(A)* a v ďalšom tento pojem zavedieme presnejšie.

Z historického hľadiska išiel vývoj od **štatistickej pravdepodobnosti** (založenej na mnohonásobnom opakovaní pokusov), cez **klasickú pravdepodobnosť** (založenej na kombinatorike), cez **geometrickú pravdepodobnosť** až po **axiomatickú pravdepodobnosť.** Dnešná podoba axiomatickej pravdepodobnosti pochádza od A.N. Kolmogorova (1934) .

Klasická definícia pravdepodobnosti

Je vybudovaná na tom, že množina elementárnych javov $\Omega = \{E_1, E_2, ..., E_n\}$ je konečná. Ku každému elementárnemu výsledku priradíme číslo $P(E_i)$, vyjadrujúce šancu – pravdepodobnosť, že tento výsledok nastane.

Pravdepodobnosťou na množine Ω nazývame každú nezápornú funkciu *P*, pre ktorú platí $P(E_1) + P(E_2) + ... + P(E_n) = 1$.

Navyše, nech všetky elementárne javy sú rovnako možné, t.j. $P(E_i) = \frac{1}{n}$, pre všetky elementárne javy. Nech jav $A \subset \Omega$ je tvorený *m* elementárnymi javmi z množiny Ω. Potom **pravdepodobnosť javu *A* je číslo $P(A)$, definované podielom**

$$P(A) = \frac{m}{n} \left(\frac{priaznivé\ prípady\ \ pre\ nastúpenie\ \ javu\ \ A}{všetky\ \ možné\ \ prípady} \right).$$

Poznámka 2.2

Číslo $100 . P(A)$ je pravdepodobnosť javu *A* vyjadrená v percentách. Tento spôsob vyjadrovania pravdepodobnosti je najčastejší v bežných praktických úlohách.

Príklad 2.2

V trojdetnej rodine, kde narodenie chlapca i dievčaťa bolo rovnako pravdepodobné, určite pravdepodobnosť nasledujúcich javov:

- *A* - deti sa narodili v poradí *CHDCH* alebo *CHCHD*
- *B* - v rodine majú aspoň 2 dievčatá
- *C* - v rodine majú všetky deti rovnakého pohlavia
- *D* - nemajú žiadne dievča
- *E* - majú aspoň jedného chlapca
- *F* - majú najviac jedného chlapca.

Všetky možnosti ako sa mohli 3 deti narodiť tvoria množinu elementárnych javov $\Omega = \{CHCHCH, CHCHD, CHDCH, DCHCH, CHDD, DDCH, DCHD, DDD\}$, ktorá má 8 prvkov. Pravdepodobnosť každého elementárneho javu z Ω je určená napríklad takto: Zo 100 manželských párov sa polovici narodí *CH* a polovici *D*. Z tých, čo už majú *CH* ako prvé dieťa, sa polovici narodí *D* a polovici *CH*. A nakoniec, z tých , čo majú už *CHD* sa zas polovici narodí *CH*. Preto podiel rodín, ktoré majú deti v poradí *CHDCH* je $\frac{1}{2} z \left(\frac{1}{2} z \frac{1}{2} \right) = \frac{1}{8} = 0{,}125$. Toto číslo dostaneme pre každý elementárny jav z Ω, t.j. všetky elementárne javy sú rovnako možné. Určíme, ktoré elementárne javy tvoria javy *A*-*F* a ich pravdepodobnosti:

$$A = \{CHDCH, CHCHD\} \Rightarrow P(A) = \frac{2}{8} = 0{,}25,$$

$$B = \{DDCH, DCHD, CHDD, DDD\} \Rightarrow P(B) = \frac{4}{8} = 0{,}5,$$

$$C = \{DDD, CHCHCH\} \Rightarrow P(C) = \frac{2}{8} = 0{,}125,$$

$$D = \{CHCHCH\} \Rightarrow P(D) = \frac{1}{8} = 0{,}125,$$

$$E = \{CHDD, DCHD, DDCH, CHCHD, CHDCH, DCHCH, CHCHCH\}$$
$$\Rightarrow P(E) = \frac{7}{8},$$

$$F = \{DDD, DDCH, DCHD, CHDD\} \Rightarrow P(F) = \frac{4}{8} = 0{,}5.$$

Príklad 2. 3

Pri riešení predchádzajúceho príkladu môžeme rozmýšľať aj takto: vytvárame usporiadané trojice prvkov *XXX*, kde na každom mieste môže byť *CH* alebo *D* (s rovnakou šancou), všetkých možností je $2 \cdot 2 \cdot 2 = 8$ (variácie s opakovaním), pre jav *A* sú vyhovujúce len dve možnosti *CHDCH, CHCHD*. Preto $P(A) = \frac{2}{8} = 0{,}25$, atď.

Štatistická pravdepodobnosť a relatívna početnosť

Vychádzame z veľkého počtu nezávislých pokusov (t.j. výsledok jedného pokusu neovplyvní výsledok nasledujúceho pokusu), vykonaných v rovnakých podmienkach a sledujeme, koľkokrát nastal jav *A*. Ak v jednej sérii *n* pokusov nastal daný jav práve *m*-krát, tak číslo

$$f = \frac{m}{n}, \quad 0 \le m \le n \tag{2.1}$$

nazývame **relatívna početnosť javu *A***.

Ak postupne zvyšujeme počet pokusov v jednotlivých sériách, dostaneme postupnosť relatívnych početností $f_1, f_2, f_3, \ldots$ a pri veľkom počte pokusov zistíme, že sa tieto relatívne početnosti od seba málo odlišujú, vykazujú istú stabilitu. Prídeme k záveru, že existuje konštanta, okolo ktorej relatívne početnosti kolíšu. Preto **pravdepodobnosť javu *A*, ozn. *P(A)* je číslo**

$$P(A) = \lim_{n\to\infty} \frac{m}{n}, \tag{2.2}$$

kde $\frac{m}{n}$ je relatívna početnosť javu *A*.

Najznámejšie odôvodnenie tejto vlastnosti pochádza od J. Bernoulliho ako jeden tvar **zákona veľkých čísiel.** Hovorí, že pravdepodobnosť toho, že sa relatívna početnosť javu *A* líši od pravdepodobnosti javu *A* o ľubovoľne malé $\varepsilon > 0$, je rovná 1, ak je počet pokusov nekonečne veľký, t.j.

$$\lim_{n\to\infty} P\left(\left| \frac{m}{n} - p \right| < \varepsilon \right) = 1, \qquad \text{kde } P(A) = p\,. \tag{2.3}$$

To nás oprávňuje pre veľké *n* nahradiť pravdepodobnosť relatívnou početnosťou,

t.j. $$P(A) \doteq \frac{m}{n}\,. \tag{2.4}$$

Príklad 2.4

Pri opakovanom hode kockou zisťujeme, či kocka nie je falošná. Sledujeme, aká je pravdepodobnosť, že padne napr. jednotka. Výsledky z 5 sérií pokusov sú v Tab. 2.1.

Tab. 2.1 Výsledky pokusov pri hode kockou

Počet hodov	Počet padnutí jednotky	Relatívna početnosť
50	5	0,1
100	13	0,13
500	88	0,176
1000	159	0,159
5000	822	0,1644

Pri hode kockou je $\Omega = \{1,2,3,4,5,6\}$, ale len jeden elementárny jav je priaznivý výsledok. Preto $P(A) = \frac{1}{6} = 0,1\overline{6}$, čo približne zodpovedá relatívnym početnostiam pri 1000 a 5000 pokusoch.

Príklad 2.5

Aby sme mohli príklad s 3 deťmi v rodine vyriešiť štatistickým prístupom, potrebovali by sme zozbierať údaje napr. z 10000 trojdetných rodín a zistiť, v koľkých rodinách sa narodili deti v poradí *CHDCH* alebo v poradí *CHCHD*, označme tento počet m. Pravdepodobnosť javu A odhadneme pomocou relatívnej početnosti $P(A) = \frac{m}{10000}$.

Geometrická definícia pravdepodobnosti

Nech Ω je množina (interval, rovinný útvar, priestorové teleso) a $G \subset \Omega$, pričom vieme vypočítať dĺžku (obsah, objem) množín G, Ω. Zistime, aká je pravdepodobnosť, že náhodne zvolený bod množiny Ω padne aj do G ? **Pravdepodobnosť** tohto javu definujeme prirodzeným spôsobom ako **podiel mier útvarov**

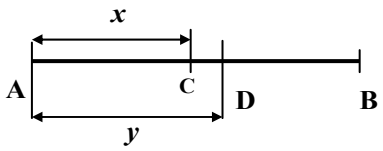
$$P(A) = \frac{\mu(G)}{\mu(\Omega)}, \tag{2.5}$$

kde $\mu(G) \geq 0$, $\mu(\Omega) > 0$ sú miery (dĺžka, obsah, objem) útvarov.

Použitie tejto definície je vhodné, ak množiny Ω, *G* majú nekonečne veľa prvkov, ale pravdepodobnosť zvolenia každého bodu v Ω je rovnaká.

Príklad 2.6

Na úsečke *AB* s dĺžkou 1 sú náhodne zvolené 2 body *C,D*. Aká je pravdepodobnosť toho, že bod *C* leží bližšie k bodu *D* ako k bodu *A*?

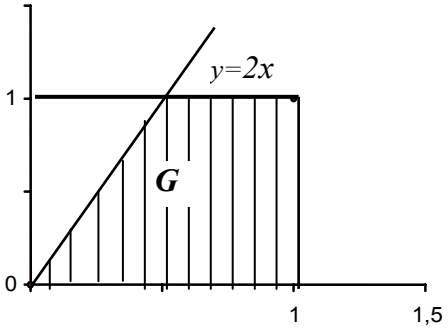


Obr. 2.2 Geometrické znázornenie príkladu

Označíme v súlade s Obr. 2.2 $|AC| = x$, $|AD| = y$. Vždy musí platiť $0 \le x \le 1$, $0 \le y \le 1$. Ak tieto všetky možné voľby x a y znázorníme ako body $[x, y]$v rovine, tak štvorec $\langle 0;1\rangle \times \langle 0;1\rangle$ na Obr. 2.3 predstavuje množinu Ω všetkých možných volieb bodov *C,D* na úsečke. Obsah tohto štvorca je $\mu(\Omega) = 1$. Bod *C* leží bližšie k bodu *D* ako k bodu *A,* ak platí $|y - x| < x$. Oborom pravdivosti tejto nerovnice bude časť Ω a predstavuje množinu *G* priaznivých výsledkov úlohy. Vyriešime túto nerovnicu a jej riešenie znázornime graficky v rovine:

1. $(y - x \ge 0) \wedge (y - x < x) \quad \Rightarrow (y \ge x) \wedge (y < 2x)$

alebo 2. $(y - x < 0) \wedge (-y + x < x) \Rightarrow (y < x) \wedge (y > 0)$



Obr. 2. 3 Grafické riešenie príkladu

Obsah vyšrafovanej oblasti je $\mu(G)=\frac{3}{4}$. Preto pravdepodobnosť toho, že bod *C* leží bližšie k bodu *D* ako k bodu *A* je $P(A)=\frac{\mu(G)}{\mu(\Omega)}=\frac{3}{4}=0{,}75$.

Axiomatická definícia pravdepodobnosti

Klasický a geometrický prístup v predchádzajúcich článkoch dávajú návod ako vypočítať pravdepodobnosť javu aj bez uskutočnenia náhodného pokusu, štatistický prístup umožňuje vypočítať pravdepodobnosť javu až po vykonaní veľkého počtu pokusov. Teoretickým zovšeobecnením týchto definícií je axiomatická definícia, ktorá však nedáva priamo návod na výpočet pravdepodobnosti javu.

Nech Ω je ľubovoľná, aspoň dvojprvková množina a *S* je neprázdna množina, obsahujúca podmnožiny množiny Ω. Nech *S* spĺňa tieto podmienky:

*S*1: ak $A \in S$, potom aj jej doplnok v Ω je prvkom *S,* t.j. $\overline{A} \in S$,

*S*2: ak $A_n \in S$, tak aj $\left(A_1 \cup A_2 \cup ... \cup A_n \cup ...\right) \in S$, pre $\forall n \in N$.

Každú funkciu *P*, ktorá je definovaná na množine *S*, s hodnotami v množine reálnych čísel, t.j. $P: S \to \mathrm{R}$ a ktorá má vlastnosti :

A1: každému javu $A \in S$ je priradené nezáporné číslo $P(A)$,

A2: $P(\Omega) = 1$,

A3: ak $A_1, A_2, ...A_n, ...$ sú disjunktné množiny z S, pre každú rôznu dvojicu množín, potom

$$P\left(A_1 \cup A_2 \cup ... \cup A_n \cup ...\right) = P(A_1) + P(A_2) + ... + P(A_n) + ... \quad , \tag{2.6}$$

voláme **pravdepodobnosť**. Prvky množiny *S* voláme **náhodné javy,** číslo *P*(*A*) voláme **pravdepodobnosť javu *A*** a usporiadanú trojicu (Ω, *S*, *P*) voláme **pravdepodobnostný priestor**.

Priamo z tejto definície vyplýva, že ak náhodný jav je zložený z niekoľkých elementárnych javov, potom pravdepodobnosť jeho nastatia je súčet pravdepodob-

ností jednotlivých elementárnych javov, lebo elementárne javy sú disjunktné a možno použiť axiómu *A3*. Napríklad, ak $A = \{E_1, E_2, E_3\}$, potom

$$P(A) = P(E_1) + P(E_2) + P(E_3).$$

Príklad 2.7

Aká je pravdepodobnosť, že na kocke padne číslo väčšie alebo rovné štyrom? Tento jav A pozostáva z troch elementárnych javov (padne štvorka, padne päťka, padne šestka), ktoré sú samozrejme disjunktné a pravdepodobnosť nastatia každej z nich je $1/6$. Preto $P(A) = 1/6 + 1/6 + 1/6 = 1/2$.

Pre ľubovoľné $A \in S, B \in S$ platia tvrdenia:

a) $P(\phi) = 0$ (2.7)

b) $P(\overline{A}) = 1 - P(A)$ (2.8)

c) $A \subset B \Rightarrow P(A) \leq P(B)$ (2.9)

d) $0 \leq P(A) \leq 1$ (2.10)

Dôkaz Pri nasledujúcich dôkazoch stačí daný jav zapísať ako zjednotenie disjunktných javov a použiť axiómu *A3*.

- $I \cup \phi = I$, $I \cap \phi = \phi$. Podľa $A3$: $P(I) = P(I \cup \phi) = P(I) + P(\phi)$.
  Podľa $A2$: $1 = 1 + P(\phi)$. Preto $0 = P(\phi)$.
- $A \cup \overline{A} = I$ a $A \cap \overline{A} = \phi$. $P(A \cup \overline{A}) = P(I)$, $P(A) + P(\overline{A}) = 1 \Rightarrow P(\overline{A}) = 1 - P(A)$.
- Ak $A \subset B$ potom $B = A \cup (\overline{A} \cap B)$ a $A \cap (\overline{A} \cap B) = \phi$. Podľa *A3*:
  $P(B) = P(A) + P(\overline{A} \cap B)$, kde $P(\overline{A} \cap B) \geq 0$. Preto $P(A) \leq P(B)$.
- $\phi \subset A \subset I$, podľa c) platí $P(\phi) \leq P(A) \leq P(I)$, t.j. $0 \leq P(A) \leq 1$.

Ďalšie vlastnosti pravdepodobnosti:

e) $P(A - B) = P(A) - P(A \cap B)$ (2.11)

Dôkaz Jav $A=(A-B)\cup(A\cap B)$ a $(A-B)\cap(A\cap B)=\phi$. Podľa *A3* platí $P(A)=P(A-B)+P(A\cap B)$. Potom $P(A-B)=P(A)-P(A\cap B)$.

f) **Veta o pravdepodobnosti zjednotenia javov** Ak javy A,B nie sú disjunktné, nedá sa aplikovať axióma *A3* z axiomatickej definície. V tomto prípade platí

$$P(A\cup B)=P(A)+P(B)-P(A\cap B). \tag{2.12}$$

Dôkaz $A\cup B=A\cup(B-A)$ a $A\cap(B-A)=\phi$. Podľa *A3* a výsledku (2.11)

$$P(A\cup B)=P(A)+P(B-A)=P(A)+P(B)-P(A\cap B).$$

g) Zovšeobecnením tejto vlastnosti dostaneme:

$$P\left(\bigcup_{i=1}^{n}A_i\right)=\sum_{i=1}^{n}P(A_i)-\sum_{\substack{i,j=1\\ i<j}}^{n}P(A_i\cap A_j)+\sum_{\substack{i,j,k=1\\ i<j<k}}^{n}P(A_i\cap A_j\cap A_k)-\ldots$$

$$\ldots+(-1)^{n-1}P(A_1\cap A_2\cap\ldots\cap A_n). \tag{2.13}$$

Poznámka 2.3

Pravdepodobnosť, že nastane aspoň jeden z javov A,B,C je

$$P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)+$$
$$+P(A\cap B\cap C).$$

h) **Podmienená pravdepodobnosť** Pravdepodobnosť nastatia náhodného javu je ovplyvnená aj tým, či nastal iný náhodný jav. Napríklad pravdepodobnosť, že športovec je basketbalista je iná v prípade, že má výšku 210 cm a iná v prípade, keď má výšku 160 cm. Alebo pravdepodobnosť, že sa učiteľ školy stane jej riaditeľom je iná v prípade, že je muž a iná, ak je to žena. Hovoríme o **podmienenej pravdepodobnosti javu *A* za podmienky, že nastal jav *B***, čo označujeme $P(A/B)$. Je definovaná vzťahom

$$P(A/B)=\frac{P(A\cap B)}{P(B)},\ \text{kde}\ P(B)>0. \tag{2.14}$$

Poznámka 2.4

Ak Ω obsahuje len konečný počet prvkov, dá sa táto vlastnosť dokázať, ak nie, treba ju brať ako definíciu (axiómu $A4$).

Na jednoduchom príklade objasníme pôvod tejto definície.

V telocvični je 50 športovcov, 30 z nich má nad 200 cm a z nich je 20 basketbalistov. Náhodne sme vybrali športovca, ktorý mal nad 200 cm. Aká je pravdepodobnosť, že je to basketbalista?

Označíme jav B – športovec je basketbalista, jav V – je vysoký nad 200 cm. Zaujíma nás pravdepodobnosť $P(B/V)$. Všetkých vysokých nad 200 cm je 30, priaznivé výsledky predstavujú tí z tejto skupiny, ktorí sú basketbalisti, tých je 20. Preto $P(B/V)=\frac{20}{30}$. Upravíme toto číslo do tvaru $P(B/V)=\frac{20/50}{30/50}$. Číslo v čitateli je $P(B\cap V)$, v menovateli $P(V)$. Odtiaľ

$$P(B/V)=\frac{P(B\cap V)}{P(V)}. \tag{2.15}$$

i) **Veta o pravdepodobnosti prieniku (súčinu) javov** Pravdepodobnosť, že javy A,B nastanú súčasne je

$$P(A\cap B)=P(B)\cdot P(A/B) \tag{2.16}$$

alebo $$P(A\cap B)=P(A)\cdot P(B/A)\ . \tag{2.17}$$

Je to jednoduchý dôsledok predchádzajúcej definície podmienenej pravdepodobnosti.

j) Jej zovšeobecnením je táto vlastnosť:

$$P(A_1\cap A_2\cap\ldots\cap A_n)=P(A_1).P(A_2/A_1).P(A_3/A_1\cap A_2)\cdot\ldots$$

$$\ldots\cdot P(A_n/A_1\cap A_2\cap\ldots\cap A_{n-1}) \tag{2.18}$$

k) **Nezávislosť javov** Ak jav A nezávisí od toho, či nastane jav B, hovoríme, že javy A, B sú **nezávislé** a platí

$$P(A/B) = P(A), \text{ kde } P(A) > 0 \text{ a } P(B) > 0. \tag{2.19}$$

Pomocou podmienenej pravdepodobnosti sa dá dokázať, že javy A, B sú **nezávislé** práve vtedy, keď $P(A \cap B) = P(A).P(B)$.

Ak sú javy nezávislé, podstatne sa zjednoduší výpočet pravdepodobnosti prieniku javov:

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n). \tag{2.20}$$

Príklad 2.8

V trojdetných (Príklad 2.2) rodinách máme zistiť pravdepodobnosť javu A, že sa deti narodia v poradí *CHDCH* alebo *CHCHD*, pričom sme predpokladali, že pravdepodobnosť narodenia chlapca aj dievčaťa sú rovnaké, bez ohľadu na pohlavie predchádzajúcich súrodencov. $P(CH) = 0{,}5$ je pravdepodobnosť narodenia chlapca, $P(A) = 0{,}5$ je pravdepodobnosť narodenia dievčaťa.

Jav A sa dá vyjadriť ako zjednotenie dvoch disjunktných javov $A_1 \cup A_2$, kde A_1 je jav, že sa deti narodia v poradí *CHDCH* a A_2 je jav, že sa narodia v poradí *CHCHD*. Preto

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

Jav A_1 je prienik (súčin) troch nezávislých javov, preto

$P(A_1) = P(CH) \cdot P(D) \cdot P(CH) = 0{,}5 \cdot 0{,}5 \cdot 0{,}5 = 0{,}125$,

$P(A_2) = P(CH) \cdot P(CH) \cdot P(A) = 0{,}125$.

Nakoniec $P(A) = 0{,}125 + 0{,}125 = 0{,}25$.

l) **Veta o úplnej pravdepodobnosti** Nech javy $H_1, H_2, ..., H_n$ tvoria úplný systém javov. Potom platí

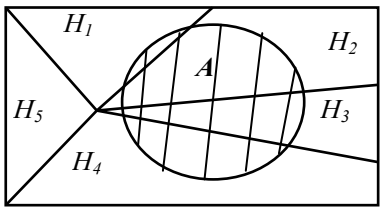
$$P(A) = \sum_{i=1}^{n} P(H_i) \cdot P(A/H_i). \tag{2.21}$$

Dôkaz Jav A je zjednotenie disjunktných javov (Obr. 2.4)

$A = (A \cap H_1) \cup (A \cap H_2) \cup ... \cup (A \cap H_n)$, preto

$$P(A) = P(A \cap H_1) + ... + P(A \cap H_n) =$$

$$= \sum_{i=1}^{n} P(A \cap H_i) = \sum_{i=1}^{n} P(H_i) \cdot P(A/H_i).$$



Obr. 2.4 Veta o úplnej pravdepodobnosti

m) **Bayesova veta** Nech javy $H_1, H_2, ..., H_n$ tvoria úplný systém javov. Potom platí

$$P(H_k/A) = \frac{P(H_k) \cdot P(A/H_k)}{P(A)}. \tag{2.22}$$

Dôkaz Vyplýva z definície podmienenej pravdepodobnosti.

$$P(H_k/A) = \frac{P(H_k \cap A)}{P(A)} = \frac{P(H_k) \cdot P(A/H_k)}{P(A)}.$$

Príklad 2.9

Pravdepodobnosť, že vodičom auta je muž, je 0,65. Pravdepodobnosť, že vodič - muž spôsobí dopravnú nehodu je 0,033 a pravdepodobnosť, že vodič - žena spôsobí dopravnú nehodu je 0,030. Na ceste sa stala dopravná nehoda. Aká je pravdepodobnosť, že nehodu spôsobila žena?

Označme jav N - stane sa dopravná nehoda a pomenujme pravdepodobnosti z príkladu:

$P(M) = 0{,}65$ je pravdepodobnosť, že vodič je muž,

$P(Z) = 0{,}35$ je pravdepodobnosť, že vodič je žena,

$P(N/Z) = 0{,}03$ je pravdepodobnosť, že nehodu spôsobí žena,

$P(N/M) = 0{,}033$ je pravdepodobnosť, že nehodu spôsobí muž.

Zaujíma nás pravdepodobnosť, že spôsobenú nehodu urobila žena, čo je

$$P(Z/N) = \frac{P(Z)\cdot P(N/Z)}{P(N)} = \frac{0{,}35\cdot 0{,}03}{0{,}35\cdot 0{,}03 + 0{,}65\cdot 0{,}033} = 0{,}3286\,,$$

kde pravdepodobnosť nehody $P(N)$ určíme pomocou vety o úplnej pravdepodobnosti.

n) **Bernoulliho schéma** Ak ten istý pokus, pri nezmenených podmienkach, opakujeme viackrát a výsledok jednotlivého pokusu je nezávislý od výsledkov predchádzajúcich pokusov, hovoríme o **opakovaných nezávislých pokusoch**. Príkladom je losovanie prvkov z nejakej množiny, kde prvky po pokuse vrátime späť.

Ak v každom pokuse nastane jav A s pravdepodobnosťou p, potom pravdepodobnosť toho, že v sérii n nezávislých pokusov nastane tento jav práve k-krát je

$$P_{k,n} = \binom{n}{k}\cdot p^k \cdot q^{n-k}, \quad \text{kde } 0 \le k \le n \quad \text{a} \quad q = 1 - p\,, \tag{2.23}$$

a kombinačné číslo $\binom{n}{k}$ je počet možností ako rozmiestniť v usporiadanej n-tici k prvkov.

Príklad 2.10

Zmeňme v rodine z Príkladu 2.2 pravdepodobnosti narodenia chlapca a dievčaťa (len pre lepšiu názornosť, podstatu riešenia to nezmení) takto: $P(CH) = 0{,}51$ a $P(D) = 0{,}49$. Predpokladali sme, že pohlavie druhého a tretieho dieťaťa v rodine

nezávisí od pohlavia skôr narodených súrodencov. Vypočítajte pravdepodobnosť javu *W*, že medzi 3 súrodencami je práve jedno dievča.

Symbolicky zapíšeme jav $W = (DCHCH) \cup (CHDCH) \cup (CHCHD)$, čo sú navzájom disjunktné javy a narodenie *CH* a *D* sú nezávislé.

Preto $P(W) = 0{,}49 \cdot 0{,}51 \cdot 0{,}51 + 0{,}51 \cdot 0{,}49 \cdot 0{,}51 + 0{,}51 \cdot 0{,}51 \cdot 0{,}49 = 3.(0{,}49)^1 \cdot (0{,}51)^2$.

Alebo $P_{2,3} = \binom{3}{2} \cdot (0{,}49)^1 \cdot (0{,}51)^2$.

o) Ak v tejto sérii pokusov výsledok každého pokusu závisí od výsledkov predchádzajúcich pokusov, hovoríme o **opakovaných závislých pokusoch.** Príkladom sú výbery prvkov z nejakej množiny, kde prvky po pokuse nevrátime späť.

p) Nech je daných *N* prvkov, z ktorých *K* prvkov má danú vlastnosť $(0 < K < N)$. Zo všetkých *N* prvkov vyberieme *n* prvkov $(0 < n < N)$. Potom pravdepodobnosť toho, že v tomto výbere má danú vlastnosť práve *k* prvkov je

$$P(A) = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}}, \text{ kde } 0 < K < N,\ 0 < n < N,\ 0 \le k \le \min\{K, n\}. \qquad (2.24)$$

Dôkaz Počet všetkých možností ako vybrať *n* prvkov z *N* prvkov je $\binom{N}{n}$, čo sú kombinácie bez opakovania. Priaznivé výsledky tvoria tie výbery, ktoré obsahujú *k* prvkov s danou vlastnosťou, zvyšných *n-k* prvkov nemá danú vlastnosť. Z *K* prvkov (nositeľov vlastnosti) sa *k* prvkov dá vybrať $\binom{K}{k}$ spôsobmi, zvyšných

n-k prvkov sa z *N–K* prvkov (nemajú danú vlastnosť) dá vybrať $\binom{N-K}{n-k}$ spô-

sobmi. Preto $P(A)=\dfrac{\binom{K}{k}.\binom{N-K}{n-k}}{\binom{N}{n}}$.

Príklady na precvičenie

2.11 Čo padne častejšie ako súčet na dvoch kockách , číslo 9 alebo 10? Ako je to na troch kockách?

2.12 V jednom úrade pracuje 7 žien a 3 muži. Je nutné znížiť stav zamestnancov o troch. Určite pravdepodobnosť toho, že pri náhodnom výbere budú prepustení 2 muži.

2.13 Na 6 lístkoch sú napísané písmená a zostavené slovo KARATE. Malé dieťa zostavilo z nich slovo RAKETA. Vie toto dieťa čítať?

2.14 V autobuse sa rozprávajú študenti o tom, ako bolo dnes na skúškach.
„Koľkí dnes boli?“
„ Dnes ôsmi na matike a štyria na fyzike .“
„ Koľkí urobili?“
„ Piati z matiky a traja z fyziky.“
„ A čo Eva?“
„ Eva urobila“.
Rozhodnite, na akej skúške bola Eva .

2.15 V antikvariáte sa cena knihy znižuje, ak má vytrhnutú aspoň jednu stranu, alebo sú jej strany popísané. Kníh, ktoré majú vytrhnuté strany je 20%,

kníh, ktoré sú popísané poznámkami je 30% a kníh bez chyby je 70%. Určite, aká je pravdepodobnosť, že náhodne vybraná kniha je popísaná, ale má všetky strany!

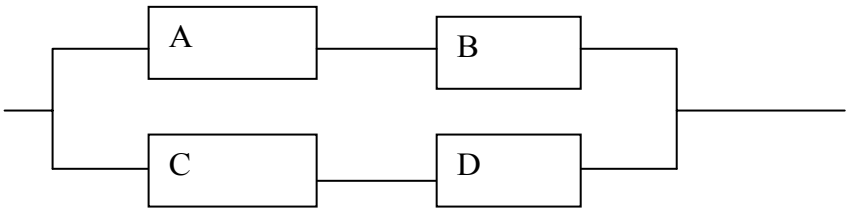
2.16 Čo je pravdepodobnejšie, vyhrať v hre s rovnocenným partnerom 3 partie zo 4 alebo 5 z 8 (remíza sa nepripúšťa)?

2.17 Dve lode musia vyložiť náklad v tom istom prístavisku. Príchody oboch lodí sú nezávislé a rovnako možné v priebehu celého dňa. Určite pravdepodobnosť toho, že niektorá z lodí bude musieť čakať na uvoľnenie prístaviska, ak jednu z lodí vykladajú 1 hodinu, druhú 2 hodiny.

2.18 Test sa skladá z 15 otázok, za každou je ponúknutých 6 možných odpovedí, z ktorých je len jedna správna. Aby študent urobil skúšku, je potrebné zaškrtnúť aspoň 10 správnych odpovedí. Aká je pravdepodobnosť, že študent skúšku urobí, ak

a) odpovede zaškrtáva úplne náhodne,
b) v každej otázke vie vždy vylúčiť 3 zlé odpovede, ale zo zvyšných troch si musí tipovať,
c) je tak pripravený na skúšku, že na 5 otázok pozná správnu odpoveď, ale na zvyšných 10 otázok si musí odpovede tipovať.

2.19 Pre sériovo - paralelnú sústavu na obrázku sú dané pravdepodobnosti bezporuchovej činnosti pre jednotlivé komponenty A, B, C, D postupne $p_1 = 0{,}85,\ p_2 = 0{,}91, p_3 = 0{,}89,\ p_4 = 0{,}95$ .Určite pravdepodobnosť bezporuchovej činnosti celej sústavy.



2.20 Pravdepodobnosť, že súčasná hospodárska depresia bude pokračovať aj na budúci rok, sa na základe odhadov ekonómov rovná 0,1. K miernemu oživeniu dôjde s pravdepodobnosťou 0,2, k výraznému oživeniu s 0,5 a k prudkému hospodárskemu rastu s pravdepodobnosťou 0,2. Pravdepodobnosť, že zisk firmy presiahne 5 miliónov korún je podľa podnikových manažérov v prvom prípade 0,05, v druhom 0,2, v treťom 0,7, vo štvrtom 0,95. Môže mať firma pri týchto podmienkach zisk väčší ako 5 miliónov ?

2.21 Nech jav A nastane v jednom pokuse s pravdepodobnosťou p. Minimálne koľko nezávislých pokusov treba uskutočniť, aby tento jav v nich nastal aspoň raz s pravdepodobnosťou P? Odvoďte vzťah!

2.22 Autobus prichádza na zastávku každé 4 minúty a električka (ktorej zastávka je vedľa autobusovej) každých 6 minút. Aká je pravdepodobnosť, že sa cestujúci dočká autobusu pred električkou.

2.23 V lotérii je n lósov, z ktorých m vyhráva. Niekto si kúpil k lósov. Aká je pravdepodobnosť, že vyhrá?

3. NÁHODNÁ PREMENNÁ

3.1 Pojem a vlastnosti náhodnej premennej

Ak ku každému výsledku náhodného pokusu priradíme nejaké reálne číslo, alebo ak výsledkom náhodného pokusu je hneď číslo, môžeme hovoriť o určitom zobrazení. Napr. po hode kockou priradíme tomuto hodu 100 násobok počtu bodov na kocke, alebo zistíme počet narodených chlapcov za týždeň, či počet chýb v diktáte, alebo čas čakania na električku. Výsledky pokusov sú náhodné, preto aj tieto číselné údaje, ktoré priradíme výsledkom pokusu majú náhodný charakter a nastávajú s istou pravdepodobnosťou. Tento typ zobrazenia sa volá **náhodná premenná (náhodná veličina)**. Presnejšie:

Ak je daný pravdepodobnostný priestor (Ω, S, P), kde Ω je konečná alebo spočítateľná množina, potom náhodná premenná (veličina) X je ľubovoľná reálna funkcia $X: \Omega \to R$.

Ak Ω je nespočítateľná, je situácia trochu zložitejšia. Napríklad: ak X je doba čakania na električku, ktorá chodí pravidelne každých 5 minút, tak nás nezaujíma, či čakáme presne 0 minút, alebo 2 minúty, ale skôr sa zaujímame, či doba čakania bude napr. menšia ako 2 minúty, alebo či budeme čakať od 1 do 2 minút.

Zaujímame sa teda o pravdepodobnosť, že $X \in \langle a,b)$, presnejšie

$$P(\{E \in \Omega; X(E) \in \langle a;b)\}).$$

Aby sa dala určiť táto pravdepodobnosť, musí byť $\{E \in \Omega,\ X(E) < x\} \in S$. Všeobecnú definíciu náhodnej premennej v tomto prípade sformulujeme nasledujúcim spôsobom. Náhodná premenná je ľubovoľné zobrazenie $X: \Omega \to R$ také, že pre každé $x \in R$ je množina tých elementárnych javov, ktoré sa zobrazia na číslo menšie ako *x,* prvkom množiny S, t.j. $\{E \in \Omega,\ X(E) < x\} \in S$.

Náhodná premenná je jednoznačne daná, ak poznáme hodnoty, ktoré nadobúda a pravdepodobnosti, s ktorými nadobúda tieto hodnoty. Tým je daný **zákon rozdelenia pravdepodobnosti náhodnej premennej**, alebo kratšie, **rozdelenie náhod-**

nej premennej. To, že náhodná premenná X nadobudne hodnotu x s pravdepodobnosťou p, označujeme $P(X = x) = p$.

Príklad 3.1

V každej trojdetnej rodine zistíme počet chlapcov medzi ich deťmi. To je náhodná premenná *X*. Počet chlapcov môže nadobudnúť len hodnoty 0, 1, 2, 3. Tento počet je náhodný, môžeme mu priradiť určitú pravdepodobnosť. Ak predpokladáme, že chlapec aj dievča sa narodia s rovnakou pravdepodobnosťou $p = \frac{1}{2}$, môžeme zistiť pravdepodobnosti, s akými nadobúda jednotlivé hodnoty táto náhodná premenná :

$P(X = 0) = \binom{3}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^3 = \frac{1}{8}$, je pravdepodobnosť, že nemajú chlapca,

$P(X = 1) = \binom{3}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^2 = \frac{3}{8}$, je pravdepodobnosť, že majú jedného chlapca.

Podobne

$P(X = 2) = \binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^1 = \frac{3}{8}$ a $P(X = 3) = \binom{3}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^0 = \frac{1}{8}$.

Náhodné premenné môžeme rozdeliť na dva základné typy :

- **Diskrétna náhodná premenná**, kde náhodná premenná nadobúda konečný počet hodnôt alebo spočítateľne veľa hodnôt (počet porúch zariadenia za deň, počet dobrých výrobkov v zásielke tovaru, počet obslúžených zákazníkov za 1 hodinu predaja).
- **Spojitá náhodná premenná,** kde náhodná premenná nadobúda všetky hodnoty z určitého intervalu (ohraničeného alebo neohraničeného). Príkladom je výsledok merania dĺžky, ak sa dopúšťame chyby $\pm\varepsilon$ cm. Výsledkom je náhodná premenná, ktorá nadobúda hodnoty z intervalu $\langle d-\varepsilon, d+\varepsilon \rangle$, kde d je skutočná dĺžka. Alebo čas, ktorý čakáme v rade, je náhodná premenná, ktorá môže nadobudnúť hodnoty z intervalu $\langle 0, t \rangle$.

3.2 Zákony rozdelenia pravdepodobnosti

Nech diskrétna náhodná premenná X nadobúda hodnoty $x_1, x_2, ..., x_n$ a nech poznáme aj pravdepodobnosti, s ktorými ich nadobúda, t.j. poznáme funkciu

$$P(X = x_i) = p_i, \quad \text{pričom } \sum_{i=1}^{n} p_i = 1, \text{ kde } i = 1, 2, ..., n .$$

Táto funkcia sa volá **pravdepodobnostná funkcia** a reprezentuje zákon rozdelenia pravdepodobnosti diskrétnej náhodnej premennej X**.**

Ak hodnoty x_i a k nim zodpovedajúce p_i usporiadame do tabuľky, dostaneme **pravdepodobnostnú tabuľku**, ako ďalší tvar zákona rozdelenia pravdepodobnosti diskrétnej náhodnej premennej *X*.

x_i	x_1	x_2	...	x_n
p_i	p_1	p_2	...	p_n

, kde $\sum_{i=1}^{n} p_i = 1$

Po grafickom znázornení bodov $[x_i, p_i]$ a ich spojení úsečkami dostaneme polygón rozdelenia pravdepodobnosti.

Príklad 3.2

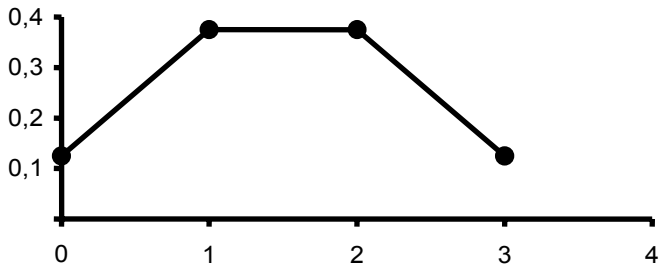
Pravdepodobnostná funkcia pre situáciu z Príkladu 3.1 má tvar

$$P(X = x) = \binom{3}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{3-x}, \text{ kde } x = 0, 1, 2, 3 .$$

Tab. 3.1 Pravdepodobnostná tabuľka k Príkladu 3.1

x_i	0	1	2	3
p_i	1/8	3/8	3/8	1/8

, kde $\sum_{i=1}^{n} p_i = 1$



Obr. 3.1 Polygón rozdelenia pravdepodobnosti k Príkladu 3.1

Ďalší zákon rozdelenia pravdepodobnosti náhodnej premennej je **distribučná funkcia**. **Je to funkcia $F : R \to R$, daná rovnosťou**

$$F(x) = P(X < x),\ x \in R. \tag{3.1}$$

Hodnota distribučnej funkcie v čísle *x* je teda pravdepodobnosť, že náhodná premenná *X* nadobudne hodnoty menšie ako zvolené *x*. Distribučnou funkciou sa dajú opísať diskrétne i spojité náhodné premenné. Ak náhodná premenná je daná pravdepodobnostnou tabuľkou, potom pre jej distribučnú funkciu platí $F(x) = \sum_{x_i < x} p_i$, čo znamená, že sčítame pravdepodobnosti tých hodnôt náhodnej premennej, ktoré sú menšie ako zvolené číslo *x*.

Vlastnosti distribučnej funkcie F náhodnej premennej X:

1. $\forall x \in R : 0 \le F(x) \le 1$ (3.2)
2. $\lim_{x \to -\infty} F(x) = 0$ (3.3)
3. $\lim_{x \to \infty} F(x) = 1$ (3.4)
4. $P(x_1 \le X < x_2) = F(x_2) - F(x_1)$, pre $x_1 < x_2$ (3.5)
5. Funkcia F je neklesajúca. (3.6)
6. Funkcia F je spojitá zľava v každom bode $x \in R$. (3.7)

Príklad 3.3

Hodnoty a graf distribučnej funkcie náhodnej premennej X z Príkladu 3.1 nájdeme osobitne v intervaloch oddelených hodnotami náhodnej premennej takto:

ak $x \le 0$, potom $F(x) = P(X < 0) = 0$;

ak $0 < x \le 1$, potom $F(x) = P(X < 1) = P(X = 0) = 0{,}125$;

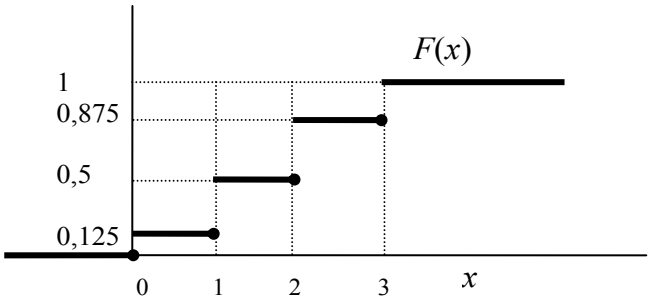
ak $1 < x \le 2$, potom $F(x) = P(X < 2) = P(X = 0) + P(X = 1) =$

$= 0{,}125 + 0{,}375 = 0{,}5$;

ak $2 < x \le 3$, potom $F(x) = P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2) =$

$= 0{,}125 + 0{,}375 + 0{,}375 = 0{,}875$;

ak $x > 3$, potom $F(x) = P(I) = 1$. Graf tejto funkcie je na Obr. 3.2.



Obr. 3.2 Graf distribučnej funkcie k Príkladu 3.3

Graf distribučnej funkcie ľubovoľnej diskrétnej premennej má tvar podobný ako na Obrázku 3.2. Je to „schodovitá“, neklesajúca funkcia, ktorá nie je spojitá v bodoch x_i. Hodnota „skoku“ v bode x_i je pravdepodobnosť p_i.

Príklad 3.4

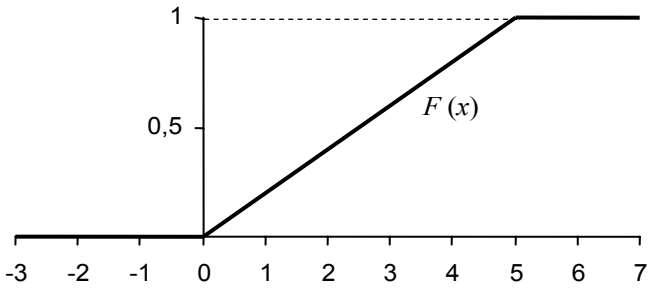
Náhodná premenná X je doba čakania na električku, ktorá chodí pravidelne každých 5 minút. Nájdite jej distribučnú funkciu, nakreslite jej graf.

Ak $x \le 0$, tak $F(x) = P(X < 0) = 0$, lebo čas čakania nemôže byť záporný.

Ak $x > 5$, potom $F(x) = P(X < x) = 1$, lebo čas čakania je najviac 5 minút.

Ak $0 < x \le 5$, potom pravdepodobnosť rastie lineárne s časom. Teda :

$$F(x)=\begin{cases} 0 & pre\ \ x\le 0 \\ x/5 & pre\ \ 0<x\le 5 \\ 1 & pre\ \ x>5 \end{cases}$$



Obr. 3.3 Graf distribučnej funkcie k Príkladu 3.4

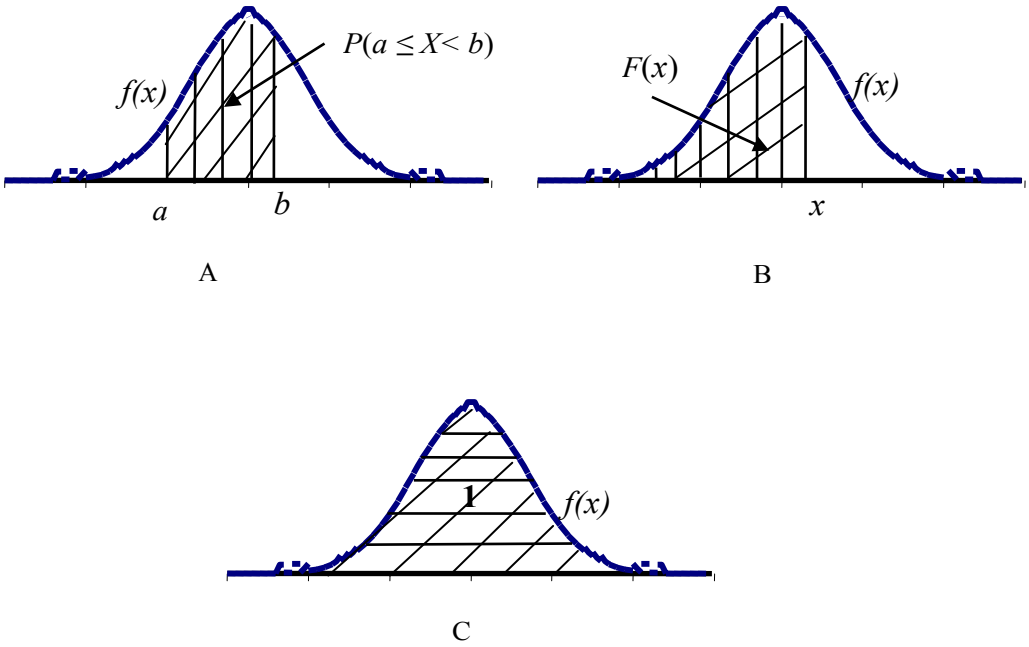
Z distribučnej funkcie $F(x)$ náhodnej premennej je odvodená definícia spojitej náhodnej premennej a ďalší zákon rozdelenia pravdepodobnosti náhodnej premennej – **hustota rozdelenia pravdepodobnosti spojitej náhodnej premennej *X***. Náhodnú premennú *X* s distribučnou funkciou *F* nazývame **spojitá**, ak existuje taká integrovateľná, nezáporná funkcia *f*, že pre všetky $x \in R$ platí

$$F(x)=P(X<x)=\int_{-\infty}^{x} f(t)dt\,. \tag{3.8}$$

Funkciu *f* nazývame **hustota** rozdelenia pravdepodobnosti spojitej náhodnej premennej *X*.

Vlastnosti funkcie hustoty pravdepodobnosti *f* spojitej náhodnej premennej *X* majú názornú geometrickú interpretáciu, jasnú z Obr. 3.4 :

1. $P(a\le X<b)=\int_{a}^{b} f(x)dx$ (Obr. 3.4A)
2. $F(x)=\int_{-\infty}^{x} f(t)dt$ (Obr. 3.4B)
3. $\int_{-\infty}^{\infty} f(x)dx=1$ (Obr. 3.4C)
4. $f(x)=F'(x)$, okrem bodov nespojitosti funkcie *f*. (3.9)



Obr. 3.4 Vlastnosti funkcie hustoty

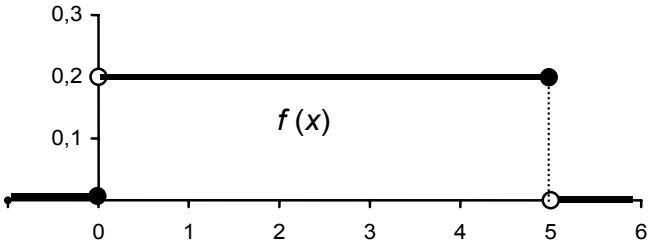
Pre spojitú náhodnú premennú platí, $P(X = x) = 0$. To znamená, pravdepodobnosť, že premenná nadobudne jednu konkrétnu hodnotu je nula. Z toho vyplýva, že platí:

$$P(a < X < b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b) = F(b) - F(a).$$

Príklad 3.5

Náhodná premenná X je doba čakania na električku, ktorá chodí pravidelne každých 5 minút. Nájdite funkciu hustoty rozdelenia pravdepodobnosti a nakreslite jej graf. V predchádzajúcom príklade sme našli distribučnú funkciu $F(x)$, podľa (3.9) jej derivovaním nájdeme funkciu hustoty *f*.

$$f(x) = \begin{cases} 0 & pre\ x \le 0 \\ 1/5 & pre\ 0 < x \le 5 \\ 0 & pre\ x > 5 \end{cases}$$



Obr. 3.5 Funkcia hustoty rozdelenia pravdepodobnosti z Príkladu 3.5

3.3 Číselné charakteristiky rozdelenia náhodnej premennej

Zákon rozdelenia pravdepodobnosti charakterizuje danú náhodnú premennú úplne. Niekedy však tento zákon nepoznáme, preto je vhodnejšie charakterizovať náhodnú premennú pomocou iných číselných údajov, ktoré vystihujú základné vlastnosti daného rozdelenia. Sú to **číselné charakteristiky rozdelenia náhodnej premennej.** Delia sa na dve veľké skupiny :

- charakteristiky polohy
- charakteristiky variability

3.3.1 Charakteristiky polohy

Opisujú „polohu“ náhodnej premennej X na číselnej osi, kde na osi sa „koncentrujú“ hodnoty náhodnej premennej. Najdôležitejšia z nich je **stredná hodnota** (**MEAN,** $E(X), \mu(X)$). Udáva akýsi stred (ťažisko) celého rozdelenia, okolo ktorého hodnoty náhodnej premennej náhodne kolíšu pri opakovaných pozorovaniach. Pre diskrétnu náhodnú premennú je definovaná vzťahom

$$E(X)=\sum_{i=1} x_i p_i \quad i=1,2,... \quad \text{(rad musí byť konvergentný)} \tag{3.10}$$

Pre spojitú náhodnú premennú je definovaná vzťahom

$$E(X)=\int_{-\infty}^{\infty} xf(x)dx \qquad \text{(integrál musí byť konvergentný).} \tag{3.11}$$

Poznámka 3.1

Stredná hodnota má taký istý rozmer ako náhodná premenná.

