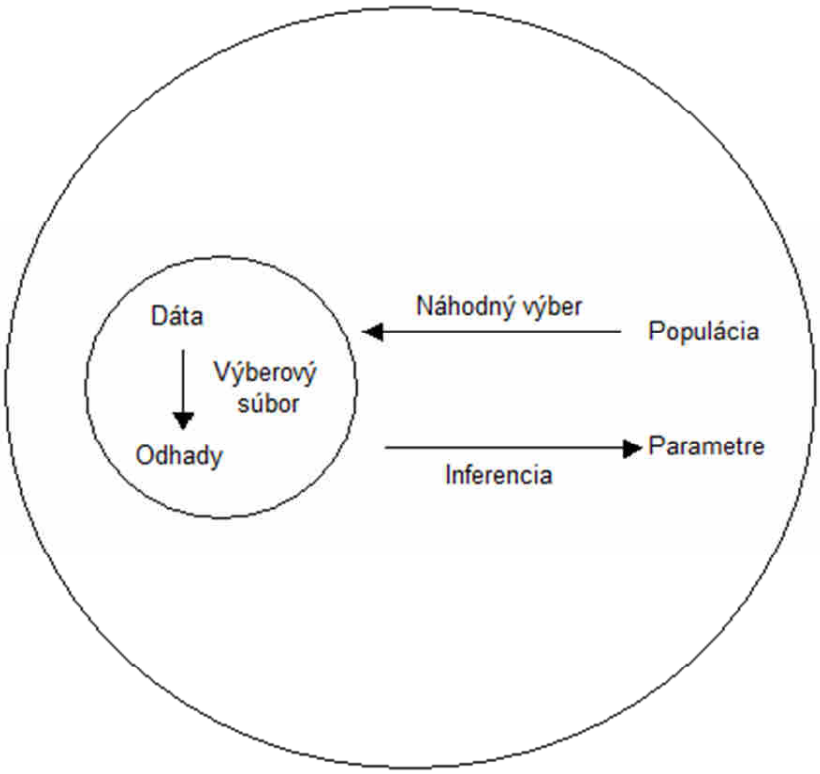
Inferenčná analýza

Úlohou inferenčnej analýzy je na základe informácii získaných z náhodných výberov urobiť závery o celej populácii (Obrázok 23).



Obrázok 23 Princíp inferenčnej analýzy

Použitie inferenčnej analýzy na nenáhodných a úplných výberoch nie je správne. V prípade úplného/totálneho výberu (cenzus) každý zistený vzťah predstavuje skutočný, t.j. významný vzťah.

Vzorky získané z nenáhodných výberov môžeme skúmať pomocou exploračnej analýzy. Závery získané takýmto spôsobom nemôžeme zovšeobecňovať na celú populáciu, ale platia iba pre analyzovanú vzorku (Hendl, 2004).

3.1 Pravdepodobnosť ako teoretický základ inferenčnej analýzy

K analýze dát môžeme pristupovať z hľadiska exploračnej a inferenčnej analýzy. Exploračnou analýzou dokážeme prehľadne zhrnúť informácie, ktoré sa týkajú práve tých objektov, ktoré sme

merali. Ak sme však dáta získali na základe dobre navrhnutého výskumného plánu, môžeme zovšeobecňovať správanie sledovaných premenných a ich parametre na celú populáciu. K tomu nám slúži inferenčná analýza, ktorá sa opiera o pravdepodobnosť. Inferenčná analýza – testovanie hypotéz a odhady parametrov vyžadujú dáta získané náhodným výberom alebo randomizáciou (náhodné priradenie jedincov do skupín – znáhodnený experiment). Pomocou pravdepodobnosti zistíme ako často nastane určitý jav, ak výber alebo experiment realizujeme viackrát. Podmienkou je, že pri získavaní využijeme náhodu (Riečan, Lamoš a Lenárt, 1992; Vrábelová a Markechová, 2001).

Štatistická definícia pravdepodobnosti:

O **náhodnom pokuse** hovoríme vtedy, ak výsledok pokusu, ktorý vykonávame nie je jednoznačne určený podmienkami, za ktorých sa uskutočňuje. Náhodným pokusom je napr. hod hracou kockou, tento pokus môže končiť šiestimi výsledkami a pred jeho realizáciou nevieme povedať, koľko bodiek bude na vrchnej stene kocky.

Náhodným javom *A* prislúchajúcim k danému náhodnému pokusu nazývame každé overiteľné tvrdenie o výsledku náhodného pokusu, napr. pri hode kockou náhodným javom *A* je tvrdenie „*Padne párny počet bodiek*“. Hovoríme, že jav *A* nastal, ak tvrdenie *A* o výsledku pokusu je pravdivé, napr. náš jav *A* pri hode kockou nastal, ak padla stena s 2 alebo 4 alebo 6 bodkami.

Budeme sa zaoberať len pokusmi, pri ktorých sa sledovaný náhodný jav pri *n*-násobnom nezávislom opakovaní pokusu (za rovnakých podmienok) vyznačuje tzv. štatistickou stabilitou, čo znamená, že relatívne početnosti javu *A* počítané pri veľkom počte opakovaní pokusu

$$f_n(A) = \frac{n_A}{n},$$

kde n_A je počet výskytov javu *A* v sérii *n* pokusov, sa príliš nemenia, kolíšu okolo nejakej konštanty. Táto konštanta sa nazýva **pravdepodobnosťou javu** *A* a označuje sa *P(A)*.

Napríklad ak vykonáme niekoľko sérií po 1000 hodoch kockou, v každej sérii spočítame koľkokrát padol párny počet bodiek a vypočítame relatívnu početnosť, tak všetky tieto relatívne početnosti budú kolísať okolo čísla ½, teda ½ je pravdepodobnosť padnutia párneho počtu bodiek pri hode kockou. Čím viac pokusov uskutočníme, tým lepší odhad pravdepodobnosti *P(A)* relatívnou početnosťou $\frac{n_A}{n}$ dostaneme.