Predstavme si diskrétnu náhodnú premennú, ktorá nadobúda všetky svoje hodnoty $x_1, x_2, ...x_n$ s rovnakou pravdepodobnosťou p. Platí:

$$1 = p_1 + p_2 + ... + p_n = p + p + ... + p = np,$$

$$p = 1/n$$

Tabuľka rozdelenia pravdepodobností má tvar:

x_i	x_1		...	x_n
p_i	$1/n$	$1/n$	...	$1/n$

V tomto prípade $E(X) = \sum_{i=1}^{n} x_i . \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$, čo je aritmetický priemer čísel $x_1, x_2, ..., x_n$. V praxi sa stredná hodnota často stotožňuje s aritmetickým priemerom. Ak hodnoty $x_1, x_2, ...x_n$ nie sú rovnako pravdepodobné, posúva sa stredná hodnota k tým hodnotám, ktoré sú pravdepodobnejšie.

Vlastnosti strednej hodnoty:

- $E(C) = C$, kde C je konštanta
- $E(AX + B) = AE(X) + B$, kde A,B sú konštanty
- $E(X + Y) = E(X) + E(Y)$, kde X,Y sú náhodné premenné

Modus náhodnej premennej *X* (**MODE,** *Mo*) je tá hodnota náhodnej premennej, ktorú premenná nadobúda s najväčšou pravdepodobnosťou. Pre spojitú náhodnú premennú X je to tá hodnota náhodnej premennej, v ktorej má hustota pravdepodobnosti $f(x)$ lokálne maximum.

Rozdelenie nemusí mať modus (antimodálne rozdelenie), ale môže mať aj viac ako jeden modus (polymodálne rozdelenie).

Medián (**MEDIAN**, *Me*) diskrétnej náhodnej premennej X je tá hodnota náhodnej premennej, pre ktorú platí

$$P(X < Me) \le \frac{1}{2} \qquad \text{a} \qquad P(X \le Me) \ge \frac{1}{2} \ . \tag{3.12}$$

Pre spojitú náhodnú premennú je medián *Me* tá hodnota náhodnej premennej, ktorá vyhovuje vzťahu

$$F(Me) = P(X < Me) = \int_{-\infty}^{Me} f(x)dx = \frac{1}{2}. \tag{3.13}$$

Príklad 3.5

S využitím pravdepodobnostnej tabuľky z Príkladu 3.1 určíme strednú hodnotu, modus a medián premennej X - počtu chlapcov v trojdetnej rodine.

x_i	0	1	2	3
p_i	1/8	3/8	3/8	1/8

$$E(X) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = 1{,}5 \ .$$

Premenná má 2 modusy, prvý $Mo = 1$, druhý $Mo = 2$. V tomto prípade určíme aj dva mediány $Me = 1$ a druhý $Me = 2$.

3.3.2 Charakteristiky variability

Tieto charakteristiky informujú o tom, ako ďaleko od strednej hodnoty sú rozptýlené hodnoty náhodnej premennej. Napríklad z dvoch strelcov, ktorí strieľali na rovnaký terč, prvý nastrieľal tromi ranami 7, 8, 9 bodov, druhý 6,8,10 bodov, sa nám zdá, že prvý strieľa lepšie. Jeho streľba je sústredenejšia, menej rozptýlená, hoci stredná hodnota je u oboch strelcov 8. Najdôležitejšia charakteristika variability je **rozptyl** (**VARIANCE**, disperzia, $D(X), \sigma^2(X)$). Je definovaný

$$D(X) = E\left\{\left[X - E(X)\right]^2\right\}, \tag{3.14}$$

teda ako stredná hodnota druhej mocniny odchýlky náhodnej premennej od jej strednej hodnoty. Špeciálne, rozptyl diskrétnej náhodnej premennej X je

$$D(X) = \sum [x_i - E(X)]^2 \cdot p_i \tag{3.15}$$

a rozptyl spojitej náhodnej premennej X je

$$D(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x) dx. \tag{3.16}$$

Vlastnosti rozptylu:

- $D(C) = 0$, kde C je konštanta
- $D(AX + B) = A^2 D(X)$, kde A, B sú konštanty
- $D(X) = E(X^2) - E^2(X)$, kde $E^2(X) = [E(X)]^2$ (3.17)

Poznámka 3.2

V praxi sa na výpočet rozptylu používa vzťah (3.17). Rozptyl má rozmer druhej mocniny rozmeru náhodnej premennej, čo nie je vo fyzikálnych úlohách vhodné, preto sa zavádza ďalšia charakteristika variability s rovnakým rozmerom ako má samotná náhodná premenná. Je to **štandardná odchýlka** (**STANDARD DEVIATION**, stredná kvadratická odchýlka, smerodajná odchýlka, $\sigma(X)$), ktorá definovaná vzťahom

$$\sigma(X) = \sqrt{D(X)}. \tag{3.18}$$

Príklad 3.6

Hrám hru s jedným protihráčom. Ak na kocke padne jednotka, súper mi zaplatí 4 Sk, ak padne dvojka alebo trojka, súper mi zaplatí 1 Sk, ak padne štvorka, súper mi zaplatí 2 Sk, v ostatných prípadoch ja platím súperovi 3 Sk. Je pre mňa táto hra výhodná? Aký mám priemerný zisk (stratu) po jednom hode? Akú štandardnú odchýlku má môj zisk?

Riešenie Náhodná premenná je môj zisk z jednej hry. Zostavíme tabuľku rozdelenia pravdepodobnosti

zisk	-3	1	2	4
p_i	2/6	2/6	1/6	1/6

$$E(X)=(-3)\cdot\frac{2}{6}+\frac{2}{6}+2\cdot\frac{1}{6}+4\cdot\frac{1}{6}=\frac{2}{6},$$

$$D(X)=E(X^2)-E^2(X),\quad E(X^2)=9\cdot\frac{2}{6}+\frac{2}{6}+4\cdot\frac{1}{6}+16\cdot\frac{1}{6}=\frac{40}{6},$$

$$D(X)=\frac{40}{6}-\left(\frac{2}{6}\right)^2=\frac{236}{36},\quad \sigma(X)=\sqrt{D(X)}=2{,}56.$$

Môj priemerný zisk z jednej hry je 2/6 Sk, čo je asi 0,33 Sk, štandardná odchýlka je veľká, až 2,56 Sk.

3.3.3 Začiatočné a centrálne momenty

Sú to všeobecnejšie číselné charakteristiky, pomocou ktorých sa dajú zaviesť ďalšie typy číselných charakteristík, ktoré názorným spôsobom popisujú nejaké rozdelenie.

Začiatočný moment k-teho rádu (ν_k) diskrétnej náhodnej premennej je definovaný vzťahom

$$\nu_k=\sum_i x_i^k p_i$$

a spojitej náhodnej premennej

$$\nu_k=\int_{-\infty}^{\infty} x^k f(x)dx.$$

Pre $k=0$ je $\nu_0=1$ a pre $k=1$ je $\nu_1=E(X)=\mu$.

Centrálny moment k-teho rádu (μ_k) diskrétnej náhodnej premennej je definovaný vzťahom

$$\mu_k = \sum_i \left[x_i - E(X)\right]^k p_i$$

a spojitej náhodnej premennej

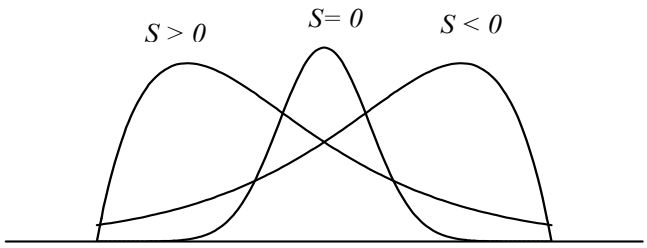
$$\mu_k = \int_{-\infty}^{\infty} \left[x - E(X)\right]^k f(x)dx .$$

Pre $k = 0,1,2$ dostaneme $\mu_0 = 1,\ \mu_1 = 0,\ \mu_2 = D(X)$.

Koeficient šikmosti (**SKEWNESS**, asymetria, $S(X)$) charakterizuje asymetriu rozdelenia. Je definovaný vzťahom

$$S(X) = \frac{\mu_3}{\sigma^3(X)} . \tag{3.19}$$

Geometrická interpretácia koeficientu v prípade spojitej premennej je jasná z Obr. 3.6. V prípade diskrétnej premennej takáto interpretácia nie je spoľahlivá.



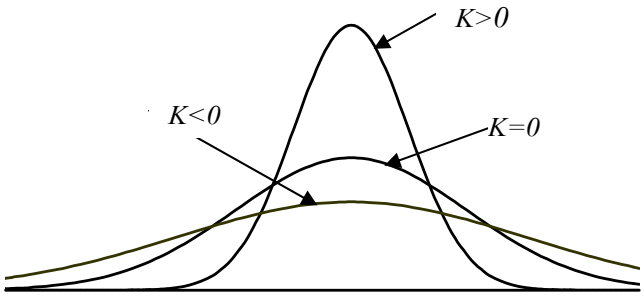
Obr. 3.6 Geometrická interpretácia koeficientu šikmosti

Koeficient špicatosti (**KURTOSIS**, exces, $K(X)$), ktorý charakterizuje špicatosť, strmosť rozdelenia je určený vzťahom

$$K(X) = \frac{\mu_4}{\sigma^4(X)} - 3 . \tag{3.20}$$

Porovnáva dané rozdelenie s grafom funkcie hustoty normovaného normálneho rozdelenia, ktoré definujeme v časti 3.5.1:

- kladná hodnota určuje strmšie rozdelenie ako normálne,
- záporná hodnota určuje plochšie rozdelenie ako normálne,
- nulovú hodnotu má rozdelenie rovnako strmé ako normálne.



Obr. 3.7 Geometrická interpretácia koeficientu špicatosti

3.3.4 Kvantily

U spojitých náhodných premenných sa v štatistike často využívajú **kvantily**. Používajú sa pri intervalovom a bodovom odhade parametrov, pri testovaní štatistických hypotéz.

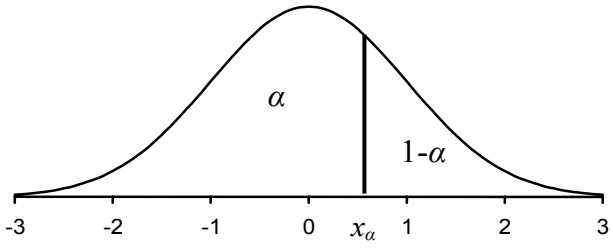
Nech $0 < \alpha < 1$ je ľubovoľné reálne číslo. $100 \cdot \alpha$ - **percentný kvantil** je tá hodnota x_α náhodnej premennej X, pre ktorú platí

$$P(X \le x_\alpha) = \alpha .$$

Z geometrického hľadiska rozdelí kolmica na os x v bode x_α obsah pod grafom funkcie hustoty na 2 časti, ľavá časť má obsah α, vpravo ležiaca časť má obsah $(1-\alpha)$ (Obr. 3.8). Niekedy sú tabelované (hlavne v staršej literatúre) tzv. **kritické hodnoty**. $100.\alpha$ **- percentná kritická hodnota** je tá hodnota x_α náhodnej premennej X, pre ktorú platí

$$P(X \ge x_\alpha) = \alpha .$$

Z geometrického hľadiska rozdelí kolmica na os x v bode x_α obsah pod grafom funkcie hustoty na 2 časti, ľavá časť má obsah $1-\alpha$, vpravo ležiaca časť má obsah α. Je jasné, že α – kvantil je $(1-\alpha)$- kritická hodnota.



Obr. 3.8 Geometrický význam kvantilu

3.4 Niektoré rozdelenia pravdepodobnosti diskrétnej náhodnej premennej

3.4.1 Diskrétne rovnomerné rozdelenie R(*n*)

Náhodná premenná môže nadobúdať hodnoty 1, 2, 3, ... , *n*. Ak pravdepodobnosť výskytu ktorejkoľvek z nich je rovnaká t.j. $1/n$, hovoríme o rovnomernom rozdelení premennej. Číslo *n* je jediný parameter rozdelenia.

Napríklad, náhodná premenná X je počet padnutých bodov na hracej kocke. Výsledkom po hode kockou je jeden zo 6 javov, každý sa stane s pravdepodobnosťou $1/6$.

Tab. 3.3 Pravdepodobnostná tabuľka rovnomerného rozdelenia

x_i	1	2	3	4	...	n
p_i	$1/n$	$1/n$	$1/n$	$1/n$	...	$1/n$

Číselné charakteristiky: $E(X)=\frac{n+1}{2}$, $D(X)=\frac{n^2-1}{12}$

3.4.2 Alternatívne rozdelenie A(*p*)

Máme jeden pokus, v ktorom skúmaný jav *A* môže nastať alebo nenastať. Napríklad pohlavie vylosovanej osoby je mužské, vybraný výrobok je dobrý, športový klub vyhral zápas, študent urobil skúšku. Teda náhodná premenná X s pravdepodobnosťou p nadobúda hodnotu $x=1$ (ak jav nastal) a s pravdepodobnosťou $q=1-p$ hodnotu $x=0$ (ak jav nenastal). Číslo *p* je jediný parameter rozdelenia.

Tab. 3.4 Pravdepodobnostná tabuľka alternatívneho rozdelenia

x_i	0	1
p_i	1-*p*	*p*

Číselné charakteristiky:

$E(X)=p$, $D(X)=pq$, $S(X)=\frac{q-p}{\sqrt{pq}}$, $K(X)=\frac{1-6pq}{pq}$

3.4.3 Binomické rozdelenie Bi(*p*, *n*)

Súvisí s vetou o opakovaní nezávislých pokusov. Napríklad, zaujímame sa o počet chlapcov v trojdetnej rodine, alebo o počet dobrých výrobkov v kontrolnej vzorke. Ak ako náhodnú premennú X označíme počet výskytov javu A v sérii *n* nezávislých pokusov, a ak v každom pokuse nastane tento jav s pravdepodobnosťou p, potom náhodná premenná *X* má binomické rozdelenie pravdepodobnosti a hodnoty $x=0,1,2,...,n$ nadobúda s pravdepodobnosťami:

$$p_x=P(X=x)=\binom{n}{x}p^x q^{n-x}, \qquad (3.21)$$

kde $p\in(0,1)$ a $q=1-p$. Čísla *p, n* sú parametre rozdelenia.

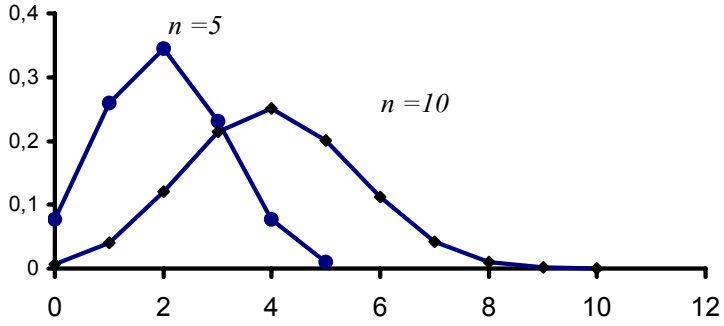
Číselné charakteristiky :

$$E(X) = np\,, \qquad D(X) = npq\,, \qquad S(X) = \frac{q-p}{\sqrt{npq}}\,, \qquad K(X) = \frac{1-6pq}{npq}$$

Príklad 3.8

Znázornime pravdepodobnostné tabuľky a polygóny binomického rozdelenia pravdepodobnosti pre

a) $n = 4;\ \ p = 0{,}4;\ \ q = 0{,}6$ b) $n = 10;\ \ p = 0{,}4;\ \ q = 0{,}6$



Obr. 3.9 Polygóny rozdelenia pravdepodobnosti k Príkladu 3.8

Tab. 3.5 Pravdepodobnostné tabuľky k Príkladu 3.8

x_i	0	1	2	3	4	5	6	7	8	9	10
a) p_i	0,078	0,259	0,346	0,230	0,077	0,010	-	-	-	-	-
b) p_i	0,006	0,040	0,121	0,215	0,251	0,201	0,111	0,042	0,011	0,002	0,000

3.4.4 Poissonovo rozdelenie Po(λ)

Je špeciálne rozdelenie pre "maličké " pravdepodobnosti, jeho pravdepodobnostná funkcia sa používa na výpočet pravdepodobností zriedkavých udalostí.

Náhodná premenná X má Poissonovo rozdelenie s parametrom $\lambda > 0$, ak nadobúda hodnoty $x = 0, 1, 2, 3, \ldots$ s pravdepodobnosťami:

$$p_x = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{3.22}$$

Číselné charakteristiky :

$E(X) = \lambda$, $D(X) = \lambda$, $S(X) = \frac{1}{\sqrt{\lambda}}$, $K(X) = \frac{1}{\lambda}$

Je to tiež rozdelenie počtu výskytov javu A za jednotku času. Ak priemerný počet objavení sa javu A za jednotku času je λ, potom pravdepodobnosť toho, že za jednotku času daný jav A nastane x - krát, má Poissonovo rozdelenie s parametrom λ. Príklady náhodných veličín s Poissonovým rozdelením:

Počet predmetov nájdených a odovzdaných denne v hypermarkete.

Počet telefonických hovorov v istom intervale.

Počet úrazov robotníkov za rok.

Poznámka 3.3

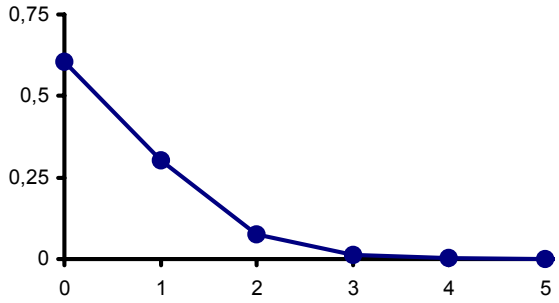
- Aj v binomickom rozdelení existujú zriedkavé javy, ale vždy sa doplňuje nastatie javu a nenastatie javu. Na mnohé javy sa však nedá aplikovať binomické rozdelenie. Napríklad, ak počas búrky bolo 10 bleskov, koľko bleskov nebolo? Ak sa od 10. hodiny do 12. hodiny narodilo 5 detí, koľko sa nenarodilo?
- Ak stredná hodnota a disperzia u neznámeho rozdelenia sú rovnaké, prichádza do úvahy práve Poissonovo rozdelenie.
- Poissonovo rozdelenie je rozdelenie „zriedkavých" javov.

Príklad 3.9

Pre $\lambda = 0{,}5$ uvádzame pravdepodobnostnú tabuľku a polygón rozdelenia pravdepodobnosti.

Tab. 3.6 Tabuľka rozdelenia pravdepodobnosti

x_i	0	1	2	3	4	5	...	...
p_i	0,607	0,303	0,076	0,013	0,002	1,6E-4	...	...



Obr. 3.10 Polygón rozdelenia pravdepodobnosti

Príklad 3.10

Povahu rozdelenia dobre ilustruje príklad z histórie. Nemecký štatistik Bartkiewicz zisťoval 10 rokov v 20 armádnych zboroch údaje o počte osôb zabitých úderom konského kopyta. Výsledky sú v tabuľke rozdelenia pravdepodobnosti, kde sme pravdepodobnosť nahradili relatívnou početnosťou:

Počet mŕtvych	**Počet prípadov**	**Relat. početnosť**	**Pravdepodobnosť**
0	109	109/200 = 0,545	0,543
1	65	65/200 = 0,325	0,331
2	22	22/200 = 0,11	0,101
3	3	3/200 = 0,015	0,021
4	1	1/200 = 0,005	0,003
	Spolu 200		

Stredná hodnota počtu mŕtvych za rok na jeden zbor je podľa tejto tabuľky 0,61. Ak sa tieto nehody dajú popísať Poissonovým rozdelením, potom $\lambda = 0{,}61$.

Z funkcie $p_x = P(X = x) = \frac{(0{,}61)^x \cdot e^{-0{,}61}}{x!}$ pre $x = 0, 1, 2, 3, 4$ vypočítame pravdepodobnosti $p_0 = 0{,}543$, $p_1 = 0{,}331$, $p_2 = 0{,}101$, $p_3 = 0{,}021$, $p_4 = 0{,}003$. Ak porovnáme posledné dva stĺpce tabuľky, vidíme, že toto rozdelenie dobre vystihlo rozdelenie počtu mŕtvych.

3.5 Niektoré rozdelenia pravdepodobnosti spojitej náhodnej premennej

3.5.1 Normálne rozdelenie (Gaussovo, Z – rozdelenie) N(μ,σ^2)

Je najdôležitejšie rozdelenie pravdepodobnosti z mnohých typov spojitých náhodných premenných, pretože mnohé náhodné premenné v technike, prírodných a spoločenských vedách majú toto rozdelenie, alebo dajú sa ním aproximovať iné spojité alebo diskrétne rozdelenia. Toto rozdelenie má aj náhodná premenná, ktorej hodnoty sú tvorené súčtom veľkého počtu malých a vzájomne nezávislých veličín. Normálne rozdelenie je závislé od dvoch parametrov μ , σ , označuje sa $N(\mu, \sigma^2)$. Určujú graf funkcie hustoty rozdelenia (Obr. 3.11, Obr. 3.12).

Spojitá náhodná premenná X má **normálne rozdelenie pravdepodobnosti** s parametrami μ, σ^2, ak jej **funkcia hustoty** má tvar:

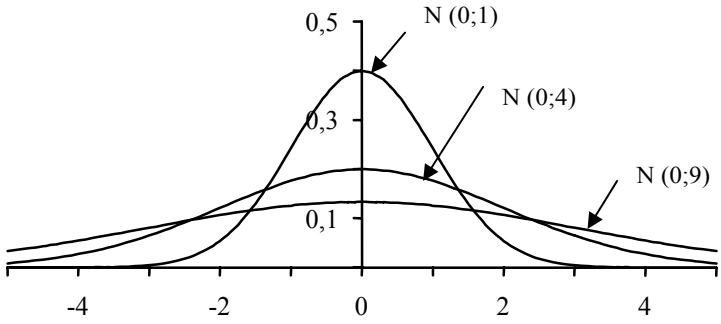
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \; x \in R \tag{3.23}$$

Distribučná funkcia rozdelenia N(μ , σ^2) má tvar

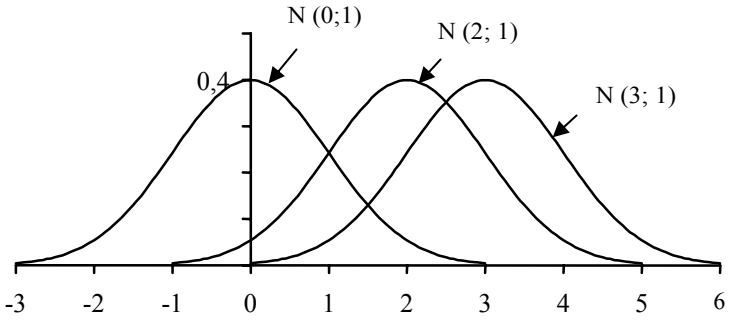
$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt \tag{3.24}$$

Číselné charakteristiky :

$$E(X)=\mu, \qquad D(X)=\sigma^2, \qquad S(X)=0, \qquad K(X)=0$$



Obr. 3.11 Graf funkcie hustoty normálneho rozdelenia pre rôzne σ

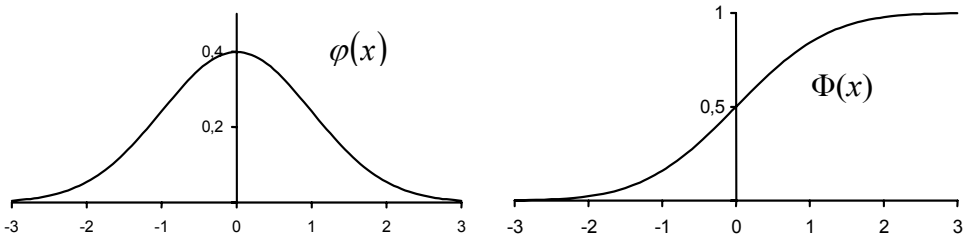


Obr. 3.12 Graf funkcie hustoty normálneho rozdelenia pre rôzne μ

Špeciálne pre $\mu=0$ a $\sigma^2=1$ dostaneme **normované (štandardizované) normálne rozdelenie,** ktoré sa označuje $N(0,1)$. Pre toto rozdelenie sa označujú funkcia hustoty $\varphi(x)$, distribučná funkcia $\Phi(x)$, graf funkcie hustoty sa volá Gaussova krivka (Obr. 3.13).

Vlastnosti normovaného normálneho rozdelenia:

- $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, $x\in R$, (3.25)
- $\varphi(x)$ je párna, t.j. $\varphi(-x)=\varphi(x)$, (3.26)
- v bode $x=0$ má maximum $\varphi(0)=\frac{1}{\sqrt{2\pi}}\doteq 0{,}4$,
- inflexné body má v číslach $x=-1,\ x=1$,
- os x je asymptota grafu funkcie $\varphi(x)$,
- $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$, (3.27)
- $\Phi(-x)=1-\Phi(x)$. (3.28)



Obr. 3.13 **F**unkcia hustoty $\varphi(x)$ a distribučná funkcia $\Phi(x)$ rozdelenia N(0,1)

V štatistických tabuľkách sa nachádzajú funkcia hustoty $\varphi(x)$, distribučná funkcia $\Phi(x)$, kvantily normovaného normálneho rozdelenia, v programových štatistických balíkoch sú zaradené funkcia hustoty, distribučná funkcia, kvantily ľubovoľného normálneho rozdelenia. Vzhľadom na koeficienty asymetrie a špicatosti sa toto rozdelenie chápe ako základné, s ktorým sa porovnávajú iné rozdelenia (pozri časť číselné charakteristiky náhodnej premennej).

Vzťahy medzi ľubovoľným $N(\mu,\sigma^2)$ a $N(0,1)$:

- $f(x)=\frac{1}{\sigma}\,\varphi\left(\frac{x-\mu}{\sigma}\right)$ (3.29)
- $F(x)=\Phi\left(\frac{x-\mu}{\sigma}\right)$ (3.30)
- $P(a\le X<b)=F(b)-F(a)=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)$ (3.31)

Platnosť týchto vzťahov dokážeme, ak vo funkcii hustoty rozdelenia $\mathrm{N}(\mu,\sigma^2)$ zvolíme substitúciu $z=\frac{x-\mu}{\sigma}$.

Príklad 3.11

Náhodná premenná má normálne rozdelenie N(1,1). Určite :

a) hodnoty $f(1),\ F(1)$,

b) pravdepodobnosť, že premenná nadobudne hodnotu z intervalu $(1,2)$,

c) pravdepodobnosť, že absolútna hodnota náhodnej premennej nadobudne hodnoty väčšie ako 2.

Riešenie Využijeme štatistické tabuľky v [4], [15], [16], prípadne štatistické funkcie ponúkané EXCELom a predchádzajúce vlastnosti normálneho rozdelenia.

a) $f(1)=\varphi\left(\frac{1-1}{2}\right)=\varphi(0)=0{,}3989$ $\qquad F(1)=\Phi\left(\frac{1-1}{2}\right)=\Phi(0)=0{,}5$

b) $P(1<X<2)=\Phi\left(\frac{2-1}{1}\right)-\Phi\left(\frac{1-1}{1}\right)=\Phi(1)-\Phi(0)=0{,}8413-0{,}5=0{,}3413$

c) $P(|X|>2)=1-P(|X|\le 2)=1-P(-2\le X\le 2)=1-\Phi\left(\frac{2-1}{1}\right)+\Phi\left(\frac{-2-1}{1}\right)=$

$=1-\Phi(1)+\Phi(-3)=1-\Phi(1)+1-\Phi(3)=1-0{,}8413+1-0{,}9987=0{,}16$

Príklad 3.12

Praktické informácie o $N(m,\sigma^2)$ dostaneme odpoveďou na otázku, aká je $P(|X-m|<\varepsilon)$, kde ε je vopred dané, ľubovoľné kladné číslo.

$$P(|X-m|<\varepsilon)=P(-\varepsilon<X-m<\varepsilon)=P(m-\varepsilon<X<m+\varepsilon)=$$

$$\Phi\left(\frac{m+\varepsilon-m}{\sigma}\right)-\Phi\left(\frac{m-\varepsilon-m}{\sigma}\right)=\Phi\left(\frac{\varepsilon}{\sigma}\right)-\Phi\left(-\frac{\varepsilon}{\sigma}\right)=\Phi\left(\frac{\varepsilon}{\sigma}\right)-1+\Phi\left(\frac{\varepsilon}{\sigma}\right)=$$

$$=2\Phi\left(\frac{\varepsilon}{\sigma}\right)-1 .$$

Zvoľme za ε postupne hodnoty σ, 2σ, 3σ. Postupne dostaneme nasledujúce tvrdenia:

$$P(|X-m|<\sigma)=2\Phi(1)-1=2.0{,}8413-1=0{,}6826$$

$$P(|X-m|<2\sigma)=2\Phi(2)-1=2.0{,}9972-1=0{,}9544$$

$$P(|X-m|<3\sigma)=2\Phi(3)-1=2.0{,}9987-1=0{,}9974$$

To znamená, že z celkovej plochy ohraničenej grafom funkcie $f(x)$ a osou x sa

- nad intervalom $(m-\sigma,m+\sigma)$ nachádza 68,26% plochy,
- nad intervalom $(m-2\sigma,m+2\sigma)$ nachádza 95,44% plochy,
- nad intervalom $(m-3\sigma,m+3\sigma)$ nachádza 99,74% plochy, t.j. takmer všetky hodnoty náhodnej premennej s $N(m,\sigma^2)$ ležia v tomto intervale. Táto posledná vlastnosť sa volá **„ pravidlo troch sigma“**. Využíva sa v praxi na prvý odhad štandardnej odchýlky pre náhodnú premennú, u ktorej predpokladáme normálne rozdelenie. Stačí, ak šírku intervalu, ktorý je určený najmenšou a najväčšou hodnotou premennej vydelíme šiestimi, výsledok je odhadom parametra σ.

EXCEL

Po voľbe ***Prilepiť funkciu/Štatistické*** sú pre normálne rozdelenie k dispozícii štatistické tabuľky, ich prehľad je v Tab. 3.6.

Tab. 3.6 Tabuľky normálneho rozdelenia v EXCELi

Označenie funkcie a zadávané parametre	**Výsledok**
NORMDIST (x, μ, σ, 1)	$F(x)$, hodnota distribučnej funkcie v čísle x pre N(μ, σ^2)
NORMDIST (x, μ, σ, 0)	$f(x)$, hodnota funkcie hustoty v čísle x pre N(μ, σ^2)
NORMINV (p, μ, σ)	$p\times100\%$-ný kvantil pre N(μ, σ^2), p- pravdepodobnosť
NORMSDIST (z)	$\Phi(z)$, hodnota distribučnej funkcie v čísle z pre N(0, 1)
NORMSINV (p)	$p\times100\%$-ný kvantil pre N(0, 1), p- pravdepodobnosť

Príklad 3.13

Vyriešime Príklad 3.11 s využitím EXCEL-u.

a) ***Prilepiť funkciu / Štatistické/*NORMDIST(1, 1, 1, 0)**.

Výsledok $F(1) = 0{,}3989$

***Prilepiť funkciu / Štatistické/*NORMDIST(1, 1, 1, 1** .

Výsledok $F(1) = 0{,}5$

b) ***Prilepiť funkciu / Štatistické/*NORMDIST(2, 1, 1, 1)**.

Výsledok $F(2) = 0{,}8413$

***Prilepiť funkciu / Štatistické/*NORMDIST(1, 1, 1, 1)**.

Výsledok $F(1) = 0{,}5$. $F(2)-F(1)=0{,}3413$

c) ***Prilepiť funkciu / Štatistické/*NORMDIST(-2, 1, 1, 1)**.

Výsledok $F(-2) = 0{,}0013$

$$P(|X|>2)=1-(F(2)-F(-2))=1-F(2)+F(-2)==1-0{,}8413+0{,}0013=0{,}16$$

Príklad 3.14

Čas potrebný na vypracovanie kontrolnej práce má normálne rozdelenie s priemernou dobou vypracovania 110 minút a štandardnou odchýlkou 20 minút.

a) Koľko percent študentov dokončí prácu do 2 hodín?

b) Koľko času by bolo treba dať na prácu, aby ju dokončilo 90 % študentov?

Riešenie Náhodná premenná je čas potrebný na vypracovanie práce.

a) Zaujíma nás $P(X \le 120) = \text{NORMDIST}(120;110;20;1)=0{,}6914$, t.j. 69% študentov spraví prácu do 2 hodín.

b) V tomto prípade nás zaujíma ten čas *t*, pre ktorý platí $P(X \le t)=0{,}9$, čo je 90%-ný kvantil tohto rozdelenia.

Preto t = NORMINV(0,9;110;20)=135,6, t.j. študenti potrebujú na prácu 136 minút.

3.6 Aproximácia diskrétnych rozdelení normálnym

V tejto časti je stručne vysvetlený význam normálneho rozdelenia ako limitného rozdelenia pre mnohé náhodné premenné, pričom sú na ne kladené len dosť všeobecné predpoklady. Základnou vetou v štatistike je **centrálna limitná veta,** ktorá hovorí:

Ak náhodné premenné $X_1, X_2, X_3, ..., X_n$ sú nezávislé náhodné premenné s rovnakým rozdelením pravdepodobnosti t.j. $E(X_i)=\mu$, $D(X_i)=\sigma^2$, potom aritmetický priemer $\overline{X}=\dfrac{\sum_{i=1}^{n} X_i}{n}$ je náhodná premenná, ktorá má pre $n \to \infty$ normálne rozdelenie s parametrami μ a σ^2 / n t.j. N($\mu,\sigma^2/n$).

Poznámka 3.4

- Ak pomocou $\overline{X}$ vytvoríme normovanú náhodnú premennú $Z=\dfrac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$,

tak táto má rozdelenie $\mathrm{N}(0,1)$.

- Aproximácia rozdelenia premennej $\overline{X}$ je tým lepšia, čím je väčšie *n*.

Najznámejší a najpoužívanejší tvar centrálnej limitnej vety je integrálna **Moivrova - Laplaceova veta**, pomocou ktorej vieme **aproximovať binomické rozdelenie normálnym**. Táto veta hovorí:

Nech náhodná premenná X má binomické rozdelenie s parametrami n a p. Potom limitným rozdelením tohto rozdelenia pre $n \to \infty$ je normálne rozdelenie s parametrami $\mu = np$ a $\sigma^2 = npq$.

Poznámka 3.5

- Ak vytvoríme pomocou *X* normovanú náhodnú premennú $\frac{X - np}{\sqrt{npq}}$, môžeme tvrdenie vety vyjadriť v tvare $\lim_{n\to\infty} P\left(\frac{X - np}{\sqrt{npq}} \le z\right) = \Phi(z)$, pre ľubovoľné $z \in R$.
- Pri dostatočne veľkom n môžeme vyjadriť pravdepodobnosť toho, že náhodná premenná s binomickým rozdelením pravdepodobnosti, nadobudne hodnoty z intervalu $\langle a,b)$ takto:

$$P(a \le X < b) \doteq F(b) - F(a) = \Phi\left(\frac{b - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - np}{\sqrt{npq}}\right). \qquad (3.32)$$

- Táto aproximácia sa používa vtedy, ak $np > 5$ aj $nq > 5$. Používa sa aj kritérium $npq > 9$.

  Ak $p = 0{,}5$, sú nerovnosti splnené už pre $n > 10$.

  Ak $p = 0{,}8$ alebo $p = 0{,}2$, požadujeme $n > 25$.

  Ak $p = 0{,}05$ alebo $p = 0{,}95$, má byť $n > 100$.

- Ak sa pri tejto aproximácii diskrétneho rozdelenia spojitým zaujímame o pravdepodobnosť konkrétnej hodnoty x binomického rozdelenia, teda $P(X = x)$, použijeme funkciu hustoty normálneho rozdelenia:

$$P(X = x) \doteq f(x) = \frac{1}{\sigma}\varphi\left(\frac{x - np}{\sqrt{npq}}\right). \tag{3.33}$$

Pri malých hodnotách *p*, prípadne pri hodnotách *p* blízkych jednotke, táto aproximácia nie je dobrá a jej zlepšenie sa dosahuje len zvyšovaním počtu pokusov v Bernoulliho schéme. V tejto situácii dobrú aproximáciu binomického rozdelenia dáva Poissonovo rozdelenie. Platí nasledujúce tvrdenie.

Nech náhodná premenná *X* má binomické rozdelenie s parametrami n a p. Potom limitným rozdelením tohto rozdelenia pre $n \to \infty$ a súčasne $p \to 0$ je Poissonovo rozdelenie s parametrom $\lambda = np$.

Poznámka 3.6

Túto aproximáciu sa odporúča použiť ak $p < 0{,}1$ alebo $p > 0{,}9$.

Príklad 3.15

Poisťovňa poistila 1000 ľudí rovnakého veku. Pravdepodobnosť úmrtia v priebehu roka je pre každého z nich 0,008. Každý poistenec zaplatil 1200 Sk. V prípade jeho úmrtia dostanú jeho príbuzní 80 000 Sk. Aká je pravdepodobnosť, že poisťovňa utrpí stratu?

Riešenie Poisťovňa vybrala na poistnom 1 200 000 Sk. Z týchto peňazí je schopná vyplatiť $1200000/80000 = 15$ pozostalých. Teda poisťovňa utrpí stratu, ak v priebehu roka zomrie viac ako 15 poistencov. Preto nás zaujíma pravdepodobnosť tohto javu. Náhodná premenná *X* je počet zomretých poistencov za rok. Riadi sa Bi(0,008;1000).

Pretože $np = 1000 \cdot 0{,}008 = 8$ a $nq = 1000 \cdot 0{,}992 = 992$ môžeme binomické rozdelenie aproximovať normálnym s parametrami

$\mu = 8;\qquad \sigma^2 = npq = 1000 \cdot 0{,}008 \cdot 0{,}992 = 7{,}936,\qquad$ t.j. $\mathrm{N}(8; 7{,}936)$.

$P(X > 15) = 1 - P(0 \le X \le 15) \doteq 1 - F(15) + F(0) =$ 1 – NORMDIST$\left(15; 8; \sqrt{7{,}936}; 1\right)$

+ NORMDIST$\left(0; 8; \sqrt{7{,}936}; 1\right) = 0{,}0089.$

Pretože pravdepodobnosť $p = 0{,}008$ je malá, môžeme binomické rozdelenie aproximovať aj Poissonovým rozdelením s parametrom $\lambda = np = 8$, t.j. Po(8).

$$P(X > 15) = 1 - P(0 \le X \le 15) \doteq e^{-8} \cdot \left(1 + 8 + \frac{8^2}{2!} + \frac{8^3}{3!} + \cdots + \frac{8^{15}}{15!}\right) = 1 - 0{,}9944 = 0{,}0056.$$

Pravdepodobnosť, že poisťovňa bude stratová je menšia ako 1% ako sme zistili oboma aproximáciami.

3.7 Rozdelenia funkcií náhodných premenných

Pri výberovom skúmaní v štatistike sa používajú ďalšie rozdelenia, ktoré vznikli ako funkcie iných rozdelení. Najčastejšie sa pri tomto skúmaní využíva

a) χ^2- rozdelenie (chí-kvadrát rozdelenie)
b) Studentovo rozdelenie (t-rozdelenie)
c) Fisherovo rozdelenie (F-rozdelenie).

3.7.1 χ^2 - rozdelenie

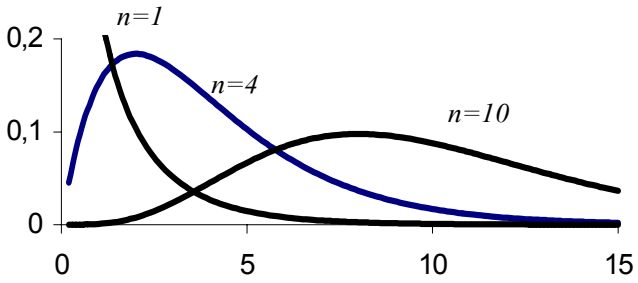
Pri konštrukcii tohto rozdelenia použijeme *n* nezávislých náhodných premenných $X_1, X_2, ..., X_n$, z ktorých každá má normované normálne rozdelenie $\mathrm{N}(0,1)$. Súčet druhých mocnín (štvorcov) týchto náhodných premenných je náhodná premenná $\chi^2 = X_1^2 + X_2^2 + ... + X_n^2$. Jej rozdelenie sa nazýva **chí - kvadrát rozdelenie**.

Poznámka 3.7

Funkčný predpis funkcie hustoty rozdelenia nebudeme uvádzať.

Vlastnosti:

1. Počet sčítancov ***n*** určuje geometrický tvar rozdelenia. Číslo ***n*** sa volá **počet stupňov voľnosti** (d.f., degrees of freedom). Je to jediný parameter tohto rozdelenia.
2. $E(X)=n$
3. $D(X)=2n$
4. Koeficient asymetrie je $8/n$, je to rozdelenie nesymetrické.
5. Koeficient špicatosti je $12/n$.
6. S rastúcim n sa oba koeficienty blížia k nule, pre $n \to \infty$ je χ^2- rozdelenie symetrické.
7. Pre $n > 30$ možno toto rozdelenie aproximovať normálnym rozdelením $\mathrm{N}(n, 2n)$.



Obr. 3.14 Funkcia hustoty χ^2- rozdelenia

3.7.2 Studentovo rozdelenie (t - rozdelenie)

Pomocou nezávislých náhodných premenných X a χ^2 , kde X má normované normálne rozdelenie $\mathrm{N}(0,1)$ a χ^2 má χ^2- rozdelenie s n stupňami voľnosti vytvoríme novú náhodnú premennú T

$$T = \frac{X}{\sqrt{\frac{\chi^2}{n}}}$$

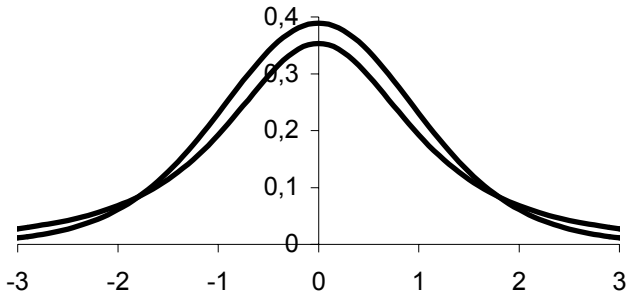
Jej rozdelenie sa nazýva **Studentovo rozdelenie s *n* stupňami voľnosti**.

Poznámka 3.8

Funkčný predpis funkcie hustoty rozdelenia nebudeme uvádzať.

Vlastnosti:

1. Číslo ***n*** – počet stupňov voľnosti je jediný parameter tohto rozdelenia, ktorý určuje aj geometrický tvar rozdelenia.
2. $E(T)=0$
3. Krivka Studentovho rozdelenia je symetrická podľa $t = 0$ a veľmi sa podobá na krivku normovaného normálneho rozdelenia, je však plochšia. To znamená, že hodnoty vzdialenejšie od nuly majú väčšiu pravdepodobnosť nastatia ako pri normálnom rozdelení.
4. Pre $n > 30$ možno Studentovo rozdelenie aproximovať normovaným normálnym rozdelením.



Obr. 3.15 Funkcia hustoty t – rozdelenia

3.7.3 Fisherovo rozdelenie (F – rozdelenie)

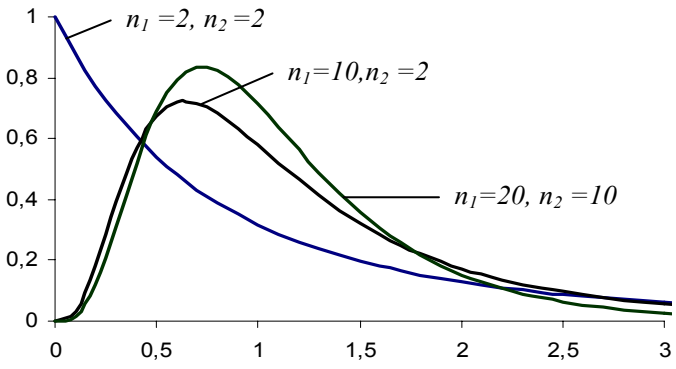
Je vytvorené pomocou dvoch náhodných premenných χ_1^2, χ_2^2, kde χ_1^2 má χ^2-rozdelenie s n_1 stupňami voľnosti a χ_2^2 má χ^2- rozdelenie s n_2 stupňami voľnosti. Ich podiel je náhodná premenná *F*,

$$F = \frac{\frac{\chi_1^2}{n_1}}{\frac{\chi_2^2}{n_2}} \quad ,$$

ktorej rozdelenie sa nazýva **Fisherovo rozdelenie s n_1 a n_2 stupňami voľnosti**.

Vlastnosti:

1. Parametrami rozdelenia sú dve čísla n_1 a n_2, ich poradie sa nemôže vymeniť.
2. Stupne voľnosti určujú geometrický tvar funkcie hustoty.



Obr. 3.16 Funkcia hustoty *F* – rozdelenia

EXCEL

Pri hľadaní kvantilov spomínaných rozdelení môžeme využiť ponuku štatistických funkcií v EXCELi. Treba podotknúť, že nie sú definované jednotne. K lepšej orientácii nám určite pomôže nasledujúca tabuľka. Okrem použitých označení štatistických funkcií, pomocou ktorých nájdeme potrebné kvantily, sú v nej zadané i hodnoty pravdepodobnosti, ktoré treba zadať pre príslušnú hodnotu hľadaného kvantilu.

Tab. 3.7 Hľadanie kvantilov v EXCELi

Funkcia	**Zadaná pravdepodobnosť**	**Nájdený kvantil**
NORMSINV	α	z_{α}
	$1-\alpha$	$z_{1-\alpha}$
TINV	α	$t_{1-\alpha/2}$
	2α	$t_{1-\alpha}$
CHIINV	α	$\chi^2_{1-\alpha}$
	$1-\alpha$	χ^2_{α}
FINV	α	$F_{1-\alpha}$
	$1-\alpha$	F_{α}

V nasledujúcej tabuľke je urobený prehľad funkcií, ktoré budeme najčastejšie používať.

Tab. 3.8 Prehľad tabuliek ponúkaných EXCELom

Funkcia	Parametre	Význam funkcie
NORMSDIST	z	Distribučná funkcia rozdelenia N(0, 1) NORMSDIST(z) = Φ(z) = $P(X< z)$
NORMDIST	$x, \mu, \sigma, 1$	Distribučná funkcia rozdelenia N(μ, σ^2) NORMDIST $(x, \mu, \sigma, 1) = F(x) = P(X < x)$
NORMDIST	$x, \mu, \sigma, 0$	Funkcia hustoty rozdelenia N(μ, σ^2) NORMDIST $(x, \mu, \sigma, 0) = f(x)$
NORMSINV	p	Kvantily rozdelenia N(0, 1), je to inverzná funkcia k distrib. funkcii rozdelenia N(0, 1), určí číslo z_p na osi tak, aby $P(X<z_p) = p$ NORMSINV(p) = $z_p \Leftrightarrow P(X<z_p) = p$
NORMINV	p, μ, σ	Kvantily rozdelenia N(μ, σ^2), je to inverzná funkcia k distrib. funkcii rozdelenia N(μ, σ^2), určí číslo u_p na osi tak, aby $P(X<u_p) = p$ NORMINV$(p, \mu, \sigma) = u_p \Leftrightarrow P(X<u_p) = p$
TDIST	x, n, 1 n – stupne voľnosti $x \geq 0$	Pre Studentove rozdelenie s n stupňami voľnosti určí pravdepodobnosť p = TDIST(x, n, 1) = $= P(X > x) = 1 - F(x)$
TDIST	x, n, 2 n – stupne voľnosti $x \geq 0$	Pre Studentove rozdelenie s n stupňami voľnosti určí pravdepodobnosť p = TDIST(x, n, 2) = $= P(\|X\| > x)$
TINV	p, n n – stupne voľnosti	Je to inverzná funkcia k funkcii TDIST(x, n, 2), určí číslo t na osi tak, aby $P(\|X\|>t) = p$ TINV(p, n) = $t \Leftrightarrow P(\|X\| >t) = p$
CHIDIST	x, n n – stupne voľnosti $x \geq 0$	Pre χ^2-rozdelenie s n stupňami voľnosti určí pravdepodobnosť p = CHIDIST(x, n) = $P(X > x) = 1 - F(x)$
CHIINV	p, n n – stupne voľnosti	Kritické hodnoty χ^2 –rozdelenia. Je to inverzná funkcia k funkcii CHIDIST, určí číslo χ^2 na osi tak, aby $P(X > \chi^2) = p$ CHIINV(p, n) = $\chi^2 \Leftrightarrow P(X > \chi^2) = p$
FDIST	x, n_1, n_2 n_1, n_2-stupne voľnosti $x \geq 0$	Pre Fisherove F-rozdelenie s n_1 a n_2 stupňami voľnosti určí pravdepodobnosť p = FDIST(x, n_1, n_2) = $P(X > x) = 1 - F(x)$
FINV	p, n_1, n_2 n_1, n_2-stupne voľnosti	Kritické hodnoty F –rozdelenia. Je to inverzná funkcia k funkcii FDIST, určí číslo f na osi tak, aby $P(X > f) = p$ FINV(p, n_1, n_2) = f $\Leftrightarrow P(X > f) = p$

Príklad 3.16 Nájdite

a) 95% kvantil normovaného normálneho rozdelenia (ozn. $z_{0,95}$),

b) 95% kvantil normálneho rozdelenia so strednou hodnotu μ =2 a disperziou σ^2 = 0,81 (ozn. $u_{0,95}$),

c) 90% kvantil Studentovho t - rozdelenia s 8-mi stupňami voľnosti (ozn. $t_{0,90;8}$),

d) 97,5% kvantil χ^2 - rozdelenia s 10-mi stupňami voľnosti (ozn. $\chi^2_{0,975;10}$),

e) 95% kvantil F - rozdelenia s n_1 = 8 a n_2 = 6 stupňami voľnosti (ozn. $F_{0,95;8;6}$),

f) 5% kvantil Studentovho t - rozdelenia s 6-mi stupňami voľnosti (ozn. $t_{0,05;6}$).

Riešenie

a) Na vstupnom paneli EXCELU vyberieme ponuku ***prilepiť funkciu*** f_x (alebo =), na ľavej strane z ponuky funkcií vyberieme ***viac funkcií,*** zvolíme ***štatistické,*** vyhľadáme **NORMSINV**, vpíšeme pravdepodobnosť 0,95. Vo výstupe máme $z_{0,95} = 1{,}644853$.

b) Postupne volíme ***=/ viac funkcií/ štatistické/*** **NORMINV** vpíšeme pravdepodobnosť 0,95, strednú hodnotu μ =2 a štandardnú odchýlku σ = 0,9. Vo výstupe máme $u_{0,95} = 3{,}48037$.

c) Funkcia **TINV** neponúka priamo kvantily rozdelenia, ale dá sa úpravou zadávanej pravdepodobnosti získať hľadaný kvantil. Ak potrebujeme *p*% kvantil, vpíšeme hodnotu pravdepodobnosti 2(1-*p*) (platí pre $p > 0{,}5$). Teda postupne volíme ***=/ viac funkcií/ štatistické/*** **TINV** vpíšeme pravdepodobnosť 2(1- 0,9) = 0,2 a 8 stupňov voľnosti. Vo výstupe je $t_{0,90;8} = 1{,}397$.

d) Funkcia **CHIINV** neponúka priamo kvantily rozdelenia, ale dá sa úpravou zadávanej pravdepodobnosti získať hľadaný kvantil. Ak potrebujeme *p*% kvantil, vpíšeme hodnotu pravdepodobnosti (1-p). Volíme ***=/ viac funkcií/ štatistické/*** **CHIINV** vpíšeme pravdepodobnosť 1 - 0,975 = 0,025 a 10 stupňov voľnosti. Vo výstupe je $\chi_{0,975;10} = 20{,}4832$.

e) Postup je rovnaký ako pri funkcii **CHIINV**. Na výpočet kvantilov tiež zmeníme zadávanú pravdepodobnosť na 1- *p*. Volíme ***=/ viac funkcií/ štatistické/***

FINV vpíšeme pravdepodobnosť 1- 0,95 = 0,05; 8 a 6 stupňov voľnosti. Vo výstupe je $F_{0,95;8;6} = 3,9715$.

f) Pre $p < 0,5$ nie je možné hneď použiť model c). Využijeme, že hustota pravdepodobnosti t-rozdelenia je symetrická podľa $t=0$ a pre kvantily platí $t_p = -t_{1-p}$. Preto: $t_{0,05;6} = -t_{0,95;6} = -1,9431$.

Príklady na precvičenie

3.17 Pre náhodnú premennú X – doba čakania na električku z Príkladov 3.4 a 3.5 nájdite strednú hodnotu a rozptyl.

3.18 Strelec strieľa na cieľ, pričom má k dispozícii 4 náboje. Každý zásah znamená zisk 5 bodov, každé minutie stratu 2 bodov. Určite tabuľku rozdelenia pravdepodobnosti počtu získaných bodov, ak pravdepodobnosť zásahu pri každom výstrele je 0,8. Určite strednú hodnotu a strednú kvadratickú odchýlku počtu získaných bodov.

3.19 Podnik vyexpedoval zásielku s 5 výrobkami. Pravdepodobnosť, že sa jeden výrobok počas prepravy poškodí je $p = 0,2$. Aká je pravdepodobnosť, že sa počas prepravy nepoškodí ani jeden výrobok? Určite strednú hodnotu a rozptyl počtu poškodených výrobkov.

3.20 Distribučná funkcia spojitej náhodnej premennej má tvar:

$$F(x) = \begin{cases} 0 & x \le 0 \\ 2cx & 0 < x \le 2 \\ 1 & x > 2 \end{cases}$$

Určite c, $E(X)$, $D(X)$, $P(1 \le X < 3)$.

3.21 Každý z 300 uchádzačov o prácu robí test, pozostávajúci z 3 častí. Uchádzač bude prijatý, ak urobí všetky 3 časti testu. Zo skúseností s takýmito testami je známe, že prvú časť testu urobí 90% , druhú 95% a tretiu 85% uchádzačov. Aká je pravdepodobnosť, že test urobí

a) práve 200 uchádzačov

b) aspoň 200 uchádzačov

3.22 Vo veľkej skupine ľudí je 5% chorých na chrípku. Aká je pravdepodobnosť, že v skupine 400 ľudí je $5 \pm 0{,}25$ % chorých na chrípku?

3.23 Telefónna ústredňa obsluhuje 3000 účastníkov. Pravdepodobnosť, že nejaký účastník bude v priebehu hodiny telefonovať, je p=0,002. Vypočítajte pravdepodobnosť, že v priebehu hodiny budú telefonovať 4 účastníci.

3.24 Pravdepodobnosť, že výrobok neprejde kontrolou je 0,05. Určite pravdepodobnosť, že medzi 500 náhodne vybranými výrobkami bude

a) práve 20 výrobkov, ktoré neprejdú kontrolou

b) od 10 do 30 výrobkov, ktoré neprejdú kontrolou.

3.25 Pravdepodobnosť, že sa na gymnázium hlási žiak s vyznamenaním zo ZŠ je 0,7. Aká je pravdepodobnosť, že medzi 500 prihlásenými je viac ako 360 žiakov s vyznamenaním.

4. VÝBEROVÉ SKÚMANIE

4.1 Náhodný výber

V úvodnej kapitole sme spomínali, že v praxi sú bežné prípady, kedy namiesto skúmania hodnôt sledovaného znaku na základnom súbore, skúmame daný znak na súbore výberovom. Je zrejmé, že sú situácie, kedy je potrebné poznať hodnotu znaku na každej štatistickej jednotke, napríklad pri voľbách nás bude zaujímať výsledok hlasovania každého voliča, ale na druhej strane, na otestovanie kvality konzervovaného výrobku nebudeme testovať všetky konzervy. Dôvody pre výberové zisťovanie sú samozrejme i iného charakteru. Môže to byť časová, finančná, resp. organizačná náročnosť skúmania základného súboru. V praxi sa z týchto príčin zvykne uprednostňovať výberové skúmanie. Pri výberovom zisťovaní nebudeme skúmať všetky štatistické jednotky základného súboru, ale iba ich časť, ktorú nazveme výberový súbor. **Výberový súbor** je každá podmnožina základného súboru. Hovoríme, že výberový súbor je **náhodný výber,** ak každá štatistická jednotka základného súboru má rovnakú šancu dostať sa do výberového súboru, t.j. o tom, či bude vybratá musí rozhodnúť náhoda. Kritérium náhodnosti je možné zabezpečiť rôznymi spôsobmi. Jeden z nich je napríklad losovanie, ktorého princíp je všeobecne známy. V prípade, že štatistické jednotky sú očíslované, je vhodné použiť **generátor náhodných čísiel,** ktorý prevzal funkciu tzv. tabuľky náhodných čísiel. Uvedenými spôsobmi dostaneme tzv. **jednoduchý náhodný výber.** Z hľadiska opakovateľnosti výberu štatistickej jednotky môžeme robiť:

- jednoduchý náhodný výber s opakovaním,
- jednoduchý náhodný výber bez opakovania.

Pri **náhodnom výbere s opakovaním** vyberáme štatistickú jednotku vždy z rovnakého základného súboru, čo zabezpečí nezávislosť jednotlivých výberov a konštantnú pravdepodobnosť vybratia každej štatistickej jednotky v priebehu výberu. V tomto prípade môže byť štatistická jednotka zaradená do výberového súboru viackrát.

Pri **náhodnom výbere bez opakovania**, vybratú jednotku do štatistického súboru nevraciame, čo znamená, že každá štatistická jednotka môže byť vo výberovom súbore len raz, avšak pravdepodobnosť výberu jednotky závisí od predchádzajúcich výberov.

V praxi sa zvyčajne nerozlišuje medzi výberom s opakovaním a bez opakovania, ak rozsah N základného súboru je veľký, prípadne ak pre rozsah n výberového súboru platí $n \leq 0{,}1\ N$.

Okrem jednoduchého výberu sa v praxi používa i **zložený výber**. V takomto prípade náhodného výberu sa jednotky základného súboru roztriedia do skupín podľa dodatočnej informácie a z týchto skupín sa rôznymi spôsobmi vyberajú štatistické jednotky. Vzhľadom na to, že sa budeme ďalej zaoberať len jednoduchými náhodnými výbermi, na získanie podrobnejších informácií o zloženom výbere odporúčame [6], [15]. Úlohy výberového skúmania môžeme rozdeliť do dvoch základných okruhov. Sú to:

- odhad parametrov základného súboru na základe parametrov vypočítaných z hodnôt výberového súboru,
- testovanie štatistických hypotéz o pravdepodobnostnom rozdelení sledovaného znaku v základnom súbore, prípadne o parametroch základných súborov na základe výsledkov výberového zisťovania.

4.2 Výberové charakteristiky

Podobne ako v prípade skúmania základného súboru, i pri výberových súboroch nás zväčša nezaujímajú konkrétne hodnoty sledovaného znaku na každej štatistickej jednotke, ale iba isté typické vlastnosti, ktoré nám poskytnú základné informácie o súbore. Zo základných parametrov, ktoré sme sledovali na základnom súbore v kapitole 1 pripomenieme aspoň:

- aritmetický priemer $\bar{x}$, ktorý budeme v základnom súbore s rozsahom N označovať μ a vypočítame

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i ,$$

- rozptyl σ^2, ktorý je v základnom súbore rozsahu N definovaný vzťahom

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (X_i - \mu)^2 ,$$

- štandardná odchýlka σ, ktorá je definovaná ako odmocnina z rozptylu

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2} ,$$

- podiel (relatívnu početnosť) štatistických jednotiek s istou vlastnosťou budeme v základnom súbore označovať π a vypočítame ho podľa vzťahu

$$\pi = \frac{K}{N},$$

kde *K* je absolútna početnosť prvkov so sledovanou vlastnosťou.

Na výpočet uvedených parametrov je potrebné poznať hodnoty znaku *X* na každej štatistickej jednotke základného súboru. Výberom štatistických jednotiek zo základného súboru dostaneme konkrétnu postupnosť hodnôt sledovaného znaku *X*. Z daného základného súboru môžeme získať veľký počet výberových súborov s vopred stanoveným rozsahom, ktoré sa navzájom líšia aspoň v jednej hodnote znaku *X*. Kvalita výberového súboru narastá s jeho rozsahom. Z hodnôt výberového súboru môžeme vypočítať **výberové charakteristiky** (štatistiky). Vzhľadom na konkrétny výberový súbor je výberová charakteristika konštantná veličina, ale z hľadiska skúmaného základného súboru je to náhodná premenná s vlastným zákonom rozdelenia, ktoré nazývame **výberové rozdelenie**. Výberové rozdelenia sú teoretickým základom pre spracovanie výsledkov výberového zisťovania. Základné informácie o najdôležitejších typoch pravdepodobnostného rozdelenia, ktoré budeme v ďalších aplikáciách používať sú zhrnuté v kapitole 3. Každé výberové rozdelenie závisí od:

- rozdelenia znaku *X,*
- funkcie výberovej charakteristiky,
- rozsahu výberového súboru.

4.2.1 Výberový priemer

Konkrétny výberový súbor je tvorený postupnosťou ($x_1, x_2,..., x_n$) hodnôt náhodnej premennej *X*. Každá z hodnôt x_i je jedna z možných hodnôt náhodnej premennej X_i, ktorá predstavuje hodnotu *X* na *i*-tej náhodne vybratej jednotke. Potom (X_1,

$X_2,..., X_n$) je náhodný výber, ktorý tvoria nezávislé náhodné premenné X_i , $i=1,2,...,n,$ s tým istým rozdelením pravdepodobnosti, t.j. s tou istou distribučnou funkciou, resp. funkciou hustoty. Náhodnú premennú

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{4.1}$$

budeme nazývať **výberový priemer.**

Ak rozdelenie pravdepodobnosti náhodnej premennej (sledovaného znaku) X v základnom súbore, z ktorého náhodný výber pochádza, má strednú hodnotu μ a rozptyl σ^2, potom pre výberový priemer platí:

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n\mu}{n} = \mu, \tag{4.2}$$

$$D\left(\overline{X}\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}, \tag{4.3}$$

čo znamená, že stredná hodnota výberového priemeru je rovnaká ako stredná hodnota rozdelenia, z ktorého náhodný výber pochádza, ale rozptyl výberového priemeru je n-krát menší, ako rozptyl náhodnej premennej X v základnom súbore.

Ak náhodný výber pochádza z normálneho rozdelenia $N(\mu,\sigma^2)$, potom výberový priemer $\overline{X}$ má rozdelenie $N(\mu,\sigma^2/n)$, teda opäť normálne rozdelenie s tou istou strednou hodnotou μ a s rozptylom σ^2/n. Ak náhodný výber pochádza z ľubovoľného rozdelenia so strednou hodnotou μ a konečným rozptylom σ^2, potom v prípade, keď rozsah n výberového súboru je dostatočne veľký, môžeme na základe centrálnej limitnej vety aproximovať rozdelenie výberového priemeru $\overline{X}$ normálnym rozdelením $N(\mu,\sigma^2/n)$.

4.2.2 Výberový rozptyl a výberová štandardná odchýlka

Nech ($X_1, X_2,..., X_n$) je náhodný výber, ktorý tvoria nezávislé náhodné premenné X_i , i=1,2,...,n, s tým istým rozdelením pravdepodobnosti. Výraz

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{4.4}$$

nazývame **výberový rozptyl** a jeho odmocninu

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{4.5}$$

nazveme **výberová štandardná odchýlka.**

Obe charakteristiky dávajú informáciu o variabilite hodnôt výberového súboru okolo výberového priemeru. Ak rozdelenie, z ktorého pochádza náhodný výber má rozptyl σ^2, potom pre strednú hodnotu výberového rozptylu platí

$$E(S^2) = \sigma^2 \tag{4.6}$$

Ak náhodný výber pochádza z normálneho rozdelenia $N(\mu,\sigma^2)$, potom náhodná premenná

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \tag{4.7}$$

má χ^2 - rozdelenie s n-1 stupňami voľnosti. (Podrobnejšie informácie v [15].)

4.2.3 Výberový podiel

Nech $(X_1, X_2,...,X_n)$ je náhodný výber z alternatívneho rozdelenia $A(\pi)$. Každá z náhodných premenných X_i, i = 1,2,...,n môže nadobúdať hodnotu 0 alebo 1. Označme

$$K = \sum_{i=1}^{n} X_i \,. \tag{4.8}$$

Podiel

$$P = \frac{K}{n} \tag{4.9}$$

nazývame **výberový podiel.** Náhodná premenná K má binomické rozdelenie $Bi(\pi,n)$ so strednou hodnotou πn a rozptylom $n\pi(1-\pi)$. Využitím vlastností strednej hodnoty a rozptylu sa dá ukázať, že pre strednú hodnotu a rozptyl výberového podielu P platí:

$$E(P) = \pi, \qquad D(P) = \frac{\pi(1-\pi)}{n}. \tag{4.10}$$

Pre dostatočne veľké n ($npq > 9$) je možné toto rozdelenie aproximovať normálnym rozdelením $N(\pi, \pi(1-\pi)/n)$.

EXCEL

Na vytvorenie jednoduchého náhodného výberu použitím EXCELu volíme ***Nástroje/analýza údajov/Vzorkovanie***. Táto procedúra ponúka dve metódy výberu - periodickú, ktorá pri voľbe periódy k vyberie každú k-tú hodnotu zo vstupnej oblasti a náhodnú, pri ktorej zadáme rozsah výberovej vzorky a každá hodnota zo vstupnej oblasti má rovnakú pravdepodobnosť, že bude vybraná.

4.3 Bodové odhady

V kapitole 3 sme sa zoznámili s niektorými základnými zákonmi rozdelenia pravdepodobnosti náhodnej premennej a mohli sme si všimnúť, že tieto zákony závisia od jedného, alebo viacerých parametrov Θ_i. Jedna z úloh štatistiky je nájsť vhodný pravdepodobnostný model rozdelenia sledovaného znaku (náhodnej premennej) X. Zvyčajne však nepoznáme parametre predpokladaných rozdelení, a preto je potrebné urobiť ich odhad práve pomocou výberových súborov.

Bodovým odhadom parametra Θ nazývame nahradenie jeho hodnoty hodnotou vhodne zvolenej výberovej charakteristiky U_n. Je zrejmé, že pre odhad parametra zvolíme výberovú charakteristiku, ktorá poskytuje najlepší odhad parametra Θ. Kvalitu odhadu zabezpečia nasledujúce kritéria:

- neskreslenosť odhadu,
- konzistentnosť odhadu,
- výdatnosť odhadu.

4.3.1 Neskreslený odhad

Výberová charakteristika U_n sa nazýva **neskreslený (nevychýlený) odhad** parametra Θ, ak platí

$$E(U_n) = \Theta\,, \tag{4.11}$$

čo znamená, že stredná hodnota odhadu je rovná skutočnej hodnote odhadovaného parametra. Rozdiel

$$E(U_n) - \Theta \tag{4.12}$$

nazývame **skreslením(vychýlením) odhadu** parametra Θ.

Výberovú charakteristiku U_n nazývame asymptoticky neskreslený odhad parametra Θ, ak platí

$$\lim_{n\to\infty} E(U_n) = \Theta. \tag{4.13}$$

Z časti 4.2.1 a 4.2.2 vieme, že pre náhodný výber ($X_1, X_2,...,X_n$), ktorý pochádza z ľubovoľného rozdelenia, ktoré má strednú hodnotu μ a konečný rozptyl σ^2 platí

$$E(\overline{X}) = \mu\,, \qquad\qquad E(S^2) = \sigma^2\,,$$

čo znamená, že výberový priemer $\overline{X}$ je neskresleným odhadom priemeru rozdelenia, z ktorého náhodný výber pochádza a výberový rozptyl S^2 je neskresleným odhadom rozptylu rozdelenia, z ktorého náhodný výber pochádza.

4.3.2 Konzistentný odhad

Výberovú charakteristiku U_n nazývame **konzistentný odhad** parametra Θ, ak platí

$$\lim_{n\to\infty} P(|U_n - \Theta| < \varepsilon) = 1, \tag{4.14}$$

čo znamená, že U_n je konzistentným odhadom parametra Θ, ak s rastúcim rozsahom n náhodného výberu rastie i pravdepodobnosť, že sa hodnota bodového odhadu bude veľmi málo líšiť od skutočnej hodnoty parametra. K tomu, aby odhad U_n bol konzistentný, stačí aby boli splnené nasledujúce podmienky:

a) U_n je neskreslený, alebo aspoň asymptoticky neskreslený odhad,

b) $\lim_{n\to\infty} D(U_n) = 0$.

4.3.3 Výdatný odhad

Výberovú charakteristiku U_n nazývame **výdatný odhad** parametra Θ, ak platí, že zo všetkých charakteristík U_n, ktoré poskytujú neskreslený bodový odhad parametra Θ má najmenší rozptyl.

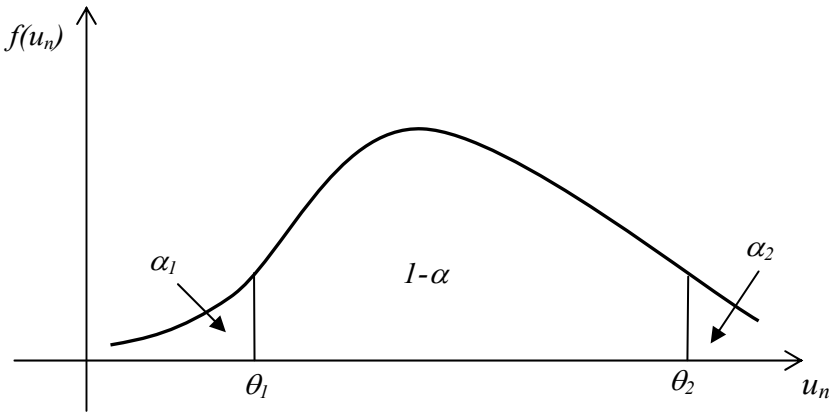
Výberové charakteristiky definované v predchádzajúcej kapitole: **výberový priemer, výberový rozptyl, výberová štandardná odchýlka a výberový podiel spĺňajú všetky tri požiadavky kladené na kvalitný odhad a môžeme ich považovať za najlepšie odhady parametrov základného súboru.**

4.4 Intervalové odhady

Pri bodovom odhade neznámeho parametra používame jednu výberovú charakteristiku U_n, ktorej hodnota u_n je vypočítaná z údajov konkrétneho výberového súboru. Iný odhad spočíva v tom, že neznámy parameter Θ odhadujeme číselným intervalom (θ_1, θ_2), v ktorom sa parameter Θ nachádza s pravdepodobnosťou, ktorá je blízka jednotke. Interval (θ_1, θ_2), pre ktorý platí

$$P(\theta_1 < \Theta < \theta_2) = 1 - \alpha \tag{4.15}$$

nazývame 100(1-α)% **dvojstranný (obojstranný) interval spoľahlivosti** pre parameter Θ. Číslo 1-α nazývame **hladina spoľahlivosti** a pravdepodobnosť α, ktorú zvyčajne volíme α = 0,01 alebo α = 0,05 nazývame **riziko odhadu**. Za predpokladu, že hladina spoľahlivosti 1-α je číslo blízke jednej, môžeme takmer s určitosťou predpokladať, že parameter základného súboru bude z intervalu spoľahlivosti. Zvyšovaním hladiny spoľahlivosti sa rozširuje interval spoľahlivosti, čo má za následok zníženie presnosti odhadu. Na druhej strane, ak hladina spoľahlivosti je nižšia, čomu zodpovedá menšia šírka intervalu, zvyšuje sa riziko odhadu. Na obrázku je znázornený dvojstranný interval spoľahlivosti parametra Θ, pričom $\alpha = \alpha_1 + \alpha_2$.

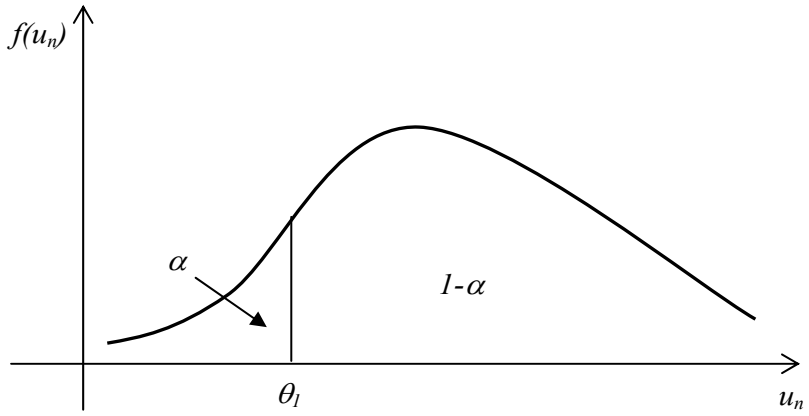


Obr. 4.1 Dvojstranný interval spoľahlivosti parametra Θ

V niektorých prípadoch nás zaujímajú intervaly typu $(-\infty, \theta_2)$ (pravostranný interval), resp. (θ_1, ∞) (ľavostranný interval), pre ktoré platí

$$P(\Theta < \theta_2) = 1 - \alpha, \qquad \text{resp.} \qquad P(\theta_1 < \Theta) = 1 - \alpha \,. \tag{4.16}$$

Tieto intervaly nazývame 100(1-α)% **jednostranné intervaly spoľahlivosti** pre parameter Θ. Na nasledujúcom obrázku je znázornený jednostranný interval spoľahlivosti pre parameter Θ.



Obr. 4.2 Jednostranný (ľavostranný) interval spoľahlivosti parametra Θ

4.4.1 Interval spoľahlivosti pre odhad priemeru μ

Nech $(X_1, X_2,..., X_n)$ je náhodný výber rozsahu n z normálneho rozdelenia $N(\mu, \sigma^2)$. Už vieme (pozri 4.2.1), že v tomto prípade výberový priemer $\overline{X}$ má rozdelenie $N(\mu, \sigma^2/n)$. Potom náhodná premenná $Z = \dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ má normované normálne rozdelenie N(0, 1). Pre takéto rozdelenie platí vzťah

$$P\left(-z_{1-\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}} \right) = 1 - \alpha , \tag{4.17}$$

kde $z_{1-\frac{\alpha}{2}}$ je kvantil rozdelenia N(0, 1).

Úpravou nerovností vo vzťahu (4.17) dostaneme pre odhad parametra μ vzťah

$$P\left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha , \tag{4.18}$$

kde $\bar{x}$ je výberový priemer vypočítaný z údajov konkrétneho výberového súboru rozsahu n. Interval

$$\left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) \tag{4.19}$$

je dvojstranný $100(1-\alpha)\%$ interval spoľahlivosti pre parameter μ rozdelenia $N(\mu, \sigma^2)$ pri známom rozptyle σ^2.

Ak použijeme vzťah

$$P\left(\frac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} < z_{1-\alpha} \right) = 1 - \alpha \tag{4.20}$$

dostaneme

$$P\left(\mu > \bar{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha\,, \tag{4.21}$$

kde $\bar{x}$ je výberový priemer vypočítaný z údajov konkrétneho výberového súboru rozsahu *n* a interval

$$\left(\bar{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}, \infty\right) \tag{4.22}$$

je jednostranný (ľavostranný) 100(1-α)% interval spoľahlivosti pre parameter μ rozdelenia N(μ, σ^2) pri známom rozptyle σ^2. Analogicky, ak vychádzame zo vzťahu

$$P\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} > -z_{1-\alpha}\right) = 1-\alpha\,, \tag{4.23}$$

dostaneme interval

$$\left(-\infty; \bar{x} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right), \tag{4.24}$$

ktorý je jednostranný (pravostranný) 100(1-α)% interval spoľahlivosti pre parameter μ rozdelenia N(μ, σ^2) pri známom rozptyle σ^2.

Vo väčšine prípadov, pri určovaní intervalu spoľahlivosti pre parameter μ základného súboru s rozdelením N(μ, σ^2), parameter σ^2 nepoznáme. Nahradíme ho preto výberovým rozptylom a namiesto výberovej charakteristiky $Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ použijeme výberovú charakteristiku $T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$, ktorá má Studentovo rozdelenie (t-rozdelenie) s *n*-1 stupňami voľnosti. Analogicky ako v predchádzajúcom prípade, získame nasledujúce 100(1-α)% intervaly spoľahlivosti pre parameter μ rozdelenia N(μ, σ^2) pri neznámom rozptyle σ^2.

$$\left(\bar{x} - t_{1-\frac{\alpha}{2},n-1} \frac{S}{\sqrt{n}}; \bar{x} + t_{1-\frac{\alpha}{2},n-1} \frac{S}{\sqrt{n}} \right) \tag{4.25}$$

$$\left(-\infty; \bar{x} + t_{1-\alpha} \frac{S}{\sqrt{n}} \right) \tag{4.26}$$

$$\left(\bar{x} - t_{1-\alpha} \frac{S}{\sqrt{n}}; \infty \right) \tag{4.27}$$

Využitím tabuľkového procesoru EXCEL, prípadne iného štatistického softvéru, nie je problém určiť potrebné kvantily $t_{1-\frac{\alpha}{2}}$, resp. $t_{1-\alpha}$ t-rozdelenia, prípadne kvantily $z_{1-\frac{\alpha}{2}}$, resp. $z_{1-\alpha}$ normovaného normálneho rozdelenia .

V minulosti sa vzťahy (4.25), (4.26), (4.27) používali pre výpočet intervalov spoľahlivosti pre tzv. malé výbery, t. j. v prípadoch, kedy $n < 30$. Ich platnosť je však všeobecná a platí pre výberové súbory akéhokoľvek rozsahu.

V prípade, že uvažujeme náhodný výber zo základného súboru s akýmkoľvek rozdelením, podľa centrálnej limitnej vety platí, že vzťahy (4.25), (4.26), (4.27) môžeme použiť, ak rozsah n výberového súboru je dostatočne veľký.

Príklad 4.1

Zaujíma nás, aký je 95% intervalový odhad počtu bodov, ktorý študenti 1. fakulty v priemere získajú na prijímacích skúškach na vojenskú akadémiu z matematiky. Ako výberový súbor použijeme súbor PRIJÍMACIE SKÚŠKY/1.fakulta/mat. Použitím tohto výberového súboru, o ktorom predpokladáme, že je vybratý zo súboru s normálnym rozdelením, urobíme 95% intervalový odhad priemeru základného súboru.

Vzhľadom na to, že rozptyl základného súboru je neznámy, použijeme na výpočet dvojstranného intervalu spoľahlivosti vzťah (4.25), prípadne niektorý zo vzťahov (4.26), (4.27), ak nás zaujíma len jednostranný interval spoľahlivosti. Použitím vzťahov (4.1) a (4.5), resp. štatistických funkcií AVERAGE a STDEV vypočítame výberový priemer a výberovú štandardnú odchýlku. Na výpočet kvantilu

$t_{1-\frac{\alpha}{2};n-1} = t_{0,95;37}$ použijeme štatistickú funkciu TINV(0,05;37) (Tabuľka 3.6), prípadne tabuľky Studentovho t-rozdelenia rozdelenia. Po dosadení do (4.25) dostaneme 95% interval spoľahlivosti

$$9{,}35 - 2{,}03\frac{6{,}58}{\sqrt{38}} < \mu < 9{,}35 + 2{,}03\frac{6{,}58}{\sqrt{38}}$$

a po vyčíslení hraníc interval (7,19; 11,51), v ktorom s pravdepodobnosťou 0,95 leží počet bodov, ktorý v priemere na prijímacích skúškach z matematiky dosiahnu študenti 1. fakulty. Podobne postupujeme i pri výpočte jednostranných intervalov spoľahlivosti.

4.4.2 Interval spoľahlivosti pre odhad rozptylu σ^2

Nech ($X_1, X_2,..., X_n$) je náhodný výber rozsahu n z normálneho rozdelenia N(μ, σ^2). Náhodná premenná

$$W = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \tag{4.28}$$

má χ^2 - rozdelenie s n-1 stupňami voľnosti. Potom platí

$$P\left(\chi^2_{\frac{\alpha}{2};n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{1-\frac{\alpha}{2};n-1}\right) = 1-\alpha\,, \tag{4.29}$$

kde $\chi^2_{\frac{\alpha}{2};n-1}$, $\chi^2_{1-\frac{\alpha}{2},n-1}$ sú kvantily χ^2 - rozdelenia. Po úpravách týchto nerovností dostaneme 100(1-α)% dvojstranný interval spoľahlivosti pre rozptyl σ^2 základného súboru v tvare

$$\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2};n-1}}\,. \tag{4.30}$$

Pripomeňme, že na rozdiel od intervalu spoľahlivosti pre priemer základného súboru, interval spoľahlivosti pre rozptyl nie je symetrický, vzhľadom na nesymetriu χ^2 - rozdelenia. Jednostranné intervaly spoľahlivosti vytvoríme analogicky ako v prípade priemeru a budú vymedzené nerovnosťou

$$\sigma^2 < \frac{(n-1)S^2}{\chi^2_{\alpha;n-1}} \tag{4.31}$$

pre pravostranný 100(1-α)% interval spoľahlivosti a nerovnosťou

$$\frac{(n-1)S^2}{\chi^2_{1-\alpha;n-1}} < \sigma^2 \tag{4.32}$$

pre ľavostranný 100(1-α)% interval spoľahlivosti pre rozptyl základného súboru. Z uvedených intervalov spoľahlivosti pre rozptyl, môžeme bezprostredne určiť intervaly spoľahlivosti pre štandardnú odchýlku. Napríklad dvojstranný interval spoľahlivosti je vymedzený nerovnosťami

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2};n-1}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2};n-1}}}\,. \tag{4.33}$$

Príklad 4.2

V tomto príklade nadviažeme na Príklad 4.1 a budeme zisťovať, aký je 95% interval spoľahlivosti pre rozptyl priemerného počtu bodov z matematiky na prijímacích skúškach na 1. fakulte vojenskej akadémie. Opäť použijeme výberový súbor PRIJÍMACIE SKÚŠKY/1.fakulta/mat. Na výpočet dvojstranného intervalu spoľahlivosti použijeme vzťah (4.30). Z príkladu 4.1 vieme, že štandardná výberová odchýlka s = 6,58. Na výpočet potrebných kvantilov χ^2- rozdelenia môžeme opäť využiť v EXCELi ponuku štatistických funkcií a pomocou funkcie CHIINV (Tabuľka 3.6) vypočítať $\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0,975} = 55,67$ a $\chi^2_{\frac{\alpha}{2}} = \chi^2_{0,025} = 22,11$.

Po dosadení do vzťahu (4.30) dostaneme

$$\frac{37.6,58^2}{55,67} < \sigma^2 < \frac{37.6,58^2}{22,11}$$

a po vyčíslení hraníc interval (28,74; 72,35), v ktorom s 95% pravdepodobnosťou leží rozptyl základného súboru.

4.4.3 Interval spoľahlivosti pre odhad podielu π

Z časti 4.2.3 vieme, že výberový podiel $P = \frac{K}{n}$, kde K je počet úspešných pozorovaní vo výberovom súbore s rozsahom n, má rozdelenie so strednou hodnotou $E(P) = \pi$ a rozptylom $D(P) = \frac{\pi(1-\pi)}{n}$, kde π je neznáma pravdepodobnosť úspešného pozorovania v základnom súbore. Pre dostatočne veľké n $\left(n > \frac{9}{\pi(1-\pi)}\right)$ je možné toto rozdelenie aproximovať normálnym rozdelením N(π, $\pi(1-\pi)/n$). Normovaním výberového podielu dostaneme náhodnú premennú

$$Z = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}},$$

ktorá má normované normálne rozdelenie N(0;1).
Vzťah

$$-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}} \quad , \tag{4.34}$$

vymedzuje 100(1-α)% interval spoľahlivosti pre premennú Z, ktorá má rozdelenie N(0,1). Ak do (4.34) dosadíme náhodnú premennú $Z = \frac{P - \pi}{\sqrt{\frac{P(1-P)}{n}}}$, v ktorej sme rozptyl nahradili bodovým odhadom rozptylu, dostaneme po vykonaní potrebných úprav dvojstranný 100(1-α)% interval spoľahlivosti pre parameter π binomického rozdelenia pre dostatočne veľké n v tvare

$$P - z_{1-\frac{\alpha}{2}}\sqrt{\frac{P(1-P)}{n}} < \pi < P + z_{1-\frac{\alpha}{2}}\sqrt{\frac{P(1-P)}{n}}\,. \tag{4.35}$$

Pre čitateľa už určite nie je problém napísať jednostranné 100(1-α)% intervaly spoľahlivosti pre podiel π.

Príklad 4.3

Zaujíma nás, aký je podiel gymnazistov medzi študentmi vojenskej akadémie. Ako výberový súbor použijeme všetkých študentov (oboch fakúlt), ktorí sú v danom roku študentmi vojenskej akadémie (pozri súbory PRIJÍMACIE SKÚŠKY/1.fakulta/stred. škola a PRIJÍMACIE SKÚŠKY/2.fakulta/stred. škola). Bude nás zaujímať výberový podiel, ktorý je najlepším bodovým odhadom podielu základného súboru, ale i 99% dvojstranný interval spoľahlivosti pre podiel základného súboru, ktorý tvoria všetci absolventi vojenskej akadémie.

Výberový podiel je definovaný ako relatívna početnosť. Vzhľadom na to, že vo výberovom súbore rozsahu 88 je 24 gymnazistov, pre výberový podiel platí $P = \frac{24}{88} = 0{,}27$, čo znamená, že bodový odhad podielu gymnazistov na vojenskej akadémii je 0,27. Všimnime si, že rozsah výberového súboru spĺňa požiadavku aproximácie rozdelenia normálnym rozdelením $\left(88 > \frac{9}{0.27(1-0.27)} = 24{,}3\right)$, a tak na výpočet 99% intervalu spoľahlivosti použijeme vzťah (4.35). Dosadením vypočítaných hodnôt dostaneme

$$0{,}27 - 2{,}576\sqrt{\frac{0{,}27(1-0{,}27)}{88}} < \pi < 0{,}27 + 2{,}576\sqrt{\frac{0{,}27(1-0{,}27)}{88}}$$

a po vyčíslení dostaneme interval (0,148; 0,392), v ktorom s 99% pravdepodobnosťou leží podiel gymnazistov zo študentov vojenskej akadémie.

EXCEL

Bodový odhad priemeru, rozptylu, resp. štandardnej odchýlky vypočítame v EXCELi pomocou funkcií AVERAGE, VAR, resp. STDEV. Pri určení intervalov spoľahlivosti je vhodné využiť štatistické funkcie ponúkané EXCELom. V Tab. 3.6 nájdeme prehľad štatistických funkcií na výpočet kvantilov pravdepodobnostných rozdelení potrebných k určeniu intervalov spoľahlivosti pre jednotlivé parametre. Tabuľka, ktorú získame po voľbe ***Nástroje/Analýza úda-***

jov/Popisná štatistika obsahuje bodový odhad priemeru, rozptylu i štandardnej odchýlky. Navyše, ak v dialógovom okne *hladina spoľahlivosti* zadáme hodnotu (1-α)100%, zobrazí sa v poslednom riadku tabuľky vypočítaná hodnota $t_{1-\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}$, potrebná na výpočet obojstranného intervalu spoľahlivosti pre priemer základného súboru (pozri (4.25)).

Príklad 4.4

Vráťme sa k Príkladu 1.1 a použime na jeho riešenie procedúru ***Nástroje/Analýza údajov/Popisná štatistika.*** V kapitole 1 sme uviedli postup, pomocou ktorého získame bodový odhad priemeru, ale napr. i rozptylu základného súboru Ak navyše v príslušnom dialógovom okienku zvolíme 95% hladinu spoľahlivosti, dostaneme v poslednom riadku tabuľky popisnej štatistiky údaj, ktorý použijeme na výpočet dvojstranného 95% intervalu spoľahlivosti. Nasledujúca tabuľka je výstupom spomínanej procedúry.

Z tabuľky sa dozvedáme, že bodový odhad priemeru základného súboru, t.j. priemerného počtu bodov, ktorý na prijímacích skúškach z matematiky dosahujú študenti 1.fakulty je približne 9,35. Bodový odhad rozptylu základného súboru je po zaokrúhlení 43,24. Na výpočet 95% intervalu spoľahlivosti použijeme údaj v poslednom riadku Tab. 4.1.

Tab. 4.1 Popisná štatistika pre súbor PRIJÍMACIE SKÚŠKY/1.fakulta/mat.

mat.	
Stř. hodnota	9.35
Chyba stř. hodnoty	1,066673
Medián	8
Modus	4
Směr. odchylka	6,575415
Rozptyl výběru	43,23608
Špičatost	-0,480817
Šikmost	0,617793
Rozdíl max-min	25
Minimum	0
Maximum	25
Součet	355,3
Počet	38
Hladina spol.(95,0%)	2,161283

Hľadaný interval má tvar

$$\left(\bar{x} - t_{1-\frac{\alpha}{2},n-1} \frac{S}{\sqrt{n}}; \bar{x} + t_{1-\frac{\alpha}{2},n-1} \frac{S}{\sqrt{n}} \right)$$

a po dosadení hodnôt z tabuľky dostaneme interval

$$(9{,}35 - 2{,}16;\ 9{,}35 + 2{,}16) = (7{,}19;\ 11{,}51),$$

v ktorom s pravdepodobnosťou 0,95 leží priemerný počet bodov nášho základného súboru. V prípade, že nás zaujíma iba pravostranný interval spoľahlivosti priemerného počtu bodov, môžeme opäť použiť procedúru ***Popisná štatistika***. Vzhľadom na to, že na výpočet hodnoty v poslednom riadku tabuľky ***Popisnej štatistiky*** je vždy po voľbe (1-α)100% hladiny použitý kvantil $t_{1-\frac{\alpha}{2}}$, ktorý je vhodný na výpočet dvojstranného intervalu spoľahlivosti, musíme na výpočet (1-α) 100% jednostranného intervalu zvoliť (1-2α)100% = 90% hladinu spoľahlivosti . Pri takejto

voľbe dostaneme na výstupe hodnotu 1,80, pomocou ktorej určíme pravostranný interval spoľahlivosti

$$\left(-\infty;\ \bar{x} + t_{1-\alpha}\frac{S}{\sqrt{n}}\right) = \left(-\infty;\ 11{,}15\right).$$

Príklady na precvičenie

4.5 Vypočítajte bodový odhad priemeru, rozptylu a štandardnej odchýlky počtu bodov, ktoré študenti dosiahli na

a) prijímacích skúškach z matematiky. Použite výberový súbor PRIJÍMACIE SKÚŠKY/2.fakulta/mat.

b) prijímacích skúškach z fyziky Použite výberový súbor PRIJÍMACIE SKÚŠKY/2.fakulta/fyz.

4.6 Urobte 90% dvojstranné a jednostranné intervaly spoľahlivosti pre priemer, rozptyl a štandardnú odchýlku z Príkladu 4.5.

4.7 Vypočítajte 95% dvojstranný interval spoľahlivosti podielu študentov zo stredných odborných učilíšť medzi študentmi vojenskej akadémie. Ako výberový súbor použijeme všetkých študentov (oboch fakúlt), ktorí sú v danom roku študentmi vojenskej akadémie. Použite súbor PRIJÍMACIE SKÚŠKY/stred. škola.

5. TESTOVANIE ŠTATISTICKÝCH HYPOTÉZ

5.1 Princíp testovania štatistických hypotéz

Štatistickou hypotézou nazývame určitý predpoklad o parametroch základného súboru, o jeho rozdelení, prípadne o parametroch viacerých súborov. Štatistické overovanie tohto predpokladu na základe výsledkov získaných z náhodného výberu nazývame **testovanie hypotéz.** Predpoklad, ktorý vyslovíme o istom parametri základného súboru, prípadne o type rozdelenia, nazývame **nulová hypotéza** a označujeme H_0. Môžeme ju formulovať v tvare

$$H_0 : \Theta = \Theta_0 , \tag{5.1}$$

ak vyjadruje predpoklad o zhode parametra Θ základného súboru so známou konštantou Θ_0, prípadne v tvare

$$H_0 : \Theta_1 = \Theta_2 , \tag{5.2}$$

ak testujeme zhodu parametra Θ v dvoch základných súboroch. Nulovú hypotézu o zhode rozdelenia pravdepodobnosti f znaku (náhodnej premennej) X v základnom súbore so známym teoretickým rozdelením g zapisujeme v tvare

$$H_0 : f(x) = g(x) . \tag{5.3}$$

Alternatívne tvrdenie, ktoré prijímame v prípade zamietnutia nulovej hypotézy, nazývame **alternatívna hypotéza.** Napríklad, v prípade nulovej hypotézy (5.1) môžeme ako alternatívnu hypotézu použiť niektoré z nasledujúcich tvrdení

$$H_1 : \Theta \neq \Theta_0 , \tag{5.4}$$

$$H_1 : \Theta > \Theta_0 , \tag{5.5}$$

$$H_1 : \Theta < \Theta_0 . \tag{5.6}$$

V prípade (5.4) alternatívna hypotéza popiera platnosť H_0 bez bližšieho špecifikovania hodnoty parametra Θ. Takto formulovanú hypotézu nazývame **dvojstranná alternatívna hypotéza** a test hypotézy nazývame **dvojstranný test**. V prípadoch (5.5) a (5.6) alternatívna hypotéza popiera platnosť nulovej hypotézy a súčasne jednostranne vymedzuje obor hodnôt parametra v základnom súbore.

Tvrdí, že hodnota parametra Θ je väčšia, prípad (5.5), resp. menšia, prípad (5.6), než hodnota predpokladaná nulovou hypotézou. Takto formulované hypotézy nazývame **jednostranné hypotézy**, konkrétne v prípade (5.5) **pravostranné** a v prípade (5.6) **ľavostranné** a test hypotézy je **jednostranným (pravostranným, ľavostranným) testom.**

Vzhľadom na to, že pri testovaní hypotéz robíme úsudky na základe údajov z výberových súborov je zrejmé, že sa môžeme dopustiť i nesprávnych rozhodnutí. Môže nastať situácia, keď zamietneme nulovú hypotézu i napriek tomu, že v skutočnosti platí. Vtedy sa dopúšťame tzv. **chyby prvého druhu** a pravdepodobnosť tejto chyby označujeme α. Pre pravdepodobnosť chyby 1. druhu platí

$$\alpha = P(zamietnutie\ \ H_0 / platí\ \ H_0).$$

Druhá možnosť chybného rozhodnutia je, že prijmeme nulovú hypotézu, i keď v skutočnosti platí alternatívna hypotéza. V takomto prípade sa dopúšťame tzv. **chyby druhého druhu** a pravdepodobnosť, že sa tejto chyby dopustíme, označujeme β. Pre pravdepodobnosť chyby 2. druhu platí

$$\beta = P(nezamietnutie\ \ H_0 / platí\ \ H_1).$$

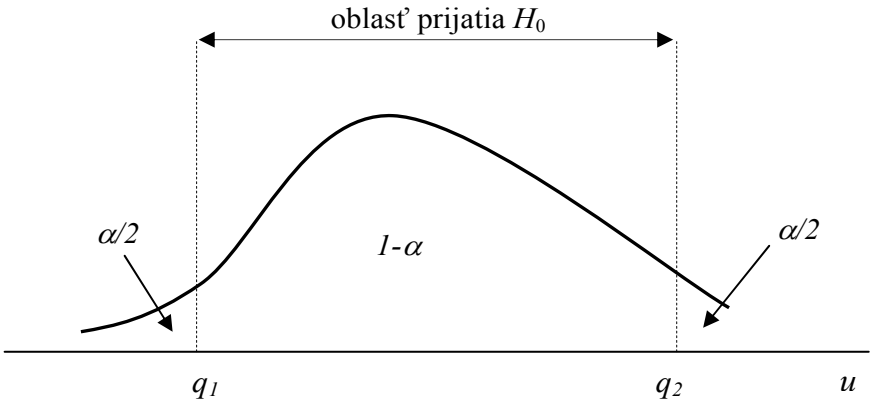
Pravdepodobnosť 1-β sa nazýva **sila testu.** Vyjadruje, s akou pravdepodobnosťou zamietneme nulovú hypotézu, ak platí alternatívna hypotéza, t.j. vyjadruje pravdepodobnosť, že sa nedopustíme chyby 2. druhu.

Pri testovaní hypotéz by bolo ideálne minimalizovať pravdepodobnosť vzniku oboch chýb. V skutočnosti to však nie je možné, pretože zmenšovaním chyby 1. druhu narastá chyba 2. druhu. Vzhľadom na to, že nás bude zaujímať praktické hľadisko testovania hypotéz, nebudeme ďalej rozoberať teoretickú stránku tohto problému. Klasický prístup testovania hypotéz je založený na tom, že vopred volíme pravdepodobnosť chyby 1. druhu, tzv. **hladinu významnosti**, v prijateľnej výške (α = 0,01, α = 0,05 alebo α = 0,1). Testovací postup je odvodený tak, aby pri danej hladine významnosti zabezpečoval minimálnu pravdepodobnosť chyby 2. druhu.

Na rozhodnutie o tom, ktorá z hypotéz H_0 a H_1 platí, použijeme **testovaciu štatistiku** Q. Testovacia štatistika je náhodná premenná, ktorá môže nadobúdať hodnoty z istého oboru hodnôt (z istej podmnožiny množiny reálnych čísiel). Je prirodzené, že súčasne s výpočtom hodnoty testovacej štatistiky pre daný výber, je potrebné určiť i pravidlo, na základe ktorého vieme rozhodnúť v prospech nulovej, prípadne alternatívnej hypotézy. Z tohto dôvodu rozdelíme obor hodnôt testovacej štatistiky na dve podmnožiny. Jednu z nich nazveme **oblasť prijatia hypotézy H_0** a druhú nazveme **kritická oblasť**, alebo **oblasť zamietnutia hypotézy H_0**, ktorá bude oblasťou prijatia alternatívnej hypotézy H_1. Obe podmnožiny sú disjunktné a hranice, ktoré ich oddeľujú, nazveme **kritické hodnoty.** Kritické hodnoty určíme ako kvantily rozdelenia použitej testovacej štatistiky Q pre zvolenú hladinu významnosti α. Ak (q_1, q_2) je oblasť prijatia H_0 a $(-\infty, q_1\rangle \cup \langle q_2, \infty)$ je oblasť prijatia dvojstrannej alternatívnej hypotézy H_1, pre q_1, q_2 platí

$$F(q_1) = \frac{\alpha}{2}, \qquad F(q_2) = 1 - \frac{\alpha}{2}\ , \tag{5.7}$$

kde $F(q)$ je distribučná funkcia náhodnej premennej Q.



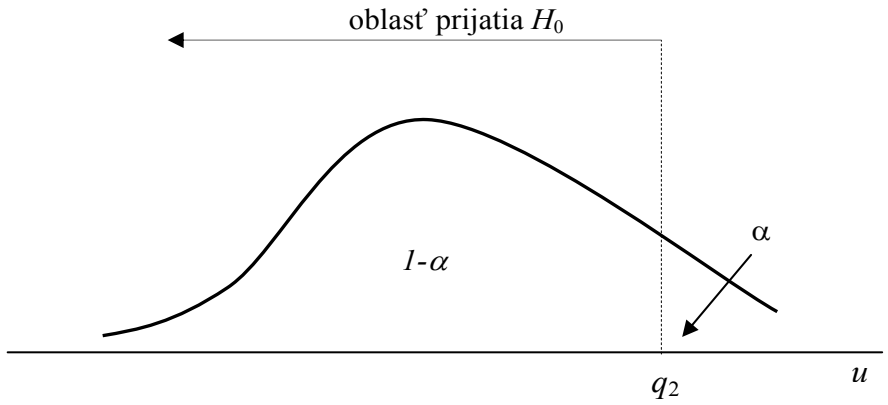
Obr. 5.1 Vymedzenie oblasti prijatia hypotézy H_0 pri dvojstrannom teste

V prípade, že alternatívna hypotéza je jednostranná, bude oblasť prijatia hypotézy H_0 interval $(-\infty, q_2)$, ak je alternatívna hypotéza pravostranná a interval (q_1, ∞), ak ide o ľavostrannú alternatívnu hypotézu. V tomto prípade pre q_1, q_2 platí

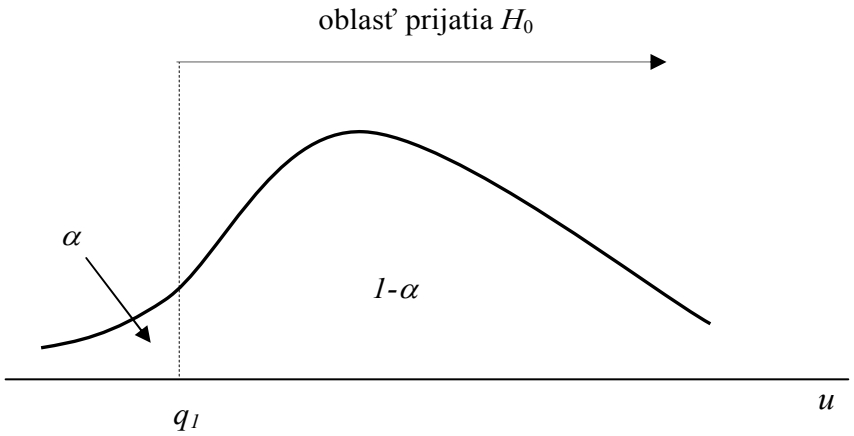
$$F(q_1) = \alpha, \qquad F(q_2) = 1 - \alpha\,, \tag{5.8}$$

kde $F(q)$ je distribučná funkcia náhodnej premennej Q.