Ak máme dva náhodné javy *A* a *B*, tak môžeme definovať ich prienik $A \cap B$, čo je jav, ktorý nastane, ak javy *A* a *B* nastanú súčasne, ich zjednotenie $A \cup B$, čo je jav, ktorý nastane, ak nastane jav *A* alebo jav *B*. Jav $\overline{A}$ označuje opačný (doplnkový) jav k javu *A* a nastane práve vtedy, ak *A* nenastane.

Jav, ktorý pri danom náhodnom pokuse nastane vždy, nazývame **istý jav** a označujeme ho Ω. Jav, ktorý nenastane nikdy sa nazýva **nemožný jav** a označuje sa $\varnothing$. Ak prienik dvoch javov je nemožný jav, dané javy nazývame nezlučiteľné (disjunktné). Zrejme

$$A \cup \overline{A} = \Omega, \; A \cap \overline{A} = \varnothing.$$

Všimnime si, že pri hode kockou môžeme „najjemnejšie“ výsledky označiť číslami 1, 2, 3, 4, 5 a 6 javu *A* vieme jednoznačne priradiť množinu výsledkov, pri ktorých jav *A* nastáva, *A* = {2, 4, 6}. Teda $\overline{A}$ = {1, 3, 5}, $A \cup \overline{A}$ = {1, 2, 3, 4, 5, 6} = Ω. Ak *B* znamená *padne stena s viac ako tromi bodkami*, tak *B* = {4, 5, 6} a $A \cup B$ = {2, 4, 5, 6}, $A \cap B$ = {4, 6}.

To platí všeobecne, pri každom pokuse definujeme množinu najjemnejších výsledkov Ω, o ktorých potrebujeme uvažovať a kde náhodné javy sú (nie nutne všetky) podmnožiny množiny Ω. Prvky množiny Ω sa nazývajú **elementárne javy**. Ak by nás pri hode kockou zaujímalo, len či padla (P) alebo nepadla (N) šestka, tak môžeme položiť Ω = {N, P} alebo Ω = {0, 1}.

Systém javov *S*, ktorý s každým javom *A* obsahuje aj jav $\overline{A}$, s každými dvoma javmi obsahuje aj $A \cap B$ a $A \cup B$ a obsahuje tiež istý jav Ω a nemožný jav $\varnothing$ sa nazýva **pole náhodných javov**.

Pravdepodobnosť definovaná na poli náhodných javov je funkcia, ktorá každému javu $A \in S$ priradí pravdepodobnosť *P(A)*.

Pravdepodobnosť náhodného javu je číslo z intervalu [0, 1], ktoré popisuje relatívnu početnosť, s akou sa jav vyskytne vo veľmi dlhom rade opakovaní situácie, kedy tento jav môže nastať.

Základné vlastnosti pravdepodobnosti:

1. Pravdepodobnosť istého javu je 1.
2. Pravdepodobnosť nemožného javu je 0.
3. Ak sa náhodný jav dá rozložiť na niekoľko disjunktných javov, potom sa jeho pravdepodobnosť rovná súčtu pravdepodobností týchto javov. Platí aj všeobecnejšie pravidlo.

Všeobecnejšie pravidlo pre dva javy:

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

Všeobecnejšie pravidlo pre tri javy:

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

Zo základných vlastností vyplývajú všetky ďalšie:

4. Pre ľubovoľný jav *A* platí

$$0 \leq P(A) \leq 1.$$

5. Ak je jav $\overline{A}$ doplnkový k javu *A*, tak

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

6. Ak je jav *A* časťou javu *B*, $A \subset B$, tak

$$P(A) \leq P(B).$$

Nezávislosť náhodných javov:

Javy *A* a *B* sa nazývajú nezávislé práve vtedy, ak $P(A \cap B) = P(A).P(B)$.

Javy A_1, A_2, ..., A_n sú nezávislé, ak pravdepodobnosť prieniku ľubovoľnej podmnožiny týchto javov sa rovná súčinu ich pravdepodobností.

3.1.1 Náhodná premenná a jej rozdelenie

Predpis, ktorý priraďuje každému výsledku náhodného pokusu určité číslo sa nazýva **náhodná premenná**. Náhodná premenná je funkcia, ktorá zobrazuje priestor výsledkov do reálnych čísiel. Výsledkom pokusu nemusí byť vždy nejaké číslo, ale vždy mu môžeme nejaké číslo priradiť (napr. hod mincou - padne znak 1, padne číslo 0). Okrem priradení nás zaujímajú pravdepodobnosti, s akými náhodná premenná nadobúda určité hodnoty. Tieto pravdepodobnosti nazývame **rozdelenie pravdepodobnosti náhodnej premennej**. Na priestore výsledkov je možné definovať viacero náhodných premenných. Funkcie náhodných premenných sú opäť náhodné premenné.

Náhodné premenné delíme na **diskrétne** a **spojité**.

Diskrétne premenné nadobúdajú navzájom izolované hodnoty (napr. počet bodiek pri hode kockou). Niekedy pozorujeme diskrétne náhodné premenné, ktoré môžu teoreticky nadobúdať nekonečne veľa hodnôt, presne toľko, koľko je prirodzených čísel (napr. počet nehôd za rok

v danom regióne). Medzi spojité náhodné premenné patria všetky bežné merania dĺžky, hmotnosti a času, taktiež sem patria výsledky psychologických, vedomostných alebo motorických testov a pod. Závisí na probléme a cieľoch, či danú premennú budeme považovať za spojitú alebo diskrétnu.

Distribučná funkcia

Z teoretického hľadiska najúplnejší popis pravdepodobnostného správania diskrétnych alebo spojitých náhodných premenných *X* predstavuje **distribučná funkcia *F*** (Obrázok 24). Distribučná funkcia v bode *x* je pravdepodobnosť, že náhodná premenná *X* nadobúda hodnotu menšiu nanajvýš rovnú *x*,

$$F(x) = P(X \leq x).$$

Distribučná funkcia je definovaná pre všetky reálne čísla *x*. Z uvedeného vyplýva, že funkcia *X* je náhodnou premennou, ak sa pravdepodobnosti $P(X \leq x)$ dajú vypočítať pre všetky reálne čísla *x*.