Na nasledujúcich obrázkoch vidíme, ako hladina významnosti α určuje hranice intervalov hodnôt testovacej štatistiky, pre ktoré prijímame, resp. zamietame H_0 pri voľbe jednostrannej alternatívnej hypotézy.



Obr. 5.2 Vymedzenie oblasti prijatia hypotézy H_0 pri pravostrannom teste



Obr. 5.3 Vymedzenie oblasti prijatia hypotézy H_0 pri ľavostrannom teste

Všeobecný postup pri testovaní hypotéz

1. Sformulujeme nulovú a alternatívnu hypotézu.
2. Vyberieme vhodnú testovaciu štatistiku Q a vypočítame jej hodnotu na základe hodnôt výberového súboru.
3. Zvolíme hladinu významnosti α a určíme kritické hodnoty q_1, q_2, ktoré vymedzujú oblasť prijatia nulovej, resp. alternatívnej hypotézy.
4. Rozhodneme o prijatí, resp. zamietnutí nulovej hypotézy na zvolenej hladine významnosti α na základe toho, či hodnota testovacej štatistiky leží v oblasti prijatia, resp. zamietnutia nulovej hypotézy.

5.2 Test hypotézy o priemere základného súboru

Podobne ako pri hľadaní intervalov spoľahlivosti, i pri testovaní hypotéz dôležitú úlohu zohráva výber vhodnej testovacej štatistiky. Významnú úlohu pri jej výbere má rozsah výberového súboru, ale aj rozdelenie sledovaného znaku (náhodnej premennej) X v základnom súbore. Túto skutočnosť zohľadníme i pri testovaní hypotéz o zhode priemeru s konštantou.

➢ **Znak má normálne rozdelenie so známym rozptylom σ^2**

Testujeme hypotézu

$$H_0 : \mu = \mu_0 .$$

Alternatívna hypotéza je v prípade dvojstranného testu

$$H_1 : \mu \neq \mu_0 .$$

V prípade jednostranného testu

$$H_1 : \mu > \mu_0 ,$$

alebo

$$H_1 : \mu < \mu_0 .$$

Za daných predpokladov použijeme testovaciu štatistiku

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} , \tag{5.9}$$

ktorá má normované normálne rozdelenie pravdepodobnosti N(0,1). Pre zvolenú hladinu významnosti α určíme hodnoty potrebných kvantilov normovaného normálneho rozdelenia, ktoré vymedzia kritickú oblasť. V prípade dvojstrannej hypotézy bude kritickou oblasťou množina

$$(-\infty; -z_{1-\alpha/2}\rangle \cup \langle z_{1-\alpha/2}; \infty) \tag{5.10}$$

a oblasť prijatia nulovej hypotézy je interval

$$\left(-z_{1-\alpha/2;}\ z_{1-\alpha/2}\right). \tag{5.11}$$

V prípade že uvažujeme jednostranné hypotézy, bude kritická oblasť pre pravostranný test vymedzená intervalom

$$\langle z_{1-\alpha}; \infty) \tag{5.12}$$

a v prípade ľavostranného testu intervalom

$$(-\infty; -z_{1-\alpha}\rangle . \tag{5.13}$$

Nulovú hypotézu zamietame, ak hodnota testovacej štatistiky, ktorá je vypočítaná z daného výberového súboru rozsahu *n*, padne do kritickej oblasti, t.j. do oblasti prijatia príslušnej alternatívnej hypotézy.

➢ **Znak má normálne rozdelenie s neznámym rozptylom σ^2**

V tomto prípade urobíme odhad neznámeho rozptylu výberovým rozptylom (4.4) a použijeme testovaciu štatistiku

$$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}, \tag{5.14}$$

ktorá má t-rozdelenie pravdepodobnosti s n -1 stupňami voľnosti. Podobne ako v predchádzajúcom prípade, jej hodnotu vypočítanú z údajov konkrétneho výberového súboru porovnáme s kritickými hodnotami pre zvolenú hladinu významnosti. V tomto prípade kritické hodnoty určíme ako kvantily t-rozdelenia.

Ak rozsah výberového súboru je dostatočne veľký ($n \geq 30$), môžeme aproximovať t-rozdelenie normálnym rozdelením a na výpočet kritických hodnôt použiť kvantily normovaného normálneho rozdelenia.

➢ **Znak má ľubovoľné rozdelenie a rozsah výberového súboru je veľký**

V takomto prípade použijeme tvrdenie centrálnej limitnej vety (kapitola 3), na základe ktorej má testovacia štatistika

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

asymptoticky normované normálne rozdelenie. V prípade, že σ^2 nepoznáme, nahradíme neznámy rozptyl výberovým rozptylom S^2, ktorý vypočítame z daného výberového súboru.

➢ **Znak má ľubovoľné rozdelenie a rozsah výberového súboru je malý**

V takomto prípade použijeme niektorý z tzv. neparametrických testov. O niektorých z nich sa viac dozvieme v kapitole 7.

V nasledujúcej tabuľke Tab. 5.4 je urobený prehľad spomínaných testovacích štatistík a príslušných kritických oblastí pre rôzne alternatívne hypotézy použité pri testovaní hypotézy o zhode priemeru s konštantou.

Tab. 5.4 Testovanie hypotézy o zhode priemeru μ s konštantou μ_0

Hypotéza H_0	**Hypotéza H_1**	**Predpoklady**	**Testovacia štatistika**	**Oblasť zamietnutia H_0**
$\mu = \mu_0$	1. $\mu \neq \mu_0$ 2. $\mu > \mu_0$ 3. $\mu < \mu_0$	znak X má v ZS rozdelenie N(μ, σ^2) σ^2 poznáme	$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$	1. $\lvert z \rvert \geq z_{1-\alpha/2}$ 2. $z \geq z_{1-\alpha}$ 3. $z \leq -z_{1-\alpha}$
		znak X má v ZS rozdelenie N(μ, σ^2) σ^2 nepoznáme, $n \geq 30$	$Z = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}$	1. $\lvert z \rvert \geq z_{1-\alpha/2}$ 2. $z \geq z_{1-\alpha}$ 3. $z \leq -z_{1-\alpha}$
		znak X má v ZS rozdelenie N(μ, σ^2) σ^2 nepoznáme, $n < 30$	$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}$	1. $\lvert t \rvert \geq t_{1-\alpha/2,n-1}$ 2. $t \geq t_{1-\alpha,n-1}$ 3. $t \leq -t_{1-\alpha,n-1}$
		znak X má v ZS ľubovoľné rozdelenie, $n \geq 30$	$Z = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}$	1. $\lvert z \rvert \geq z_{1-\alpha/2}$ 2. $z \geq z_{1-\alpha}$ 3. $z \leq -z_{1-\alpha}$

Poznámka 5.1

Testovanie hypotéz o zhode priemeru s konštantou je možné robiť i pomocou intervalov spoľahlivosti pre priemer základného súboru (pozri časť 4.2.1). Pri testovaní hypotéz, rozhodnutie o prijatí, resp. zamietnutí nulovej hypotézy H_0 na zvolenej hladine významnosti α robíme na základe toho, či hodnota príslušnej testova-

cej štatistiky leží v oblasti prijatia hypotézy H_0. Napríklad, v prípade dvojstranného testu prijímame H_0 , ak testovacia štatistika

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

leží v intervale $\left(-z_{1-\alpha/2;}\, z_{1-\alpha/2}\right)$, t.j. ak

$$-z_{1-\alpha/2} < \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < z_{1-\alpha/2},$$

pričom $\bar{x}$ je výberový priemer vypočítaný z konkrétneho výberového súboru. Úpravou týchto nerovností dostaneme vzťah

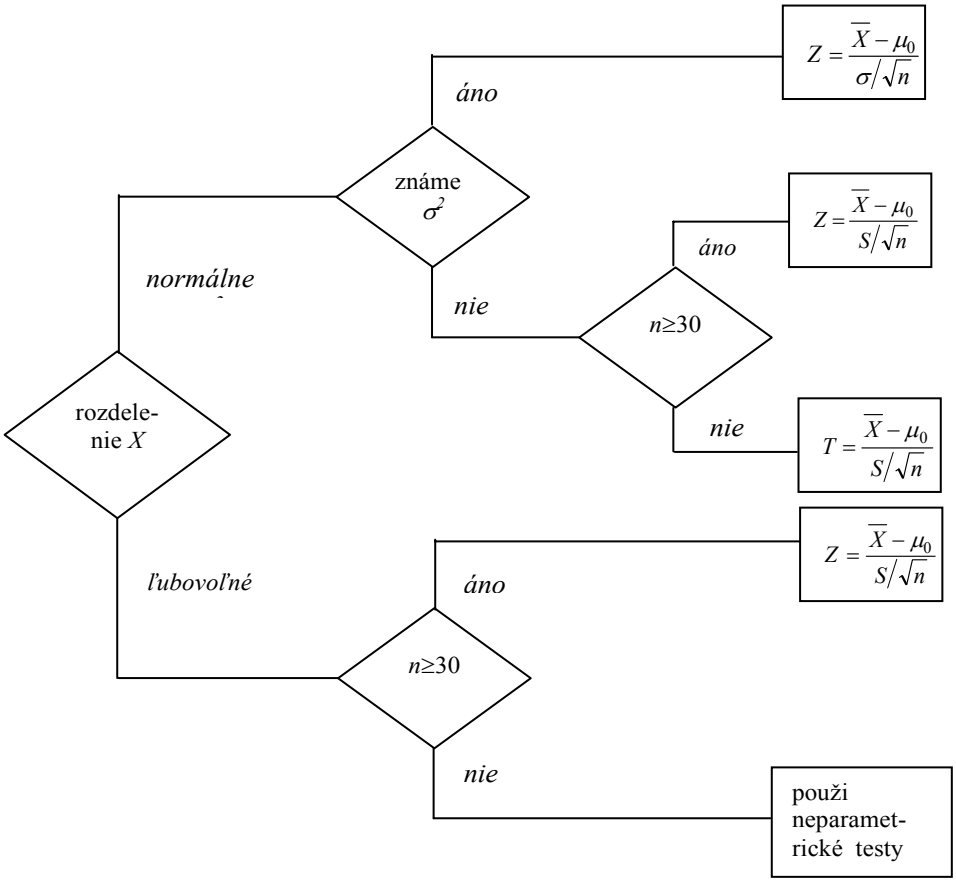
$$\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu_0 < \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

Interval $\left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}};\ \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$ je ale 100(1-α)% obojstranný interval spoľahlivosti pre priemer základného súboru pri známom σ^2. Znamená to teda, že ak konštanta μ_0 je z intervalu spoľahlivosti pre priemer μ, môžeme na hladine významnosti α prijať hypotézu o rovnosti priemeru μ s konštantou μ_0. V opačnom prípade prijímame obojstrannú alternatívnu hypotézu.

Podobnú úvahu použijeme i pri jednostranných testoch. Konkrétne, ak volíme alternatívnu hypotézu $H_1 : \mu > \mu_0$, prijímame H_0, ak $\mu_0 \in \left(\bar{x} - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}; \infty\right)$. Ak alternatívna hypotéza je $H_1 : \mu < \mu_0$, prijímame H_0, ak $\mu_0 \in \left(-\infty;\ \bar{x} + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right)$.

V prípade, že rozptyl základného súboru nie je známy, alebo rozsah výberového súboru je malý, postupujeme pri výpočte intervalov spoľahlivosti podľa kapitoly 4 (časť 4.4.1).

Nasledujúca schéma je návod na postup pri výbere vhodnej testovacej štatistiky pri testovaní hypotézy o zhode priemeru s konštantou.



Obr. 5.4 Výber testovacej štatistiky pri testovaní hypotéz o zhode priemeru μ s konštantou μ_0

Príklad 5.1

Zaujíma nás, či počet bodov, ktorý na prijímacích skúškach z matematiky v priemere získajú študenti 1. fakulty je menší ako 13. Ako výberový súbor použijeme súbor PRIJÍMACIE SKÚŠKY/1.fakulta/mat. o ktorom predpokladáme, že má normálne rozdelenie. O tomto výberovom súbore už máme isté informácie

z predchádzajúcej kapitoly (Tab. 4.1), a tak na základe zisteného výberového priemeru sformulujeme nulovú a alternatívnu jednostrannú hypotézu v tvare:

$$H_0: \ \mu = 13 ,$$

$$H_1: \ \mu < 13 .$$

Vzhľadom na to, že rozptyl základného súboru nie je známy, použijeme testovaciu štatistku (5.14), ktorá má t-rozdelenie s n-1 stupňami voľnosti. Použitím výberového priemeru a výberovej štandardnej odchýlky, hodnoty ktorých získame z tabuľky popisnej štatistiky, prípadne použitím štatistických funkcií AVERAGE a STDEV, dostaneme hodnotu testovacej štatistiky

$$t = \frac{9{,}35 - 13}{\frac{6{,}575}{\sqrt{38}}} = -3{,}422 .$$

Našu úlohu budeme riešiť na hladine významnosti α = 0,05. Použitím štatistickej funkcie TINV(0,1;37) získame kritickú hodnotu $t_{1-\alpha,n-1} = t_{0{,}95,37} = 1{,}687$, pomocou ktorej určíme oblasť zamietnutia nulovej hypotézy, ktorou je interval $(-\infty; t_{\alpha;n-1}) = (-\infty; -1{,}687)$. Vzhľadom na to, že testovacia štatistika leží v tejto oblasti, zamietame na hladine α = 0,05 hypotézu H_0 a prijímame alternatívnu hypotézu, ktorá hovorí, že študenti 1. fakulty získajú na prijímacej skúške z matematiky v priemere menej ako 13 bodov.

Na základe Poznámky 5.1, úlohu môžeme riešiť i využitím jednostranného intervalu spoľahlivosti pre priemer základného súboru. Ľahko vypočítame (pozri 4.4.1), že 95% pravostranný interval spoľahlivosti pre priemer základného súboru je $\left(-\infty; \bar{x} + t_{1-\alpha;n-1}\frac{S}{\sqrt{n}}\right) = (-\infty; 11{,}15)$. Hypotézu H_0 prijímame na hladine významnosti α = 0,05, ak μ_0 leží v tomto intervale. Vzhľadom na to, že μ_0 =13 v tomto intervale neleží, zamietame hypotézu H_0 a prijímame alternatívnu hypotézu H_1, pričom chyba, ktorej sa dopustíme v prípade že H_0 je pravdivá, je 5% .

5.3 Test hypotézy o rozptyle základného súboru

V tejto časti budeme predpokladať, že znak X má v základnom súbore rozdelenie $N(\mu, \sigma^2)$. Bude nás zaujímať, či hodnota σ^2 je rovná konštante σ_0^2. Testujeme teda hypotézu

$$H_0 : \sigma^2 = \sigma_0^2 .$$

Alternatívna hypotéza je v prípade dvojstranného testu

$$H_1 : \sigma^2 \neq \sigma_0^2 .$$

V prípade jednostranného testu

$$H_1 : \sigma^2 > \sigma_0^2 ,$$

alebo

$$H_1 : \sigma^2 < \sigma_0^2 .$$

Použijeme testovaciu štatistiku

$$W = \frac{(n-1)S^2}{\sigma_0^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})}{\sigma_0^2} , \qquad (5.15)$$

kde S^2 je výberový rozptyl. Štatistika (5.15) má χ^2 – rozdelenie s n-1 stupňami voľnosti. Na vymedzenie oblasti zamietnutia H_0 preto použijeme kvantily χ^2 – rozdelenia pre zvolenú hladinu významnosti. Oblasti zamietnutia hypotézy H_0 pre rôzne alternatívne hypotézy sú uvedené v nasledujúcej tabuľke.

Tab. 5.2 Testovanie hypotézy o zhode rozptylu σ^2 s konštantou σ_0^2

Hypotéza H_0	**Hypotéza H_1**	**Predpoklady**	**Testovacia štatistika**	**Oblasť zamietnutia H_0**
$\sigma^2 = \sigma_0^2$	1. $\sigma^2 \neq \sigma_0^2$ 2. $\sigma^2 > \sigma_0^2$ 3. $\sigma^2 < \sigma_0^2$	znak X má v ZS rozdelenie $N(\mu, \sigma^2)$	$W = \frac{(n-1)S^2}{\sigma_0^2}$	1. $w \leq \chi^2_{\alpha/2,n-1}$ alebo $w \geq \chi^2_{1-\alpha/2,n-1}$ 2. $w \geq \chi^2_{1-\alpha,n-1}$ 3. $w \leq \chi^2_{\alpha,n-1}$

Poznámka 5.2

Podobne ako v prípade testovania hypotézy o zhode priemeru s konštantou, pri ktorom sme využili intervaly spoľahlivosti pre priemer, môžeme postupovať i pri testovaní hypotézy o zhode rozptylu s konštantou. Nulovú hypotézu o zhode σ^2 s konštantou σ_0^2 v prípade dvojstrannej alternatívnej hypotézy prijímame, ak $\sigma_0^2 \in \left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2};n-1}}, \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2};n-1}} \right)$. Ak alternatívna hypotéza je $H_1 : \sigma^2 > \sigma_0^2$ ($H_1 : \sigma^2 < \sigma_0^2$), tak H_0 prijímame, ak $\sigma_0^2 > \frac{(n-1)S^2}{\chi^2_{1-\alpha;n-1}}$ $\left(\sigma_0^2 < \frac{(n-1)S^2}{\chi^2_{\alpha;n-1}} \right)$.

Príklad 5.2

Môžeme na hladine významnosti α = 0,05 tvrdiť, že rozptyl priemerného počtu bodov z matematiky na prijímacích skúškach na 1. fakulte je 30? Opäť použijeme výberový súbor PRIJÍMACIE SKÚŠKY/1.fakulta/mat. Testujeme hypotézu

$$H_0 : \sigma^2 = 30 .$$

Ako alternatívnu hypotézu zvolíme dvojstrannú hypotézu

$$H_1 : \sigma^2 \neq 30 .$$

Hodnotu testovacej štatistiky, ktorá má χ^2 – rozdelenie s n-1 stupňami voľnosti, získame dosadením rozsahu výberového súboru (n = 38) a výberového rozptylu ($s^2 = 43{,}236$) do testovacej štatistiky (5.15), a dostaneme

$$w = \frac{37{,}43{,}236}{30} = 53{,}325 .$$

Použitím štatistickej funkcie CHIINV, konkrétne CHIINV(0,975;37) získame kritickú hodnotu $\chi^2_{\alpha/2,n-1} = \chi^2_{0{,}025;37} = 22{,}11$ a voľbou CHIINV(0,025;37) dostaneme $\chi^2_{1-\alpha/2,n-1} = \chi_{0.975;37} = 59{,}19$, ktoré vymedzujú oblasť na prijatie H_0, t.j. interval $(22{,}11; 59{,}19)$. Vzhľadom na to, že vypočítaná testovacia štatistika w =53,325 je z tohto intervalu, nie je dôvod zamietnuť nulovú hypotézu a môžeme na hladine významnosti α = 0,05 tvrdiť, že rozptyl priemerného počtu bodov z matematiky na prijímacích skúškach na 1. fakulte je 30.

Na základe Poznámky 5.2 môžeme našu úlohu riešiť i pomocou intervalov spoľahlivosti. Konkrétne nás bude zaujímať, či σ_0^2 leží v 95% dvojstrannom intervale spoľahlivosti pre rozptyl základného súboru. V Príklade 4.2 sme vypočítali, že spomínaný interval spoľahlivosti je $(28{,}74; 72{,}35)$. Vzhľadom na to, že $\sigma_0^2 = 30$, leží v tomto intervale prijímame hypotézu H_0, čo je v súlade s predchádzajúcim záverom testovania.

5.4 Test hypotézy o podiele základného súboru

Podobne ako v predchádzajúcich častiach, budeme testovať hypotézu o zhode podielu π základného súboru s konštantou π_0. Nulová hypotéza má tvar

$$H_0 : \pi = \pi_0 .$$

Alternatívna hypotéza je v prípade dvojstranného testu

$$H_1 : \pi \neq \pi_0 ,$$

a pre jednostranný test je

$$H_1 : \pi > \pi_0 ,$$

alebo

$$H_1 : \pi < \pi_0 .$$

Z predchádzajúcich kapitol vieme, že ak rozsah výberového súboru je dostatočne veľký, môžeme rozdelenie výberového podielu aproximovať normálnym rozdelením $\mathrm{N}\left(\pi, \sqrt{\pi(1-\pi)/n}\right)$. To nám umožní použiť testovaciu štatistiku

$$Z = \frac{P - \pi_0}{\sqrt{\frac{P(1-P)}{n}}}, \tag{5.16}$$

ktorá má rozdelenie N(0,1).

Oblasti zamietnutia hypotézy H_0 pre rôzne alternatívne hypotézy sú uvedené v nasledujúcej tabuľke.

Tab. 5.3 Testovanie hypotézy o zhode podielu π s konštantou π_0

Hypotéza H_0	**Hypotéza H_1**	**Predpoklady**	**Testovacia štatistika**	**Oblasť zamietnutia H_0**
$\pi = \pi_0$	1. $\pi \neq \pi_0$ 2. $\pi > \pi_0$ 3. $\pi < \pi_0$	znak X má v ZS ľubovoľné rozdelenie $n > \frac{9}{P(1-P)}$	$Z = \frac{P - \pi_0}{\sqrt{\frac{P(1-P)}{n}}}$	1. $\lvert z \rvert > z_{1-\alpha/2}$ 2. $z > z_{1-\alpha}$ 3. $z < -z_{1-\alpha}$

Príklad 5.3

Zaujíma nás, či podiel gymnazistov medzi študentmi vojenskej akadémie je 40%. Ako výberový súbor použijeme všetkých študentov (oboch fakúlt), ktorí sú v danom roku študentmi vojenskej akadémie (pozri súbory PRIJÍMACIE SKÚŠKY/1.fakulta/stred. škola a PRIJÍMACIE SKÚŠKY/ 2.fakulta/stred. škola). Nulová hypotéza má tvar

$$H_0 : \pi = 0{,}4 .$$

Vzhľadom na to, že výberový podiel je $P = 0{,}27$ (Príklad 4.3), zvolíme alternatívnu hypotézu v tvare

$$H_1 : \pi < 0{,}4 .$$

Uvedené hypotézy budeme testovať na hladine $\alpha = 0{,}05$. Rozsah výberového súboru je 88, takže je vhodné použiť testovaciu štatistiku (5.16), ktorá má rozdelenie N(0,1). Dosadením výberového podielu a rozsahu výberového súboru do vzťahu (5.16) dostaneme

$$z = \frac{0{,}27 - 0{,}4}{\sqrt{\frac{0{,}27\left(1 - 0{,}27\right)}{88}}} = -2{,}96 .$$

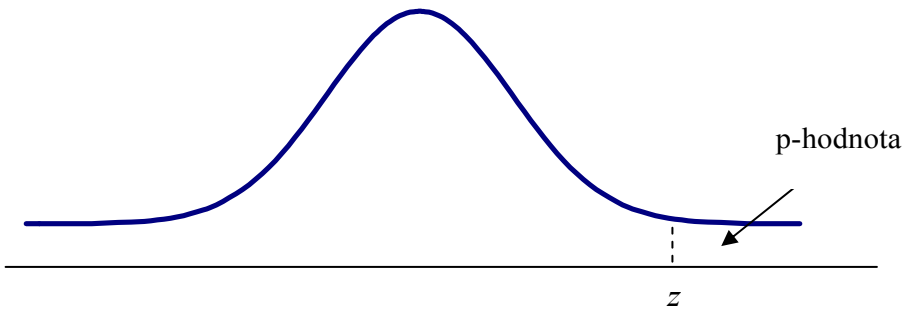
Na vymedzenie oblasti prijatia, resp. zamietnutia nulovej hypotézy je potrebné určiť kritickú hodnotu, ktorou je v tomto prípade 5% kvantil normovaného normálneho rozdelenia. Použitím štatistickej funkcie NORMSINV vypočítame jeho hodnotu, ktorá je $z_{0,05} = -1{,}645$. Vzhľadom na to, že vypočítaná testovacia štatistika je menšia ako kritická hodnota, zamietame hypotézu H_0 a na hladine významnosti $\alpha = 0{,}05$ prijímame alternatívnu hypotézu, ktorá hovorí, že podiel gymnazistov medzi študentmi vojenskej akadémie je menší ako 40%.

5.5 p-hodnota

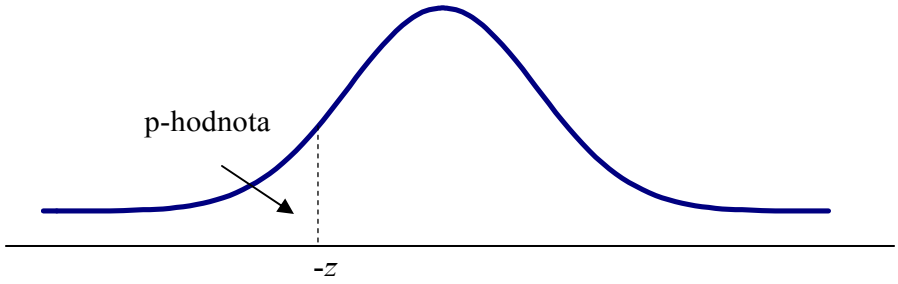
Pri testovaní hypotéz robíme rozhodnutie o prijatí, respektíve zamietnutí nulovej hypotézy na základe toho, či vypočítaná testovacia štatistika leží v oblasti prijatia nulovej hypotézy, pričom veľkosť tejto oblasti závisí od vopred zvolenej hladiny významnosti

α. Položme si teraz otázku. Pre aké hladiny významnosti môžeme nulovú hypotézu prijať, prípadne aká je najnižšia možná hladina významnosti, na ktorej už H_0 zamietame? Naznačme riešenie tohto problému na príklade.

Testujme hypotézu $H_0 : \mu = \mu_0$, pri alternatívnej hypotéze $H_1 : \mu > \mu_0$. Nech z je hodnota použitej testovacej štatistiky vypočítanej z daného výberového súboru, ktorá má normované normálne rozdelenie N(0,1). Táto hodnota je súčasne i hodnota niektorého z kvantilov rozdelenia N(0,1). Označme p plochu pod grafom funkcie hustoty, ktorá sa nachádza vpravo od hodnoty testovacej štatistiky. Z Obr. 5.6 je zrejmé, že hodnota testovacej štatistiky bude v oblasti prijatia H_0, ak pre hladinu významnosti α platí, že $\alpha < p$. V prípade, že $\alpha \geq p$ zamietame H_0 a prijímame alternatívnu hypotézu H_1. Hodnota p nazývaná **p-hodnota** je teda najnižšia hladina na zamietnutie nulovej hypotézy. Čím menšia je p-hodnota, tým viac sme presvedčení, že H_0 nie je správna a mala by byť prijatá alternatívna hypotéza H_1. Analogicky budeme postupovať i pri určení p-hodnoty v prípade ľavostranného testu, t.j. v prípade, že $H_1 : \mu < \mu_0$.



Obr. 5.6 p-hodnota pravostranného testu



Obr. 5.7 p-hodnota ľavostranného testu

Uvažujme teraz dvojstrannú alternatívnu hypotézu $H_1 : \mu \neq \mu_0$. Opäť označme hodnotu testovacej štatistiky z. V tomto prípade je p-hodnota súčet 2 plôch, z ktorých jedna je vpravo od z a druhá vľavo od $-z$. Ak robíme testovanie hypotézy na hladine $\alpha \geq p$, hodnota testovacej štatistiky sa dostane do oblasti zamietnutia H_0.

Uvedené úvahy teda naznačujú ďalší možný spôsob rozhodovania sa o výsledku pri testovaní hypotéz. Je to porovnávanie p-hodnoty a hladiny významnosti α. Použijeme nasledujúce pravidlá.

- Nulovú hypotézu H_0 zamietame pre každú hladinu významnosti α, pre ktorú platí $\alpha \geq p$.
- Nulovú hypotézu H_0 prijímame pre každú hladinu významnosti α, pre ktorú platí $\alpha < p$.

Väčšina štatistických programových systémov vypočíta p-hodnotu, a tak pri počítačovom spracovávaní údajov a ich analýze je tento spôsob najrýchlejší. Stačí totiž, ak porovnáme p-hodnotu so zvolenou hladinou významnosti α. Vidíme, že p-hodnota poskytuje viac informácií o výsledku testu. Ak totiž poznáme p-hodnotu, vieme povedať, na akej hladine významnosti môžeme prijať H_0, prípadne pre akú hladinu už musíme nulovú hypotézu zamietnuť.

V príkladoch, ktoré sme použili pri objasňovaní p-hodnoty, sme použili testovaciu štatistiku, ktorá má rozdelenie N(0,1). Analogicky by sme však postupovali i v prípade

testovacej štatistiky s t-rozdelením, χ^2-rozdelením, prípadne s inými typmi rozdelenia pravdepodobnosti. Pri počítačovom spracovaní údajov štatistické programy počítajú p-hodnotu príslušného rozdelenia.

EXCEL

Na testovanie hypotéz o zhode priemeru s konštantou je možné v EXCELi použiť štatistickú funkciu ZTEST. Číslo, ktoré získame na výstupe tejto funkcie je hodnotou distribučnej funkcie rozdelenia N(0,1) v hodnote testovacej štatistiky *z*. Doplnok tejto hodnoty do 1 je vlastne p-hodnota pre jednostranný pravostranný test, na základe ktorej sa vieme rozhodnúť, či prijímame nulovú, alebo alternatívnu hypotézu. Pripomíname, že p-hodnota pre dvojstrannný test je dvojnásobok p-hodnoty pre test jednostranný.

5.6 Test hypotézy o zhode priemerov dvoch základných súborov

Tento test patrí medzi najčastejšie používané štatistické testy, pretože umožňuje porovnávať dva základné súbory a na základe toho robiť o nich úsudky. Budeme teda uvažovať dva základné súbory, na ktorých sledujeme hodnoty znaku *X*. Z každého z nich vytvoríme náhodný výberový súbor, ktorý použijeme na overovanie hypotézy o zhode priemerov v základných súboroch. I v tomto prípade, pri voľbe vhodnej testovacej štatistiky významnú úlohu zohráva rozdelenie pravdepodobnosti znaku *X* v oboch základných súboroch, rozsah výberových súborov, prípadne rovnosť rozptylov v základných súboroch. Pri voľbe testovacej štatistiky treba prihliadať i na to, či výberové súbory sú nezávislé, čo znamená, že hodnoty znaku *X* sledované na prvkoch jedného výberového súboru nezávisia od hodnôt tohto znaku v druhom výberovom súbore. Tieto skutočnosti zohľadníme pri testovaní hypotéz o zhode priemerov dvoch základných súborov.

- **Znak *X* má v oboch základných súboroch normálne rozdelenie, rozptyly oboch základných súborov sú známe, výberové súbory sú nezávislé**

Nech teda sledovaný znak X má v prvom i v druhom základnom súbore normálne rozdelenie so známymi rozptylmi σ_1^2, σ_2^2 . Testujeme nulovú hypotézu

$$H_0 : \mu_1 = \mu_2 .$$

Ako testovacie kritérium použijeme testovaciu štatistiku

$$Z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{5.17}$$

kde $\overline{X_1}$, $\overline{X_2}$ sú priemery výberových súborov s rozsahmi n_1 *a* n_2. Táto testovacia štatistika má normované normálne rozdelenie N(0,1). Kritické hodnoty teda nájdeme ako príslušné kvantily rozdelenia N(0,1).

- **Znak *X* má v oboch základných súboroch normálne rozdelenie, rozsahy výberových súborov sú veľké, rozptyly základných súborov nepoznáme, výberové súbory sú nezávislé**

Na testovanie nulovej hypotézy $H_0 : \mu_1 = \mu_2$ použijeme testovaciu štatistiku (5.17), v ktorej sme neznáme rozptyly σ_1^2, σ_2^2 nahradili ich bodovými odhadmi. Testovacia štatistika bude mať potom tvar

$$Z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{5.18}$$

a má normované normálne rozdelenie N(0,1).

Štatistiku (5.18) je vhodné použiť i v prípade, keď nie je splnená podmienka normálneho rozdelenia pravdepodobnosti sledovaného znaku. V tomto prípade je normalita rozdelenia testovacej štatistiky (5.18) dôsledkom platnosti centrálnej limitnej vety (kapitola 3).

- **Znak *X* má v oboch základných súboroch normálne rozdelenie, rozptyly nepoznáme, môžeme predpokladať $\sigma_1^2 = \sigma_2^2$, výberové súbory sú nezávislé**

Na testovanie nulovej hypotézy $H_0 : \mu_1 = \mu_2$ oproti ľubovoľnej alternatívnej hypotézy použijeme testovaciu štatistiku

$$T = \frac{\overline{X_1} - \overline{X_2}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} , \tag{5.19}$$

kde S_p^2 je združený výberový rozptyl definovaný vzťahom

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} , \tag{5.20}$$

pričom S_1^2, S_2^2 sú výberové rozptyly. Testovacia štatistika má t-rozdelenie s n_1+n_2 -2 stupňami voľnosti. Kritické hodnoty určíme na zvolenej hladine významnosti α ako príslušné kvantily t-rozdelenia s n_1+ n_2 -2 stupňami voľnosti.

- **Znak *X* má v oboch základných súboroch normálne rozdelenie, rozptyly nepoznáme, môžeme predpokladať $\sigma_1^2 \neq \sigma_2^2$, výberové súbory sú nezávislé**

V takomto prípade na testovanie nulovej hypotézy $H_0 : \mu_1 = \mu_2$ použijeme testovaciu štatistiku v tvare

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \tag{5.21}$$

ktorá má t-rozdelenie s počtom stupňov voľnosti ν_r, ktorý nazývame redukovaný počet stupňov voľnosti, pre ktorý platí vzťah

$$v_r = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2}. \qquad (5.22)$$

V prípade, že ν_r nie je celé číslo, bude počet stupňov voľnosti najbližšie menšie celé číslo. Vzťah (5.22) sa nazýva Welchova aproximácia.

- **Výberové súbory sú závislé (párové), párové rozdiely majú normálne rozdelenie**

Pod párovými výberovými súbormi budeme rozumieť súbory, ktoré sú tvorené tými istými štatistickými jednotkami, na ktorých sledujeme ten istý znak *X,* ale v rôznych situáciách. V tomto prípade však už nemôžeme hovoriť o nezávislosti súborov. Napríklad nás zaujíma, či počet predaných produktov vo vybranej sieti predajní sa zvýšil po reklamnej kampani na daný produkt. Budeme sledovať zmenu hodnoty znaku na každej jednotke, t.j. rozdiel medzi počtom predaných produktov pred a po reklamnej kampani v jednotlivých predajniach. Nulovú hypotézu v takomto prípade je vhodné zapísať v tvare

$$H_0 : \mu_D = 0,$$

pričom μ_D je definovaný ako $\mu_D = \mu_1 - \mu_2$. Vidíme, že test o zhode priemerov dvoch základných súborov sme vlastne zredukovali na test jedného parametra. Testovaným znakom *D* je rozdiel medzi dvoma pozorovaniami daného znaku na tej istej štatistickej jednotke. Pri tomto teste budeme vychádzať z rovnakých predpokladov ako pri testoch v časti 5.2. Použijeme testovaciu štatistiku

$$T = \frac{\overline{D}}{\frac{S_D}{\sqrt{n}}}, \qquad (5.23)$$

kde $\overline{D}$ je priemer znaku *D,* S_D je štandardná odchýlka rozdielov a *n* je počet párových pozorovaní, ktorá má t-rozdelenie pravdepodobnosti s *n* -1 stupňami voľnosti.

V prípade, že počet pozorovaní je dostatočne veľký, môžeme namiesto t-rozdelenia použiť normované normálne rozdelenie.

V Tab. 5.4 sú uvedené oblasti zamietnutia nulovej hypotézy pre jednotlivé typy alternatívnych hypotéz pre všetky spomínané prípady testovania zhody priemerov v dvoch základných súboroch.

Príklad 5.4

Zaujíma nás, či vstupná úroveň študentov z matematiky je porovnateľná na oboch fakultách vojenskej akadémie. Použijeme súbory PRIJÍMACIE SKÚŠKY/1.fakulta/mat. a PRIJÍMACIE SKÚŠKY/2.fakulta/mat., o ktorých predpokladáme, že sú vybrané zo súborov s normálnym rozdelením. Nulová hypotéza H_0 bude predpokladať rovnosť priemerov základných súborov, pričom prvý je tvorený študentmi 1. fakulty vojenskej akadémie a druhý študentmi 2. fakulty, t.j.

$$H_0 : \mu_1 = \mu_2 .$$

Alternatívnu hypotézu budeme formulovať v tvare

$$H_1 : \mu_1 \neq \mu_2 .$$

Vzhľadom na to, že rozsahy oboch výberových súborov sú väčšie ako 30 a rozptyly základných súborov nepoznáme, môžeme použiť výberovú štatistiku (5.18).

Po dosadení výberových priemerov (4.1), výberových rozptylov (4.4) a rozsahov oboch výberových súborov dostaneme hodnotu testovacej štatistiky

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{9{,}35 - 15{,}12}{\sqrt{\frac{43{,}23}{38} + \frac{34{,}56}{50}}} = -4{,}27 .$$

Tab. 5.4 Testovanie hypotéz o zhode priemerov dvoch základných súborov

Hypotéza H_0	**Hypotéza H_1**	**Predpoklady**	**Testovacia štatistika**	**Oblasť zamietnutia H_0**
$\mu_1 = \mu_2$	1. $\mu_1 \neq \mu_2$ 2. $\mu_1 > \mu_2$ 3. $\mu_1 < \mu_2$	σ_1, σ_2 poznáme znak X má normálne rozdelenie v oboch ZS výberové súbory sú nezávislé	$Z = \dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	1. $\lvert z \rvert > z_{1-\alpha/2}$ 2. $z > z_{1-\alpha}$ 3. $z < -z_{1-\alpha}$
		σ_1, σ_2 nepoznáme $n_1 \geq 30$, $n_2 \geq 30$ výberové súbory sú nezávislé	$Z = \dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$	
		σ_1, σ_2 nepoznáme $\sigma_1 = \sigma_2$ znak X má normálne rozdelenie v oboch ZS výberové súbory sú nezávislé	$T = \dfrac{\overline{X_1} - \overline{X_2}}{S_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ s_P pozri (5.20)	1. $\lvert t \rvert > t_{1-\alpha/2}$ 2. $t > t_{1-\alpha}$ 3. $t < -t_{1-\alpha}$ $d.f. = n_1 + n_2 - 2$
		σ_1, σ_2 nepoznáme $\sigma_1 \neq \sigma_2$ znak X má normálne rozdelenie v oboch ZS výberové súbory sú nezávislé	$T = \dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$	1. $\lvert t \rvert > t_{1-\alpha/2}$ 2. $t > t_{1-\alpha}$ 3. $t < -t_{1-\alpha}$ *d.f.* pozri (5.22)
$\mu_D = 0$	1. $\mu_D \neq 0$ 2. $\mu_D > 0$ 3. $\mu_D < 0$	σ_1, σ_2 nepoznáme D - rozdiel medzi dvoma pozorovaniami má normálne rozdelenie výberové súbory sú závislé	$t = \dfrac{\overline{D}}{s_D / \sqrt{n}}$	1. $\lvert t \rvert > t_{1-\alpha/2}$ 2. $t > t_{1-\alpha}$ 3. $t < -t_{1-\alpha}$ $d.f. = n - 1$

Zvolenú hypotézu budeme testovať na hladine významnosti $\alpha = 0{,}05$, čo znamená že kritické hodnoty pre oblasť prijatia, resp. zamietnutia H_0 budú určené kvantilmi normovaného normálneho rozdelenia $z_{1-\alpha/2} = z_{0{,}975} = 1{,}96$ a $z_{\alpha/2} = z_{0{,}025} = -1{,}96.$ Oblasť prijatia hypotézy H_0 je teda interval $(-1{,}96;1{,}96)$. Vzhľadom na to, že testovacia štatistika neleží v tomto intervale, zamietame nulovú hypotézu a prijímame hypotézu alternatívnu, ktorá tvrdí, že vstupná úroveň študentov z matematiky nie je porovnateľná na oboch fakultách vojenskej akadémie.

Vzhľadom na výrazný rozdiel priemerov výberových súborov je vhodné formulovať alternatívnu hypotézu v tvare

$$H_1 : \mu_1 < \mu_2 .$$

Kritická hodnota tohto ľavostranného testu je určená kvantilom $z_\alpha = z_{0{,}05} = -1{,}64$ a oblasťou prijatia H_0 je interval $(-1{,}64;\infty)$. I v tomto prípade, testovacia štatistika leží v oblasti zamietnutia H_0, čo znamená, že vstupná úroveň študentov na 1. fakulte je štatisticky významne nižšia ako na 2. fakulte.

5.7 Test hypotézy o zhode rozptylov dvoch základných súborov

Pri popise konštrukcie testu zhody dvoch priemerov základných súborov sme videli, že dôležitú úlohu pri voľbe testovacej štatistiky zohráva predpoklad rovnosti rozptylov v základných súboroch. Predpoklad o zhode rozptylov na základe výberových súborov overíme nasledujúcim testom.

Uvažujme nezávislé náhodné výbery s rozsahmi n_1, n_2, ktoré sme vytvorili z dvoch základných súborov, v ktorých má sledovaný znak normálne rozdelenie. Z týchto výberov sme vypočítali výberové rozptyly s_1^2, s_2^2 (4.4). Nulovú hypotézu budeme formulovať v tvare

$$H_0 : \sigma_1^2 = \sigma_2^2 .$$

Na overenie tejto hypotézy použijeme testovaciu štatistiku

$$F = \frac{S_1^2}{S_2^2}, \tag{5.23}$$

ktorá má Fisherovo rozdelenie pravdepodobnosti so stupňami voľnosti $\nu_1 = n_1$-1 a $\nu_2 = n_2$-1. Vymedzenie kritických oblastí pre jednotlivé typy alternatívnych hypotéz je v nasledujúcej tabuľke.

Tab. 5.5 Testovanie hypotézy o zhode rozptylov dvoch základných súborov

Hypotéza H_0	**Hypotéza H_1**	**Predpoklady**	**Testovacia štatistika**	**Oblasť zamietnutia H_0**
$\sigma_1^2 = \sigma_2^2$	1. $\sigma_1^2 \neq \sigma_2^2$ 2. $\sigma_1^2 > \sigma_2^2$ 3. $\sigma_1^2 < \sigma_2^2$	znak X má normálne rozdelenie v oboch ZS	$F = \frac{S_1^2}{S_2^2}$	1. $f < F_{\alpha/2,(\nu_1,\nu_2)}$ alebo $f > F_{1-\alpha/2,(\nu_1,\nu_2)}$ 2. $f > F_{1-\alpha,(\nu_1,\nu_2)}$ 3. $f < F_{\alpha,(\nu_1,\nu_2)}$ $\nu_1 = n_1 - 1$ $\nu_2 = n_2 - 1$

Príklad 5. 5

Pýtame sa, či rozptyl bodov, ktoré študenti získajú na prijímacích skúškach z matematiky je porovnateľný na oboch fakultách vojenskej akadémie. Opäť použijeme súbory PRIJÍMACIE SKÚŠKY/1.fakulta/mat. a PRIJÍMACIE SKÚŠKY/2.fakulta /mat., o ktorých predpokladáme, že sú vybrané zo súborov s normálnym rozdelením. Nulová hypotéza H_0 hovorí o rovnosti rozptylov základných súborov, pričom prvý je tvorený študentmi 1. fakulty a druhý študentmi 2. fakulty, t.j.

$$H_0 : \sigma_1^2 = \sigma_2^2.$$

Alternatívnu hypotézu volíme v tvare

$$H_1 : \sigma_1^2 \neq \sigma_2^2,$$

pričom hypotézu budeme testovať na hladine významnosti $\alpha = 0{,}05$. Do vzťahu na výpočet testovacej štatistiky (5.23) dosadíme výberové rozptyly (4.4) vypočítané z daných výberových súborov a dostaneme

$$f = \frac{S_1^2}{S_2^2} = \frac{43{,}24}{34{,}56} = 1{,}25.$$

Oblasť prijatia hypotézy H_0 je interval, ktorého hranice sú určené hodnotami kvatilov Fisherovho rozdelenia pravdepodobnosti so stupňami voľnosti $\nu_1 = n_1\text{-}1$ a $\nu_2 = n_2\text{-}1$. Na ich výpočet použijeme v EXCELi štatistickú funkciu FINV. Konkrétne, voľbou FINV(0,975;37;49) dostaneme kvantil $F_{\alpha/2,(\nu_1,\nu_2)} = F_{0.025(37,49)} = 0{,}54$ a kvantil $F_{1-\alpha/2(\nu_1,\nu_2)} = F_{0.975(37,49)} = 1{,}82$ voľbou FINV(0,025;37;49). Vzhľadom na to, že $1{,}25 \in (0{,}54;1{,}82)$, nemáme dôvod zamietnuť na hladine významnosti $\alpha = 0{,}05$ hypotézu o rovnosti rozptylov základných súborov.

Testovanie hypotéz o zhode priemerov, prípadne rozptylov dvoch základných súborov je možné jednoducho realizovať s využitím štatistických procedúr ponúkaných EXCELom. Na záver tejto kapitoly uvedieme príklady, pri riešení ktorých budú tieto procedúry použité.

5.8 Test hypotézy o zhode podielov dvoch základných súborov

Metódy, ktoré využívame na testovanie podielov dvoch základných súborov sú založené na aproximácii binomického rozdelenia rozdelením normálnym. Už vieme (kapitola 3), že táto aproximácia vyžaduje dostatočne veľké výberové súbory. Uvažujme teda nezávislé náhodné výbery s dostatočne veľkými rozsahmi n_1, n_2, z ktorých vypočítame príslušné výberové podiely P_1, P_2. Testujeme hypotézu o zhode podielov v oboch základných súboroch v tvare

$$H_0 : \pi_1 = \pi_2 .$$

Vhodná testovacia štatistika, ktorú použijeme na jej testovanie má tvar

$$Z = \frac{P_1 - P_2}{\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}}. \tag{5.24}$$

Pre dostatočne veľké n_1, n_2 môžeme rozdelenie tejto testovacej štatistiky aproximovať rozdelením N(0,1) a kritické hodnoty vypočítať ako príslušné kvantily normovaného normálneho rozdelenia. Vymedzenie oblastí zamietnutia H_0 nájdeme v nasledujúcej tabuľke.

Tab. 5.5 Testovanie hypotézy o zhode podielov dvoch základných súborov

Hypotéza H_0	**Hypotéza H_1**	**Predpoklady**	**Testovacia štatistika**	**Oblasť zamietnutia H_0**
$\pi_1 = \pi_2$	1. $\pi_1 \neq \pi_2$ 2. $\pi_1 < \pi_2$ 3. $\pi_1 > \pi_2$	znak X má normálne rozdelenie v oboch ZS	$Z=\frac{P_1-P_2}{\sqrt{\frac{P_1(1-P_1)}{n_1}+\frac{P_2(1-P_2)}{n_2}}}$	1. $\lvert z\rvert > z_{1-\alpha/2}$ 2. $z < -z_{1-\alpha}$ 3. $z > z_{1-\alpha}$

Príklad 5.6

Zaujíma nás, či je rozdiel v percentuálnom zastúpení žien študujúcich na jednotlivých fakultách vojenskej akadémie. Ako výberové súbory použijeme študentov 1. a 2. fakulty, ktorí v danom roku študovali na vojenskej akadémii. Zistili sme, že medzi 85 študentmi 1. fakulty bolo 12 žien a na 2. fakulte bolo z celkového počtu 69 študentov 14 žien. Budeme testovať hypotézu o zhode podielov v oboch základných súboroch v tvare

$$H_0 : \pi_1 = \pi_2 .$$

K hypotéze H_0, ktorú budeme testovať na hladine významnosti $\alpha = 0{,}1$ zvolíme alternatívnu hypotézu v tvare

$$H_1 : \pi_1 \neq \pi_2 .$$

Na výpočet testovacej štatistiky použijeme vzťah (5.24), do ktorého dosadíme výberové podiely a rozsahy jednotlivých výberových súborov. Dostaneme

$$z = \frac{P_1 - P_2}{\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}} = \frac{0{,}143 - 0{,}203}{\sqrt{\frac{0{,}143(1-0{,}143)}{85} + \frac{0{,}203(1-0{,}203)}{69}}} = -0{,}976 .$$

Oblasť prijatia hypotézy H_0 je vymedzená kvantilmi $z_{\frac{\alpha}{2}} = z_{0,05} = -1{,}645$ a $z_{1-\frac{\alpha}{2}} = z_{0,95} = 1{,}645$. Je zrejmé, že $-0{,}976 \in (-1{,}645; 1{,}645)$, čo znamená, že na hladine významnosti $\alpha = 0{,}1$, nie je dôvod zamietnuť hypotézu H_0, takže môžeme tvrdiť, že percentuálne zastúpenie žien študujúcich na 1. a 2. fakulte je porovnateľné.