Vlastnosti distribučnej funkcie:

1. $0 \leq F(x) \leq 1$,
2. ak $x \to -\infty$, potom $F(x) \to 0$,
3. ak $x \to +\infty$, potom $F(x) \to 1$,
4. *F*(*x*) je funkcia neklesajúca, t.j. ak $x_i < x_j$, potom $F(x_i) \leq F(x_j)$,
5. *F*(*x*) nemusí byť spojitá, ak je *F(x)* spojitá funkcia, potom príslušná náhodná premenná je spojitá.

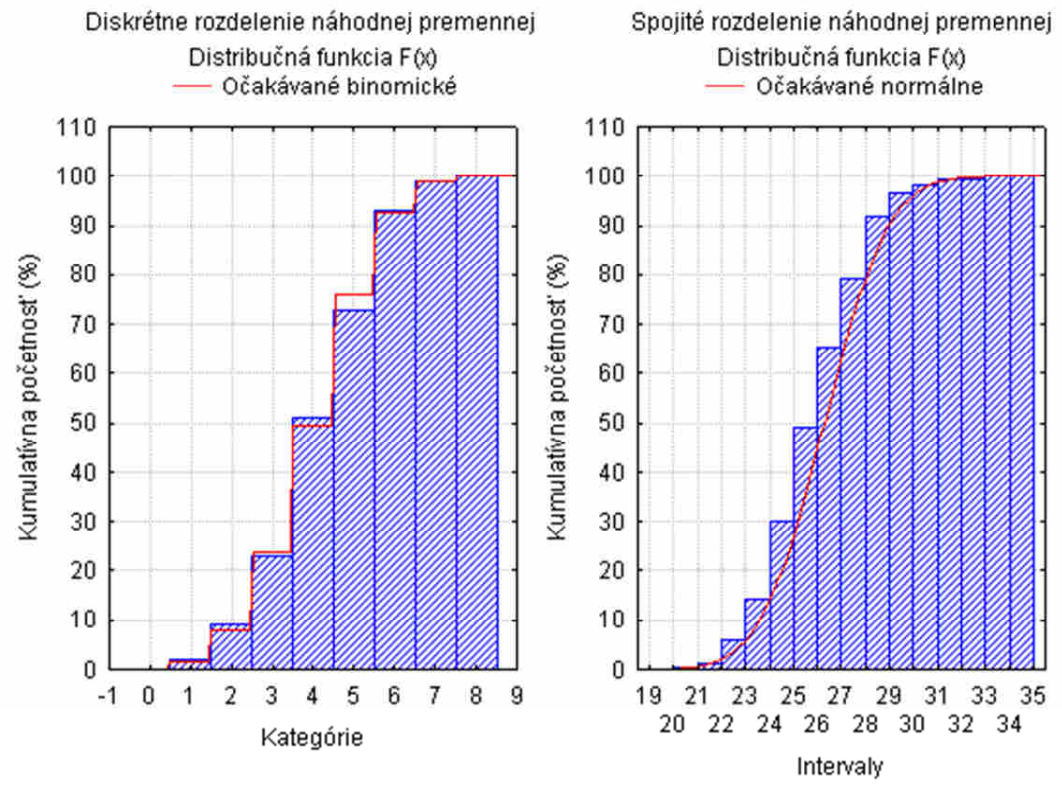
Pre počítanie pravdepodobnosti platia vzorce:

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1),$$

$$P(X > x) = 1 - F(x).$$

Výberovým ekvivalentom teoretickej distribučnej premennej funkcie *F*(*x*) je **výberová distribučná funkcia** $\widehat{F}(x)$, ktorá popisuje rozdelenie hodnôt výberu. Empirická distribučná funkcia $\widehat{F}(x)$ je definovaná v bode *x* relatívnym počtom meraní, ktoré sú menšie alebo rovné *x*,

$$\widehat{F}(x) = \frac{počet\_x_i \leq x}{n}.$$



Obrázok 24 Distribučná funkcia diskrétnej a spojitej náhodnej premennej

Pravdepodobnostná funkcia a hustota pravdepodobnosti

U diskrétnej náhodnej premennej správanie popisuje **pravdepodobnostná funkcia**

$$p(x) = P(X = x)\,.$$

Táto funkcia priraďuje diskrétnym hodnotám náhodnej premennej pravdepodobnosti. Diskrétna náhodná premenná X nadobúda hodnoty $x_1, x_2, \ldots, x_m$ s pravdepodobnosťou $p_1, p_2, \ldots, p_m$, pričom platí, že súčet všetkých p_i sa rovná 1. Týmto spôsobom je popísané rozdelenie diskrétnej náhodnej premennej (Obrázok 25).

Ak poznáme pravdepodobnostnú funkciu, vieme vypočítať distribučnú funkciu a naopak.

Pre spojitú náhodnú premennú existuje ekvivalent pravdepodobnostnej funkcie. Ak má $F(x)$ pre všetky x deriváciu, nazývame túto deriváciu **hustotou pravdepodobnosti** $f(x)$ náhodnej premennej X (Obrázok 25). Môžeme ju interpretovať ako približnú pravdepodobnosť, že hodnota náhodnej premennej bude ležať v intervale jednotkovej dĺžky okolo hodnoty x. Hustota pravdepodobnosti $f(x)$ vykazuje podobné vlastnosti ako $p(x) = P(X = x)$ u diskrétnych náhodných premenných:

$$P(x_1 < X \le x_2) = \int_{x_1}^{x_2} f(x)\,dx,$$

$$\int_{-\infty}^{+\infty} f(x)dx = 1,$$

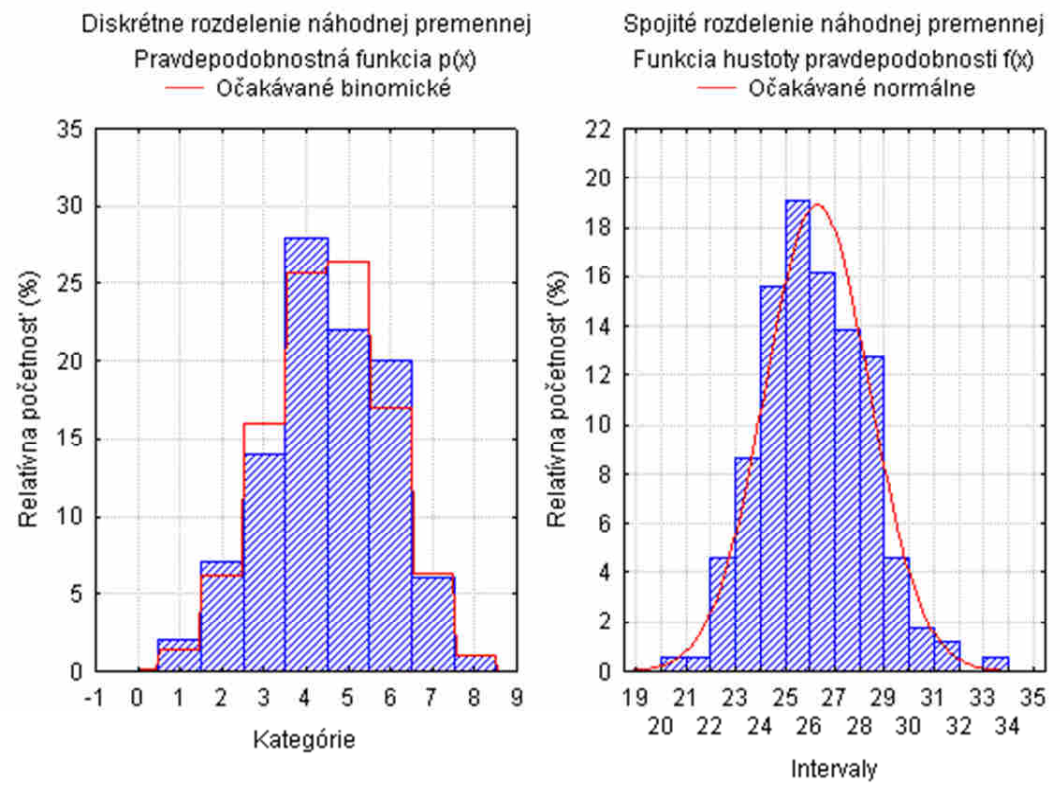
kde $f(x) \ge 0$ pre každé *x*.

Distribučná funkcia spojitej premennej, ak existuje jej hustota, sa vypočíta nasledovne:

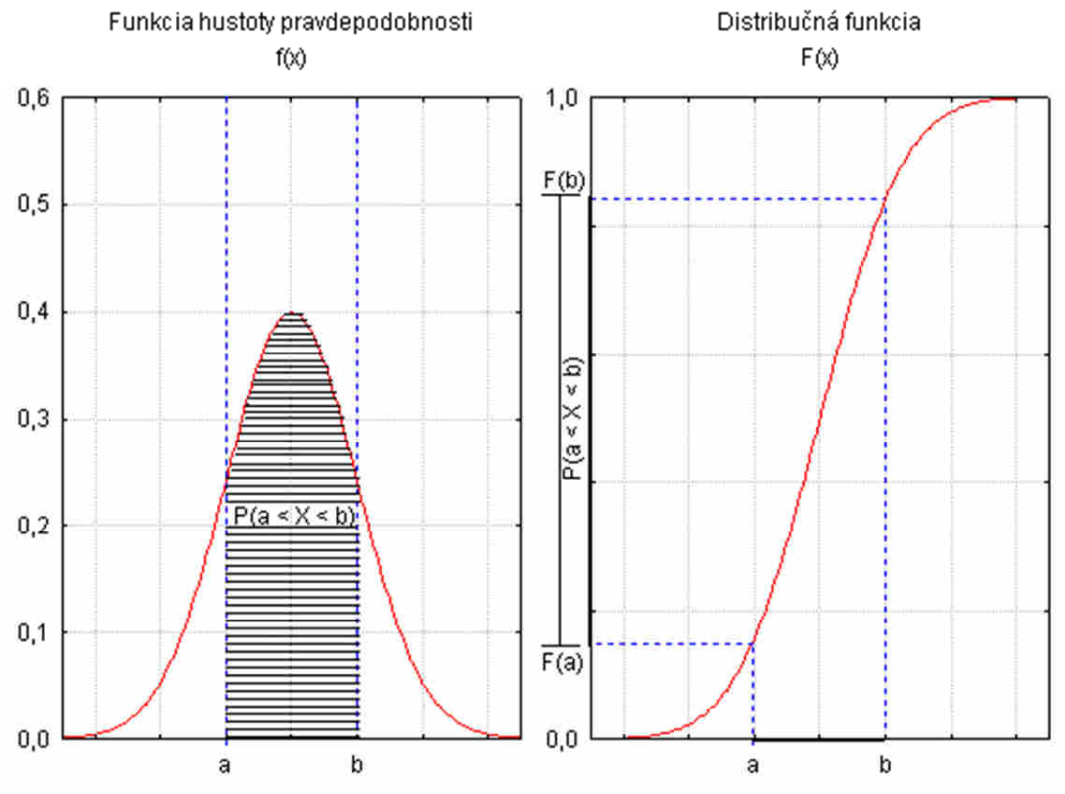
$$F(x) = \int_{-\infty}^{x} f(x)dx.$$

Pre diskrétnu náhodnú premennú vypočítame hodnotu jej distribučnej funkcie ako súčet pravdepodobností jednotlivých hodnôt x_i, ktoré sú menšie alebo rovné *x*, t.j.