EXCEL

Na testovanie hypotéz o zhode priemerov, prípadne rozptylov dvoch súborov poskytuje EXCEL niekoľko testov. Voľbou ***Nástroje/Analýza údajov*** si môžeme z ponuky vybrať test v závislosti od toho, aké predpoklady spĺňajú naše výberové súbory. ***Dvojvýberový z-test na strednú hodnotu*** použijeme v prípade, keď oba výberové súbory majú rozsah väčší ako 30 a sú vybrané z nezávislých súborov. Pri voľbe tohto testu sa objaví dialógové okno, do ktorého vkladáme požadované informácie, ako je vstupná oblasť, rozptyly (v prípade, že rozptyly nie sú známe zadáme ich bodové odhady), hladina významnosti a pod. Výstupná tabuľka dvojvýberového z-testu na strednú hodnotu obsahuje okrem bodových odhadov stredných hodnôt oboch výberových súborov a ich rozsahov i hodnotu testovacej štatistiky *z*. Symbolom *z krit*(2) je vo výstupnej tabuľke označený kvantil normovaného normálneho rozdelenia $z_{1-\alpha/2}$, pomocou ktorého vymedzíme kritickú oblasť pre dvojstranný test. Symbolom *z krit(1)* je označený kvantil z_α, potrebný pre vymedzenie kritickej oblasti jednostranného testu. Test je založený na porovnaní hodnôt z a *z krit*(2), resp. *z krit*(1). Ak $z > z\,krit(2)$ alebo $z < -z\,krit(2)$ neprijímame H_0 a prijímame hypotézu H_1. V opačnom prípade nie je dôvod H_0 na danej hla-

dine významnosti zamietnuť. V prípade, že volíme jednostrannú alternatívnu hypotézu H_1: $\mu_1 > \mu_2$ ($\mu_1 < \mu_2$), neprijímame H_0, ak platí $z > z\,krit(1)$ ($z < -z\,krit(1)$). Vo výstupnej tabuľke je symbolom $P(Z <= z)(2)$ resp. $P(Z <= z)(1)$ označená p-hodnota pre dvojstranný, resp. jednostranný test.

EXCEL nám ďalej ponúka ***dvojvýberový t-test s rovnosťou rozptylov***, prípadne ***dvojvýberový t-test s nerovnosťou rozptylov***, ktoré sa zvyknú používať v prípade, keď rozsahy výberových súborov sú malé, ale vhodné sú i pre výberové súbory ľubovoľných rozsahov. Už z názvu týchto testov je zrejmé, že skôr, ako sa rozhodneme pre niektorý z nich, je potrebné testovať hypotézu o zhode rozptylov. Pre takýto prípad si z ponuky analytických nástrojov vyberieme ***F-test.***

Vo vstupnom dialógovom okne ***F-testu***, podobne ako v predchádzajúcom teste, vyplníme požadované informácie. Výstupná tabuľka obsahuje bodové odhady stredných hodnôt a rozptylov oboch súborov, rozsah výberových súborov, stupne voľnosti, testovaciu štatistiku F, ale iba kritickú hodnotu $F\,krit(1)$ a p-hodnotu $P(F <= f)(1)$. Znamená to, že dáva informácie len pre jednostranný test. V prípade, že chceme použiť tento test pre dvojstrannú alternatívnu hypotézu, odporúčame zadať hladinu významnosti $\alpha/2$, čím získame kritickú hodnotu potrebnú pri rozhodovaní v dvojstrannom teste.

Dvojvýberový t-test s rovnosťou rozptylov a ***Dvojvýberový t-test s nerovnosťou rozptylov*** sú vhodné pre výberové súbory ľubovoľného rozsahu, ktoré sú vybrané z dvoch nezávislých súborov s normálnym rozdelením pravdepodobnosti a s rovnakými rozptylmi, resp. rôznymi rozptylmi. Vo vstupnom dialógovom okne opäť zadáme adresy buniek oboch výberových súborov, zvolíme výstupnú oblasť, hypotetický rozdiel stredných hodnôt a hladinu významnosti. Na výpočet testovacej štatistiky *t stat* je v EXCELi použité Studentovo t-rozdelenie pravdepodobnosti. Vo výstupnej tabuľke nájdeme podobne ako v predchádzajúcom teste bodové odhady stredných hodnôt, rozsahy výberových súborov a navyše počet stupňov voľnosti (vo výstupnej tabuľke označený ako *rozdiel*), ktorý bol použitý na výpočet kritickej hodnoty *t krit*(2), resp. *t krit*(1). Symboly $P(T <= t)(2)$ resp. $P(T <= t)(1)$ majú ten istý význam ako v predchádzajúcich testoch.

Párový t-test na strednú hodnotu volíme v prípade, že testujeme závislé, tzv. párové súbory. Vo vstupnom dialógovom okne zadáme adresy buniek oboch výberových súborov, ktoré musia mať v tomto teste rovnaký rozsah, výstupnú oblasť i hladinu významnosti. Výstupná tabuľka udáva bodové odhady stredných hodnôt a rozptylov oboch základných súborov, rozsah súborov, *Pearsonov korelačný koeficient*, ktorý udáva silu lineárnej závislosti medzi oboma súbormi, *hypotetický rozdiel*, ktorý volíme s ohľadom na nulovú hypotézu, počet stupňov voľnosti *(rozdiel),* testovaciu štatistiku *t stat*, kritické hodnoty *t krit*(2), resp. *t krit*(1*)*, ktoré použijeme podľa voľby alternatívnej hypotézy a už spomínané p hodnoty $P(T <= t)(2)$ resp. $P(T <= t)(1)$, ktoré udávajú najnižšiu hladinu významnosti pre zamietnutie nulovej hypotézy.

Príklad 5.7

Vráťme sa k Príkladu 5.4 a k Príkladu 5.5 a riešme ich použitím procedúr ponúkaných EXCELom.

Zaujíma nás teda, či študenti oboch fakúlt vojenskej akadémie majú porovnateľnú vstupnú úroveň z matematiky. Predpokladáme, že oba zvolené výberové súbory sú vybrané zo súborov s normálnym rozdelením, sú nezávislé, a tak je vhodné použiť niektorý z dvojvýberových t-testov. Predtým je však potrebné urobiť F-test na testovanie zhody rozptylov základných súborov. Na hladine významnosti $\alpha = 0{,}05$ začneme s testovaním hypotézy

$$H_0 : \sigma_1^2 = \sigma_2^2,$$

pričom alternatívnu hypotézu zvolíme v tvare

$$H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Voľbou ***Nástroje/Analýza údajov/Dvojvýberový F-test pre rozptyl*** získame vstupné dialógové okno ***F-testu***. Pripomíname, že naša alternatívna hypotéza je dvojstranná a F-test dáva informáciu len o teste jednostrannom. Ak chceme získať kritickú hodnotu, ktorá prislúcha dvojstrannej alternatívnej hypotéze, do okienka z názvom *alfa* zadáme hodnotu 0,025. Po vyplnení požadovaných informácii získame na výstupe nasledujúcu tabuľku.

Tab. 5.6 Dvojvýberový F-test pre rozptyl

	mat. 1.F	*mat. 2. F*
Stř. hodnota	9,35	15,12
Rozptyl	43,24	34,56
Pozorování	38	50
Rozdíl	37	49
F	1,25	
P(F<=f) (1)	0,23	
F krit (1)	1,82	

Ak porovnáme hodnoty v Tab. 5.6 s výsledkami v Príklade 5.5 vidíme, že testovacia štatistika F je v oboch prípadoch 1,25 a hodnota v poslednom riadku tabuľky je kvantil $F_{1-\alpha/2(\nu_1,\nu_2)} = F_{0.975(37,49)} = 1{,}82$. Vo výstupe uvedeného testu nájdeme vždy iba jednu z kritických hodnôt, pričom hodnota kvantilu $F_{\alpha/2,(\nu_1,\nu_2)}$ bude v poslednom riadku tabuľky v prípade, keď testovacia štatistika $F < 1$ (stačí zameniť poradie súborov). Vzhľadom na to, že testovacia štatistika je menšia ako kritická hodnota a určite väčšia ako $F_{\alpha/2,(\nu_1,\nu_2)}$, prijímame hypotézu H_0 o rovnosti rozptylov v oboch základných súboroch. Hodnota, označená ako *P(F<=f) (1)* v predposlednom riadku Tab. 5.6 udáva p-hodnotu jednostranného testu. V prípade dvojstranného testu je táto hodnota dvojnásobná. Takže i na základe p-hodnoty, ktorá je väčšia ako hladina významnosti $\alpha = 0{,}05$, prijímame hypotézu o rovnosti rozptylov v základnom súbore.

Na základe tohto výsledku, na testovanie hypotézy o zhode priemerov použijeme ***Dvojvýberový t-test s rovnosťou rozptylov***. Budeme teda na hladine významnosti $\alpha = 0{,}05$ testovať nulovú hypotézu

$$H_0 : \mu_1 = \mu_2 ,$$

ktorou vyslovujeme predpoklad o porovnateľnej vstupnej úrovni študentov oboch fakúlt z matematiky a alternatívnu hypotézu budeme formulovať v tvare

$$H_1 : \mu_1 \neq \mu_2 .$$

Výberom tohto testu z ponuky *Analýzy údajov* sa opäť objaví dialógové okno, do ktorého vkladáme požadované informácie. Na výstupe dostaneme nasledujúcu tabuľku.

Tab. 5.7 Dvojvýberový t-test s rovnosťou rozptylov

	mat. 1.F	*mat. 2.F*
Stř. hodnota	9,35	15,12
Rozptyl	43,24	34,56
Pozorování	38	50
Společný rozptyl	38,29	
Hyp. rozdíl stř. hodnot	0	
Rozdíl	86	
t stat	-4,33	
P(T<=t) (1)	1,99E-05	
t krit (1)	1,66	
P(T<=t) (2)	3,97E-05	
t krit (2)	1,99	

Výstupná tabuľka tohto testu obsahuje informácie pre jednostranné i dvojstranné testovanie. Vzhľadom na našu dvojstrannú alternatívnu hypotézu, budú pre nás zaujímavé údaje označené symbolom *(2)* v posledných dvoch riadkoch tabuľky. V poslednom riadku je kvantil $t_{1-\alpha/2;n-1} = t_{0,975;85}$, ktorý je kritickou hodnotou na vymedzenie oblasti prijatia H_0. Vidíme, že testovacia štatistika označená vo výstupnej tabuľke ako *t stat* neleží v oblasti prijatia nulovej hypotézy, ktorou je interval $(-1{,}99;1{,}99)$. Znamená to, že na hladine významnosti $\alpha = 0{,}05$ zamietame hypotézu H_0. V predposlednom riadku tabuľky je symbolom *P(T<=t)(2)* označená p-hodnota dvojstranného testu, na základe ktorej vieme povedať, že nulovú hypotézu zamietame na každej hladine významnosti, ktorá je väčšia ako 3,97E-05. Veľmi nízka p-hodnota naznačuje, že máme významný dôvod na zamietnutie nulovej hypotézy.

Našu úlohu môžeme vzhľadom na rozsah výberových súborov riešiť i použitím ***Dvojvýberového z-testu na strednú hodnotu***. Vo vstupnom dialógovom okne tohto testu opäť zadáme údaje použité v predchádzajúcom teste, pričom do okienka

známy rozptyl vložíme bodové odhady rozptylov základných súborov. Na výstupe dostaneme nasledujúcu tabuľku.

Tab. 5.8 Dvojvýberový z-test na strednú hodnotu

	mat. 1.F	*mat. 2.F*
Stř. hodnota	9,35	15,12
Známý rozptyl	43,23	34,56
Pozorování	38	50
Hyp. rozdíl stř. hodnot	0	
z	-4,27	
P(Z<=z) (1)	9,92745E-06	
z krit (1)	1,64	
P(Z<=z) (2)	1,98549E-05	
z krit (2)	1,96	

Interpretácia hodnôt v tabuľke je analogická ako v predchádzajúcom prípade. Všimnime si, že údaje v tabuľke sú totožné s hodnotami vypočítanými v Príklade 5.4.

Príklad 5.8

Môžeme na hladine významnosti $\alpha = 0{,}05$ tvrdiť, že hodnotenie študentov na skúške z matematiky v 3. semestri sa výrazne nelíši od hodnotenia v 2. semestri?

Ako výberový súbor použijeme hodnotenie študentov 2. fakulty v 2. a v 3. semestri, ktoré nájdeme v súbore PRIJÍMACIE SKÚŠKY/2.fakulta. Vzhľadom na to, že výberové súbory sú závislé (párové) použijeme *dvojvýberový párový t-test*, pričom predpokladáme, že párové rozdiely majú normálne rozdelenie. Na hladine významnosti $\alpha = 0{,}05$ budeme testovať nulovú hypotézu

$$H_0 : \mu_2 = \mu_3 \ \ (\mu_2 - \mu_3 = 0),$$

ktorou vyslovujeme predpoklad o porovnateľných výsledkoch na skúške z matematiky v 2. i v 3. semestri . Alternatívnu hypotézu budeme formulovať v tvare

$$H_1 : \mu_2 \neq \mu_3 \ \ (\mu_2 - \mu_3 \neq 0).$$

Voľbou ***Nástroje/Analýza údajov/Dvojvýberový párový t-test na strednú hodnotu*** a vložením požadovaných údajov do dialógového okna získame na výstupe nasledujúcu tabuľku.

Tab. 5.9 Dvojvýberový párový t-test na strednú hodnotu

	mat.(upr) 2	*mat.(upr) 3*
Stř. hodnota	2,56	2,58
Rozptyl	0,62	0,37
Pozorování	50,00	50,00
Pears. korelace	0,82	
Hyp. rozdíl stř. hodnot	0,00	
Rozdíl	49,00	
t stat	-0,22	
P(T<=t) (1)	0,41	
t krit (1)	1,68	
P(T<=t) (2)	0,83	
t krit (2)	2,01	

Z prvého riadku Tab. 5.9 je zrejmé, že bodový odhad priemeru celkového hodnotenia na skúškach v 1. semestri sa málo líši od priemeru celkového hodnotenia v 3. semestri. O tom, či môžeme na zvolenej hladine významnosti prijať nulovú hypotézu, však rozhodneme až na základe porovnania testovacej štatistiky *(t stat)* a kritickej hodnoty *(t krit 2).* Z tabuľky je zrejmé, že testovacia štatistika, ktorej hodnota je –0,22 leží v oblasti prijatia H_0, ktorou je interval $(-2{,}01;\ 2{,}01)$. Správnosť rozhodnutia o prijatí hypotézy o zhode priemerov nám potvrdzuje i p-hodnota *(P(T<=t) (2)),* ktorá je výrazne väčšia ako zvolená hladina významnosti $\alpha = 0{,}05$.

5.9 Analýza rozptylu (ANOVA)

V prípade, že chceme porovnávať priemery viacerých základných súborov, je vhodné použiť aparát analýzy rozptylu ANOVA (ANalysis Of VAriance). Názov vyplýva zo skutočnosti, že pri tejto metóde vlastne testujeme vplyv úrovní daného faktora na variabilitu hodnôt analyzovanej premennej. Variabilita hodnôt sledovaného znaku môže byť výsledkom viacerých príčin. Zvykneme ich rozdeľovať na príčiny systematické a náhodné.

Základné súbory, ktorých priemery porovnávame, zväčša vzniknú rozdelením jedného základného súboru podľa úrovní istého faktora. Napríklad, súbor študentov danej školy, na ktorom sledujeme úspešnosť študentov na prijímacích skúškach, môžeme rozdeliť podľa faktora, ktorým je príslušnosť k jednej z fakúlt školy. Jednotlivé fakulty sú teda úrovne daného faktora. Môžeme overiť, či vstupná úroveň študentov je na jednotlivých fakultách porovnateľná. V mnohých prípadoch je prirodzené očakávať, že hodnota sledovaného znaku závisí od úrovní viacerých faktorov. V predchádzajúcom príklade by sme napríklad mohli sledovať i vplyv absolvovanej strednej školy, prípadne pohlavia na výsledky prijímacích skúšok. V závislosti od toho, či sledujeme vplyv jedného, alebo viacerých faktorov, hovoríme o **jednofaktorovej** alebo **viacfaktorovej analýze rozptylu**. Vzhľadom na zameranie tejto knižky a rozsah problematiky, budeme sa zaoberať iba jednofaktorovou analýzou rozptylu. Podrobnejšie o viacfaktorovej analýze sa môžete dočítať napr. v [6].

Jednofaktorová analýza rozptylu

Uvažujme výberové súbory, na ktorých budeme sledovať vplyv úrovne jedného faktora na hodnotu sledovaného znaku. Na zvolenej hladine významnosti budeme testovať hypotézy o zhode priemerov v jednotlivých základných súboroch, z ktorých boli uskutočnené náhodné výbery. Nulová hypotéza bude mať tvar

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k , \quad \text{pre } k > 2 \tag{5.25}$$

a alternatívna hypotéza bude

$$H_1 : \text{ Aspoň dva priemery sa nerovnajú.} \tag{5.26}$$

Pre použitie metódy analýzy rozptylu je potrebné, aby boli splnené tri základné predpoklady:

- všetky porovnávané výberové súbory pochádzajú zo základných súborov s normálnym rozdelením,
- výbery sú nezávislé,
- rozptyly všetkých základných súborov sa rovnajú (homoskedasticita).

Predtým, ako budeme aplikovať metódy analýzy rozptylu, je teda potrebné overiť splnenie jednotlivých predpokladov. Na overenie prvého predpokladu použijeme metódy uvedené v nasledujúcej kapitole. V prípade, že základné súbory nemajú normálne rozdelenie, je vhodné použiť Kruskalov –Wallisov test, ktorý je popísaný v kapitole 7. Podmienku nezávislosti náhodných výberov vieme zväčša logicky posúdiť. Na overenie rovnosti rozptylov základných súborov je možné použiť napr. Bartlettov test homoskedasticit (pozri [6]). Overenie tohto predpokladu je však významné v prípadoch, keď medzi rozsahmi jednotlivých výberových súborov sú veľké rozdiely. V takýchto prípadoch hovoríme o **nevyvážených modeloch**. V prípade **vyváženého modelu**, t.j. keď rozsahy všetkých výberových súborov sú rovnaké, nesplnenie podmienky homoskedasticity nemá významný vplyv na skreslenie výsledkov testu. Preto je vhodné použiť na testovanie vplyvu úrovne daného faktora vyvážený model.

Cieľom analýzy rozptylu je porovnať priemery $k\ (k > 2)$ základných súborov. Po uskutočnení výberu z každého základného súboru vypočítame pre každý z nich výberový priemer

$$\overline{x_i} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \tag{5.27}$$

a celkový výberový priemer

$$\overline{x} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}}{n}, \tag{5.28}$$

pričom x_{ij} je *j*-ta štatistická jednotka v *i*-tom výberovom súbore, n_i je rozsah *i*-teho výberového súboru a $n = n_1 + ... + n_k$. Pre každú zo skupín vypočítame variabilitu hodnôt sledovaného znaku vo vnútri skupiny, ktorú charakterizujeme vzťahom

$$SSE = \sum_{i=1}^{k} \sum_{i=1}^{n_i} \left(x_{ij} - \overline{x_i} \right)^2 . \tag{5.29}$$

Variabilitu *SSE* (Sum of Squares for Error) nazývame **vnútroúrovňová variabilita**. Nie je spôsobená rozdielmi medzi jednotlivými úrovňami, ale iba náhodnými vplyvmi. Variabilitu, ktorá je spôsobená rozdielmi medzi úrovňami faktora označujeme *SST* (Sum of Squares for Treatment), budeme nazývať **medziúrovňová variabilita,** alebo **variabilita vysvetlená faktorom** a vypočítame ju podľa vzťahu

$$SST = \sum_{i=1}^{k} n_i \left(\overline{x_i} - \overline{x} \right). \tag{5.30}$$

Celkovú variabilitu výberových údajov, ktorú budeme označovať SSTotal (Total Sum of Squares), môžeme vyjadriť ako súčet vnútroúrovňovej a medziúrovňovej variability. Platí totiž vzťah

$$SSTotal = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2 = \sum_{i=1}^{k} n_i \left(\overline{x_i} - \overline{x} \right)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x_i} \right)^2 . \tag{5.31}$$

Na určenie testovacej štatistiky budeme potrebovať priemery štvorcov jednotlivých odchýlok, ktoré dostaneme vydelením súčtu štvorcov odchýlok príslušnými stupňami voľnosti. Počet stupňov voľnosti pre jednotlivé priemery dostaneme, ak od počtu sčítancov odpočítame počet použitých štatistík vypočítaných z výberových súborov. Pre výpočet priemerov štvorcov vnútroúrovňových a medziúrovňových odchýlok platia nasledujúce vzťahy

$$MSE = \frac{SSE}{n-k} \qquad\qquad MST = \frac{SST}{k-1} \tag{5.32}$$

Testovacia štatistika *F*, ktorú definujeme vzťahom

$$F = \frac{MST}{MSE}, \tag{5.33}$$

má Fisherovo F-rozdelenie s počtom stupňov voľnosti k-1 a n-k. Pre zvolenú hladinu významnosti α zamietame nulovú hypotézu, ak platí

$$F \geq F_{1-\alpha;k-1,n-k},$$

pričom $F_{1-\alpha;k-1,n-k}$ je príslušný kvantil F rozdelenia. V nasledujúcej tabuľke je urobený prehľad údajov potrebných k jednofaktorovej analýze rozptylu.

Tab. 5.9 ANOVA

Zdroj variability	**Súčet štvorcov odchýlok**	**Počet stupňov voľnosti**	**Priemer štvorcov odchýlok**	**Testovacia štatistika**
Faktor	*SST*	*k-1*	$MST = SST/\ k-1$	$F = MST/MSE$
Náhoda	*SSE*	*n-k*	$MSE = SSE/\ n-k$	
Spolu	*SSTotal*	*n-1*		

EXCEL

V ponuke EXCELu nájdeme niekoľko procedúr na analýzu rozptylu. Sú to:
ANOVA: jeden faktor
ANOVA: dva faktory bez opakovania
ANOVA: dva faktory s opakovaním
Zameriame sa na prvú procedúru, ktorá súvisí s jednofaktorovou analýzou.

Vstupné výberové údaje vložíme do stĺpcov vedľa seba, pričom jednotlivé stĺpce obsahujú hodnoty analyzovaného znaku pre konkrétnu úroveň daného faktora. Po vyplnení dialógového okna, v ktorom špecifikujeme naše požiadavky na vstupnú oblasť, hladinu významnosti a výstupnú oblasť, dostaneme na výstupe dve tabuľky, z ktorých jedna obsahuje vybrané popisné štatistiky a druhá je klasickou tabuľkou analýzy rozptylu. Okrem údajov, ktoré sa nachádzajú v Tab.5.9 obsahuje

i p-hodnotu pre daný test, na základe ktorej, podobne ako v predchádzajúcich testoch, vieme rozhodnúť na akej hladine významnosti môžeme prijať, resp. už musíme zamietnuť nulovú hypotézu. Interpretáciu jednotlivých hodnôt v spomínaných výstupných tabuľkách procedúry ANOVA si ukážeme na nasledujúcom príklade.

Príklad 5.9

Môžeme na hladine významnosti $\alpha = 0{,}05$ tvrdiť, že ceny áut rovnakého typu a rovnakého veku sú v troch sledovaných autobazároch porovnateľné?

Ako výberový súbor použijeme súbor AUTOBAZÁR, v ktorom sa nachádzajú ceny ojazdených áut ŠKODA FELICIA v troch sledovaných autobazároch. Z každého autobazáru náhodne vyberieme 8 áut (rovnaký rozsah všetkých výberových súborov zabezpečí vyváženosť modelu), pri ktorých je uvedený vek 4 roky. Ceny týchto áut v jednotlivých predajniach sú uvedené v nasledujúcej tabuľke.

Tab. 5.10 Ceny 4-ročných áut náhodne vybraných z troch autobazárov

cena 1	**cena 2**	**cena 3**
148	137,7	129
149	144	135
158	145	145
168	157	154,9
179,9	159	155
180	159	159
189	166	159
190	179	166

Predpokladajme, že všetky tri súbory pochádzajú zo súborov s normálnym rozdelením (s testami na overenie tohto predpokladu sa zoznámite v nasledujúcej kapitole 6).

Nulová hypotéza bude mať tvar

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

a alternatívna hypotéza bude

H_1 : Aspoň dva priemery sa nerovnajú.

Na riešenie našej úlohy vyberieme z ponuky *Analýza údajov* procedúru ***ANOVA: jeden faktor***. Po vyplnení dialógového okna, v ktorom špecifikujeme vstupnú oblasť (je potrebné, aby hodnoty sledovaných znakov boli v troch susedných stĺpcoch), hladinu významnosti $\alpha = 0{,}05$ a výstupnú oblasť, dostaneme na výstupe nasledujúcu tabuľku.

Tab. 5.9 ANOVA: jeden faktor

Faktor				
Výběr	*Počet*	*Součet*	*Průměr*	*Rozptyl*
cena1	8	1361,9	170,24	290,22
cena2	8	1246,7	155,84	178,24
cena3	8	1202,9	150,36	165,28

ANOVA						
Zdroj variability	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
Mezi výběry	1686,3	2	843,13	3,991	0,034	3,467
Všechny výběry	4436,2	21	211,25			
Celkem	6122,5	23				

Z prvej tabuľky (Faktor) získame informácie o rozsahoch jednotlivých súborov (Počet), celkový súčet cien automobilov (Súčet), bodové odhady priemernej ceny áut (Priemer) a rozptyly cien v jednotlivých autobazároch. Nulovú hypotézu o rovnosti priemerov zamietame, ak platí

$$F \geq F_{1-\alpha;k-1,n-k} = F\,krit\,.$$

Z druhej tabuľky (ANOVA) sa dozvedáme, že na zvolenej hladine významnosti $\alpha = 0{,}05$ musíme zamietnuť nulovú hypotézu, pretože $F > F\,krit$. a prijímame alternatívnu hypotézu, t.j. ceny automobilov nie sú v sledovaných autobazároch porovnateľné. Potvrdzuje to i p-hodnota (0,034), ktorá je menšia ako zvolená

hladina významnosti. Môžeme však tvrdiť, že v prípade hladiny významnosti menšej ako p-hodnota, napr. $\alpha = 0{,}01$, nie je dôvod zamietnuť nulovú hypotézu.

Príklady na precvičenie

5.10 Môžeme na hladine významnosti $\alpha = 0{,}1$ tvrdiť, že počet bodov, ktorý na prijímacích skúškach z fyziky v priemere získajú záujemcovia o štúdium na vojenskej akadémii je 12?

a) Ako výberový súbor použite súbor PRIJÍMACIE SKÚŠKY/1.fakulta/fyzika, pričom predpokladáme, že je vybraný zo súboru s normálnym rozdelením.

b) Na testovanie hypotézy použite interval spoľahlivosti pre priemer vypočítaný v Príklade 4.6.

5.11 Môžeme na základe intervalu spoľahlivosti pre rozptyl z Príkladu 4.6 tvrdiť, že rozptyl priemerného počtu bodov z fyziky, ktorý na prijímacích skúškach získajú záujemcovia o štúdium na vojenskej akadémii je 45?

5.12 Môžeme na hladine významnosti $\alpha = 0{,}1$ tvrdiť, že podiel absolventov stredných odborných škôl medzi študentmi vojenskej akadémie je 60%. Ako výberový súbor použijeme všetkých študentov (oboch fakúlt), ktorí sú v danom roku študentmi vojenskej akadémie (použi súbory PRIJÍMACIE SKÚŠKY/1.fakulta/stred. škola a PRIJÍMACIE SKÚŠKY/2.fakulta/stred. škola).

5.13 Je vstupná úroveň študentov z fyziky porovnateľná na oboch fakultách vojenskej akadémie? Použite výberové súbory PRIJÍMACIE SKÚŠKY/1.fakulta/fyzika a PRIJÍMACIE SKÚŠKY/ 2.fakulta/fyzika, o ktorých predpokladáme, že sú vybrané zo súborov s normálnym rozdelením.

Určite p-hodnotu vhodného jednostranného testu a hladiny významnosti na ktorých prijímame, resp. zamietame nulovú hypotézu.

5.14 Môžeme na hladine významnosti $\alpha = 0{,}01$ tvrdiť, že podiel absolventov stredných odborných škôl študujúcich na jednotlivých fakultách vojenskej akadémie je porovnateľný? Ako výberové súbory použijeme súbory PRIJÍMACIE SKÚŠKY/1.fakulta/stred.škola a PRIJÍMACIE SKÚŠKY/2.fakulta/stred.škola.

5.15 Môžeme na hladine významnosti $\alpha = 0{,}1$ tvrdiť, že priemerné hodnotenie študentov na skúškach v 3. semestri sa výrazne nelíši od hodnotenia v 2. semestri? Ako výberový súbor použijeme hodnotenie študentov 2. fakulty v 2. a v 3. semestri, ktoré sa nachádza v súbore PRIJÍMACIE SKÚŠKY/2.fakulta/priemer.

5.16 Môžeme na hladine významnosti $\alpha = 0{,}05$ tvrdiť, že ceny 3-ročných áut rovnakého typu sú v autobazároch v priemere porovnateľné? Výberové súbory vytvorte zo súboru AUTOBAZÁR, v ktorom sa nachádzajú ceny 3-ročných áut ŠKODA FELICIA v troch sledovaných autobazároch.

6. VYŠETROVANIE NORMALITY ROZDELENIA

6.1 χ^2- test dobrej zhody

V doterajších kapitolách o bodových a intervalových odhadoch parametrov základného súboru, aj o testovaní parametrov základného súboru sme predpokladali, že rozdelenie pravdepodobnosti v základnom súbore bolo známe, predpokladali sme väčšinou normálne rozdelenie. Teraz sa budeme venovať postupom, ktoré umožnia preveriť predpoklady o **type rozdelenia** skúmanej náhodnej premennej *X*. Voľba použitého testu je ovplyvnená informáciami o základnom súbore, ako i rozsahom výberového súboru.

Prvú, orientačnú predstavu o rozdelení si urobíme podľa histogramu, ktorý získame z dostatočne veľkého výberového súboru. Nulová hypotéza potom môže vyzerať:

H_0 : Náhodná premenná má normálne rozdelenie.

Alebo pre iný typ rozdelenia :

H_0 : Náhodná premenná má rovnomerné (resp. Poissonovo, binomické) rozdelenie.

Testom, ktoré umožňujú testovať hypotézu o type rozdelenia (nemusí to byť len normálne rozdelenie), hovoríme **testy zhody.** Najpoužívanejší je **Pearsonov** χ^2 **- test dobrej zhody** (čítaj chí-kvadrát test)**.** Dá sa použiť pre spojité i diskrétne rozdelenia. Myšlienka testu je nasledujúca. Náhodný výber s rozsahom n sa roztriedi do r tried (skupín) a posúdi sa miera zhody napozorovaných výsledkov s výsledkami, ktoré očakávame, ak by daná nulová hypotéza platila. Porovnávame tak **empirické početnosti** s **teoretickými** početnosťami. Testovacou štatistikou je náhodná premenná

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} \tag{6.1}$$

n - je rozsah výberového súboru,

n_i - je empirická početnosť i - tej triedy I_i (intervalu),

r - je počet tried I_i (intervalov), $i = 1, 2, ..., r$,

p_i - je pravdepodobnosť, že X nadobudne hodnotu z i-tej triedy I_i (intervalu), za predpokladu, že platí nulová hypotéza,

np_i - je teoretická (očakávaná) početnosť v i-tej triede I_i (intervale).

Náhodná premenná (6.1) má pre $n \to \infty$ asymptoticky χ^2 - rozdelenie s (r-s-1) stupňami voľnosti, pričom s je počet parametrov, ktoré je potrebné odhadnúť pomocou výberových údajov, aby sme vedeli vypočítať p_i. V našom prípade je $s = 2$, lebo μ a σ^2 musíme odhadnúť.

Z tvaru štatistiky je zrejmé, že jej veľké hodnoty signalizujú nezhodu medzi nameranými a teoretickými početnosťami. Preto nulovú hypotézu na zvolenej hladine významnosti α zamietame, ak je hodnota testovacej štatistiky väčšia ako $100(1-\alpha)$-percentný kvantil χ^2 -rozdelenia s (r-s-1) stupňami voľnosti. Stručne je tento test zachytený v Tab. 6.1.

Tab. 6.1 Pearsonov χ^2 - test dobrej zhody

Hypotézy	**Použité rozdelenie**	**Testovacia štatistika**	**Oblasť zamietnutia H_0**
H_0 : Premenná má dané rozdelenie. H_1 : Premenná nemá dané rozdelenie.	Chí-kvadrát	$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$	$\chi^2 > \chi^2_{1-\alpha}$ $d.f. = r - s - 1$

V praxi sa test používa pre $n \geq 50$ a triedy (intervaly) treba zvoliť tak, aby pre všetky $i = 1, 2, ..., r$ platilo $np_i \geq 5$. Ak to tak nie je, vhodne zlúčime susedné triedy tak, aby v každej platilo $np_i \geq 5$. Každé zlúčenie znižuje počet tried, čím sa znižuje aj počet stupňov voľnosti χ^2 - rozdelenia.

Príklad 6.1

Hracou kockou bolo urobených 1800 hodov, výsledky sú zaznamenané v prvých dvoch stĺpcoch tabuľky 6.2. Zistite, či kocka je homogénna, t.j. či všetky čísla padajú s rovnakou pravdepodobnosťou..

Počet bodov na kocke je diskrétna náhodná premenná *X*, ktorá nadobúda hodnoty $i = 1, 2, ...,6$. Ak platí nulová hypotéza, t.j. náhodná premenná *X* má diskrétne rovnomerné rozdelenie, tak sú všetky hodnoty nadobúdané s rovnakou pravdepodobnosťou $p_1 = p_2 = ... = p_6 = 1/6$. Teoretické početnosti, vypočítané pomocou týchto pravdepodobností sú v treťom stĺpci Tab. 6.2.

Zvolíme hladinu významnosti $\alpha = 0{,}05$. Hodnota testovacej štatistiky je $\chi^2 = 0{,}84665$, je to súčet štvrtého stĺpca tabuľky. Táto malá hodnota naznačuje, že nulová hypotéza bude platná. Presnejšie, porovnáme ju s 95 % - kvantilom $\chi^2_{0{,}95;5} = 11{,}07$. Stupne voľnosti sme určili $d.f. = 5$ (počet skupín je 6, počet odhadovaných parametrov je 0). Pretože $0{,}84655 < 11{,}07$, **nemáme dôvod na zamietnutie nulovej hypotézy**.

Tab. 6.2 Empirické a teoretické početnosti k Príkladu 6.1

Počet bodov na kocke i	**Empirické početnosti** n_i	**Teoretické početnosti** np_i	$(n_i - np_i)^2 / np_i$
1	299	300	0,00333
2	295	300	0,08333
3	303	300	0,03000
4	307	300	0,16333
5	289	300	0,40333
6	307	300	0,16333
spolu	1800	1800	0,84665

Pomocou toho istého testu u spojitej náhodnej premennej *X* chceme overiť, či má normálne rozdelenie. Preto určíme nulovú a alternatívnu hypotézu takto:

H_0: Náhodná premenná *X* má v základnom súbore normálne rozdelenie.

H_1: Náhodná premenná *X* nemá normálne rozdelenie.

Empirické údaje $x_1, x_2, ..., x_n$ roztriedime do r disjunktných intervalov tvaru $\langle a_{i-1}, a_i \rangle$ a $p_i = P(a_{i-1} < X \le a_i) = F(a_i) - F(a_{i-1})$, kde $F(x)$ je distribučná funkcia normálneho rozdelenia, ktoré predpokladá nulová hypotéza, pričom jeho parametre odhadneme. Tým získame všetky údaje na výpočet hodnoty testovacej štatistiky (6.1). Ďalší postup je jasný.

EXCEL

V ponuke EXCELu nenájdeme priamo χ^2-test dobrej zhody, ale pri realizácii tohto testu je vhodné využiť ponuku rôznych štatistických funkcií a procedúr ponúkaných EXCELom. Môžme sa o tom presvedčiť pri riešení nasledujúceho príkladu.

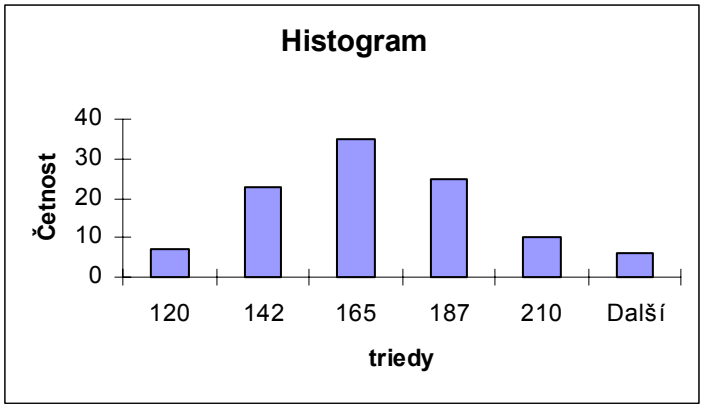
Príklad 6.2

Na súbore AUTOBAZÁR pre všetky 3 predajne spolu zistite, či cena áut má v základnom súbore normálne rozdelenie. Zvoľte hladinu významnosti $\alpha = 0{,}01$ a $\alpha = 0{,}05$.

Na konštrukciu Tab. 6.3 sme použili EXCEL. Z popisnej štatistiky sme získali bodový odhad strednej hodnoty $\mu = 159{,}988$ tis. Sk a štandardnej odchýlky ceny auta $\sigma = 33{,}027$ tis. Sk. Zostrojili sme histogram bez zadania hraníc tried, potom sme šírku a hranice tried upravili tak, aby v každej teoretickej triede bolo aspoň 5 hodnôt. Získané údaje sú uvedené v Tab. 6.3. Príslušný histogram je na Obr. 6.1.

Tab. 6.3 Empirické a teoretické početnosti k Príkladu 6.2

triedy		**početnosti**	$F(a_i) - F(a_{i-1}) = p_i$	$n.p_i$	$(n_i - np_i)^2 / np_i$
$-\infty$	120	7	0,113	11,977	2,068
120	142	23	0,180	19,081	0,805
142	165	35	0,267	28,335	1,568
165	187	25	0,233	24,695	0,004
187	210	10	0,142	15,024	1,680
210	$+\infty$	6	0,065	6,887	0,114
	Hodnota testovacej štatistiky:				6,239



Obr.6.1 Histogram ceny áut súboru AUTOBAZÁR

Z tvaru histogramu môžeme predpokladať, že cena áut predávaných v autobazároch má normálne rozdelenie, čo potvrdíme χ^2 -testom.

Pravdepodobnosti p_i sme našli takto:

$$p_1 = P(-\infty < X \le 120) = F(120) - F(-\infty) = 0{,}113 - 0 = 0{,}113$$

$$p_2 = P(120 < X \le 142) = F(142) - F(120) = 0{,}239 - 0{,}113 = 0{,}180$$

.....

$$p_6 = P(210 < X < +\infty) = 1 - F(210) = 1 - 0{,}935 = 0{,}065$$

Na výpočet hodnôt distribučnej funkcie sme použili EXCEL a funkciu NORMDIST. Súčet posledného stĺpca tabuľky je podľa (6.1) hodnota testovacej štatistiky $\chi^2 = 6{,}239$. Pomocou funkcie CHIINV nájdeme kvantily $\chi^2_{0{,}95} = 7{,}815$ a $\chi^2_{0{,}99} = 11{,}345$ pre 3 stupne voľnosti ($d.f. = 6-2-1 = 3$). Pretože hodnota testovacej štatistiky $\chi^2 = 6{,}239$ je menšia ako oba príslušné kvantily, na oboch hladinách významnosti prijímame hypotézu, že náhodný výber je z normálneho rozdelenia s parametrami (zaokrúhlene) $\mu = 160000$ Sk a $\sigma = 33000$ Sk .

6.2 Shapiroov – Wilkov test normality

Ak náhodný výber je malého rozsahu $n \in \langle 7{,}30 \rangle$, nedá sa použiť Pearsonov test dobrej zhody, preto použijeme iný, napríklad Shapiroov – Wilkov test normality. Nulovú a alternatívnu hypotézu môžeme sformulovať takto :

H_0: Náhodný výber pochádza zo súboru s normálnym rozdelením.

H_1: Náhodný výber nepochádza zo súboru s normálnym rozdelením.

Pri aplikácii tohto testu postupujeme tak, že hodnoty znaku z náhodného výberu $x_1, x_2, ..., x_n$ usporiadame do neklesajúcej postupnosti $x_1^* \le x_2^* \le x_3^* \le ... \le x_n^*$. Hodnotu x_i^* nazývame *i*-ta poriadková štatistika.

Testovacou štatistikou v tomto teste je náhodná premenná

$$W = \frac{\left[\sum_{i=1}^{m} a_i^{(n)} \left(x_{n-i+1}^* - x_i^*\right)\right]^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \tag{6.2}$$

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- $m = \begin{cases} \frac{n}{2}, & ak\ n\ je\ párne \\ \frac{n-1}{2}, & ak\ n\ je\ nepárne \end{cases}$

- $a_i^{(n)}$ - sú koeficienty z tabuľky Shapiroovho - Wilkovho testu, ktoré sú uvedené v prílohe ako Tab.1.

Vypočítanú hodnotu testovacej štatistiky w porovnáme s tabuľkovou hodnotou kvantilu W_α (nachádza sa v prílohe ako Tab.2) pre dané n a hladinu významnosti $\alpha = 0{,}01$ alebo $\alpha = 0{,}05$. Hypotézu, že základný súbor má normálne rozdelené hodnoty sledovaného znaku nezamietame, ak platí $w > W_\alpha$. V prípade, že $w \le W_\alpha$, nemôžeme rozdelenie znaku v základnom súbore považovať za normálne.

6.3 D´ Agostinov test normality

Tento test je určený pre výberový súbor s rozsahom $n \in \langle 31, 100 \rangle$. Nulová a alternatívna hypotéza sú rovnaké ako v predchádzajúcom teste:

H_0: Náhodný výber pochádza zo súboru s normálnym rozdelením.

H_1: Náhodný výber nepochádza zo súboru s normálnym rozdelením.

Pri aplikácii tohto testu postupujeme tak, že hodnoty znaku z náhodného výberu $x_1, x_2, ..., x_n$ usporiadame do neklesajúcej postupnosti $x_1^* \le x_2^* \le x_3^* \le ... \le x_n^*$.

Testovacou štatistikou v tomto teste je náhodná premenná

$$Y = \frac{\sqrt{n}\left(D - 0{,}2820948\right)}{0{,}0299860}\ , \tag{6.3}$$

ktorá je definovaná pomocou náhodnej premennej D

$$D = \frac{\sum_{i=1}^{m} b_i^{(n)}\left(x_{n-i+1}^* - x_i^*\right)}{\sqrt{n^3 \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

- $m = \begin{cases} \frac{n}{2}, & ak\ n\ je\ párne \\ \frac{n-1}{2}, & ak\ n\ je\ nepárne \end{cases}$
- $b_i^{(n)} = \frac{n+1}{2} - i$.

Ak pre hodnotu náhodnej premennej *y* platí $Y_{\frac{\alpha}{2}} < y < Y_{1-\frac{\alpha}{2}}$, tak nezamietame nulovú hypotézu, že náhodný výber pochádza zo základného súboru s normálnym rozdelením. Kvantily $Y_{\frac{\alpha}{2}}$ a $Y_{1-\frac{\alpha}{2}}$ sú tabuľkové hodnoty pre dané n a hladinu významnosti $\alpha = 0{,}01$ alebo $\alpha = 0{,}05$. Nachádzajú sa v prílohe ako Tab.3.

EXCEL

Podobne ako v predchádzajúcom prípade, spomínaný test nenájdeme v ponuke EXCELu. Opäť ale odporúčame jeho použitie pri testovaní rozdelenia znaku.

Príklad 6.3

Na teste zo základov špeciálnej pedagogiky dosiahli 20 poslucháči bodové hodnotenie uvedené v druhom stĺpci tabuľky (maximum bolo 30 bodov). Pomocou Shapiroovho -Wilkovho testu zistite, či tento výberový súbor pochádza zo základného súboru s normálnym rozdelením. Zvoľte hladinu významnosti $\alpha = 0{,}05$.

Príklad vyriešime s použitím EXCELu, ktorý použijeme len na pohodlné vypočítanie testovacej štatistiky (pozri Tab.6.4).V treťom stĺpci sú body zoradené zostupne, vo štvrtom vzostupne, z nich je však ponechaných len $m = 20/2 = 10$ údajov, v šiestom sú hodnoty koeficientov $a_i^{(20)}$ z tabuľky Shapiroovho - Wilkovho testu pre $n = 20$.

Priemerný počet získaných bodov je 13,2. Hodnota testovacej štatistiky je $w = 643{,}324 / 673{,}2 = 0{,}956$. Porovnáme ju s kvantilom $W(\alpha)$ z tabuliek pre danú hladinu významnosti, ktorý je $W(0{,}05) = 0{,}905$. Pretože $W(0{,}05) < w$, nulovú hypotézu nezamietame a tvrdíme, že body z kontrolného testu majú v základnom súbore normálne rozdelenie na danej hladine významnosti.

Tab. 6.4 Riešenie Príkladu 6.3

i	x_i	x_{20-i+1}	$x_{20-i+1} - x_i$	x_i	$a_i(20)$	$a_i(20)\,(x_{20-i+1} - x_i)$	$(x_i - \bar{x})^2$
1	9	29	3	26	0,4734	12,3084	17,64
2	14	19	6	13	0,3211	4,1743	0,64
3	10	19	6	13	0,2565	3,3345	10,24
4	19	19	8	11	0,2085	2,2935	33,64
5	19	17	9	8	0,1686	1,3488	33,64
6	8	16	10	6	0,1334	0,8004	27,04
7	6	16	10	6	0,1013	0,6078	51,84
8	16	15	10	5	0,0711	0,3555	7,84
9	29	14	11	3	0,0422	0,1266	249,64
10	16	14	13	1	0,014	0,014	7,84
11	17				**súčet stĺpca**	**25,3638**	14,44
12	10						10,24
13	3						104,04
14	19			**2. mocnina súčtu stĺpca**		**643,32235**	33,64
15	6						51,84
16	15						3,24
17	14						0,64
18	11						4,84
19	10						10,24
20	13						0,04
					súčet stĺpca		**673,2**

Príklad 6.4

V Príklade 3.10 sme skúmali počet úmrtí po úraze konským kopytom na výberovom súbore s rozsahom $n = 200$. Pomocou χ^2-testu overte, či táto náhodná premenná má Poissonovo rozdelenie.

Čiastkové výsledky sú uvedené v Tab. 6.5. V štvrtom stĺpci sú uvedené očakávané teoretické početnosti v jednotlivých triedach, za predpokladu, že náhodná premenná má Poissonovo rozdelenie. V posledných dvoch triedach sú tieto početnosti malé, preto posledné dve triedy pripojíme k tretej triede. Dostaneme tak tri nové triedy, empirické a teoretické početnosti v nich sú v piatom a šiestom stĺpci tabuľky. Tieto hodnoty použijeme na výpočet hodnoty testovacej štatistiky (posledný stĺpec).

Tab. 6.5 Riešenie Príkladu 6.4

x_i	n_i	p_i	np_i	n_i	np_i	$(n_i - np_i)^2 / np_i$
0	109	0,543	108,6	109	108,6	0,001473
1	65	0,331	66,2	65	66,2	0,021752
2	22	0,101	20,2	26	25	0,04
3	3	0,021	4,2			
4	1	0,003	0,6	hodnota testovacej štatistiky		0,063226

Náhodná premenná (6.1) má χ^2 - rozdelenie s jedným stupňom voľnosti, ktorý je určený číslom $r - s - 1$, v našom prípade ostali 3 triedy a pri výpočte pravdepodobnosti p_i sme použili jeden odhadnutý parameter $\lambda = 0{,}61$ (Príklad 3.10).

Pomocou funkcie CHIINV nájdeme 99%-ný a 95%-ný kvantil χ^2 - rozdelenia CHIINV(0,01;1) = 6,634891, CHIINV(0,05;1) = 3,841455. Hodnota testovacej štatistiky je menšia ako tieto čísla, preto prijímame na hladinách významnosti 0,01 aj 0,05 nulovú hypotézu, že počet mŕtvych zabitých konským kopytom má Piossonovo rozdelenie.

Príklady na precvičenie

6.5 Pomocou D´ Agostinovho testu normality overte, či cena áut v 1. predajni v súbore AUTOBAZÁR má normálne rozdelenie. Zvoľte $\alpha = 0{,}05$ aj $\alpha = 0{,}01$.

6.6 Pomocou D´ Agostinovho testu normality overte, či cena áut v 3. predajni v súbore AUTOBAZÁR má normálne rozdelenie.

6.7 Vhodným testom overte, že body z matematiky na 1. fakulte, aj na 2. fakulte nachádzajúce sa v súbore PRIJÍMACIE SKÚŠKY pochádzajú zo súborov s normálnym rozdelením. Zvoľte $\alpha = 0{,}05$.

6.8 Pearsonovým χ^2 - testom na hladine významnosti $\alpha = 0{,}05$ zistite, či získané body z matematiky na prijímacích skúškach, ktoré dostanete spojením súborov 1.fakulta a 2.fakulta do jedného výberového súboru pochádzajú zo súboru s normálnym rozdelením.

7. NEPARAMETRICKÉ TESTY

V kapitole 5 sme na testovanie parametrov základných súborov používali testovacie štatistiky, ktoré vo väčšine prípadov vyžadovali splnenie istého predpokladu o rozdelení znaku v základnom súbore. V mnohých situáciách tento predpoklad nie je splnený, vtedy je vhodné použiť tzv. neparametrické testy, ktoré tvoria veľmi širokú a rôznorodú oblasť štatistických testov. Pozornosť však budeme venovať iba niektorým z nich. Testy, ktorými sa budeme v tejto kapitole zaoberať, nevyžadujú normálne rozdelenie sledovaného znaku v základnom súbore, čo je ich hlavná prednosť. Táto prednosť však na druhej strane má za následok stratu na sile testu. Jednu skupinu týchto testov tvoria testy dobrej zhody, s ktorými sme sa zoznámili v predchádzajúcej kapitole. V tejto kapitole sa budeme zaoberať metódami na porovnávanie priemerov základných súborov.

7.1 Testy o zhode priemerov dvoch základných súborov

7.1.1 Mannov–Whitneyho U–test

V prípade, že nie je splnený predpoklad o normalite rozdelenia znaku v základných súboroch je Mann–Whitneyov U–test, nazývaný i Wilcoxonov test sumy poradí (Wilcoxon rank sum test), neparametrickou alternatívou k t-testu pre nezávislé výbery na porovnávanie priemerov dvoch základných súborov. Vyžadujeme iba splnenie predpokladu, aby náhodné výberové súbory boli nezávislé. Nulovú hypotézu formulujeme v tvare

$$H_0 : \mu_1 = \mu_2 .$$

Porovnávaním stredných hodnôt (priemerov) základných súborov overujeme, či oba výbery pochádzajú z toho istého rozdelenia pravdepodobnosti.

K výpočtu testovacej štatistiky je potrebné, aby všetky prvky oboch výberových súborov boli označené poradovými číslami od najmenšieho po najväčšie. Pri rovnosti hodnôt znaku priradíme priemerné poradové číslo. Označenie výberov volíme tak, aby platilo $n_1 \geq n_2$. Súčet poradí hodnôt znaku v prvom súbore ozna-

číme R_1 a v druhom R_2 (na kontrolu správnosti výpočtu poradia môžeme použiť vzťah $R_1 + R_2 = (n_1 + n_2)(n_1 + n_2 + 1)/2$). Označme

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \qquad \text{a} \qquad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 . \tag{7.1}$$

Testovacia štatistika pri tomto označení bude mať tvar

$$U = \min(U_1, U_2) . \tag{7.2}$$

Hodnoty kvantilov rozdelenia testovacej štatistiky U, ktoré predstavujú kritické hodnoty pre prijatie, resp. zamietnutie H_0, sú tabelované a uvedené v tabuľkovej prílohe. V prípade, že testovacia štatistika (7.2) je menšia ako kritická hodnota pre zvolenú hladinu významnosti α, zamietame nulovú hypotézu a prijímame alternatívnu hypotézu. Na rozdiel od kritérií pre parametrické testy musí byť vypočítaná testovacia štatistika U menšia ako kritická hodnota $U_{\alpha/2}$ v prípade obojstrannej alternatívnej hypotézy a v prípade jednostrannej alternatívnej hypotézy menši ako U_α. Súvisí to s tým, že na porovnanie používame menšiu hodnotu z vypočítaných hodnôt U_1, U_2.

Ak rozsahy výberových súborov sú väčšie ako 10, je vhodné použiť aproximáciu normálnym rozdelením a testovaciu štatistiku (7.2) nahradiť nasledujúcou štatistikou

$$Z = \frac{U_1 - \mu(U_1)}{\sigma(U_1)} , \tag{7.3}$$

pričom $\mu(U_1) = \frac{n_1 n_2}{2}$ je priemer U_1 a $\sigma(U_1) = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$ je štandardná odchýlka U_1. V tomto prípade je oblasť zamietnutia nulovej hypotézy vymedzená príslušnými kvantilmi normovaného normálneho rozdelenia.

V nasledujúcej tabuľke Tab. 7.1 nájdeme oblasti zamietnutia nulovej hypotézy pre jednotlivé formulácie alternatívnej hypotézy v prípade ľubovoľných rozsahov výberových súborov.

Tab. 7.1 Mannov–Whitneyov U–test (Wilcoxonov test sumy poradí)

Hypotéza H_0	**Hypotéza H_1**	**Predpoklady**	**Testovacia štatistika**	**Oblasť zamietnutia H_0**
$\mu_1 = \mu_2$	1. $\mu_1 \neq \mu_2$ 2. $\mu_1 > \mu_2$ 3. $\mu_1 < \mu_2$	výberové súbory sú nezávislé, $n_1,\ n_2 \leq 10$ $n_1 \geq n_2$	$U = \min(U_1, U_2)$ $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$ $U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$ R_1, R_2 – súčet poradí	1. $u \leq U_{\alpha/2}$ 2. $u \leq U_{\alpha}$ 3. $u \leq U_{\alpha}$
		výberové súbory sú nezávislé $n_1, n_2 > 10$ $n_1 \geq n_2$	$Z = \frac{U_1 - \mu(U_1)}{\sigma(U_1)}$ $\mu(U_1) = \frac{n_1 n_2}{2}$ $\sigma(U_1) = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$	1. $\lvert z \rvert \geq z_{1-\alpha/2}$ 2. $z \geq z_{1-\alpha}$ 3. $z \leq -z_{1-\alpha}$

Príklad. 7.1

Na dvoch výberových súboroch, z ktorých jeden pozostával zo 14 náhodne vybratých mužov a druhý z 11 náhodne vybratých žien sa testoval ich postoj k trestu smrti. V Tab. 7.2 sú uvedené výsledky tohto testu. Môžeme na hladine významnosti $\alpha = 0{,}1$ tvrdiť, že medzi postojom mužov a žien k trestu smrti nie je významný rozdiel?

Tab. 7.2 Postoj mužov a žien k trestu smrti

Muži	67	72	48	30	92	15	5	87	91	54	66	72	98	97
Ženy	20	40	37	42	51	15	68	35	12	31	85			

Riešenie začneme formuláciou nulovej a alternatívnej hypotézy.

H_0 : $\mu_M = \mu_{\check{Z}}$ (Priemerný počet bodov v testoch je porovnateľný v prípade mužov a žien.)

H_1 : $\mu_M \neq \mu_{\check{Z}}$ (Priemerný počet bodov v testoch je v prípade mužov a žien rozdielny.)

Hodnoty v Tab. 7.2 v prípade mužov i žien usporiadame vzostupne a priradíme im poradové čísla.

Tab. 7.3 Výpočtová tabuľka

Muži	5	15	30	34	48	66	67	67	72	72	77	91	92	98
Ženy	15	20	31	35	37	40	48	68	71	85	92			
Muži (poradie)	1	2,5	5	7	11,5	13	14,5	14,5	18,5	18,5	20	22	23,5	25
Ženy (poradie)	2,5	4	6	8	9	10	11,5	16	17	21	23,5			

Súčet poradí pre mužov je $r_M = 196,5$ a pre ženy $r_{\check{Z}} = 128,5$. Vypočítame

$$u_M = 14.11 + \frac{14.15}{2} - 196,5 = 62,5 \qquad \text{a} \qquad u_{\check{Z}} = 14.11 + \frac{11.12}{2} - 128,5 = 91,5 .$$

Testovacia štatistika $u = \min(u_M, u_{\check{Z}}) = 62,5$. Kritická hodnota je v prípade dvojstranného testu pre $\alpha = 0,05$ a dané rozsahy výberových súborov rovná 40 (Príloha Tab. 5). Pretože $u = 62,5 > 40$, nemôžeme zamietnuť hypotézu H_0 a konštatujeme, že postoj mužov a žien k trestu smrti je na základe výsledkov testu porovnateľný.

Vzhľadom na to, že oba výberové súbory majú rozsah väčší ako 10 môžeme na výpočet použiť testovaciu štatistiku (7.3), ktorá má normované normálne rozdelenie. Hodnota testovacej štatistiky vzhľadom na väčší rozsah výberového súboru mužov bude mať tvar

$$z = \frac{u_M - \mu(u_M)}{\sigma(u_M)} = \frac{62,5 - \frac{14.11}{2}}{\sqrt{\frac{14.11.(14+11+1)}{12}}} = -0,79 .$$

Kritická hodnota je $z_{1-\alpha/2} = z_{0,975} = 1{,}96$, čo znamená, že oblasť prijatia H_0 je interval $(-1{,}96;\ 1{,}96)$. Vzhľadom na to, že testovacia štatistika leží v tomto intervale, nemáme dôvod zamietnuť hypotézu, že postoj mužov a žien k trestu smrti je porovnateľný.

7.1.2 Wilcoxonov test

Wilcoxonov test (**Wilcoxon signed-rank test**) je vhodné použiť v prípade, keď nie je splnený predpoklad o normálnom rozdelení rozdielov znaku v základných súboroch a výberové súbory sú párové. Je neparametrickou alternatívou t-testu pre závislé (párové) výberové súbory. Nulová hypotéza je opäť formulovaná v tvare

$$H_0 : \mu_1 = \mu_2 .$$

Náhodným výberom sme získali z dvoch základných súborov dva párové výberové súbory. Pre každý pár vypočítame rozdiel

$$D = X_1 - X_2 .$$

Absolútnym hodnotám rozdielov D priradíme poradové čísla na základe ich veľkosti. Pri rovnosti hodnôt znaku priradíme priemerné poradové číslo. Dvojicu, medzi ktorou je nulový rozdiel, vylúčime z výberových súborov. Testovacia štatistika T_W je definovaná

$$T_W = \min\left(\sum(+), \sum(-)\right), \tag{7.4}$$

kde $\sum(+)$ je súčet poradových čísiel, ktoré prislúchajú ku kladným hodnotám rozdielov D a $\sum(-)$ je súčet poradových čísiel prislúchajúcich k záporným hodnotám rozdielov D. Pre overenie správnosti použijeme vzťah $\sum(+) + \sum(-) = \frac{n(n+1)}{2}$.

Kritické hodnoty pre vymedzenie oblasti zamietnutia nulovej hypotézy určíme ako príslušné kvantily T_W rozdelenia, ktoré sú tabelované (Tabuľková príloha, Tab. 6).

Ak počet párov pozorovaní po vylúčení rovnakých dvojíc je dostatočne veľký (n>25), pravdepodobnostné rozdelenie Wilcoxonovej testovacej štatistiky T_W sa približuje k normálnemu rozdeleniu a v takomto prípade môžeme testovaciu štatis-

tiku (7.4) aproximovať testovacou štatistikou, ktorá má normované normálne rozdelenie. Testovacia štatistika má tvar

$$Z = \frac{T_W - \mu(T_W)}{\sigma_{T_W}}, \tag{7.5}$$

kde $\mu(T_W) = \frac{n(n+1)}{4}$ je priemer T_W a $\sigma(T_W) = \sqrt{\frac{n(n+1)(2n+1)}{24}}$ je štandardná odchýlka T_w.

V nasledujúcej tabuľke nájdeme oblasti zamietnutia nulovej hypotézy pre jednotlivé formulácie alternatívnej hypotézy.

Tab. 7.4 Wilcoxonov test (Wilcoxonov znamienkovo-poradový test)

Hypotéza H_0	**Hypotéza H_1**	**Predpoklady**	**Testovacia štatistika**	**Oblasť zamietnutia H_0**
$\mu_1 = \mu_2$	1. $\mu_1 \neq \mu_2$ 2. $\mu_1 > \mu_2$ 3. $\mu_1 < \mu_2$	výberové súbory sú párové $n \leq 25$	$T_W = \min\left(\sum(+), \sum(-)\right)$ $\sum(+)$ súčet poradí $\sum(-)$ súčet poradí	1. $t_W \leq T_{\alpha/2}$ 2. $t_W \leq T_\alpha$, $t_W = \sum(-)$ 3. $t_W \leq T_\alpha$, $t_W = \sum(+)$
		výberové súbory sú párové $n > 25$	$Z = \frac{T_W - \mu(T_W)}{\sigma(T_W)}$ $\mu(T_W) = \frac{n(n+1)}{4}$ $\sigma(T_W) = \sqrt{\frac{n(n+1)(2n+1)}{24}}$	1. $\lvert z \rvert \geq z_{1-\alpha/2}$ 2. $z \geq z_{1-\alpha}$, $t_W = \sum(+)$ 3. $z \leq -z_{1-\alpha}$, $t_W = \sum(+)$

Ani jeden z uvedených neparametrických testov nie je v ponuke EXCELUu. I napriek tomu je realizácia týchto testov použitím rôznych funkcií ponúkaných EXCELom nenáročná.

Príklad 7.2

V Tab. 7.5 sú výsledky testov logického myslenia náhodne vybratých študentov. V prvom riadku, označenom Test 1, sú výsledky testov, ktoré boli robené v tichom, pokojnom prostredí. V druhom riadku, označenom Test 2, sú výsledky testov tej istej vzorky študentov, ale v tomto prípade test robili v hlučnom prostredí. Môžeme na hladine významnosti $\alpha = 0{,}05$ tvrdiť, že hluk má negatívny vplyv na výsledky testov?

Tab. 7.5 Výsledky testov logického myslenia

Test 1	67	78	81	72	75	92	84	83	77	65	71	79	80
Test 2	68	81	85	60	75	81	73	78	84	56	61	64	63

Riešenie začneme formulovaním nulovej a v tomto prípade jednostrannej alternatívnej hypotézy.

$H_0: \ \mu_1 = \mu_2$ (Priemerný počet bodov získaných v oboch testoch je porovnateľný.)

$H_1 : \ \mu_1 > \mu_2$ (Priemerný počet bodov v testoch je v pokojnom prostredí vyšší ako v hlučnom prostredí.)

Pre každý pár vypočítame rozdiel $D = X_1 - X_2$ a absolútnym hodnotám rozdielov priradíme poradové čísla na základe ich veľkosti. Pri rovnosti hodnôt znaku priradíme priemerné poradové číslo. Pozri nasledujúcu tabuľku.

Tab. 7.6 Výpočtová tabuľka

Študent	Test 1	Test 2	D	|D|	Poradie (+)	Poradie (-)
1	67	68	-1	1		1
2	78	81	-3	3		2
3	81	85	-4	4		3
4	72	60	12	12	10	
5	75	75	0	0	-	
6	92	81	11	11	8,5	
7	84	73	11	11	8,5	
8	83	78	5	5	4	
9	77	84	-7	7		5
10	65	56	9	9	6	
11	71	61	10	10	7	
12	79	64	15	15	11	
13	80	63	17	17	12	
súčet					67	11

Určíme súčet poradových čísiel v stĺpcoch Poradie (+) a Poradie (-).
Pre kontrolu výpočtov môžeme použiť vzťah

$$\sum(+) + \sum(-) = 78 = \frac{n(n+1)}{2} = \frac{12.13}{2} = 78 .$$

Vzhľadom na to, že sme volili jednostrannú alternatívnu hypotézu $\mu_1 > \mu_2$, testovacou štatistikou bude pre nás súčet poradí záporných hodnôt , takže $t_W = 11$. Testovaciu štatistiku porovnáme s kritickou hodnotou $T_{0,05} = 17$. Vzhľadom na to, že $t_W = 11 < 17 = T_{0,05}$, zamietame H_0 a na hladine $\alpha = 0{,}05$ prijímame hypotézu, že hlučné prostredie má negatívny vplyv na výsledky testu.

7.2 Testy o zhode priemerov *k* základných súborov ($k \geq 3$)

7.2.1 Kruskalov - Wallisov test

Kruskalov - Wallisov test je neparametrickou alternatívou k jednofaktorovej ANOVE. Je zovšeobecnením Mannovho - Whitneyho U-testu pre prípad k výberov, pričom $k \geq 3$. Použijeme ho v prípade, keď nie sú splnené predpoklady o tom, že výberové súbory pochádzajú zo súborov s normálnym rozdelením.

Majme k dispozícii k $(k \geq 3)$ nezávislých výberových súborov s rozsahmi $n_1, ..., n_k$. Zaujíma nás, či všetky tieto výberové súbory pochádzajú z toho istého rozdelenia. Pomocou Kruskalovho - Walisovho testu budeme overovať rovnosť priemerov základných súborov. Neurčíme však ním, ktoré základné súbory sa navzájom líšia. Na tento účel je potrebné použiť iné typy testov. Nulovú a alternatívnu hypotézu budeme formulovať v tvare

H_0: Rozdelenie sledovaného znaku je vo všetkých základných súboroch rovnaké (všetky základné súbory majú rovnaké priemery)

H_1: Rozdelenie sledovaného znaku nie je vo všetkých základných súboroch rovnaké (nie všetky základné súbory majú rovnaké priemery)

Postup pri tomto teste je podobný postupu použitému v U-teste. Všetkým hodnotám v jednotlivých výberových súboroch zoradených podľa veľkosti priradíme poradové číslo a pre každý výberový súbor určíme súčet poradí $R_1, ..., R_k$. Označme $n = n_1 + ... + n_k$. Kruskalovu - Wallisovu testovaciu štatistiku vypočítame podľa vzťahu

$$K = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) . \qquad (7.6)$$

V prípade, že výberové súbory majú rozsah aspoň 5, je vhodné rozdelenie štatistiky K aproximovať χ^2 – rozdelením s k-1 stupňami voľnosti (ak pracujeme so súbormi, ktorých rozsah je menší ako 5, je možné použiť tabelované hodnoty K štatistiky). Kritická hodnota je potom príslušný kvantil χ^2 – rozdelenia. Hypotézu H_0 zamietame, ak $k \geq \chi^2_{1-\alpha;k-1}$.

Príklad 7.3

Sledujeme aktivitu 3 žiaričov s rôznymi rádionuklidmi. Zaujíma nás, či všetky tri žiariče majú v priemere rovnakú aktivitu. Na testovanie hypotézy o zhode priemerov použijeme tri výberové súbory, ktoré obsahujú hodnoty aktivity v kBq namerané na jednotlivých žiaričoch. Namerané hodnoty sa nachádzajú v nasledujúcej tabuľke. Úlohu budeme riešiť na hladine významnosti $\alpha = 0{,}05$.

Tab. 7.7 Nameraná aktivita 3 žiaričov v kBq

meranie	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
žiarič 1	2,2	1,5	3,5	3,2	1,4	2,8	2,4	1,8	
žiarič 2	3,1	3,1	2,5	2,8	3,2	2,8			
žiarič 3	1,6	1,8	2,6	2,8	1,6	3,8	3,5	3,2	2,4

Riešenie úlohy začneme formulovaním nulovej a alternatívnej hypotézy.

H_0: Priemerná aktivita žiarenia je rovnaká pre všetky tri žiariče.

H_1: Priemerná aktivita žiarenia nie je rovnaká pre všetky tri žiariče.

Na riešenie našej úlohy je potrebné zoradiť všetky hodnoty v jednotlivých výberových súboroch podľa veľkosti, priradiť im príslušné poradové číslo a pre každý výberový súbor určiť súčet poradí. Spomínané kroky nájdeme zrealizované v nasledujúcej tabuľke.

Tab. 7.8 Výpočtová tabuľka

meranie	**žiarič 1**	**žiarič 2**	**žiarič 3**	**poradie 1**	**poradie 2**	**poradie 3**
1	1,4	2,5	1,6	1	10	3,5
2	1,5	2,8	1,6	2	13,5	3,5
3	1,8	2,8	1,8	5,5	13,5	5,5
4	2,2	3,1	2,4	7	16,5	8,5
5	2,4	3,1	2,6	8,5	16,5	11
6	2,8	3,2	2,8	13,5	19	13,5
7	3,2		3,2	19		19
8	3,5		3,5	21,5		21,5
9			3,8			23
súčet				78	89	109

Kruskalovu-Wallisovu testovaciu štatistiku vypočítame dosadením vypočítaných súčtov a rozsahov výberových súborov do vzťahu (7.6) . Dostaneme

$$k = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{23.24} \left(\frac{78^2}{8} + \frac{89^2}{6} + \frac{109^2}{9} \right) - 3{,}24 = 1{,}93 .$$

Vzhľadom na to, že všetky tri výberové súbory majú rozsah väčší ako 5, je vhodné rozdelenie štatistiky *K* aproximovať χ^2 – rozdelením. Kritickú hodnotu určíme ako 95% kvantil χ^2 – rozdelenia s 2 stupňami voľnosti. Na jej výpočet môžeme v EXCELi použiť štatistickú funkciu CHIINV(0,05;2) = 5,99. Hypotézu H_0 zamietame, ak $k \geq \chi^2_{1-\alpha;k-1}$. Keďže vypočítaná testovacia štatistika je menšia ako kritická hodnota, nie je dôvod zamietnuť nulovú hypotézu a na hladine významnosti $\alpha = 0{,}05$ môžeme tvrdiť, že všetky tri žiariče majú v priemere rovnakú aktivitu.

7.2.2 Friedmanov test

V prípade, že chceme porovnať niekoľko základných súborov u ktorých nemožno predpokladať normálne rozdelenie, na základe závislých (párových) výberových súborov je vhodné použiť **Friedmanov test**. Podobne ako všetky metódy uvedené v tejto časti i tento test je založený na poradiach. V podstate sa jedná o Wilcoxonov test rozšírený na viac ako dva súbory.

Budeme testovať nasledujúce hypotézy.

H_0: Rozdelenie sledovaného znaku je vo všetkých základných súboroch sú rovnaké (všetky základné súbory majú rovnaké priemery).

H_1: Rozdelenie sledovaného znaku nie je vo všetkých základných súboroch sú rovnaké (nie všetky základné súbory majú rovnaké priemery).

Údaje je vhodné usporiadať do tabuľky, v ktorej stĺpce predstavujú jednotlivé súbory (jednotlivé úrovne faktora) a riadky hodnoty sledovaného znaku pre jednotlivé úrovne faktora pre konkrétnu štatistickú jednotku. Hodnote sledovaného znaku v prípade každej z *n* štatistických jednotiek (každého riadku) priradíme poradie, t.j. číslo od 1 po *k*. Potom vypočítame súčet poradí $R_1,\ldots, R_k$ pre každý súbor (stĺpec). Je prirodzené oča-

kávať, že v prípade platnosti nulovej hypotézy by mali byť hodnoty $R_1,..., R_k$ približne rovnaké. Na testovanie použijeme Friedmanovu testovaciu štatistiku

$$Q = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1). \tag{7.7}$$

Uvedená štatistika má so vzrastajúcim rozsahom (pre $k = 3,\ n \geq 10,\ k = 4,\ n \geq 5$) približne χ^2 – rozdelenie s *k*-1 stupňami voľnosti. Kritické hodnoty určíme ako príslušné kvantily χ^2 – rozdelenia. Hypotézu H_0 zamietame, ak $Q \geq \chi^2_{1-\alpha;k-1}$. Pre menšie rozsahy je možné použiť tabelované hodnoty *Q* štatistiky, ktoré však nie sú obsahom tejto učebnice.

Príklad 7.4

Môžeme prijať hypotézu, že rozdelenie znaku, ktorým sú študijné výsledky študentov v jednotlivých semestroch, je rovnaké vo všetkých základných súboroch? V nasledujúcej tabuľke sú priemerné známky 12 náhodne vybratých študentov v troch po sebe nasledujúcich semestroch. Hypotézu budeme testovať na hladine významnosti $\alpha = 0{,}1$.

Tab. 7.9 Priemerné známky študentov v dvoch semestroch

študent	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
priemer 1	1,2	1,4	1,6	1,8	1,8	1,9	2,0	2,1	2,1	2,5	2,5	2,9
priemer 2	1,8	1,8	2,0	2,0	2,0	2,2	2,2	2,5	2,5	2,5	2,2	2,8
priemer 3	1,4	1,2	2,0	2,1	1,6	1,6	1,8	2,2	2,2	2,3	2,3	2,4

Nulovú a alternatívnu hypotézu budeme formulovať v tvare

H_0: Rozdelenie sledovaného znaku je vo všetkých základných súboroch rovnaké (všetky základné súbory majú rovnaké priemery).

H_1: Rozdelenie sledovaného znaku nie je vo všetkých základných súboroch rovnaké (nie všetky základné súbory majú rovnaké priemery).

V nasledujúcej tabuľke sme údaje usporiadali do stĺpcov, ktoré predstavujú konkrétne výberové súbory, t.j. priemerné hodnotenie každého z 12 študentov v 3 semestroch. Každej hodnote sledovaného znaku v prípade každej z 12 štatistických jednotiek priradíme poradie, t.j. číslo od 1 po 3. V prípade rovnosti znaku priradíme priemerné poradové číslo.

Tab. 7.10 Výpočtová tabuľka

študent	**priemer 1**	**priemer 2**	**priemer 3**	**poradie 1**	**poradie 2**	**poradie 3**
1	1,2	1,8	1,4	1	3	2
2	1,4	1,8	1,2	2	3	1
3	1,6	2,0	2	1	2,5	2,5
4	1,8	2,0	2,1	1	2	3
5	1,8	2,0	1,6	2	3	1
6	1,9	2,2	1,6	2	3	1
7	2,0	2,2	1,8	2	3	1
8	2,1	2,5	2,2	1	3	2
9	2,1	2,5	2,2	1	3	2
10	2,5	2,5	2,3	2,5	2,5	1
11	2,5	2,2	2,3	3	1	2
12	2,9	2,8	2,4	3	2	1
súčet				21,5	31	19,5

Potom vypočítame súčet poradí R_1, R_2 a R_3 pre každý súbor (stĺpec), ktoré sú $R_1 = 21{,}5,\quad R_2 = 31,\quad R_3 = 19{,}5.$ Tieto hodnoty spolu s rozsahom $n = 12$ dosadíme do vzťahu (7.7) a dostaneme

$$q = \frac{12}{nk(k+1)}\sum_{i=1}^{k} R_i^2 - 3n(k+1) = \frac{12}{12.3.4}\left(21{,}5^2 + 31^2 + 19{,}5^2\right) - 3.12.4 = 6{,}29\,.$$

Kritickú hodnotu určíme ako kvantil $\chi^2_{1-\alpha;k-1} = \chi_{0.9;2} = 5{,}99.$ Hypotézu H_0 zamietame, ak vypočítaná testovacia štatistika je väčšia ako kritická hodnota, čo je v našom prípade splnené. Znamená to teda, že prijímame alternatívnu hypotézu, t.j. priemerné hodnotenie študentov v jednotlivých semestroch nie je rovnaké. pričom chyba, ktorej sa dopúšťame je 1%.

Príklady na precvičenie

7.5. Sú výrazné rozdiely vo výsledkoch psychodiagnostického vyšetrenie absolventov gymnázií a absolventov stredných odborných učilísk, ktorí majú záujem o štúdium na vojenskej akadémii? Ako výberové súbory použite absolventov gymnázií a stredných odborných učilísk, ktorí robili prijímacie skúšky na 1. fakultu. Použite súbor PRIJÍMACIE SKÚŠKY/1.fakulta/psych. Hypotézu testujte na hladine významnosti $\alpha = 0{,}1$.

7.6. Môžeme na hladine významnosti $\alpha = 0{,}05$ tvrdiť, že hodnotenia absolventov gymnázií je z matematiky v 3. semestri horšie ako v 2. semestri? Výberové súbory vytvorte z absolventov gymnázií zo súborov PRIJÍMACIE SKÚŠKY/2.fakulta/ mat.2 (upr) a PRIJÍMACIE SKÚŠKY/2.fakulta/ mat.3 (upr).

7.7. Použitím vhodného neparametrického testu riešte Príklad 5.9 z predchádzajúcej kapitoly na hladine významnosti $\alpha = 0{,}01$.

7.8. Testujte na hladine významnosti $\alpha = 0{,}1$ či použitie rôzneho typu konzervačného prípravku má vplyv na trvanlivosť potravinárskeho výrobku.

výrobok	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
prípr. 1	3,5	5,0	4,0	8,0	5,5	12,5	5,5	10,5	13,0	5,5	15,0	3,5	12,5	5,5
prípr. 2	3,3	3,5	3,0	8,0	4,0	12,0	5,5	11,0	13,5	4,0	14,0	3,0	12,0	5,5
prípr. 3	3,5	5,5	3,5	9,5	4,5	13,0	6,0	12,0	13,0	4,5	14,5	3,0	12,5	6,0

8. SKÚMANIE ZÁVISLOSTI DVOCH KVANTITATÍVNYCH ZNAKOV

8.1 Štatistická závislosť

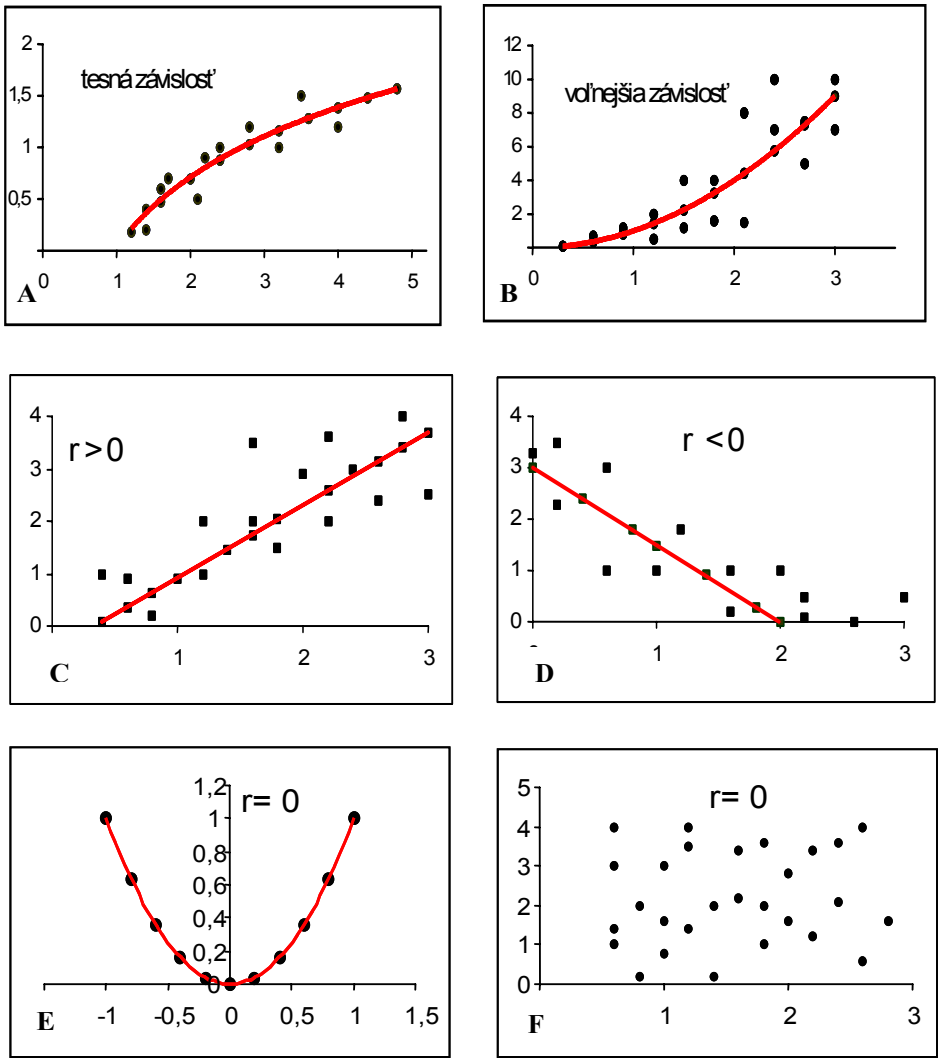
Dôležitá úloha všetkých technických, ekonomických i sociálnych oborov je hľadať a skúmať závislosť medzi premennými. Doteraz sme pracovali s funkčnými vzťahmi, kde závislá premenná y je jednoznačne určená funkciou $y = f(x)$ alebo $y = f(x_1, x_2, ..., x_n)$.

Často však, v dôsledku pôsobenia náhodných faktorov, alebo nezohľadňovania nejakého faktora, či v dôsledku nepresnosti merania má závisle premenná Y a jej pozorované hodnoty $y_1, y_2, ..., y_n$ povahu náhodnej veličiny, ktorá má isté rozdelenie pravdepodobnosti. Takáto závislosť sa volá **stochastická (štatistická) závislosť**. Nezávislé premenné môžu byť nenáhodné (fixné) alebo tiež náhodné veličiny. V tejto časti sa budeme zaoberať **jednoduchou (párovou) regresiou,** kde uvažujeme len jednu nezávislú premennú X s hodnotami $x_1, x_2, ..., x_n$.

Uvažujme závislosť ceny ojazdeného auta v autobazáre od veku auta. Zistíme, že autá s rovnakým vekom majú rôznu cenu. Preto cenu napríklad štvorročného auta považujeme za náhodnú premennú, jej rozdelenie sa volá podmienené rozdelenie. Kedy teda považujeme náhodné veličiny za štatisticky závislé? Rozdelenie početností jednej veličiny Y (kvantitatívneho znaku), ktoré zodpovedá istej, konkrétnej hodnote druhej veličiny X (kvantitatívneho znaku) sa volá **podmienené rozdelenie početností.** Ak pri zmenách hodnôt jedného znaku dochádza ku zmenám podmieneného rozdelenia početností druhého znaku, považujeme znaky za štatisticky **závislé.** A naopak, ak pri zmenách jedného znaku sa nemení rozdelenie druhého znaku, považujeme ich za **nezávislé.** O štatistickej závislosti možno hovoriť aj u kvalitatívnych znakov.

Elementárny spôsob grafického znázornenia závislosti dvoch kvantitatívnych znakov je **bodový diagram**. Zo znázornenia bodov $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$

v rovine, kde (x_i, y_i) sú konkrétne hodnoty premenných X , Y namerané na i-tej štatistickej jednotke, možno zistiť charakteristické rysy závislosti. Obr. 8.1A ukazuje, že s narastajúcimi hodnotami premennej X rastú aj hodnoty premennej Y a navyše, že sa tento rast postupne spomaľuje. Schematicky znázorňuje túto tendenciu krivka preložená medzi bodmi. Voláme ju **regresná krivka**. Na Obr.8.1B s narastajúcim X rastú aj hodnoty Y, ale rast sa postupne zrýchľuje. Závislosti znázornené na obrázkoch majú teda rôzny priebeh.



Obr. 8.1 Rôzne druhy závislostí

Obrázky sa líšia ešte z iného hľadiska. Na Obr.8.1B sú jednotlivé body rozptýlené okolo regresnej krivky oveľa viac ako na Obr.8.1A. Medzi *X* a *Y* na Obr.8.1B je **voľnejšia závislosť** ako na Obr.8.1A. Obe závislosti sa líšia **silou** závislosti.
Pri skúmaní závislosti teda treba riešiť dve úlohy, ktoré spolu úzko súvisia:

- Posúdiť tesnosť závislosti pomocou nejakej charakteristiky, ktorá popisuje do akej miery premenná *X* vysvetľuje variabilitu premennej *Y* (**korelačná analýza)**.
- Charakterizovať priebeh tejto závislosti, to znamená, odhadnúť funkčný vzťah, podľa ktorého sa mení závislá premenná pri zmenách nezávisle premennej (**regresná analýza)**.

Podľa toho, koľko nezávislých premenných berieme do úvahy pri riešení týchto úloh, hovoríme

- o jednoduchej (párovej) korelácii a regresii, ak pracujeme len s jednou nezávislou premennou,
- o viacnásobnej (mnohonásobnej) korelácii a regresii, ak je počet nezávislých premenných väčší ako jeden.

Použitie viacnásobnej regresie síce vedie k presnejším odhadom, ale veľký počet premenných sťažuje analýzu úlohy i interpretáciu výsledkov. Preto v modeli treba uvažovať len tie premenné, ktoré majú zásadný vplyv na závislú premennú.
V celej tejto kapitole ide len o zisťovanie matematických súvislostí, ktoré nemôžeme zamieňať za vzťah príčiny a následku, lebo ani vysoký stupeň štatistickej závislosti nehovorí nič o príčinnej súvislosti javov. Väčšinou túto zdanlivú súvislosť spôsobuje tretí faktor, na ktorom sú oba pôvodné javy závislé. Pri zlej interpretácii môžeme dostať komické tvrdenia. Napríklad zistený vzťah medzi nízkou augustovou spotrebou plynu v kotolniach a vysokým predajom opaľovacích krémov ovplyvňuje tretí faktor - počasie.

8.2 Korelačná analýza

Vzťah medzi *X, Y* môže mať rôznu intenzitu, od úplnej nezávislosti až po úplnú funkčnú závislosť. Stupeň štatistickej závislosti sa dá popísať rôznymi mierami, my sa budeme venovať len **kovariancii a korelačnému koeficientu premenných *X, Y*.** Obe charakteristiky sú miery **lineárnej závislosti** premenných *X, Y*. **Kovariancia medzi *X, Y*** vo výberovom súbore s rozsahom *n* je číslo

$$\operatorname{cov} xy = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right). \tag{8.1}$$

Vzťah sa dá upraviť na jednoduchší tvar

$$\operatorname{cov} xy = \frac{1}{n}\sum_{i=1}^{n}\left(x_i y_i - \bar{x}\, y_i - x_i\, \bar{y} + \bar{x}\, \bar{y}\right) = \frac{1}{n}\sum x_i y_i - \frac{\bar{x}}{n}\sum y_i - \frac{\bar{y}}{n}\sum x_i + \frac{\bar{x}\,\bar{y}}{n}\sum 1 =$$

$$= \overline{xy} - \bar{x}\,\bar{y}. \tag{8.2}$$

Vlastnosti kovariancie:

- $\operatorname{cov} xy$ môže nadobudnúť ľubovoľnú reálnu hodnotu.
- $\operatorname{cov} xy = \operatorname{cov} yx$.
- Ak $\operatorname{cov} xy > 0$, premenné *X, Y* sú **priamo** lineárne závislé (Obr. 8.1C).
- Ak $\operatorname{cov} xy < 0$, premenné *X, Y* sú **nepriamo** lineárne závislé (Obr. 8.1D).
- Ak *X,Y* sú nezávislé, potom $\operatorname{cov} xy = 0$ (Obr. 8.1F).
- Kovariancia je mierou lineárnej závislosti, nehovorí nič o iných typoch závislosti. To, že $\operatorname{cov} xy = 0$ (hovoríme aj, že *X, Y* sú nekorelované) ešte neznamená, že *X, Y* sú nezávislé. Aj v prípade nulovej kovariancie môžu byť znaky **nelineárne** funkčne závislé (Obr. 8.1E).
- Nevýhodou kovariancie je, že jej hodnoty sú závislé na mierke, v ktorej sú vyjadrené *X , Y*. Preto vznikla veličina, ktorá tento nedostatok nemá, a to korelačný koeficient.

Korelačný koeficient je v základnom súbore označovaný $\rho_{x,y}$ a definovaný

$$\rho_{x,y} = \frac{\operatorname{cov} xy}{\sigma_x . \sigma_y}, \tag{8.3}$$

Ak použijeme namiesto základného súboru výberový súbor a kovarianciu výberového súboru a štandardné odchýlky výberového súboru

$$s_x = \sqrt{\frac{1}{n}\sum\left(x_i - \overline{x}\right)^2} \quad \text{a} \quad s_y = \sqrt{\frac{1}{n}\sum\left(y_i - \overline{y}\right)^2},$$

dostaneme bodový odhad (ale skreslený) korelačného koeficientu, ktorý sa volá **výberový korelačný koeficient** $r_{x,y}$:

$$r_{x,y} = \frac{\sum\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum\left(x_i - \overline{x}\right)^2} . \sqrt{\sum\left(y_i - \overline{y}\right)^2}} \tag{8.4}$$

Vlastnosti korelačného koeficientu:

- $\left|r_{xy}\right| \leq 1$.
- $r_{xy} = r_{yx}$, preto sa používa stručné označenie len r (alebo len ρ).
- Ak $r_{xy} > 0$, premenné *X,Y* sú priamo lineárne závislé (Obr. 8.1C).
- Ak $r_{xy} < 0$, premenné *X, Y* sú nepriamo lineárne závislé (Obr. 8.1D).
- Korelačný koeficient je mierou sily **lineárnej závislosti**, nehovorí nič o iných typoch závislosti. V prípade nulového korelačného koeficientu znaky sú **lineárne nezávislé**, môžu byť ale až nelineárne funkčne závislé, čo ilustruje Obr. 8.1E.
- Keď medzi premennými *X*, *Y* je funkčný **lineárny** vzťah $Y = B_0 + B_1 X$ ($B_1 \neq 0$), potom $r_{xy} = 1$ pre $B_1 > 0$, ($r_{xy} = -1$ pre $B_1 < 0$).
- Interpretácia konkrétnej hodnoty korelačného koeficientu závisí od povahy experimentálnych údajov a od rozsahu výberového súboru. Absolútna hodnota

korelačného koeficientu blízka jednotke znamená silnú závislosť, blízka nule slabú závislosť.

- Hodnota korelačného koeficientu je nezávislá na merných jednotkách.

Ak je výberový korelačný koeficient blízky nule, chceme overiť, či je nenulový len v dôsledku náhodného výberu. Uvedieme len jeden z mnohých testov pre testovanie korelačného koeficientu.

T-test **lineárnej nezávislosti premenných *X, Y*** overuje platnosť $H_0 : \rho = 0$ oproti alternatívnej hypotéze $H_1 : \rho \neq 0$.

Ekvivalentne možno formulovať test takto:

H_0 : Znaky sú lineárne nezávislé.

H_1 : Znaky sú lineárne závislé.

Tab. 8.1 T-test lineárnej nezávislosti

Hypotézy	**Použité rozdelenie**	**Testovacia štatistika**	**Oblasť zamietnutia H_0**
H_0: $\rho = 0$ H_1: $\rho \neq 0$	Studentovo	$T = r\sqrt{\frac{n-2}{1-r^2}}$	$\lvert t \rvert > t_{1-\alpha/2,}$ $d.f. = n-2$

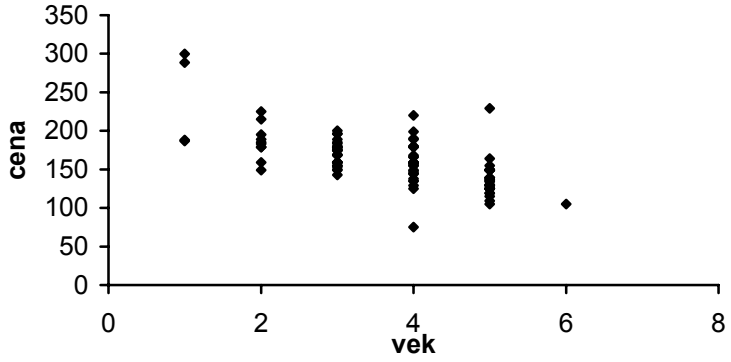
Príklad 8.1

V súbore **Autobazár** sú údaje o veku a cenách áut z 3 predajní autobazáru. Znázornite bodovým diagramom závislosť ceny od veku. Vyšetrite pomocou korelačného koeficientu a kovariancie závislosť ceny auta od veku, použite údaje zo všetkých 3 predajní. Na hladine spoľahlivosti $\alpha = 0{,}05$ otestujte nulovú hypotézu $H_0 : \rho = 0$, oproti alternatívnej hypotéze $H_1 : \rho \neq 0$.

EXCEL

Použitie EXCELu pri riešení korelačnej úlohy budeme ilustrovať na riešení Príkladu 8.1.

Po voľbe ***Vložiť graf/ Závislosť*** vytvoríme bodový diagram (Obr. 8.2). Z grafu vidieť, že s narastajúcim vekom mierne klesá cena áut. Po voľbe ***Nástroje/Analýza údajov/Korelácia*** a zadaní údajov sa objaví výstupná **korelačná matica.** Na jej uhlopriečke sú $r_{xx} = 1$ a $r_{yy} = 1$, a okrem toho výberový korelačný koeficient $r_{xy} = -0{,}6748$, čo predstavuje nepriamu miernu lineárnu závislosť, t.j. s narastajúcim vekom klesá cena auta.



Obr. 8.2 Bodový graf závislosti ceny áut od veku

Po voľbe ***Nástroje/Analýza údajov/Kovariancia*** ako výstup dostaneme ***kovariančnú maticu***, na jej uhlopriečke sú hodnoty $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ a $s_y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ a $\operatorname{cov} xy = -25{,}044$. Rovnaké výsledky sa dajú získať aj postupom ***Prilepiť funkciu/štatistické/*** **CORREL (COVAR)**.

Tab. 8.2 Korelačná matica

	cena	**vek**
cena	1	
vek	-0,67487	1

Tab. 8.3 Kovariančná matica

	cena	**vek**
cena	1080,48391	
vek	-25,044	1,27456

Na záver testujme hypotézu $H_0: \rho = 0$ oproti alternatívnej hypotéze $H_1: \rho \neq 0$.

Hodnota testovacej štatistiky je $t = -0{,}67487\sqrt{\frac{106-2}{1-(-0{,}67487)^2}} = -9{,}3265$. Porovnáme ju s kvantilom $t_{0{,}975;104} = 1{,}983035$ Studentovho rozdelenia. Platí $-9{,}3265 < -1{,}9830$, preto zamietame nulovú hypotézu a tvrdíme, že na hladine významnosti $\alpha = 0{,}05$ je $\rho \neq 0$, alebo že lineárna závislosť znakov je štatisticky významná.

8.3 Regresná analýza

Jednoduchá (párová) lineárna regresia

Úlohou regresnej analýzy pri skúmaní štatistickej závislosti *Y* na *X* je nájsť vhodný matematický model (funkciu), v ktorom je vyjadrená predstava o tejto závislosti. Ak by sa nám podarilo odstrániť spolupôsobenie vedľajších vplyvov na vzťah medzi *X* a *Y,* ležali by všetky body (x_i, y_i) na krivke s rovnicou $y = \eta(x)$, čo je deterministický model. Na premennú *Y* však vplývajú okrem *X* aj iné faktory, preto body (x_i, y_i) neležia na krivke, ale kolíšu okolo nej. To sa snažíme zachytiť aj v matematickom modeli. Preto každú hodnotu závisle premennej *Y* rozložíme na dve zložky, na deterministickú a náhodnú, t.j.

$$y_i = \eta(x_i) + \varepsilon_i, \quad i = 1,2,....,n.$$

Funkcia $\eta = \eta(x)$ sa volá **regresná funkcia.** Môže to byť napr. priamka $y = B_0 + B_1 x$, parabola $y = B_0 + B_1 x + B_2 x^2$ a iné známe funkcie. Náš model, ktorý zachytáva **lineárnu závislosť** *X*, *Y* bude **lineárna funkcia – regresná priamka.** Lineárny vzťah medzi *Y* a *X* v základnom súbore možno vyjadriť modelom

$$y_i = B_0 + B_1 x_i + \varepsilon_i \quad i = 1, 2, ... \tag{8.5}$$

kde y_i – *i*-ta hodnota premennej *Y* v základnom súbore,

B_0 – priesečník osi *y* s regresnou priamkou,

B_1 – **regresný koeficient** v základnom súbore, ktorý udáva o koľko sa zmení y, ak sa x zmení o jednu jednotku (je to smernica regresnej priamky),

x_i – *i*-ta hodnota premennej *X* v základnom súbore,

ε_i – *i*-ta **náhodná chyba** premennej *Y*.

Časť $B_0 + B_1 x_i = \eta_i$ je **deterministická časť modelu,** voláme ju **regresná funkcia**. Je to nám nedostupná teoretická priamka - **regresná priamka** v základnom súbore, okolo ktorej kolíšu skutočné hodnoty *Y* pre dané hodnoty *X*. Pretože k dispozícii máme len výberový súbor s rozsahom *n*, preložíme bodmi výberového súboru **vyrovnávajúcu regresnú priamku**, ktorú môžeme považovať za bodový odhad regresnej priamky v základnom súbore. Označíme ju vzťahom

$$\tilde{y}_i = b_0 + b_1 x_i, \; i = 1, 2, ..., n \tag{8.6}$$

kde $\tilde{y}_i$ - **očakávaná (vyrovnaná) hodnota premennej *Y*** pre danú hodnotu premennej *X,*

x_i - *i*-ta hodnota premennej X,

b_0 - bodový odhad koeficientu B_0,

b_1 - bodový odhad koeficientu B_1, volá sa **výberový regresný koeficient.**

Na výpočet neznámych koeficientov b_0 a b_1 v rovnici vyrovnávajúcej regresnej priamky sa používa **metóda najmenších štvorcov**. Označme rozdiely (chyby) medzi nameranými hodnotami y_i a medzi vyrovnanými hodnotami $\tilde{y}_i$, t.j. $y_i - \tilde{y}_i = e_i$ ako **rezíduá (reziduálne odchýlky)**. Sú to bodové odhady náhodných chýb ε_i regresného modelu. „ Najlepšie preložená“ priamka medzi bodmi (x_i, y_i) je tá, ktorá minimalizuje súčet štvorcov reziduálnych odchýlok

$$\sum_{i=1}^{n} e_i^2 = \sum (y_i - \tilde{y}_i)^2 . \tag{8.7}$$

To je podstata **metódy najmenších štvorcov**. Pri hľadaní koeficientov b_0 a b_1 využijeme skutočnosť, že hľadáme minimum funkcie dvoch premenných

$$f(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum (y_i - \tilde{y}_i)^2 = \sum [y_i - (b_0 + b_1 x)]^2 . \tag{8.8}$$

Vieme, že extrém funkcie tohto typu môže existovať len v stacionárnom bode funkcie, t.j. musí platiť

$$\frac{\partial f}{\partial b_0} = 0 \quad \text{a} \quad \frac{\partial f}{\partial b_1} = 0 .$$

Teda

$$\sum -2(y_i - b_0 - b_1 x_i) = 0 \tag{8.9}$$

$$\sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0 \tag{8.10}$$

Po úprave rovnice (8.9) dostaneme

$$\sum y_i - b_1 \sum x_i = b_0 n$$

odtiaľ $$b_0 = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n} = \bar{y} - b_1 \bar{x} .$$

Úpravou rovnice (8.10) dostaneme

$$\sum x_i y_i - b_0 \sum x_i = b_1 \sum x_i^2$$

$$\sum x_i y_i - (\bar{y} - b_1 \bar{x}) \sum x_i = b_1 \sum x_i^2$$

$$\sum x_i y_i - \bar{y}\sum x_i = b_1(\sum x_i^2 - \bar{x}\sum x_i).$$

Po vynásobení poslednej rovnice výrazom $1/n$ a úprave

$$\frac{\sum x_i y_i}{n} - \bar{x}.\bar{y} = b_1\left[\frac{\sum x_i^2}{n} - \left(\bar{x}\right)^2\right] \quad \Rightarrow \quad b_1 = \frac{\overline{xy} - \bar{x}.\bar{y}}{s_x^2} = \frac{\mathrm{cov}\, xy}{s_x^2} = r\frac{s_y}{s_x}$$

$$b_1 = r\frac{s_y}{s_x} \tag{8.11}$$

$$b_0 = \bar{y} - b_1\bar{x}\ . \tag{8.12}$$

Vyrovnávajúca regresná priamka má rovnicu $\tilde{y} = b_0 + b_1 x$

$$\tilde{y} = \left(\bar{y} - \bar{x}\, r\frac{s_y}{s_x}\right) + \left(r\frac{s_x}{s_y}\right)x$$

čo po úprave je

$$\tilde{y} - \bar{y} = r\frac{s_y}{s_x}\left(x - \bar{x}\right) \tag{8.13}$$

$$\tilde{y} - \bar{y} = b_1\left(x - \bar{x}\right) \tag{8.14}$$

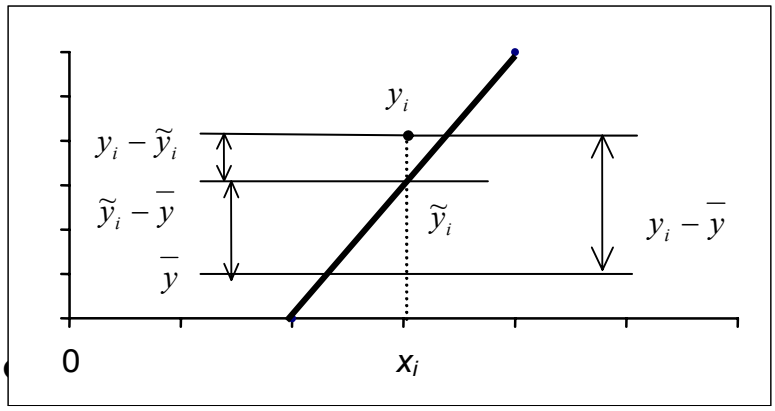
Nebudeme sa zdržiavať dôkazom, že v tomto stacionárnom bode má funkcia skutočne lokálne minimum. Teoretickú regresnú priamku sme odhadli priamkou $\tilde{y} = b_0 + b_1 x$, ktorú považujeme za **bodový odhad** neznámej regresnej priamky .