$$F(x) = \sum_{x_i \le x} p_i\ .$$



Obrázok 25 Pravdepodobnostná funkcia a funkcia hustoty



Obrázok 26 Vzťah medzi funkciou hustoty pravdepodobnosti a distribučnou funkciou

Parametre rozdelenia náhodnej premennej

Pravdepodobnostné správanie náhodnej premennej je dokonale popísané rozdelením premennej. Niekedy však postačuje uviesť charakteristiky správania náhodnej premennej. Medzi základné charakteristiky, parametre rozdelenia, patrí **stredná hodnota** $E(X)$ a **rozptyl** $Var(X)$ (Obrázok 27, Obrázok 28). V prípade diskrétnej premennej platí

$$E(X) = \mu = \sum_{i=1}^{m} x_i p_i \,,$$

$$Var(X) = \sigma^2 = E(X - E(X))^2 = \sum_{i=1}^{m} (x_i - E(X))^2 p_i = \sum_{i=1}^{m} (x_i - \mu)^2 p_i \,.$$

Veľmi často sa uvádza aj **smerodajná (štandardná) odchýlka**, ktorá sa vypočíta ako druhá odmocnina z rozptylu.

Všeobecne sa dá zapísať stredná hodnota $E(g(X))$ funkcie g diskrétnej náhodnej premennej X ako

$$E(g(X)) = \sum_{i=1}^{m} g(x_i) p_i \,.$$

Stredná hodnota sa nazýva aj očakávaná hodnota, je to priemerná hodnota náhodnej premennej pripadajúca na jeden pokus, ak by sa pokus opakoval nekonečne veľakrát.

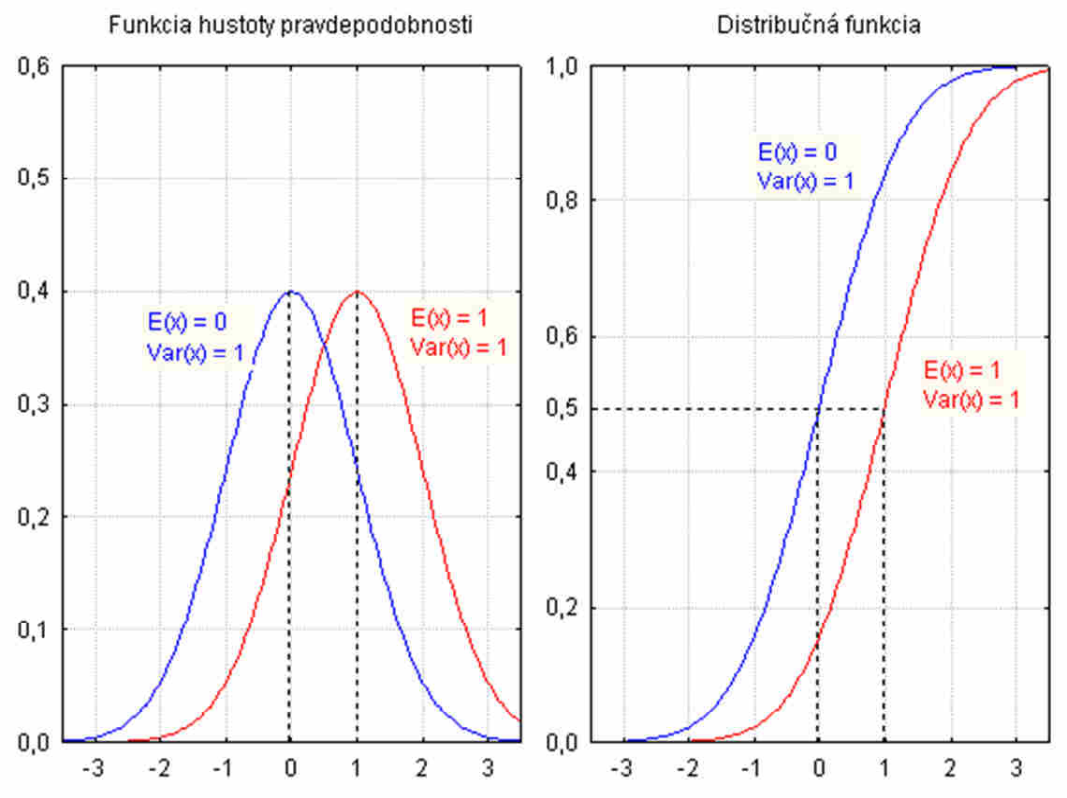
Parametre rozdelenia náhodnej premennej nie sú zvyčajne známe. Na druhej strane je pre posúdenie správania náhodnej premennej dôležité mať o týchto parametroch informácie. V takomto prípade o nich usudzujeme pomocou dát, ktoré sa získali z výskumného plánu.

Hustota sa využíva pre výpočet strednej hodnoty, rozptylu alebo strednej hodnoty funkcie $g(X)$ spojitej náhodnej premennej. Postupuje sa podobne ako u diskrétnej náhodnej premennej:

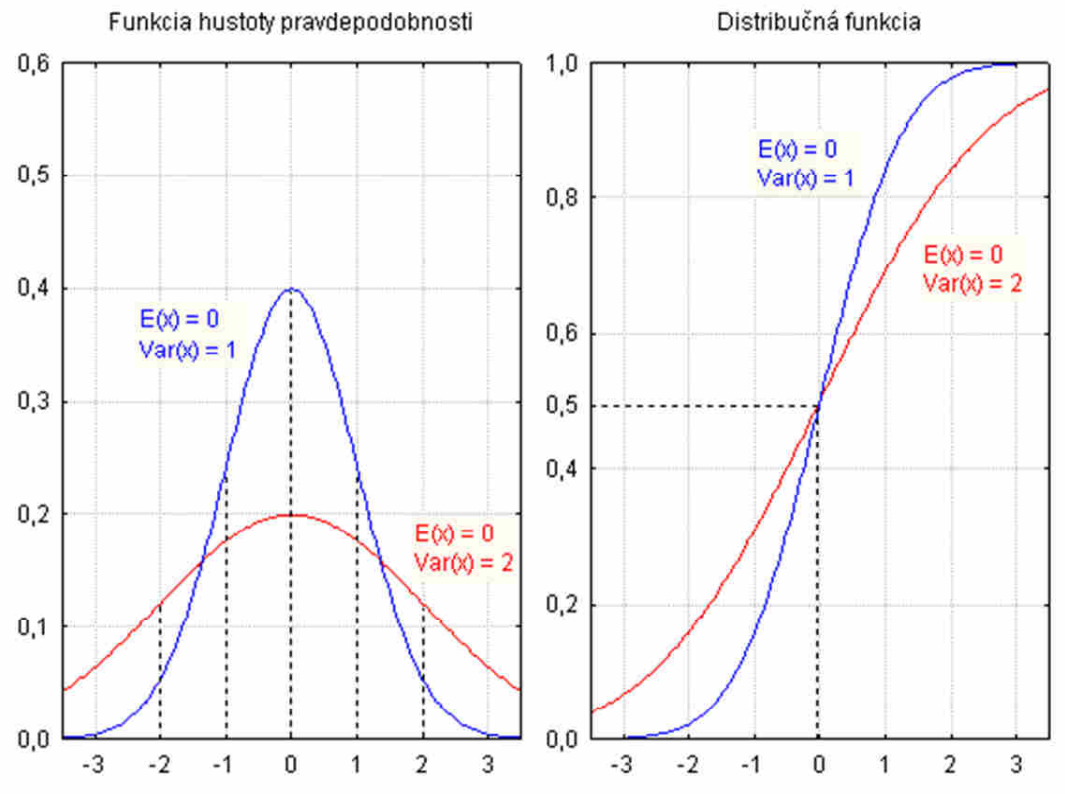
$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx\,,$$

$$Var(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x)dx\,,$$

$$E(g(X)) = \int_{-\infty}^{+\infty} g(x) f(x)dx\,.$$



Obrázok 27 Rozdelenie náhodnej premennej pre strednú hodnotu 0 a 1



Obrázok 28 Rozdelenie náhodnej premennej pre rozptyl 1 a 2

3.1.2 Normálne rozdelenie náhodnej premennej

Normálne rozdelenie je jedným z najdôležitejších rozdelení. Normálnym rozdelením sa riadi veľa náhodných premenných. Normálne rozdelenie je spojité, jednovrcholové rozdelenie, symetrické okolo strednej hodnoty μ. Stredná hodnota tohto rozdelenia je rovná modusu a mediánu.

Normálne rozdelenie je najpoužívanejším rozdelením pre modelovanie náhodného správania sa premenných z nasledujúcich dôvodov:

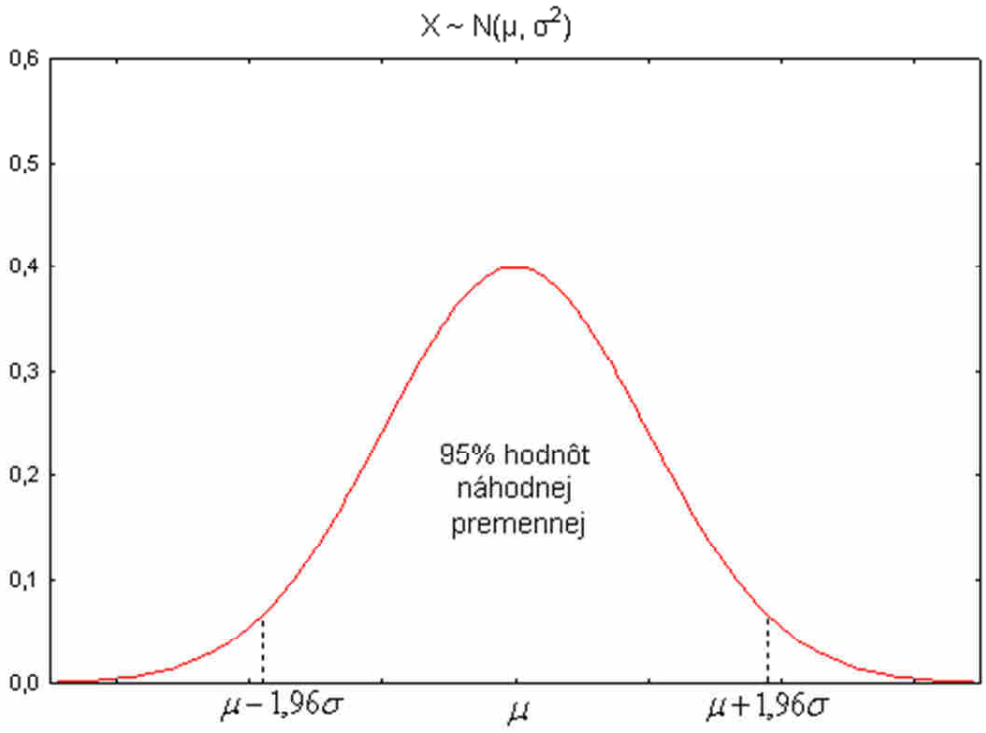
1. väčšinu náhodných premenných môžeme aproximatívne modelovať pomocou tohto rozdelenia (za určitých podmienok dobre aproximuje rad iných (i diskrétnych) pravdepodobnostných rozdelení),
2. niektoré premenné sa dajú transformovať na premennú, ktorá má normálne rozdelenie,
3. existuje množstvo štatistických procedúr, ktoré boli odvodené pre toto rozdelenie,
4. platí centrálna limitná veta, ktorá umožňuje použitie procedúr, ktoré boli navrhnuté na základe normálneho rozdelenia aj pre premenné, ktoré sa týmto rozdelením neriadia.

Hustota pravdepodobnosti (Gaussova krivka) má zvonovitý tvar, maximum nadobúda v strednej hodnote. Smerodajná odchýlka σ (odmocnina z rozptylu) určuje ako sú po oboch stranách od hodnoty μ vzdialené inflexné body, t.j. ako je krivka roztiahnutá do šírky. Normálne rozdelenie je jednoznačne určené strednou hodnotou μ a rozptylom σ^2, ktoré predstavujú jeho parametre. Ak tieto dve charakteristiky poznáme, môžeme určiť tvar celého rozdelenia.