Poznámka 8.1

Na konštrukciu koeficientov b_0 a b_1 nemôžeme použiť len súčet chýb e_i, lebo vždy platí $\sum_{i=1}^{n} e_i = 0$, aj pre zle zvolenú regresnú priamku. Všimnime si ešte dve vlastnosti regresnej priamky. Regresná priamka prechádza bodom $\left(\bar{x}, \bar{y}\right)$ a regresný koeficient má vždy rovnaké znamienko ako korelačný koeficient.

8. 4 Skúmanie štatistickej významnosti modelu

Po nájdení rovnice regresnej priamky treba overiť, či tento model je „kvalitný“, či dobre vystihuje závislosť medzi *X, Y*. Pri riešení regresnej úlohy prichádza často do úvahy viacero typov regresných funkcií (kvadratická, logaritmická), preto sa skúma, ktorá z týchto funkcií „lepšie prilieha“ výberovým údajom. To sa dá merať rôznymi charakteristikami: **reziduálny súčet štvorcov, reziduálny rozptyl, štandardná odchýlka reziduí, koeficient determinácie** alebo preveriť rôznymi **testami**.



Obr. 8.3 Rozklad celkovej variability premennej Y

Na Obr. 8.3 je jasný vzťah:

$$\left(y_i - \bar{y}\right) = \left(\tilde{y}_i - \bar{y}\right) + \left(y_i - \tilde{y}_i\right), \tag{8.15}$$

t.j. odchýlka od celkového priemeru = odchýlka vysvetlená regresiou + odchýlka nevysvetlená regresiou (reziduálna). Prekvapivo platí aj

$$\sum \left(y_i - \bar{y}\right)^2 = \sum_i \left(\tilde{y}_i - \bar{y}\right)^2 + \sum_i \left(y_i - \tilde{y}_i\right)^2, \tag{8.16}$$

$$SSY = SSR + SSE \tag{8.17}$$

SSY - je **celková variabilita** premennej *Y* (celkový súčet štvorcov, sum of squares total),

SSR - je **variabilita vysvetlená regresným modelom** (sum of squares due to regression),

SSE - je **variabilita nevysvetlená regresným modelom, reziduálny súčet štvorcov** (sum of squares due to error).

Dokážeme vlastnosť (8.16). Po umocnení výrazu (8.15) a sčítaní pre všetky $i = 1,2,...,n$ dostaneme

$$\sum \left(y_i - \bar{y}\right)^2 = \sum_i \left(\tilde{y}_i - \bar{y}\right)^2 + \sum_i \left(y_i - \tilde{y}_i\right)^2 + 2\sum_{i=1} \left(\tilde{y}_i - \bar{y}\right)\left(y_i - \tilde{y}_i\right) .$$

Hodnota posledného sčítanca je nula, lebo

$$\sum \left(y_i - \tilde{y}_i\right)\tilde{y}_i - \bar{y}\sum \left(y_i - \tilde{y}_i\right) = \sum \left(y_i - b_0 - b_1 x_i\right)\left(b_0 + b_1 x_i\right) - \bar{y}\sum \left(y_i - b_0 - b_1 x_i\right)$$

$$= b_0 \sum \left(y_i - b_0 - b_1 x_i\right) + b_1 \sum \left(y_i - b_0 - b_1 x_i\right)\left(x_i\right) - \bar{y}\sum \left(y_i - b_0 - b_1 x_i\right) = 0 ,$$

pričom sme použili vzťahy (8.9) a (8.10), t.j. parciálne derivácie

$$\frac{\partial f}{\partial b_0} = 0 \text{ a } \frac{\partial f}{\partial b_1} = 0 .$$

Porovnanie zložiek SSY, SSR, SSE je jedna možnosť, ako posúdiť štatistickú významnosť modelu ako celku:

- Pri funkčnej závislosti je *SSE* = 0, *SSY* = *SSR*, lebo všetky body y_i ležia na vyrovnávajúcej priamke.
- Pri nezávislosti je *SSR* = 0, *SSY* = *SSE*, lebo vyrovnávajúca priamka je rovnobežná s osou *x* a prechádza napríklad bodom $\left(x_1, \bar{y}\right)$.
- Závislosť *X, Y* je tým silnejšia, čím je väčší podiel variability *SSR* na celkovej variabilite *SSY*. Sila tejto lineárnej závislosti sa meria **výberovým koeficientom determinácie,** ktorý je definovaný

$$r^2 = \frac{SSR}{SSY}; \qquad r^2 \in \langle 0,1 \rangle . \tag{8.18}$$

Lineárny vzťah medzi *X,Y* je tak vysvetlený na $\frac{SSR}{SSY}.100$ %, preto je z viacerých modelov „kvalitnejší“ model s vyšším koeficientom determinácie. Výberový koeficient determinácie r^2 je bodovým odhadom **koeficientu determinácie** ρ^2 v

základnom súbore, ale skresleným. Neskreslený odhad dáva **korigovaný koeficient determinácie**

$$r_{adj}^2 = 1 - \left(1 - r^2\right)\frac{n-1}{n-2}. \tag{8.19}$$

Koeficient determinácie $r^2 = \frac{SSR}{SSY}$ (8.18) je druhá mocnina korelačného koeficientu r (8.4), ktorý bol definovaný v časti 8.2. Dokážeme toto tvrdenie. Využijeme rovnicu vyrovnávajúcej regresnej priamky (8.14)

$$\tilde{y}_i - \bar{y} = b_1\left(x_i - \bar{x}\right).$$

Po umocnení a sčítaní pre $i = 1, 2, ..., n$ platí

$$\sum\left(\tilde{y}_i - \bar{y}\right)^2 = b_1^2 \sum\left(x_i - \bar{x}\right)^2.$$

Po dosadení tohto vzťahu do $SSY = SSR + SSE$ dostaneme:

$$\sum\left(y_i - \bar{y}\right)^2 = b_1^2 \sum\left(x_i - \bar{x}\right)^2 + \sum\left(y_i - \tilde{y}_i\right)^2. \tag{8.20}$$

Podľa (8.11), kde r je korelačný koeficient, platí $b_1 = r\frac{s_y}{s_x}$ t.j.

$$b_1^2 = r^2\frac{s_y^2}{s_x^2} = r^2\frac{\sum\left(y_i - \bar{y}\right)^2}{\sum\left(x_i - \bar{x}\right)^2}$$

a po dosadení do (8.20)

$$SSY = r^2\sum\left(y_i - \bar{y}\right)^2 + SSE$$

$$SSY = r^2 SSY + SSE$$

$$r^2 = \frac{SSY - SSE}{SSY} = \frac{SSR}{SSY}.$$

Cieľom metódy najmenších štvorcov bolo minimalizovať variabilitu nevysvetlenú regresným modelom, hodnotu $SSE = \sum\left(y_i - \tilde{y}_i\right)^2$, ktorá sa volá aj **reziduálny súčet štvorcov**. Z dvoch modelov, ktoré by teoreticky prichádzali do úvahy, je lepší ten, kde je menší SSE. Mierou variability hodnôt y_i okolo vyrovnávajúcej regresnej priamky je **štandardná odchýlka rezíduí**

$$s_{rez} = \sqrt{\frac{\sum (y_i - \tilde{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}. \qquad (8.21)$$

Je to neskreslený bodový odhad štandardnej odchýlky náhodných chýb ε_i v základnom súbore. Jej druhá mocnina s_{rez}^2 sa nazýva **reziduálny rozptyl**.

Poznámka 8.2

Koeficient determinácie, štandardná odchýlka rezíduí, korigovaný koeficient determinácie tvoria výstup EXCELu po procedúre ***Regresia***.

8.5 Testy hypotéz používané pri voľbe regresnej funkcie

a) test linearity (celkový F-test)

Na začiatku našich úvah sa pýtame, či vôbec medzi premennými *X* a *Y* existuje lineárna závislosť. Ak empirické údaje zobrazíme bodovým diagramom a body $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ ležia v páse, ktorý sa dá približne ohraničiť dvomi priamkami, ktoré nie sú rovnobežné s osou *x*, môžeme predpokladať lineárnu závislosť medzi *X* a *Y*. Preto sformulujeme nulovú a alternatívnu hypotézu takto:

H_0 : Lineárny model nie je štatisticky významný (t.j. *X,Y* nie sú lineárne závislé).

H_1 : Lineárny model je štatisticky významný (t.j. *X,Y* sú lineárne závislé).

Na overenie platnosti H_1 použijeme známu analýzu rozptylu tak, že odhad s_y^2 celkového rozptylu σ^2 závisle premennej *Y* rozložíme na dve zložky:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{SSE}{n-1} = \frac{1}{n-1}\left[\sum_i (y_i - \tilde{y}_i)^2 + \sum_i (\tilde{y}_i - \bar{y})^2\right]$$

$$= \frac{1}{n-1}\left[SSE + SSR\right], \text{ t.j.}$$

$$(n-1)s_y^2 = SSE + SSR$$

Náhodné premenné $\frac{(n-1)s_y^2}{\sigma^2}$, $\frac{SSE}{\sigma^2}$, $\frac{SSR}{\sigma^2}$ majú χ^2 – rozdelenia postupne s $(n-1)$, $(n-2)$ a 1 stupňom voľnosti. Podiel rozptylov $F = \frac{SSR/1}{SSE/n-2} = \frac{MSR}{MSE}$ má Fisherovo rozdelenie s $(1, n-2)$ stupňami voľnosti, kde

MSR - priemerný štvorec regresie (mean square of regression),

MSE - priemerný štvorec chýb (mean square of errors).

Podstata testu je v tom, že sme našli náhodnú premennú, ktorá je funkciou *SSR* a *SSE* a ktorej rozdelenie poznáme. Model je tým lepší, čím je väčšie číslo *F*, preto veľké hodnoty testovacej štatistiky *F* hovoria v prospech alternatívnej hypotézy, teda padnú do oboru zamietnutia H_0.

Záver: F – test je len jednostranný test (pravostranný). Nulovú hypotézu zamietame, ak pri zvolenej hladine významnosti α je hodnota testovacej štatistiky $F > F_{1-\alpha(1,n-2)}$, kde $F_{1-\alpha}$ je príslušný kvantil F-rozdelenia s $(1, n-2)$ stupňami voľnosti. V tomto prípade teda **prijímame alternatívnu hypotézu o lineárnom vzťahu medzi *X* a *Y***. Nájdená regresná priamka je vhodný typ funkcie na vyjadrenie priebehu závislosti.

Tab. 8.4 Celkový F-test

Hypotézy	Použité rozdelenie	Testovacia štatistika	Oblasť zamietnutia H_0
H_0 : *X,Y* sú lineárne nezávislé. H_1 : *X,Y* sú lineárne závislé.	Fisherovo	$F = \frac{SSR/1}{SSE/n-2}$	$F > F_{1-\alpha}$ *d.f.* $= (1, n-2)$

b) t-test o lineárnej nezávislosti *X, Y*

Tento test je založený na nasledujúcej myšlienke. Regresný koeficient B_1 je smernica regresnej priamky a vyjadruje priemernú zmenu *Y* pri zmene *X* o jednu jed-

notku. Ak $B_1 = 0$, regresná priamka je rovnobežná s osou *x*, teda aj po zmene nezávisle premennej *X* sa nemenia hodnoty *Y* (presnejšie podmienené stredné hodnoty). Preto sa nedá hovoriť o lineárnej závislosti *X, Y*.

Ak je výberový regresný koeficient b_1 blízky nule, treba overiť hypotézu, či koeficient B_1 je rôzny od nuly, t.j. overiť hypotézu, či medzi *X* a *Y* existuje lineárna závislosť.

H_0: $B_1 = 0$ (t.j. *X,Y* sú lineárne nezávislé.)

H_1: $B_1 \neq 0$ (t.j. *X,Y* sú lineárne závislé.)

Na testovanie použijeme testovaciu štatistiku $T = \frac{b_1 - B_1}{s(b_1)}$, ktorá má Studentovo rozdelenie s $(n-2)$ stupňami voľnosti a $s(b_1) = \frac{s_{rez}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$ je štandardná odchýlka koeficientu b_1. Ak platí nulová hypotéza, vypočítame hodnotu testovacej štatistiky $T = \frac{b_1}{s(b_1)}$.

Záver: Nulovú hypotézu zamietame, ak pri zvolenej hladine významnosti α je hodnota testovacej štatistiky $|t| > t_{1-\alpha/2,(n-2)}$, kde $t_{1-\alpha/2}$ je kvantil Studentovho rozdelenia s $(n-2)$ stupňami voľnosti. V tomto prípade teda **prijímame alternatívnu hypotézu o lineárnom vzťahu medzi *X* a *Y***.

Tab. 8.5 T-test o lineárnej nezávislosti

Hypotézy	**Použité rozdelenie**	**Testovacia štatistika**	**Oblasť zamietnutia H_0**
H_0: $B_1 = 0$ H_1: 1. $B_1 \neq 0$ 2. $B_1 > 0$ 3. $B_1 < 0$	Studentovo	$T = \frac{b_1}{s(b_1)}$ $s(b_1) = \frac{s_{rez}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$	1. $\|t\| > t_{1-\alpha/2,}$ 2. $t > t_{1-\alpha}$ 3. $t < -t_{1-\alpha}$ $d.f. = n-2$

Podobne sa dá testovať hypotéza $H_0 : B_0 = 0,\ H_1 : B_0 \neq 0$.

Test lineárnej závislosti vieme urobiť tromi ekvivalentnými spôsobmi, posledné dva sú aj výstupom EXCELu :

- testovať korelačný koeficient
- celkový F-test
- testovať regresný koeficient

EXCEL

Na riešenie regresnej úlohy ponúka EXCEL nasledujúce prostriedky.

Po voľbe ***Nástroje/ Analýza údajov / Regresia*** a zadaní údajov najskôr pre závislú *Y*, potom pre nezávislú premennú *X* sa v tabuľkách objavia údaje Tab. 8.6, pričom niektoré sú zle pomenované.

Tab. 8.6 Regresná štatistika

pomenovanie	skutočný význam
Násobné R	$\lvert r \rvert$ - absolútna hodnota *r*
Hodnota spoľahlivosti	r^2 - koef. determinácie
Nastavená hodnota spoľahlivosti	upravený koef. determinácie
Chyba strednej hodnoty	s_{rez}
Pozorovania	*n*

Tabuľka ANOVA poskytuje rozklad celkového rozptylu na dve zložky a celkový F-test .

Tab. 8.7 ANOVA

	stupne voľnosti	*SS*	*MS*	F	významnosť F p- hodnota
Regresia	1	*SSR*	*MSR*=*SSR*/1	hodnota testovacej štatistiky	H_0 : Lineárny model nie je štatisticky významný.
Rezíduá	*n*-2	*SSE*	*MSE*=*SSE*/*n*-2 s^2_{rez}		
Celkom	*n*-1	*SSY*			

V poradí tretia Tab. 8.8 okrem koeficientov regresnej priamky obsahuje aj t-test pre nulovosť regresného koeficientu B_1 (druhý riadok) a koeficientu B_0 (prvý riadok).

Tab. 8.8 Testovanie koeficientov regresnej priamky

	koeficienty	chyba strednej hodnoty	t-stat	p-hodnota	dolný 95%	horný 95%	dolný 99%	horný 99%
hranice	b_0	$s(b_0)$	$b_0/s(b_0)$	H_0: B_0=0	intervaly spoľahlivosti pre B_0			
X	b_1	s(b_1)	$b_1/s(b_1)$	H_0: B_1=0	intervaly spoľahlivosti pre B_1			

Posledná tabuľka obsahuje aj pre každý prvok x_i výberového súboru vypočítanú očakávanú hodnotu $\tilde{y}_i$ a aj rezíduum $e_i = y_i - \tilde{y}_i$.

Príklad 8.2

V súbore **Autobazár** sú údaje o veku a cenách 106 áut z 3 predajní autobazáru. Vyšetrite lineárnu závislosť ceny auta od veku, použite údaje zo všetkých 3 predajní, nájdite rovnicu regresnej priamky, na hladine významnosti $\alpha = 0{,}05$ otestujte štatistickú významnosť lineárneho modelu.

V príklade po voľbe ***Nástroje/ Analýza údajov / Regresia*** dostaneme nasledujúce výstupné tabuľky:

Regresní statistika	
Násobné R	0,675
Hodnota spolehlivosti R	0,455
Nastavená hodnota spolehlivosti F	0,450
Chyba stř. hodnoty	24,489
Pozorování	106

	Koeficienty	Chyba stř. hodnoty	t stat	Hodnota P	Dolní 95%	Horní 95%	Dolní 99%	Horní 99%
Hranice	233,951	8,279	28,257	1,325E-50	217,533	250,369	212,226	255,676
vek	-19,649	2,107	-9,327	2,1494E-15	-23,827	-15,471	-25,178	-14,121

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	1	52163,363	52163	86,98364	2,14942E-15
Rezidua	104	62367,931	599,7		
Celkem	105	114531,294			

Z tabuliek vyplýva:

- Absolútna hodnota korelačného koeficientu je $|r| = 0{,}675$, regresný koeficient je (–19,649). Korelačný koeficient má rovnaké znamienko ako regresný koeficient, preto je korelačný koeficient $r = -0{,}675$, čo interpretujeme ako nepriamu, miernu lineárnu závislosť.
- Koeficient determinácie je $r^2 = 0{,}455$, tzn. len 45,5 % variability ceny áut sa dá vysvetliť lineárnym vzťahom s vekom áut.
- Neskreslený odhad koeficientu determinácie v základnom súbore je číslo 0,4502.
- p-hodnota pre celkový F-test je $2{,}14 \cdot 10^{-15}$, čo je veľmi malé číslo. Na všetkých bežných hladinách významnosti zamietame nulovú hypotézu, prijímame alternatívnu hypotézu, že daný model je štatisticky významný, t.j. premenné sú lineárne závislé.
- **Rovnica vyrovnávajúcej regresnej priamky** je $y = -19{,}649x + 233{,}95$.

- $s_{rez} = 24{,}489$, tzn. skutočné ceny áut sa odchyľujú od hodnôt regresnej priamky približne o $\pm 24{,}5$ tisíc korún.
- p-hodnota pri t-teste hypotézy $H_0 : B_1 = 0$ oproti $H_1 : B_1 \neq 0$ je to isté malé číslo $2{,}14 \cdot 10^{-15}$, preto na každej bežnej hladine významnosti prijímame alternatívnu hypotézu, že premenné sú lineárne závislé.

8.6 Použitie regresnej priamky

Regresnú priamku $\tilde{y} = b_0 + b_1 x$ považujeme za **bodový odhad** strednej hodnoty závisle premennej Y a môžeme ju použiť na

- bodový odhad hodnoty Y pre jednu konkrétnu hodnotu X,
- bodový odhad priemernej hodnoty Y pre istú úroveň znaku X, ale len na intervale $\langle x_{min}, x_{max} \rangle$.

Napríklad, môžeme očakávať, že cena jedného 2 ročného auta je

$$y = 233{,}95 - 19{,}65.2 = 194{,}65 \text{ tisíc Sk.}$$

Je to zároveň priemerná cena všetkých dvojročných áut.

Poznámka 8.3

Vieme však určiť aj :

- 100(1-α)% interval spoľahlivosti pre koeficient B_0:

$$b_0 - t_{1-\alpha/2,(n-2)}.s(b_0) < B_0 < b_0 + t_{1-\alpha/2,(n-2)}.s(b_0),$$

- 100(1-α)% interval spoľahlivosti pre koeficient B_1:

$$b_1 - t_{1-\alpha/2,(n-2)}.s(b_1) < B_1 < b_1 + t_{1-\alpha/2,(n-2)}.s(b_1),$$

- 100(1-α)% interval spoľahlivosti pre **priemernú hodnotu *Y* v základnom súbore pre danú konkrétnu hodnotu** x_i, označíme ju $\mu(y/x_i)$ a platí

$$(b_0 + b_1 x_i) - t_{1-\alpha/2,(n-2)}.s_i < \mu(y/x_i) < (b_0 + b_1 x_i) + t_{1-\alpha/2,(n-2)}.s_i,$$

kde $s_i = s_{rez}.\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}}$ je štandardná odchýlka vyrovnaných hodnôt $\tilde{y}_i = b_0 + b_1 x_i$. Šírka tohto intervalu je iná pre každé x_i a rozširuje sa so vzďaľovaním x_i od $\bar{x}$.

- 100(1-α)% interval spoľahlivosti pre **individuálnu hodnotu *Y* v základnom súbore pre danú konkrétnu hodnotu** x_i **,** označíme ju $Y(x_i)$ a platí

$$(b_0 + b_1 x_i) - t_{1-\alpha/2,(n-2)}.s_i < Y(x_i) < (b_0 + b_1 x_i) + t_{1-\alpha/2,(n-2)}.s_i ,$$

kde $s_i = s_{rez}.\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}+1}$ je štandardná odchýlka individuálnych hodnôt premennej *Y*. Šírka tohto intervalu je väčšia ako pre odhad stabilnejšej priemernej hodnoty.

EXCEL

Z výstupných tabuliek pre ***Regresiu*** sa dá pre náš Príklad 8.2 určiť :

- 95% interval spoľahlivosti pre B_0 : $217{,}53 < B_0 < 250{,}36$,
- 95% interval spoľahlivosti pre B_1 : $-23{,}83 < B_1 < -15{,}47$,
- Podobne z tabuliek sú známe 99% intervaly spoľahlivosti pre B_0 a B_1.

95% interval spoľahlivosti pre priemernú cenu 2 – ročného auta musíme dopočítať bez EXCELu.

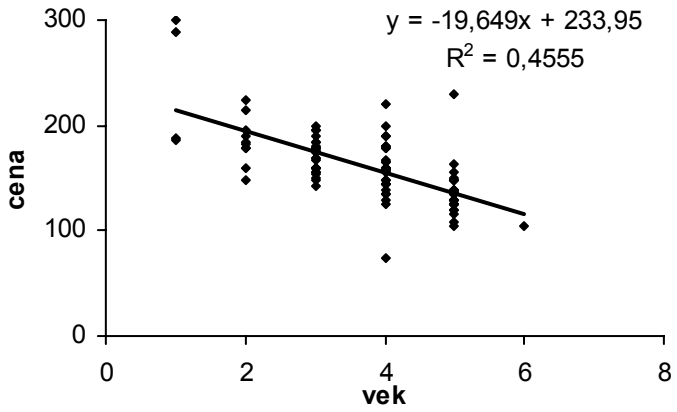
$$\bar{x} = 3{,}764,\ (2-\bar{x})^2 = 3{,}112,\ \sum_i (x_i-\bar{x})^2 = 135{,}104\ \sqrt{\frac{1}{n}+\frac{(2-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}} = 0{,}180$$

$$s_{rez} = 24{,}489,\quad t_{1-\alpha/2,(n-2)} = 1{,}983,\quad \tilde{y}_2 = b_0 + b_1.2 = 194{,}652$$

maximálne prípustná chyba odhadu je teda $t_{1-\alpha/2,(n-2)}.s_i = 8{,}750$, preto 95% interval spoľahlivosti pre priemernú cenu 2 – ročného auta je $194{,}652 \pm 8{,}750$ tisíc Sk.

Poznámka 8.4

Úpravami bodového diagramu v EXCELi sa dá do grafu vložiť regresná krivka, rovnica regresnej krivky i koeficient determinácie. Klikneme pravým tlačidlom na niektorý bod grafu, zvolíme ***Pridať trendovú čiaru***, na záložke ***Typ/lineárny*** a na záložke ***Možnosti/zobraziť v grafe rovnicu regresnej priamky a R^2***.



Obr. 8.4 Úpravy bodového grafu

8.7 Predpoklady pre použitie metódy najmenších štvorcov

Metóda najmenších štvorcov dáva neskreslený odhad regresnej priamky pri splnení istých predpokladov o rozdelení pravdepodobnosti náhodných chýb ε_i v modeli $y_i = B_0 + B_1 x_i + \varepsilon_i$.

Sú to tieto predpoklady:

- Stredná hodnota náhodných chýb je nula.
- Rozptyl náhodných chýb je konštantný.
- Rozdelenie pravdepodobnosti náhodných chýb je normálne .
- Náhodné chyby sú medzi sebou vzájomne nezávislé.

Splnenie týchto predpokladov sa dá overiť až po zvolení regresného modelu, lebo až vtedy sú známe rezíduá, ktoré sú odhadmi náhodných chýb. Ich splnenie sa približne overí graficky zostrojením histogramu rozdelenia rezíduí a z bodového diagramu hodnôt $(\tilde{y}_i , e_i)$. Podrobnejšie sa tomuto problému nebudeme venovať (pozri [6]).

8.8 Iné typy regresných funkcií

Lineárna regresná funkcia je vďaka ľahkej interpretácii preferovaná pred inými typmi, ale niekedy z povahy problému vyplýva, že pre popis danej závislosti by bola vhodnejšia iná regresná funkcia. Uvedieme niektoré iné modely.

Parabolická regresia Regresná funkcia je tvaru $\eta = B_0 + B_1 x + B_2 x^2$, bodové odhady koeficientov získame priamo použitím metódy najmenších štvorcov, t.j. hľadaním minima funkcie troch premenných $\sum_{i=1}^{n} e_i^2 = \sum (y_i - \tilde{y}_i)^2$, kde

$$\tilde{y}_i = b_0 + b_1 x_i + b_2 x_i^2 .$$

Zovšeobecnením môže byť polynomická regresia vyššieho stupňa, v praxi sa stretávame s polynómami maximálne 3. a 4. stupňa.

Hyperbolická regresia Regresná funkcia je tvaru $\eta = B_0 + \frac{B_1}{x}$. Bodové odhady koeficientov získame tiež priamo použitím metódy najmenších štvorcov.

Logaritmická regresia Regresná funkcia je tvaru $\eta = B_0 + B_1 \log x$. Bodové odhady koeficientov získame priamo použitím metódy najmenších štvorcov.

Exponenciálna regresia Regresná funkcia je tvaru $\eta = B_0 \cdot B_1^x$. Bodové odhady koeficientov sa nedajú získať priamo použitím metódy najmenších štvorcov. Vhodnou úpravou (transformáciu) regresnej funkcie ju upravíme na taký tvar, kde funkcie jej parametrov sa dajú odhadnúť metódou najmenších štvorcov. V tomto prípade logaritmovaním dostaneme :

$$\ln \eta = \ln B_0 + x \ln B_1$$

a budeme hľadať minimum funkcie $\sum (\ln y_i - \ln b_0 - x_i \ln b_1)^2$.

Známym spôsobom použijeme parciálne derivácie tejto funkcie a ako riešenie sústavy získame $\ln b_0$ a $\ln b_1$.

Poznámka 8.5

Podobne postupujeme aj v prípade iných typov funkcií, ktorým sa však nebudeme venovať.

Pri výbere vhodného typu regresnej funkcie sa v EXCELi orientujeme podľa hodnoty koeficientu determinácie, ktorý je definovaný nezávisle na type regresnej funkcie.

Príklad 8.3

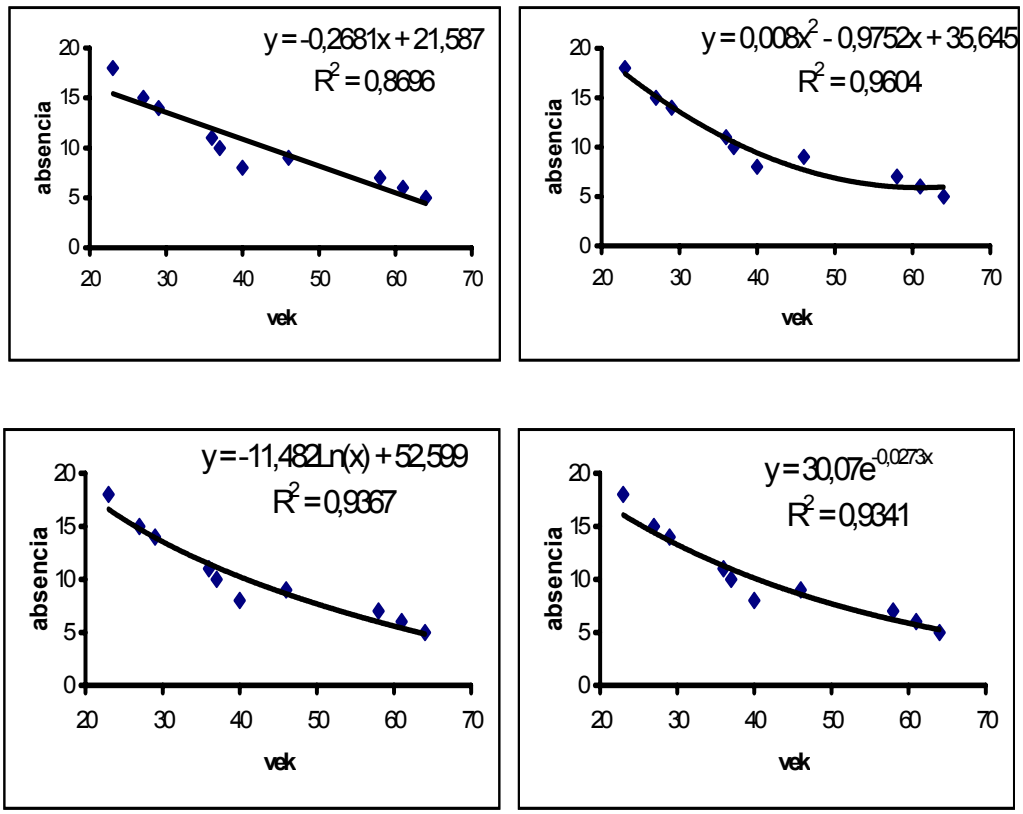
Pracovník personálneho oddelenia cíti, že existuje vzťah medzi počtom dní absencie v práci a vekom pracovníka. Vyšetrite túto závislosť na základe údajov v Tab. 8.9. Pre názornosť príkladu sme použili nevyhovujúci, veľmi malý rozsah výberu.

Tab. 8.9

vek	27	61	37	23	46	58	29	36	64	40
absencia	15	6	10	18	9	7	14	11	5	8

EXCEL

S vyžitím EXCELu vieme vložiť do bodových diagramov graf regresnej čiary, EXCEL poskytuje na výber lineárnu, logaritmickú, exponenciálnu, polynomickú (ľubovoľného stupňa) alebo mocninovú krivku s jej analytickým vyjadrením i koeficientom determinácie. V tomto príklade vidíme, že z viacerých možných riešení je najvhodnejšia parabola s najväčším koeficientom determinácie. Sám riešiteľ úlohy sa musí rozhodnúť, ktorý z týchto možných modelov je pre jeho potreby vyhovujúci.



Obr. 8.5 Rôzne modely regresných funkcií k Príkladu 8.3

Ak sa podrobnejšie zaujímame aj o celkový F-test štatistickej významnosti modelu, t-testy nulovosti koeficientov regresnej funkcie, o intervalové odhady koeficientov regresnej funkcie použijeme v EXCELi funkciu ***Regresia***, kde do vstupnej tabuľky pre

nezávislú premennú vložíme viac stĺpcov . Výstupné tabuľky pre náš príklad sú uvedené pre parabolickú regresiu, interpretácia tabuliek je rovnaká ako pri lineárnej regresii.

Tab. 8.10 Kvadratická regresia k Príkladu 8.3

abs	vek	vek na druhú
15	27	729
6	61	3721
10	37	1369
18	23	529
9	46	2116
7	58	3364
14	29	841
11	36	1296
5	64	4096
8	40	1600

VÝSLEDEK

Regresní statistika	
Násobné R	0,9800
Hodnota spolehlivosti R	0,9604
Nastavená hodnota spolehlivosti R	0,9491
Chyba stř. hodnoty	0,9516
Pozorování	10

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	2	153,7611	76,8805	84,8979	1,23507E-05
Rezidua	7	6,3389	0,9056		
Celkem	9	160,1000			

	Koeficienty	*Ch.stř. hodnoty*	*t stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	35,645	3,638	9,799	0,0000	27,043	44,247
vek	-0,975	0,178	-5,484	0,0009	-1,396	-0,555
vek na druhu	0,008	0,002	4,006	0,0052	0,003	0,013

Príklady na precvičenie

8.4 Zostrojte bodový diagram závislosti znakov *Y* na *X* pre dané výberové súbory a dané premenné, vložte do grafu rovnicu regresnej priamky, hodnotu koeficientu determinácie, vytvorte korelačnú maticu, pomocou t-testu na

hladine významnosti $\alpha = 0{,}05$ otestujte lineárnu závislosť premenných, do grafu vložte iný typ regresnej krivky.

a) PRIJÍMACIE SKÚŠKY/1.fakulta, *X*-matematika, *Y*- fyzika

Určite, aké priemerný počet bodov dosiahne študent z fyziky, ak z matematiky získal 15 bodov (21 bodov).

b) PRIJÍMACIE SKÚŠKY/2.fakulta, *X*-matematika, *Y*- fyzika.

Určite, aké priemerný počet bodov dosiahne študent z fyziky, ak z matematiky získal 15 bodov (21bodov).

c) PRIJÍMACIE SKÚŠKY/1.fakulta, *X*-matematika (upr.) *Y*- priemer

Určite, aké priemernú známku dosiahne študent po 2.semestri, ak zo skúšky z matematiky (upír) mal známku 3,2 (2).

d) PRIJÍMACIE SKÚŠKY/2.fakulta, *X*-matematika, *Y*- priemer.

Určite, aké priemernú známku dosiahne študent po 2.semestri, ak zo skúšky z matematiky (upr) mal známku 3,2 (2).

8.5 Pomocou nástroja ***Regresia*** vyšetrite lineárnu závislosť znakov *X, Y* v základných súboroch, určite korelačný koeficient, koeficient determinácie, nájdite rovnicu regresnej priamky, na hladine významnosti $\alpha = 0{,}05$ pomocou celkového F-testu, alebo t-testu posúďte významnosť štatistického modelu, nájdite príslušné intervaly spoľahlivosti pre koeficienty regresnej priamky.

a) PRIJÍMACIE SKÚŠKY/1.fakulta, *X*-matematika, *Y*- fyzika,

b) PRIJÍMACIE SKÚŠKY/2.fakulta, *X*-matematika, *Y*- fyzika.

9. SKÚMANIE ZÁVISLOSTI DVOCH KVALITATÍVNYCH ZNAKOV

V marketingových, pedagogických, sociologických i psychologických výskumoch sa často pracuje s kvalitatívnymi znakmi, napr. národnosť, pohlavie, farba očí, typ bývania, názor na trest smrti. Preto túto časť venujeme otázke vyšetrenia závislosti **(asociácie, kontingencie)** dvoch znakov takéhoto charakteru. O asociácii hovoríme v prípade, ak oba znaky majú len po dva varianty hodnôt (tzv. alternatívne, resp. dichotomické znaky), o kontingencii vtedy, ak aspoň jeden znak má aspoň tri možné varianty nadobudnutých hodnôt. Je výhodné, ak je výberový súbor roztriedený podľa variant oboch znakov do **kontingenčnej** alebo **asociačnej** tabuľky.

9.1 χ^2 - test nezávislosti

Populárny a jednoduchý test, ktorým sa dá overiť nezávislosť dvoch znakov je χ^2 **- test nezávislosti**. Dá sa použiť aj v prípade dvoch kvantitatívnych znakov, aj v prípade kombinácie kvantitatívneho i kvalitatívneho znaku, no v praxi sa rozšíril hlavne pre overovanie závislosti kvalitatívnych znakov. Pre dva kvantitatívne údaje použijeme regresnú analýzu.

Nulovú a alternatívnu hypotézu môžeme sformulovať takto:

H_0: Medzi kvalitatívnymi znakmi nie je závislosť.

H_1: Medzi kvalitatívnymi znakmi je závislosť.

Základná myšlienka testu nezávislosti, testovacia štatistika aj jej interpretácia je rovnaká ako pri Pearsonovom χ^2 - teste zhody a tou je porovnanie napozorovaných n_{ij} (empirických) početností a za predpokladu platnosti nulovej hypotézy očakávaných e_{ij} (teoretických) početností. Ak súbor s rozsahom n je roztriedený podľa *r* variantov znaku *A* a *s* variantov znaku *B* do kontingenčnej tabuľky (Tab. 9.1), potom vnútorné pole tabuľky má $r.s$ políčok, ktoré obsahujú empirické po-

četnosti n_{ij}, kde n_{ij} znamená počet štatistických jednotiek, ktoré majú *i*-ty variant znaku *A* ($i = 1,2,...,r$) a súčasne *j*-ty variant znaku *B* ($j = 1,2,...s$).

Tab. 9.1 Kontingenčná tabuľka pre *r = 4* a *s = 3*

	b_1	b_2	b_3	$\sum$
a_1	n_{11}	n_{12}	n_{13}	(a_1)
a_2	n_{21}	n_{22}	n_{23}	(a_2)
a_3	n_{31}	n_{32}	n_{33}	(a_3)
a_4	n_{41}	n_{42}	n_{43}	(a_4)
$\sum$	(b_1)	(b_2)	(b_3)	n

Označme symbolom e_{ij} teoretický počet štatistických jednotiek, ktoré nadobúdajú *i*-ty variant znaku *A* a súčasne *j*-ty variant znaku *B*, za predpokladu, že znaky sú nezávislé. Pre pravdepodobnosť nezávislých javov platí

$$P(A \cap B) = P(A).P(B),$$

preto môžeme pravdepodobnosť, že jeden prvok nadobudne *i*-ty variant znaku *A* (označíme ho a_i) a súčasne *j*-ty variant znaku *B* (označíme ho b_j) vypočítať (presnejšie je odhadnúť) ako súčin pravdepodobnosti *i*-teho variantu znaku *A* a pravdepodobnosti *j*-teho variantu znaku *B*, t.j.

$$P(A = a_i \wedge B = b_j) = P(A = a_i).P(B = b_j).$$

Pravdepodobnosti $P(A = a_i)$ a $P(B = b_j)$ odhadneme pomocou relatívnych početností $P(A = a_i) = \frac{(a_i)}{n}$, $P(B = b_j) = \frac{(b_j)}{n}$, kde *n* - je rozsah výberového súboru,

(a_i) - je počet jednotiek, ktoré majú *i*-ty variant znaku *A* (súčet *i*-teho riadku),

(b_j) - je počet jednotiek, ktoré majú *j*-ty variant znaku *B* (súčet *j*-teho stĺpca).

Preto $$P(A = a_i \wedge B = b_j) = \frac{(a_i)}{n} . \frac{(b_j)}{n}.$$

Teoretické absolútne početnosti potom odhadneme:

$$e_{ij} = n.\frac{(a_i)}{n}.\frac{(b_j)}{n} = \frac{(a_i).(b_j)}{n} \quad . \tag{9.1}$$

Z týchto teoretických početností zostavíme „teoretickú tabuľku“ (Tab. 9.2) podobnú kontingenčnej tabuľke a použijeme ju na výpočet hodnoty testovacej štatistky.

Tab. 9.2 Tabuľka teoretických početností

	b_1	b_2	b_3	$\sum$
a_1	e_{11}	e_{12}	e_{13}	(a_1)
a_2	e_{21}	e_{22}	e_{23}	(a_2)
a_3	e_{31}	e_{32}	e_{33}	(a_3)
a_4	e_{41}	e_{42}	e_{43}	(a_4)
$\sum$	(b_1)	(b_2)	(b_3)	n

Na overenie nulovej hypotézy sa používa testovacia štatistika

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad , \tag{9.2}$$

ktorú poznáme aj z vyšetrovania normality rozdelenia. Testovacia štatistika (9.2) má χ^2 - rozdelenie s $(r-1).(s-1)$ stupňami voľnosti. Ak sa teoretické početnosti, ktoré predpokladajú nezávislosť, „veľmi“ odlišujú od pozorovaných početností, hodnota testovacej štatistiky bude veľká a nulovú hypotézu treba zamietnuť. Ak rozdiely medzi nimi sú „malé“ (štatisticky nevýznamné), môžeme ich vysvetliť náhodným vplyvom, nemáme dôvod zamietnuť nulovú hypotézu. Rovnako ako pri testoch zhody porovnáme hodnotu testovacej štatistiky χ^2 s $(1-\alpha).100$ percentným kvantilom χ^2 - rozdelenia s $(r-1).(s-1)$ stupňami voľnosti, alebo podľa hodnoty testovacej štatistiky určíme p - hodnotu a porovnáme ju so zvolenou hladinou významnosti.

Záver testu: Ak $\chi^2 > \chi^2_{1-\alpha}$, zamietame H_0 na zvolenej hladine významnosti α (Tab. 9.3).

Tab. 9.3 χ^2 - test nezávislosti

Hypotézy	**Rozdelenie**	**Testovacia štatistika**	**Oblasť zamietnutia H_0**
H_0 : Medzi *A,B* nie je závislosť. H_1 : Medzi *A, B* je závislosť.	Chí-kvadrát	$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$	$\chi^2 > \chi^2_{1-\alpha}$ $d.f. = (r-1).(s-1)$

Chí-kvadrát test nezávislosti sa dá použiť, ak pre všetky políčka „teoretickej tabuľky“ platí $e_{ij} \geq 5$. Ak to tak nie je, zlúčime riadok (stĺpec) tabuľky s touto nízkou početnosťou so susedným riadkom (stĺpcom) tak, aby vyhovoval podmienke. Samozrejme tým znížime počet stupňov voľnosti testovacej štatistiky. To je jediné obmedzenie použitia tohto testu.

Príklad 9.1

Na výberovom súbore s rozsahom 750 ľudí sa skúmal vzťah medzi zaradením do sociálnej skupiny (znak *A* - riadok) a typom bývania (znak *B* - stĺpec) . Sociálna skupina má 4 varianty: a_1 - zamestnanec štátnej alebo verejnej služby, a_2 - ostatní zamestnanci, a_3 - podnikateľ, a_4 - dôchodca. Typ bývania má 5 variantov: b_1 - vlastný dom, b_2 - vlastný byt, b_3 - nájom u súkromnej osoby, b_4 - družstevný byt, b_5 - služobný byt (firemný, vojenský). Na hladine významnosti $\alpha = 0{,}05$ zistite, či možno hovoriť o existencii závislosti medzi týmito znakmi. Empirické údaje sú uvedené bez zátvoriek v nasledujúcej kontingenčnej tabuľke .

Tab. 9.4 Kontingenčná tabuľka a teoretické početnosti k Príkladu 9.1

	b_1	b_2	b_3	b_4	b_5	$\sum$
a_1	40 (44,8)	40(28)	60(44,8)	0(20,5)	0(01,9)	140
a_2	20(92,8)	90(58)	110(92,8)	60(42,5)	10(3,9)	290
a_3	140(48)	10(30)	0(48)	0(22)	0(2)	150
a_4	40(54,4)	10(34)	70(54,4)	50(24,9)	0(2,7)	170
$\sum$	240	150	240	110	10	750

Pomocou vzťahu (9.1) a tejto tabuľky, v ktorej sú empirické početnosti n_{ij} postupne vypočítame teoretické početnosti e_{ij}, ktoré sú uvedené v okrúhlych zátvorkách.

Napr. $e_{11} = 750.\frac{140}{750}.\frac{240}{750} = 44{,}8$, $e_{12} = 750.\frac{140}{750}.\frac{150}{750} = 28$ atď.

Teoretické početnosti v poslednom stĺpci sú menšie ako 5, preto zlúčime štvrtý a piaty stĺpec do jedného. V hranatých zátvorkách Tab. 9.5 sú pre všetky *i, j.* vypočítané hodnoty $\frac{(n_{ij} - e_{ij})^2}{e_{ij}}$. Hodnotu testovacej štatistiky vypočítame ako súčet hodnôt v hranatých zátvorkách $\chi^2 = 429{,}176$. Je to veľké číslo, ktoré hovorí v neprospech myšlienky, že znaky sú nezávislé.

Tab. 9.5 Pokračovanie Príkladu 9.1

	b_1	b_2	b_3	b_4+b_5	$\sum$
a_1	40 (44,8) [0,51]	40 (28) [5,14]	60 (44,8) [5,16]	0 (22,4) [22,4]	140
a_2	20 (92,8) [57,11]	90 (58) [17,66]	110 (92,8) [3,19]	70 (46,4) [12]	290
a_3	140 (48) [176,3]	10 (30) [13,33]	0 (48) [48]	0 (24) [24]	150
a_4	40 (54,4) [3,81]	10 (34) [16,94]	70 (54,4) [4,47]	50 (27,2) [19,11]	170
$\sum$	240	150	240	120	750

Presne: V tabuľkách nájdeme 95%-ný kvantil χ^2- rozdelenia s $(4-1).(4-1)=9$ stupňami voľnosti t.j. $\chi^2_{0,95}=16,92$. Pretože 429,176 >16,92 zamietame nulovú hypotézu a prijímame alternatívnu, čo znamená, že medzi sociálnou skupinou a typom bývania ľudí je štatisticky významná závislosť.

Iná možnosť je využiť p-hodnotu. K vypočítanej testovacej štatistike $\chi^2=429,176$ určíme p-hodnotu p $=8,085.10^{-87}$ (v EXCELi funkcia CHI-DIST(429,176; 9)). Táto skoro nulová hodnota je menšia ako všetky bežné hladiny významnosti, preto zamietame *H*o a prijímame alternatívnu hypotézu, t.j. znaky sú závislé.

EXCEL

Test nezávislosti má EXCEL v ponuke svojich testov, po voľbe ***Prilepiť funkciu/ štatistické / CHITEST***. Daný príklad vyriešime jednoducho po zadaní polí empirických a teoretických početností (tab.9.6), ale pole teoretických početností si musíme vytvoriť sami. Výstupom je jediný údaj, p - hodnota vypočítaná k hodnote testovacej štatistiky .

Bližším skúmaním jednotlivých sčítancov tvoriacich testovaciu štatistiku vidíme, že najviac do nej prispievajú podnikatelia bývajúci vo vlastných domoch (n_{31}) a ostatní zamestnanci bývajúci vo vlastných domoch (n_{21}).

Tab. 9.6 Tvar tabuliek pred použitím CHITESTu

empirické početnosti

	1	2	3	4	
1	40	40	60	0	140
2	20	90	110	70	290
3	140	10	0	0	150
4	40	10	70	50	170
	240	150	240	120	750

teoretické početnosti

	1	2	3	4
1	44,8	28	44,8	22,4
2	92,8	58	92,8	46,4
3	48	30	48	24
4	54,4	34	54,4	27,2

Pre **alternatívne znaky (dichotomické)**, dostaneme špeciálny prípad kontingenčnej tabuľky. Výsledkom triedenia je štvorpolíčková tabuľka typu 2x2, ktorá sa volá **asociačná tabuľka**.

Tab. 9.7 Asociačná tabuľka

	b_1	b_2	$\sum$
a_1	n_{11}	n_{12}	(a_1)
a_2	n_{21}	n_{22}	(a_2)
$\sum$	(b_1)	(b_2)	n

Príklad 9.2

Na základe výskumu u 81 osôb je zostavená asociačná tabuľka, v ktorej sú osoby roztriedené podľa pohlavia a podľa toho, či majú vysokoškolské vzdelanie. Na 5% hladine významnosti overte, či tieto znaky sú závislé.

Tab. 9.8 Početnosti k Príkladu 9.2

empirické početnosti

	muži	ženy	
má VŠ	12	10	22
nemá VŠ	26	33	59
	38	43	81

teoretické početnosti

	muži	ženy	
má VŠ	10,32	11,68	22
nemá VŠ	27,68	31,32	59
	38	43	81

Tabuľku teoretických početností sme doplnili podľa vzťahu (9.1). Využijeme voľbu ***Prilepiť funkciu/ štatistické / CHITEST*** a po zadaní polí empirických a teoretických početností z Tab. 9.8 dostaneme príslušnú p-hodnotu $p = 0{,}401$. Pretože $0{,}05 < 0{,}401$, nemáme dôvod zamietnuť nulovú hypotézu o nezávislosti pohlavia a absolvovania vysokoškolského štúdia.

9.2 Miery intenzity závislosti dvoch kvalitatívnych znakov

Ak pomocou testu nezávislosti zistíme asociáciu medzi znakmi, môžeme sa rovnako ako pri kvantitatívnych znakoch pýtať na silu tejto asociácie. V rôznych štatistických programoch existuje množstvo koeficientov na jej vyjadrenie, my spomenieme len niektoré. Treba zdôrazniť, že sú to koeficienty vhodné na skúmanie sily závislosti kvalitatívnych znakov, nie sú vhodné pre kvantitatívne a ordinálne znaky.

Pearsonov kontingenčný koeficient

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad , \qquad C \in \langle 0,1)$$

- Hodnoty blízke jednotke znamenajú silnejšiu asociáciu medzi znakmi.
- Hodnoty bližšie nule znamenajú slabšiu závislosť medzi znakmi.
- Je závislý od spôsobu triedenia.
- V štvorcovej tabuľke $k \times k$ je $C_{\max} = \sqrt{\frac{k-1}{k}}$.