Pre každé normálne rozdelenie s parametrami μ a σ^2 platí:

- interval $\mu \pm \sigma$ obsahuje 68,3% populácie,
- interval $\mu \pm 2\sigma$ obsahuje 95,5% populácie,
- interval $\mu \pm 3\sigma$ obsahuje 99,7% populácie,
- 95% populácie je obsiahnuté v intervale $\mu \pm 1{,}96\sigma$,
- 99% populácie je obsiahnuté v intervale $\mu \pm 2{,}58\sigma$.

Pre premennú *X* s normálnym rozdelením je možné histogram výsledkov z veľkého počtu *n* nezávislých pozorovaní vyrovnať krivkou.



Obrázok 29 Normálne rozdelenie náhodnej premennej X

Obsah plochy pod krivkou hustoty normálneho rozdelenia je rovná jednej. Pravdepodobnosť, že náhodná premenná nadobúda hodnoty z určitého intervalu, je rovná obsahu plochy pod

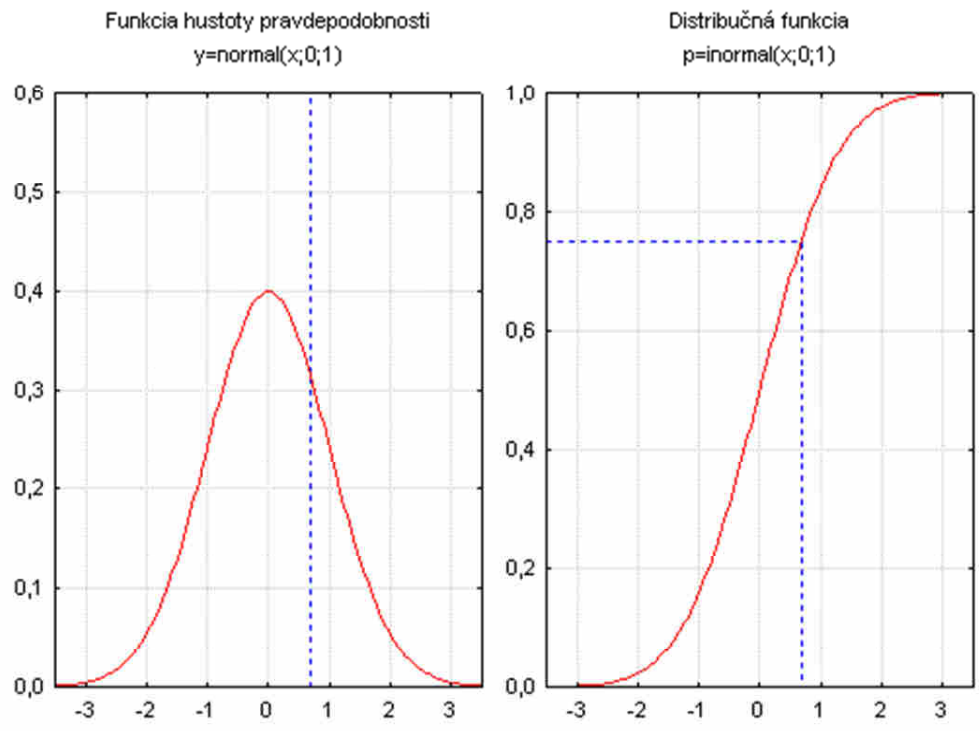
hustotou nad týmto intervalom. Napríklad pre interval s hranicami $\mu \pm 1{,}96\sigma$ má táto plocha obsah 0,95 (Obrázok 29). Náhodná premenná *X* nadobúda teda hodnoty z tohto intervalu s 95% pravdepodobnosťou a iba s 5% pravdepodobnosťou ležia jej hodnoty mimo uvedeného intervalu.

Štandardizované normálne rozdelenie

Pre premennú *X* s normálnym rozdelením s parametrami μ a σ^2, $X \sim N(\mu, \sigma^2)$, je možné vypočítať hodnotu distribučnej funkcie $F(x) = P(X \leq x)$ pre ľubovoľné *x* pomocou hustoty normálneho rozdelenia. V štatistických tabuľkách nájdeme vypočítané hodnoty distribučnej funkcie iba pre normálne rozdelenie so strednou hodnotou 0 a rozptylom 1. Premennú s týmto normálnym rozdelením označíme *Z*, t.j. $Z \sim N(0, 1)$, a rozdelenie nazývame štandardizované (normované) normálne rozdelenie (Obrázok 30). Distribučnú funkciu štandardizovaného normálneho rozdelenia označíme $\Phi(z) = P(Z \leq z)$. Ľubovoľnú premennú $X \sim N(\mu, \sigma^2)$ môžeme transformovať na veličinu $Z = \frac{X - \mu}{\sigma}$, ktorá má štandardizované normálne rozdelenie, t.j. $Z \sim N(0, 1)$.

Tabuľky rozdelenia $N(0, 1)$ sa používajú v dvoch prípadoch:

- pri výpočte kvantilu pre danú hladinu pre rozdelenie $N(\mu, \sigma^2)$,
- pri určení hodnoty distribučnej funkcie rozdelenia $N(\mu, \sigma^2)$.



Obrázok 30 Štandardizované normálne rozdelenie N(0, 1), p = 0,75

Centrálna limitná veta

Mimoriadne postavenie normálneho rozdelenia spočíva okrem iného v tom, že súčet nezávislých ľubovoľne rozdelených náhodných premenných je približne normálne rozdelený, tým lepšie, čím je viac sčítancov. Toto tvrdenie o asymptotickom správaní súčtu náhodných premenných, ktoré presne vyjadruje centrálna limitná veta, je základom pre skutočnosť, že množstvo rozdelení výberových štatistík je možné aproximovať (približne popísať) pri väčšom rozsahu výberu normálnym rozdelením.