Špeciálne pre alternatívne znaky A, B uvedieme dva koeficienty:

Koeficient Phí :

$$\phi = \sqrt{\frac{\chi^2}{n}} = \frac{|n_{21}n_{12} - n_{11}n_{22}|}{\sqrt{(a_1)(a_2)(b_1)(b_2)}} \quad , \qquad \phi \geq 0$$

- Je najvhodnejší z koeficientov, ktoré spomíname.
- Často sa kontingenčná tabuľka upraví pozlučovaním vhodných riadkov i stĺpcov až na štvorpolíčkovú tabuľku, pre ktorú sa určí ϕ .

Koeficient korelácie kvalitatívnych znakov

$$R = \frac{nn_{11} - (a_1)(b_1)}{\sqrt{(a_1)(a_2)(b_1)(b_2)}}, \quad R \in \langle -1,1 \rangle$$

- Hodnoty v absolútnej hodnote blízke jednotke znamenajú silnú závislosť, nulová hodnota znamená nezávislosť.

- Pre ordinálne znaky alebo pre rôzne kombinácie kvalitatívnych, ordinálnych alebo kvantitatívnych znakov sú vypracované špeciálne koeficienty na meranie intenzity závislosti, ale nebudeme sa im venovať.

9.3 Poradová korelácia

Výsledky prieskumov sú často prezentované rôznymi poradiami (napr. poradie žiakov v triede podľa sympatií, poradie skupiny osôb podľa vhodnosti na nejaké povolanie, poradie predmetov podľa nevyhnutnosti na prežitie v krutej zime atď.), pričom vzniká potreba porovnať mieru zhody (odlišnosti) poradí zostavených rôznymi respondentmi. Stupeň závislosti medzi dvomi poradiami *X, Y* vyjadríme Spearmanovým koeficientom poradovej korelácie

$$R = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)}, \qquad -1 \le R \le 1 \tag{9.3}$$

kde n - rozsah súboru ,

$d_i = \left(x_i - y_i\right)$- rozdiel dvojice poradí .

- Jeho interpretácia je rovnaká ako pre korelačný koeficient.
- Niekedy sa na poradia pretransformujú metrické alebo ordinárne údaje, a to hlavne vtedy, ak treba vypočítať koreláciu medzi rôzne škálovanými premennými.

Príklad 9.3

Do finále Miss 2003 sa dostalo 10 dievčat. Pred samotnou súťažou bývalá Miss 2002 zostavila očakávané výsledné poradie súťažiacich takto: K, L, M, H, T, R, E, J, B, S a riaditeľ reklamnej agentúry takto: T, H, L, E, B, K, M, R, J, S. (Dievčatá sú označené začiatočným písmenom svojho mena).

1. Vypočítajte stupeň zhody názorov týchto dvoch posudzujúcich.
2. Súťažiace získali vo finále počty bodov dané tabuľkou. Porovnajte správnosť prognóz Miss 2002 i riaditeľa.

K	L	M	T	H	R	B	E	J	S
48	36	36	32	32	31	29	29	29	11

Obe úlohy sme vyriešili naraz s využitím EXCELu. Pri riešení druhej úlohy sme zo získaných bodov vyrobili poradie súťažiacich. Pri rovnosti získaných bodov sme vytvorili priemerné poradie, napr. 29 bodov získali 3 súťažiace, ktoré obsadili 7.,8., 9.miesto, preto im priradíme poradie (7+8+9)/3 = 8. Transformovaním bodov na poradia strácame informácie, lebo dve susedné poradia neodrážajú skutočný rozdiel v získaných bodoch. Napríklad medzi posledným a predposledným poradím je rozdiel až 18 bodov, no medzi susednými poradiami 4,5 a 6 je rozdiel len jeden bod.

Tab. 9.9 Výpočet poradového korelačného koeficientu

	A	*B*	*C*	mocnina rozdielu poradí		
meno	poradie Miss	poradie riaditeľ	poradie skutočné	$(A-B)^2$	$(B-C)^2$	$(A-C)^2$
T	5	1	4,5	16	12,25	0,25
H	4	2	4,5	4	6,25	0,25
L	2	3	2,5	1	0,25	0,25
E	7	4	8	9	16	1
B	9	5	8	16	9	1
K	1	6	1	25	25	0
M	3	7	2,5	16	20,25	0,25
R	6	8	6	4	4	0
J	8	9	8	1	1	0
S	10	10	10	0	0	0
			súčet stĺpca	92	94	3

Na zistenie stupňa zhody poradí použijeme Spaermanov poradový korelačný koeficient (9.3) $R_1 = \frac{6,92}{10 \cdot 99} = 0,4424$, čo vyjadruje slabú zhodu hodnotení Miss 2002 a riaditeľa. $R_2 = 0,4303$, čo vyjadruje slabú zhodu hodnotení riaditeľa a skutočným poradím. Nakoniec, $R_3 = 0,9818$, čo je vysoká zhoda medzi hodnotením, ktoré urobila Miss 2002 a skutočným poradím po súťaži.

9.4 Použitie asociačných a kontingenčných tabuliek v iných testoch

a) testovanie rozdielu početností dvoch nezávislých súborov

Alternatívny znak *A* rozdelí prvky vo výberových nezávislých súboroch vždy na dve skupiny, podľa toho, či prvok daný znak *A* má alebo nemá. Pýtame sa, či rozdiely v početnostiach skupín so znakom *A* pre prvý a druhý súbor sú náhodné, alebo štatisticky významné? Postup ilustrujeme na údajoch z Príkladu 9.2.

Príklad 9.4

Z 38 opýtaných mužov má 12 vysokoškolské vzdelanie, zo 43 opýtaných žien má 10 vysokoškolské vzdelanie. Zistite, či medzi mužmi a ženami je rovnaký počet osôb s vysokoškolským vzdelaním (VŠ). Nulovú a alternatívnu hypotézu môžeme formulovať takto:

H_0: Podiel osôb s VŠ je v oboch súboroch rovnaký.

H_1: Podiel osôb s VŠ nie je v oboch súboroch rovnaký.

alebo:

H_0: Rozdelenia v oboch súboroch sú rovnaké.

H_1: Rozdelenia v oboch súboroch nie sú rovnaké.

Údaje spracujeme do empirickej asociačnej Tab. 9.10 , ktorá je rovnaká ako Tab. 9.8.

Tab. 9.10 Tabuľka početností k Príkladu 9.4

empirické početnosti

	muži	ženy	
má VŠ	12	10	22
nemá VŠ	26	33	59
	38	43	81

teoretické početnosti

	muži	ženy	
má VŠ	10,32	11,68	22
nemá VŠ	27,68	31,32	59
	38	43	81

Tabuľku teoretických početností vyplníme nasledujúcim spôsobom:

Za predpokladu platnosti H_0 (t.j. pravdepodobnosť, že žena má VŠ, je rovnaká, ako pravdepodobnosť, že muž má VŠ) odhadneme pravdepodobnosť, že niekto má VŠ relatívnou početnosťou $p = \frac{22}{81}$ a nemá VŠ $q = \frac{59}{81}$. Absolútne očakávané početnosti dostaneme vynásobením relatívnej početnosti rozsahom súboru:

počet mužov s VŠ je $e_{11} = \frac{22}{81} \cdot 38 = 10{,}32$,

počet mužov bez VŠ je $e_{21} = \frac{59}{81} \cdot 38 = 27{,}68$,

počet žien s VŠ je $e_{12} = \frac{22}{81} \cdot 43 = 11{,}68$,

počet žien bez VŠ je $e_{22} = \frac{59}{81} \cdot 43 = 31{,}32$.

Vidíme, dostávame tie isté tabuľky ako pri teste nezávislosti v Príklade 9.2.

Pozorované a teoretické hodnoty porovnáme pomocou χ^2 - testu, p-hodnota je 0,401, z čoho vyplýva, že napr. na hladine významnosti $\alpha = 0{,}1$ aj $\alpha = 0{,}05$ nemáme dôvod zamietnuť nulovú hypotézu o tom, že podiel mužov i žien s VŠ je rovnaký (to isté potvrdil aj test nezávislosti).

Poznámka 9.1

Tá istá úvaha sa dá použiť aj pre znak, ktorý má viac ako 2 varianty. Tento test sa dá použiť aj pre kvantitatívne údaje, ak nie sú splnené predpoklady pre t-test a údaje sú roztriedené do tried.

b) testovanie rozdielu početností dvoch závislých súborov (párové hodnoty)

Alternatívny znak *A* rozdelí prvky vo výberových závislých súboroch tiež na dve skupiny, podľa toho, či prvok daný znak *A* má (1) alebo nemá (0). Na rozdiel od bodu a) máme teda na každej štatistickej jednotke danú usporiadanú dvojicu typu $(1,1),(0,0),(1,0),(0,1)$, podľa toho, či mala alebo nemala znak pri dvoch zisťovaniach. Pýtame sa, či rozdiely v početnostiach skupín so znakom *A* pre prvý a druhý súbor sú náhodné, alebo štatisticky významné? Postup ilustrujeme na príklade.

Príklad 9.5

Na skupine 140 pacientov sa zisťovala účinnosť 2 liekov na ich chorobu. Výsledky pre prvý a druhý liek sú spracované v asociačnej Tabuľke 9.11. Vyšetrite na hladine významnosti $\alpha = 0{,}05$, či oba lieky sú rovnako účinné. Nulovú a alternatívnu hypotézu môžeme sformulovať takto:

H_0: Lieky sú rovnako účinné.

H_1: Lieky nie sú rovnako účinné.

O porovnaní účinnosti nám nedávajú žiadnu informáciu usporiadané dvojice (1,1) a (0,0), teda tie, kde nedošlo k zmene $n_{11} = 30$ a $n_{22} = 48$ z asociačnej tabuľky. Výpovednú hodnotu majú len usporiadané dvojice (1,0) a (0,1), teda tie, kde došlo k zmene. To sú údaje z asociačnej tabuľky $n_{12} = 36$ a $n_{21} = 26$. Ak ich účinnosť je rovnaká, tak teoreticky predpokladáme, že počet dvojíc (1,0) a (0,1) je rovnaký. Preto $e_{12} = e_{21} = \frac{n_{12} + n_{21}}{2}$. V našom prípade je to 31. Takto vyplníme tabuľku teoretických početností a na porovnanie oboch tabuliek použijeme χ^2 - test.

Výsledkom CHITESTu v EXCELi je p-hodnota $p = 0{,}204$, čo znamená, že na danej hladine významnosti prijímame nulovú hypotézu o rovnakej účinnosti oboch liekov.

Tab. 9.11 Asociačná tabuľka k Príkladu 9.5

empirické početnosti

		1. liek		
2. liek		účinný	neú-činný	spolu
	účinný	30	36	66
	neúčin-ný	26	48	74
		56	84	140

teoretické početnosti

		1. liek		
2. liek		účinný	neú-činný	spolu
	účinný	30	31	61
	nečin-ný	31	48	79
		61	79	140

Príklady na precvičenie

9.6 Vyšetrite závislosť výsledkov psychotestov od typu absolvovanej strednej školy pre súbor, ktorý dostanete zlúčením súborov PRIJÍMACIE SKÚŠKY/1.fakulta a PRIJÍMACIE SKÚŠKY/2.fakulta.

9.7 Vyšetrite závislosť úspešnosti na prijímacej skúške z matematiky od typu absolvovanej strednej školy pre súbor, ktorý dostanete zlúčením súborov PRIJÍMACIE SKÚKY/1.fakulta a PRIJÍMACIE SKÚŠKY/2.fakulta. Študent na skúške uspel, ak získal aspoň 12 bodov.

9.8 Navrhnite, ako by ste si pripravili, vykonali a vyhodnotili test
a) účinnosti jedného lieku proti vysokej hladine cholesterolu
b) účinnosti jedného lieku proti bolesti hlavy
c) účinnosti vakcíny proti chrípke.

TABUĽKOVÁ PRÍLOHA

Tab. 1 Koeficienty Shapiroovho - Wilkovho testu

i \ n	7	8	9	10	11	12	13	14
1	0,6233	0,6052	0,5888	0,5739	0,5601	0,5475	0,5359	0,5251
2	0,3031	0,3164	0,3244	0,3291	0,3315	0,3325	0,3325	0,3318
3	0,1401	0,1734	0,1976	0,2141	0,2260	0,2347	0,2412	0,2460
4	0	0,0561	0,0947	0,1224	0,1429	0,1586	0,1707	0,1802
5	0	0	0	0,0399	0,0695	0,0922	0,1099	0,1240
6	0	0	0	0	0	0,0303	0,0539	0,0727
7	0	0	0	0	0	0	0	0,0240

i \ n	15	16	17	18	19	20	21	22
1	0,5150	0,5056	0,4968	0,4886	0,4808	0,4734	0,4643	0,4590
2	0,3306	0,3290	0,3273	0,3253	0,3232	0,3211	0,3185	0,3156
3	0,2495	0,2521	0,2540	0,2553	0,2561	0,2565	0,2758	0,2571
4	0,1878	0,1939	0,1988	0,2027	0,2059	0,2085	0,2119	0,2131
5	0,1353	0,1447	0,1524	0,1587	0,1641	0,1686	0,1736	0,1764
6	0,0880	0,1005	0,1109	0,1197	0,1271	0,1334	0,1399	0,1443
7	0,0433	0,0593	0,0725	0,0837	0,0932	0,1013	0,1092	0,1150
8	0	0,0196	0,0359	0,0496	0,0612	0,0711	0,0804	0,0878
9	0	0	0	0,0163	0,0303	0,0422	0,0530	0,0618
10	0	0	0	0	0	0,0140	0,0263	0,0368
11	0	0	0	0	0	0	0	0,0122

i \ n	23	24	25	26	27	28	29	30
1	0,4542	0,4493	0,4450	0,4407	0,4366	0,4328	0,4291	0,4254
2	0,3126	0,3098	0,3069	0,3043	0,3018	0,2992	0,2968	0,2944
3	0,2563	0,2554	0,2543	0,2533	0,2522	0,2510	0,2499	0,2487
4	0,2139	0,2145	0,2148	0,2151	0,2152	0,2151	0,2150	0,2148
5	0,1787	0,1807	0,1822	0,1836	0,1848	0,1857	0,1864	0,1870
6	0,1480	0,1512	0,1539	0,1563	0,1584	0,1601	0,1616	0,1630
7	0,1201	0,1245	0,1283	0,1316	0,1346	0,1372	0,1395	0,1415
8	0,0941	0,0997	0,1046	0,1089	0,1128	0,1162	0,1192	0,1219
9	0,0696	0,0764	0,0823	0,0876	0,0923	0,0965	0,1002	0,1036
10	0,0459	0,0539	0,0610	0,0672	0,0728	0,0778	0,0822	0,0962
11	0,0228	0,0321	0,0403	0,0476	0,0540	0,0598	0,0650	0,0697
12	0	0,0107	0,0200	0,0284	0,0358	0,0424	0,0483	0,0537
13	0	0	0	0,0094	0,0178	0,0253	0,0320	0,0381
14	0	0	0	0	0	0,0084	0,0159	0,0227
15	0	0	0	0	0	0	0	0,0076

Tab. 2 Kvantily Shapiro – Wilkovej náhodnej premennej *W*

n \ α	0,01	0,05
7	0,730	0,803
8	0,749	0,818
9	0,764	0,829
10	0,781	0,842
11	0,792	0,850
12	0,805	0,859
13	0,814	0,866
14	0,825	0,874
15	0,835	0,881
16	0,844	0,887
17	0,851	0,892
18	0,858	0,897

n \ α	0,01	0,05
19	0,863	0,901
20	0,868	0,905
21	0,873	0,908
22	0,878	0,911
23	0,881	0,914
24	0,884	0,916
25	0,888	0,918
26	0,891	0,920
27	0,894	0,923
28	0,986	0,924
29	0,898	0,926
30	0,900	0,927

Tab. 3 Kvantily D´Agostiniovej náhodnej premennej *D*

n \ α	0,005	0,025	0,975	0,995
30	-4,19	-2,91	0,830	0,941
32	-4,16	-2,88	0,862	0,983
34	-4,12	-2,86	0,891	1,02
36	-4,09	-2,85	0,917	1,05
38	-4,06	-2,83	0,941	1,08
40	-4,03	-2,81	0,964	1,11
42	-4,00	-2,80	0,986	1,14
44	-3,98	-2,78	1,01	1,17
46	-3,95	-2,77	1,02	1,19
48	-3,93	-2,75	1,04	1,22
50	-3,91	-2,74	1,06	1,24
60	-3,81	-2,68	1,13	1,34
70	-3,73	-2,64	1,19	1,42
80	-3,67	-2,60	1,24	1,48
90	-3,61	-2,57	1,28	1,54
100	-3,57	-2,54	1,31	1,59

Tab. 4 Kvantily $U(n_1, n_2, 0{,}01)$ pre Mann–Whitneyov test pre nezávislé výbery

n_1	n_2																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
4	-	-	-																
5	-	-	-	0															
6	-	-	0	1	2														
7	-	-	0	1	3	4													
8	-	-	1	2	4	6	7												
9	-	0	1	3	5	7	9	11											
10	-	0	2	4	6	9	11	13	16										
11	-	0	2	5	7	10	13	16	18	21									
12	-	1	3	6	9	12	15	18	21	24	27								
13	-	1	3	7	10	13	17	20	24	27	31	34							
14	-	1	4	7	11	15	18	22	26	30	34	38	42						
15	-	2	5	8	12	16	20	24	29	33	37	42	46	51					
16	-	2	5	9	13	18	22	27	31	36	41	45	50	55	60				
17	-	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70			
18	-	2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81		
19	0	3	7	12	17	22	28	33	39	45	51	57	63	69	74	81	87	93	
20	0	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	105
21	0	3	8	14	19	25	32	38	44	51	58	64	71	78	84	91	98	105	112
22	0	4	9	14	21	27	34	40	47	54	61	68	75	82	89	96	104	111	118
23	0	4	9	15	22	29	35	43	50	57	64	72	79	87	94	102	109	117	125
24	0	4	10	16	23	30	37	45	52	60	68	75	83	91	99	107	115	123	131
25	0	5	10	17	24	32	39	47	55	63	71	79	87	96	104	112	121	129	138
26	0	5	11	18	25	33	41	49	58	66	74	83	92	100	109	118	127	135	144
27	1	5	12	19	27	35	43	52	60	69	78	87	96	105	114	123	132	142	151
28	1	5	12	20	28	36	45	54	63	72	81	91	100	109	119	128	138	148	157
29	1	6	13	21	29	38	47	56	66	75	85	94	104	114	124	134	144	154	164
30	1	6	13	22	30	40	49	58	68	78	88	98	108	119	129	139	150	160	170

Tab. 5 Kvantily $U(n_1, n_2, 0{,}05)$ pre Mann–Whitneyov test pre nezávislé výbery

n_1	n_2																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
4	-	-	0																
5	-	0	1	2															
6	-	1	2	3	5														
7	-	1	3	5	6	8													
8	0	2	4	6	8	10	13												
9	0	2	4	7	10	12	15	17											
10	0	3	5	8	11	14	17	20	23										
11	0	3	6	9	13	16	19	23	26	30									
12	1	4	7	11	14	18	22	26	29	33	37								
13	1	4	8	12	16	20	24	28	33	37	41	45							
14	1	5	9	13	17	22	26	31	36	40	45	50	55						
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64					
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75				
17	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87			
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99		
19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	
20	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
21	3	8	15	22	29	36	43	50	58	65	73	80	88	96	103	111	119	126	134
22	3	9	16	23	30	38	45	53	61	69	77	85	93	101	109	117	125	133	141
23	3	9	17	24	32	40	48	56	64	73	81	89	98	106	115	123	132	140	149
24	3	10	17	25	33	42	50	59	67	76	85	94	102	111	120	129	138	147	156
25	3	10	18	27	35	44	53	62	71	80	89	98	107	117	126	135	145	154	163
26	4	11	19	28	37	46	55	64	74	83	93	102	112	122	132	141	151	161	171
27	4	11	20	29	38	48	57	67	77	87	97	107	117	127	137	147	158	168	178
28	4	12	21	30	40	50	60	70	80	90	101	111	122	132	143	154	164	175	186
29	4	13	22	32	42	52	62	73	83	94	105	116	127	138	149	160	171	182	193
30	5	13	23	33	43	54	65	76	87	98	109	120	131	143	154	166	177	189	200

Tab. 6 Kvantily T_W pre Wilcoxonov test pre závislé výbery

n	$T_{W\ 0,005}$	$T_{W\ 0,01}$	$T_{W\ 0,025}$	$T_{W\ 0,05}$
4	-	-	-	-
5	-	-	-	0
6	-	-	0	2
7	-	0	2	3
8	0	1	3	5
9	1	3	5	8
10	3	5	8	10
11	5	7	10	13
12	7	9	13	17
13	9	12	17	21
14	12	15	21	25
15	15	19	25	30
16	19	23	29	35
17	23	27	34	41
18	27	32	40	47
19	32	37	46	53
20	37	43	52	60
21	42	49	58	67
22	48	55	65	75
23	54	62	73	83
24	61	69	81	91
25	68	76	89	100
26	75	84	98	110
27	83	92	107	119
28	91	101	116	130
29	100	110	126	140
30	109	120	137	151

Tab. 7 Súbor PRIJÍMACIE SKÚŠKY

1. fakulta								
hodnotenie z prijímacej skúšky					hodnotenie 2. sem.		hodnotenie 3. sem.	
por. číslo	stred. škola	mat.	fyzika	psych.	mat.(upr)	prie-mer	mat.(upr)	prie-mer
1	G	25,0	22,0	1	1,0	1,3	1,5	1,5
2	P	14,0	12,0	3	2,0	2,1	2,0	2,5
3	U	1,5	1,0	2	3,5	3,0	4,0	3,5
4	G	21,0	25,5	2	3,0	2,3	3,0	2,5
5	P	4,0	0,0	3	3,5	3,0	4,0	3,5
6	P	16,0	20,0	3	3,5	3,0	3,0	3,0
7	P	6,0	10,0	3	4,0	3,0		
8	P	15,0	26,0	3	3,0	2,9	2,5	3,0
9	U	2,5	4,5	3	4,0	3,0		
10	P	11,5	11,0	2	2,0	2,1	2,0	2,5
11	OA	12,5	7,0	3	3,2	2,3	3,0	2,3
12	OA	8,5	7,0	3	3,0	2,0	3,5	2,5
13	G	22,0	15,0	1	2,0	2,1	2,5	2,0
14	U	3,0	5,5	3	3,0	2,6	2,5	2,0
15	P	2,5	9,0	3	4,0	3,0		
16	OA	6,5	0,0	3	3,0	3,0	4,0	3,5
17	P	10,5	3,0	3	3,5	3,0	3,0	2,8
18	G	16,5	9,0	1	1,0	1,3	1,5	1,3
19	G	15,0	3,5	2	2,0	2,0	2,5	2,2
20	P	15,0	19,0	1	2,0	1,8	2,2	2,0
21	P	19,5	12,0	1	1,0	1,5	1,2	1,2
22	P	12,0	15,0	2	2,0	2,3	1,5	2,0
23	G	4,0	16,0	3	3,2	3,0	3,0	3,2
24	G	8,0	18,0	3	3,2	2,4	3,2	3,0
25	P	18,5	14,0	2	1,0	1,8	1,5	2,5
26	P	8,5	16,5	3	3,0	2,0	3,0	2,6
27	U	5,5	12,0	3	3,0	2,6	3,2	2,5
28	U	5,0	7,0	3	4,0	3,0		
29	U	0,0	2,5	3	4,0	3,0		
30	P	3,0	13,0	3	3,2	2,9	3,5	3,2
31	P	9,0	8,0	3	3,0	1,8	3,0	2,4
32	P	1,0	10,0	3	3,0	2,8	3,2	2,6
33	G	7,0	19,0	3	3,2	2,9	3,2	3,0
34	P	6,0	9,0	3	3,0	2,3	2,5	3,0
35	U	7,3	5,0	1	3,0	2,3	2,2	2,0
36	U	4,0	0,0	3	2,0	2,4	2,0	1,8
37	U	8,0	8,0	3	3,2	2,8	3,5	2,2
38	P	0,5	9,0	3	4,0	3,0		

Súbor PRIJÍMACIE SKÚŠKY (pokračovanie)								
2. fakulta								
hodnotenie z prijímacej skúšky					**hodnotenie 2. sem.**		**hodnotenie 3. sem.**	
por. číslo	**stred.škola**	**mat.**	**fyzika**	**psych.**	**mat.(upr)**	**prie-mer**	**mat.(upr)**	**prie-mer**
1	P	17,0	12,5	3	3,2	2,5	3,0	2,2
2	P	20,0	15,0	3	3,0	1,9	2,5	2,5
3	P	9,0	12,0	3	3,5	2,9	3,0	2,2
4	G	30,0	22,5	1	3,5	2,5	3,0	2,0
5	P	12,0	15,0	3	3,0	2,1	3,2	2,0
6	G	15,0	25,5	2	3,0	1,2	3,0	2,5
7	P	20,0	10,0	3	2,0	1,4	2,5	1,8
8	P	11,0	20,0	3	2,0	1,6	2,0	1,8
9	P	20,0	10,0	3	3,0	2,1	2,5	2,5
10	P	6,0	0,0	2	2,0	1,8	2,5	2,0
11	P	14,0	4,5	3	3,2	1,8	3,0	2,8
12	G	20,5	11,0	3	3,2	2,0	3,0	2,8
13	P	17,0	25,0	2	3,0	1,6	3,5	3,0
14	G	18,0	12,0	3	3,0	1,8	3,0	2,8
15	P	14,0	15,0	3	3,2	1,9	3,0	2,5
16	P	16,5	18,0	3	3,0	1,6	2,5	2,0
17	U	9,0	3,0	3	3,2	2,1	2,5	2,2
18	G	26,0	21,0	1	1,0	1,4	1,5	1,2
19	G	18,5	7,0	3	2,0	1,3	2,0	1,5
20	G	25,0	21,0	1	1,0	1,5	1,0	1,2
21	U	6,0	7,0	3	3,0	1,9	3,0	2,5
22	U	11,0	2,0	2	3,2	1,6	3,0	2,5
23	G	24,5	19,5	1	1,0	1,3	1,2	1,2
24	P	14,0	18,0	1	2,0	1,3	1,5	1,5
25	P	9,5	10,0	3	2,0	1,9	2,0	1,6
26	P	6,0	8,5	3	3,2	2,1	3,0	2,5
27	P	16,0	11,0	3	3,2	2,3	3,0	2,4
28	P	13,0	11,5	3	3,0	1,8	3,2	3,0
29	G	15,0	11,5	3	3,2	2,1	3,2	3,0
30	P	9,0	5,0	3	3,2	2,0	3,0	2,8
31	G	13,0	10,0	3	3,0	1,2	3,0	2,8
32	P	10,0	8,0	2	3,0	2,0	2,5	2,2
33	P	12,5	15,5	3	3,2	1,8	2,5	2,2
34	U	11,5	13,0	2	3,2	1,9	3,0	2,5
35	G	16,0	20,0	1	2,0	1,9	3,0	2,4
36	G	19,5	10,0	2	2,0	1,6	3,0	2,4
37	G	3,5	1,5	2	3,5	2,5	2,5	2,0
38	P	19,0	22,0	1	2,0	2,0	2,2	2,2

pokračovanie								
39	P	12,0	15,5	2	2,0	2,3	2,2	2,0
40	P	15,0	10,0	2	3,0	2,4	3,2	2,8
41	P	13,0	6,0	3	2,0	1,8	2,0	1,5
42	P	18,0	23,0	1	1,0	1,6	1,2	1,2
43	P	9,0	19,0	3	3,2	2,4	3,0	2,5
44	G	16,5	12,5	3	2,0	1,9	2,5	2,1
45	P	11,5	15,0	2	2,0	1,9	2,2	2,0
46	G	9,0	12,0	3	3,5	2,0	3,5	2,8
47	P	22,5	12,5	1	1,0	2,0	2,0	1,5
48	P	13,0	1,0	2	2,0	1,8	2,0	1,6
49	G	28,0	25,5	1	1,0	1,2	2,0	1,6
50	P	21,0	14,0	3	2,5	1,8	3,0	2,5

LEGENDA

G - gymnázium, **P -** stredná odborná škola, **U** - stredné odborné učilisko
mat. - počet bodov z prijímacej skúšky z matematiky
fyzika - počet bodov z prijímacej skúšky z fyziky
psych. - hodnotenie z psychotestov
mat.(upr) - hodnotenie z matematiky v závislosti od počtu opravných termínov
priemer - priemerná známka zo všetkých skúšok v danom semestri

Tab. 8 Súbor AUTOBAZÁR

	Predajňa 1		Predajňa 2		Predajňa 3	
Por. číslo	**Cena**	**vek**	**Cena**	**vek**	**Cena**	**vek**
1	129,9	5	153	3	159	4
2	179,9	4	159	3	115	5
3	225	2	179	4	159,5	3
4	299,9	1	179	3	166	4
5	215	2	188	1	199	4
6	168	4	183	2	154,9	4
7	176	3	187	1	139	5
8	196,5	3	149	5	179	4
9	189	4	179	2	129	4
10	149	5	180	3	125	4
11	200	3	159	4	135	4
12	159	3	129	5	145	4
13	155	3	179	3	159	4
14	149,9	3	149	5	155	4
15	174,9	3	143	3	149	3
16	220	4	145	4	175	3
17	168	3	139	5	169	3
18	195	2	144	4	189	3
19	159	3	139	5	155	3
20	183,5	3	229	5	185	3
21	149	5	137,7	4	159	3
22	159	2	149	2	169	3
23	180	4	130	5	185	2
24	125	5	178	3	189	2
25	149	4	149	5	135	5
26	174,5	3	150	5	105	5
27	196	3	137	5	129	5
28	119	5	105	6	125	5
29	75	4	166	4	125	5
30	148	4	157	4	155	5
31	136	5	288,6	1	109	5
32	135	4	164	5	125	5
33	179	2	159	4	129	5
34	158	4	149	5	125	5
35	134	5			119	5
36	190	4			125	5

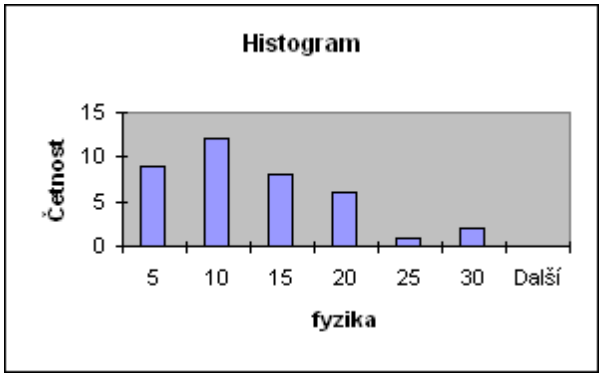
V tabulkách sú uvádzané údaje o veku (v rokoch) a cenách (v tisícoch korún) ojazdených osobných áut ŠKODA FELICIA z troch predajní.

VÝSLEDKY PRÍKLADOV NA PRECVIČENIE

1. ZÁKLADNÉ POJMY

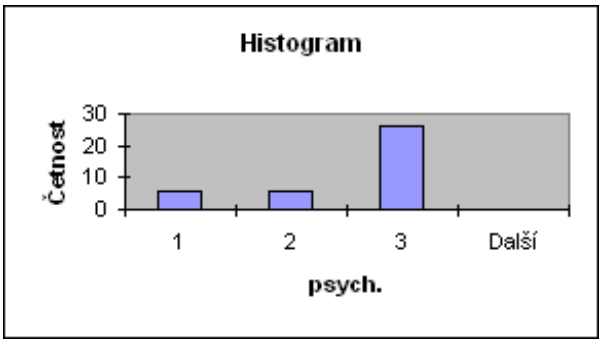
1.6 Triedenie a histogram podľa zadaných hraníc

fyzika	*Četnost*
5	9
10	12
15	8
20	6
25	1
30	2
Další	0



1.7 Triedenie do troch tried a histogram

psych.	*Četnost*
1	6
2	6
3	26
Další	0



1.8 Kontingenčná tabuľka

2. fak.triedenie v %	stred.škola			
psych.	G	P	U	Celkový súčet
1	12,00%	8,00%	0,00%	20,00%
2	6,00%	14,00%	4,00%	24,00%
3	14,00%	38,00%	4,00%	56,00%
Celkový súčet	32,00%	60,00%	8,00%	100,00%

1.9 Kontingenčná tabuľka

1. fak. triedenie	stred.škola				
psych.	G	OA	P	U	Celkový súčet
1	3		2	1	6
2	2		3	1	6
3	3	3	13	7	26
Celkový súčet	8	3	18	9	38

1.10 a) Kontingenčná tabuľka

Priemer z mat.	stred.škola				
psych.	G	OA	P	U	Celkový súčet
1	21,17		17,25	7,30	17,55
2	18,00		14,00	1,50	13,25
3	6,33	9,17	7,38	4,00	6,56
Celkový súčet	14,81	9,17	9,58	4,09	9,35

1.11 a) Kontingenčná tabuľka

Priemer z fyzika	stred.škola				
psych.	G	OA	P	U	Celkový súčet
1	15,33		15,50	5,00	13,67
2	14,50		13,33	1,00	11,67
3	17,67	4,67	11,19	5,64	9,69
Celkový súčet	16,00	4,67	12,03	5,06	10,63

1.12 a) Výpočet parametrov použitím ***Popisnej štatistiky*** a použitím štat. funkcií

mat.1. fakulta		**štat. funkcie**
Stř. hodnota	9,35	9,35
Chyba stř. hodnoty	1,07	
Medián	8	8
Modus	4	4
Směr. odchylka	**6,58**	**6,49**
Rozptyl výběru	**43,24**	**42,10**
Špičatost	-0,48	-0,48
Šikmost	0,62	0,62
Rozdíl max-min	25	
Minimum	0	
Maximum	25	
Součet	355,3	
Počet	38	

2. PRAVDEPODOBNOSŤ

2.11 Na dvoch súčet 9 ($p = 4/7$), na troch súčet 10 ($p = 24/37$).

2.12 $p = 0{,}175$

2.13 Pravdepodobnosť náhodného zostavenia slova RAKETA je 0,0027, preto dieťa vie čítať s pravdepodobnosťou $p = 1 - 0{,}0027 = 0{,}9973$.

2.14 S pravdepodobnosťou $p = 0{,}625$ bola na matematike.

2.15 $p = 0{,}1$

2.16 $P_{3,4} = 0{,}25$, $P_{5,8} = 0{,}21875$

2.17 $p = 0{,}12066$

2.18 a) 0,0000219

b) 0,0085 c) 0,0154

2.19 $P(S) = P[(A \cap B) \cup (C \cap D)] = 0{,}965$

2.20 $p = 0{,}585$

2.21 $P = 1 - P_{0,n} = 1 - \binom{n}{0} \cdot p^0 \cdot q^n$, odtiaľ $n \geq \dfrac{\ln(1-P)}{\ln(1-p)}$.

2.22 $p = 2/3$

2.23 Vyhrať aspoň na jednom lóse $p = 1 - P_{0,n} = 1 - \dfrac{\binom{n-m}{k}}{\binom{n}{k}}$.

3. NÁHODNÁ PREMENNÁ

3.17 $E(X) = 2{,}5$, $D(X) = 2{,}08$, $\sigma(X) = 1{,}44$.

3.18

body	-8	-1	6	13	20
p_i	0,0016	0,0256	0,1536	0,4096	0,4096

$E(X) = 14{,}4$, $D(X) = 31{,}36$, $\sigma(X) = 5{,}6$.

3.19

počet	0	1	2	3	4	5
p_i	0,3277	0,4096	0,2048	0,0512	0,0064	0,0003

$E(X)=1 \quad D(X)=0{,}8 \quad P(X=0)=0{,}3277$.

3.20 $c=0{,}25$; $E(X)=1$; $D(X)=1/3$; $P(1\le X<3)=0{,}5$.

3.21 Pravdepodobnosť, že uchádzač urobí test je $p=0{,}72675$.

Rozdelenie Bi(0,72675; 300) aproximujeme N(218,025; 59,57).

$P(X=200)\doteq 0{,}003382$; $P(200\le X\le 300)\doteq 0{,}99$.

3.22 Rozdelenie Bi(0,05; 400) aproximujeme N(22; 19). $P(19\le X\le 21)\doteq 0{,}1815$.

3.23 Rozdelenie Bi(0,002; 3000) aproximujeme Po(6). $P(X=4)\doteq 0{,}428$.

3.24 a) $P(X=20)\doteq 0{,}04836$ b) $P(10\le X\le 30)\doteq 0{,}8465$.

3.25 $P(360\le X\le 500)\doteq 0{,}165$.

4. VÝBEROVÉ SKÚMANIE

4.5 a) $\bar{x}=15{,}12$; $\sigma=5{,}88$; $\sigma^2=34{,}56$

b) $\bar{x}=13$; $\sigma=6{,}62$; $\sigma^2=43{,}89$

4.6 a) 90% intervaly spoľahlivosti pre priemer

$(13{,}73;16{,}51)$, $(0;16{,}2)$, $(14{,}04;\infty)$

90% intervaly spoľahlivosti pre rozptyl

$(25{,}52;49{,}90)$, $(0;45{,}2)$, $(27{,}29;\infty)$

b) 90% intervaly spoľahlivosti pre priemer

$(11{,}43;14{,}57)$, $(0;14{,}27)$, $(11{,}78;\infty)$

90% intervaly spoľahlivosti pre rozptyl

$(32{,}42;63{,}38)$, $(0;58{,}41)$, $(34{,}66;\infty)$

4.7 90% interval spoľahlivosti pre podiel $(0{,}07;0{,}22)$

5. TESTOVANIE ŠTATISTICKÝCH HYPOTÉZ

5.10 a) $\bar{x} = 12 \in (8{,}74;12{,}52) \Rightarrow$ s tvrdením môžeme na zvolenej hladine významnosti súhlasiť

b) $\bar{x} = 12 \in (11{,}43;14{,}57) \Rightarrow$ s tvrdením môžeme na zvolenej hladine významnosti súhlasiť

5.11 $(45 \in 32{,}42;63{,}38) \Rightarrow$ s tvrdením môžeme na zvolenej hladine významnosti súhlasiť

5.12 $z = -1{,}027;\ z_{krit}(2) = -1{,}64,\quad z \in (-1{,}64;1{,}64) \Rightarrow$ s tvrdením môžeme na zvolenej hladine významnosti súhlasiť

5.13 H_0: Vstupná úroveň študentov z fyziky je na oboch fakultách porovnateľná.

H_1: Vstupná úroveň študentov z fyziky je lepšia na 2. fakulte.

p(1) = 0,052 (použitý je dvojvýberový z-test na strednú hodnotu)

p(1) = 0,053 (použitý je dvojvýberový t-test)

H_0 zamietame na každej hladine významnosti $\alpha >$ p(1).

5.14 z_{stat} = -1,185, $z_{krit\ (2)}$= -2,576 $\Rightarrow$ s tvrdením môžeme na zvolenej hladine významnosti súhlasiť

5.15 použijeme dvojvýberový párový t-test

t_{stat} = -4,48, $t_{krit(2)}$= 2,01 $\Rightarrow$ zamietame hypotézu o porovnateľných výsledkoch v 2. a 3. semestri

6. VYŠETROVANIE NORMALITY ROZDELENIA

6.5 H_0 : Cena áut v 1. predajni znak má normálne rozdelenie.

H_1 : Cena áut v 1. predajni nemá normálne rozdelenie.

D=0,26311; Y= -3,79719

a) kvantily -4,09; 1,05, pre α=0,01 H_0 nezamietame.

b) kvantily -2,85; 0,917, pre α=0,05 H_0 zamietame.

6.6 *D=0,390; Y=21,79*

H_0 zamietame na všetkých bežných hladinách významnosti.

6.7 použijeme χ^2 – test dobrej zhody, akceptujeme vytvorené triedy, zlúčime

triedy s malou početnosťou

a) H_0: Získané body z mat. na 1. fak. majú normálne rozdelenie.

H_1: Získané body z mat.na 1. fak. nemajú normálne rozdelenie.

$\chi^2 = 1{,}7164$, $\chi^2(0{,}05;2)=5{,}9915$. $\chi^2 < \chi^2(0{,}05;2)$, nezamietame nulovú hypotézu.

b) $\chi^2 = 0{,}3553$, $\chi^2(0{,}05;2)=5{,}9915$. $\chi^2 < \chi^2(0{,}05;2)$, nezamietame nulovú hypotézu.

6.8 Použijeme χ^2 – test dobrej zhody, akceptujeme vytvorené triedy, zlúčime triedy s malou početnosťou.

H_0: Získané body z mat. na oboch fakultách majú normálne rozdelenie.

H_1: Získané body z mat. na oboch fakultách nemajú normálne rozdelenie.

$\chi^2 = 2{,}1588$, $\chi^2(0{,}05;7)=14{,}067$. $\chi^2 < \chi^2(0{,}05;2)$, nezamietame nulovú hypotézu.

7. NEPARAMETRICKÉ TESTY

7.5 H_0: $\mu_G = \mu_U$, H_1: $\mu_G \neq \mu_U$,

$u = 21$, $U_{0{,}05}(8,9) = 15$, $u > U_{\alpha/2} \Rightarrow H_0$ prijímame

7.6 H_0: $\mu_2 = \mu_3$, H_1: $\mu_2 < \mu_3$,

$t_W = 13$, $T_W(0{,}05;\ 9)=8$, $t_W > T_W \Rightarrow H_0$ prijímame

H_0: Všetky priemery sa rovnajú.

H_1: Aspoň dva priemery sa nerovnajú.

$k = 5{,}18$ (Kruskalova-Wallisova testovacia štatistika),

$\chi^2(0{,}99;2) = 9{,}21$, $k > \chi^2 \Rightarrow H_0$ zamietame

H_0: Konzervačný prípravok nemá vplyv na trvanlivosť výrobku.

H_1: Konzervačný prípravok má vplyv na trvanlivosť výrobku.

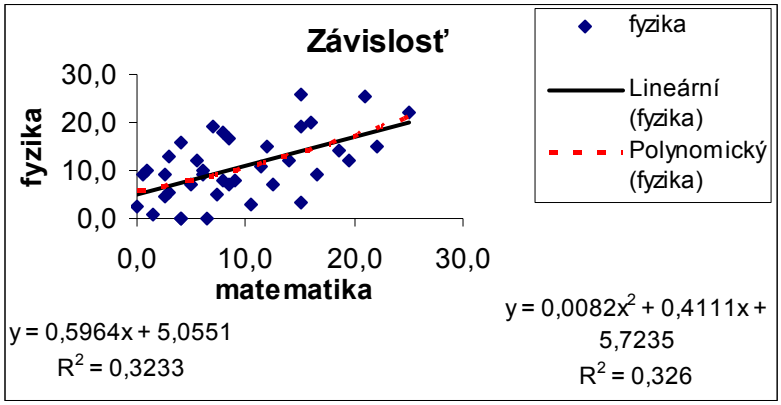
$q = 3{,}38$ (Friedmanova testovacia štatistika), $\chi^2(0{,}9;2) = 4{,}61$

$q > \chi^2 \Rightarrow H_0$ zamietame

8. SKÚMANIE ZÁVISLOSTI DVOCH KVANTITATÍVNYCH ZNAKOV

8.4 a) korelačná matica

	mat.	*fyzika*
mat.	1	
fyzika	0,568553	1



$r = 0{,}5686$, $r^2 = 0{,}3233$, rovnica regresnej priamky $y = 0{,}5964x+5{,}0551$

$y(15)$=14 bodov, $y(21)$=17,58 bodov.

b) $r = 0{,}5640$, $r^2 = 0{,}3181$, rovnica regresnej priamky $y = 0{,}6356x+3{,}3897$

$y(15) = 12{,}92$ bodov, $y(21) = 16{,}74$ bodov.

c) $r^2 = 0{,}7203$, rovnica regresnej priamky $y = 0{,}5142x+0{,}9889$

$y(3{,}2) = 2{,}63$ –priemerná známka, $y(2) = 2{,}01$- priemerná známka

d) $r^2 = 0{,}2979$, rovnica regresnej priamky y = $0{,}2624x+1{,}1775$

$y(3{,}2) = 2{,}12$ –priemerná známka, $y(2) = 1{,}80$- priemerná známka

8.5 a) $r = 0{,}5686$, $r^2 = 0{,}3233$, rovnica regresnej priamky $y = 0{,}5964x+5{,}0551$,

p -hodnota pre celkový F- test je p = 0,000196 $\Rightarrow$ lineárny model je štatisticky významný, t.j. znaky sú lineárne závislé,

$1{,}735 \le B_0 \le 8{,}375$; $0{,}305 \le B_1 \le 0{,}888$

Regresní statistika	
Násobné R	0,5686
Hodnota spolehlivosti R	0,3233
Nastavená hodnota spolehlivosti R	0,3045
Chyba stř. hodnoty	5,7526
Pozorování	38

ANOVA					
	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	1	569,04	569,04	17,20	0,000196008
Rezidua	36	1191,31	33,09		
Celkem	37	1760,34			

	Koeficienty	*Chyba stř. hodnoty*	*t stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	5,055	1,64	3,0883	0,00386	1,735	8,375
mat.	0,596	0,14	4,1468	0,00020	0,305	0,888

b) $r = 0{,}5640$, $r^2 = 0{,}3181$, rovnica regresnej priamky $y = 0{,}6356x+3{,}3897$
p – hodnota pre celkový F – test je 1,99.E – 05, t.j. na všetkých bežných hladinách významnosti je lineárny model štatisticky významný.
- $0{,}986 \leq B_0 \leq 7{,}765$; $0{,}366 \leq B_1 \leq 0{,}906$.

9. SKÚMANIE ZÁVISLOSTI DVOCH KVALITATÍVNYCH ZNAKOV

9.6 použijeme χ^2 – test závislosti

	empirické početnosti			
	psychotesty			celkom
stredná škola	1	2	3	
G	9	5	10	24
OA+P+U	7	13	44	64
celkom	16	18	54	88

	teoretické početnosti			
	psychotesty			
stredná škola	1	2	3	celkom
G	4,363636	4,909091	14,72727	24
OA+P+U	11,63636	13,09091	39,27273	64
celkom	16	18	54	88

p- hodnota χ^2 - test závislosti je 0,01190, na hladine významnosti $\alpha = 0{,}05$ zamietame nulovú hypotézu o nezávislosti znakov.

9.7

	úspešnosť		
stred.škola	Neupel	Spravil	celkom
G	5	19	24
OA+P+U	35	29	64
celkom	40	48	88

	úspešnosť		
	Neuspel	Spravil	celkom
G	10,91	13,09	24
OA+P+U	29,09	34,91	64
celkom	40	48	88

p - hodnota χ^2 - test závislosti je 0,0045, na hladine významnosti $\alpha = 0{,}05$ zamietame nulovú hypotézu o nezávislosti znakov.

9.8 a) Je to kvantitatívny znak, u jedného pacienta zistíme hladinu cholesterolu pred liečbou a po istom období používania lieku. Na výberový súbor použijeme párový t-test na testovanie zhody stredných hodnôt (alebo jeho neparametrickú alternatívu).

b) Bolesť hlavy sa nedá kvantitatívne odmerať, je to kvalitatívny znak. Pacientovi dáme istý čas striedavo užívať placebo a skutočný liek, pacient nevie, čo je liek. Pacient má potom označiť, čo bol liek. Štatistický znak je počet správnych označení v skupine n pacientov.

Použijeme test o zhode podielu s konštantou. $H_0 : \pi = 0{,}5$; $H_1: \pi > 0{,}5$.

c) V jednom výberovom súbore budú chorí, ale neočkovaní pacienti, v druhom tiež chorí, ale očkovaní pacienti. Sledujeme priebeh chrípky (ľahký, ťažký) u oboch skupín. Testujeme pomocou χ^2- testu závislosť priebehu chrípky od očkovania.

LITERATÚRA

[1] ANDĚL, J.: Statistické metody, Matfyzpress, Praha 1998

[2] BAKYTOVÁ, H., UGRON, M., KONTŠEKOVÁ, O.: Základy štatistiky, ALFA Bratislava 1975

[3] BAKYTOVÁ, H. a kol.: Štatistika pre ekonómov, ES EU, Bratislava 1994

[4] HINDLS, R., HRONOVÁ, S., SEGER, J.: Statistika pro ekonomy, Profesional Publishing, Praha 2003

[5] CHAJDIAK, J.: Štatistika v Exceli, STATIS, Bratislava 2002

[6] CHAJDIAK, J., RUBLÍKOVÁ, E., GUDABA, M.: Štatistické metódy v praxi, STATIS, Bratislava 1997

[7] KUBÁČEK, L., PÁZMAN, A.: Štatistické metódy v meraní, VEDA, Bratislava 1979

[8] KUBÁČKOVÁ, L.: Metódy spracovania experimentálnych údajov, VEDA, Bratislava 1990

[9] LIKEŠ, J., CYHELSKÝ, L., HINDLS, R.: Úvod do štatistiky a pravdepodobnosti, Statistika A VŠE, Praha 1993

[10] LYSÁ, Ľ.: Štatistika v praxi vojenského manažéra, VA Liptovský Mikuláš 2001

[11] LYSÁ, Ľ., DROPPA, M.: Kvantitatívne metódy v manažmente, VA Liptovský Mikuláš 2004

[12] LEVIN, R.I., RUBIN, D.S.: Statistics for management, Prentice Hall, New Jersey 1994

[13] MAGERA, I.: Excel pro verze 2002, 2000 a 97, Computer Press, Praha 2001

[14] MELOUN, M., MILITKÝ, J.: Statistické zpracování experimentálnich dat, PLUS, Praha 1994

[15] PACÁKOVÁ, V. a kol.: Štatistika pre ekonómov, Edícia ekonómia, Bratislava 2003

[16] POTOCKÝ, R. a kol.: Zbierka úloh z pravdepodobnosti a matematickej štatistiky, ALFA, Bratislava 1991

[17] RIEČAN, B., LAMOŠ, F., LENÁRT, C.: Pravdepodobnosť a matematická štatistika, ALFA, Bratislava 1984

[18] WIMMER, G.: Štatistické metódy v pedagogike, GAUDEAMUS, Hradec Králové 1993

[19] WONNACOTT, T.H., WONNACOT, R.J.: Statistika pro obchod a hospodářství, Victoria Publishing, Praha 1993