Uvedieme jednu z formulácií vety:

Ak majú prvky X_i postupnosti nezávislých náhodných premenných rovnaké rozdelenie so strednou hodnotou μ a smerodajnou odchýlkou σ, potom rozdelenie náhodnej premennej Z_n

$$Z_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

sa s rastúcim n blíži k normálnemu rozdeleniu so strednou hodnotou μ a smerodajnou odchýlkou $\frac{\sigma}{\sqrt{n}}$.

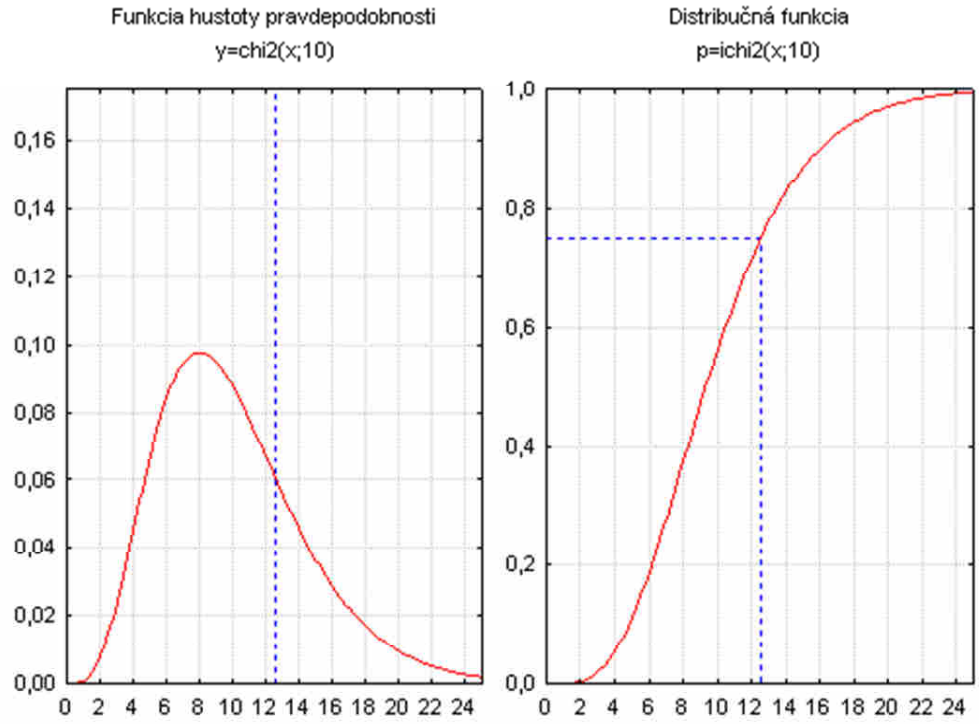
Z tejto vety vyplýva, že aritmetický priemer ako náhodná premenná je za veľmi malých obmedzení asymptoticky normálne rozdelený.

3.1.3 Prehľad rozdelení odvodených od normálneho rozdelenia

Chí-kvadrát o *k* stupňoch voľnosti:

$$X_k^2 = \sum_{j=1}^{k} Z_j^2 \sim \chi^2(k),$$

kde $Z_1, Z_2, \ldots, Z_k \sim N(0, 1)$ sú nezávislé náhodné premenné (Obrázok 31).

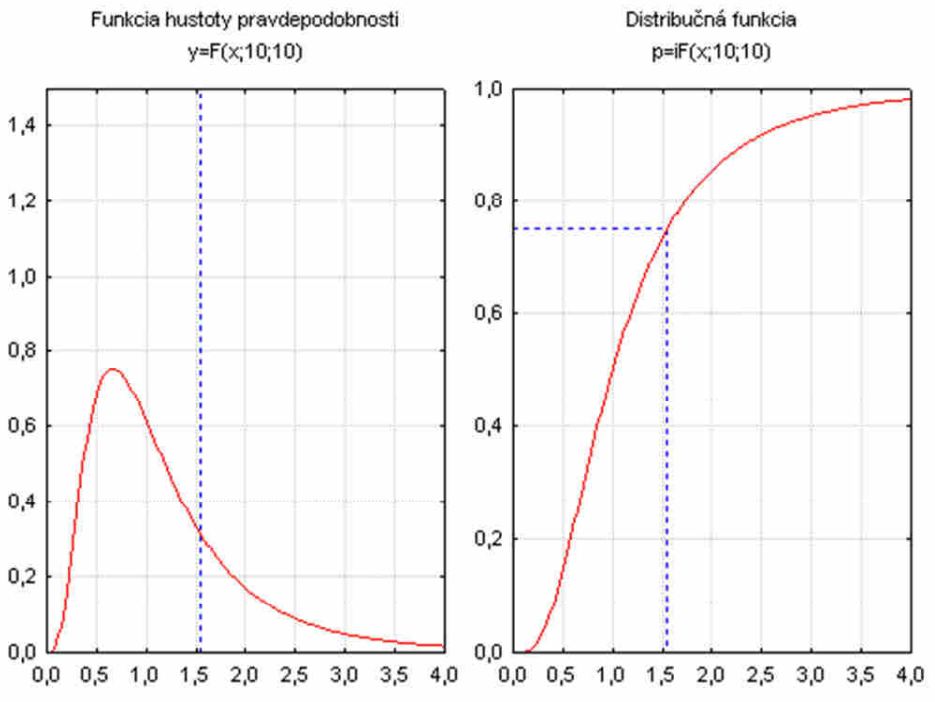


Obrázok 31 Chí-kvadrát rozdelenie $\chi^2(10)$ *, p = 0,75*

(Fisherovo-Snedecorovo) F-rozdelenie s *k* a *m* stupňami voľnosti:

$$F_{k,m} = \frac{X_k^2 / k}{X_m^2 / m} \sim F(k, m),$$

kde $X_k^2 \sim \chi^2(k)$, $X_m^2 \sim \chi^2(m)$ sú nezávislé náhodné premenné (Obrázok 32).

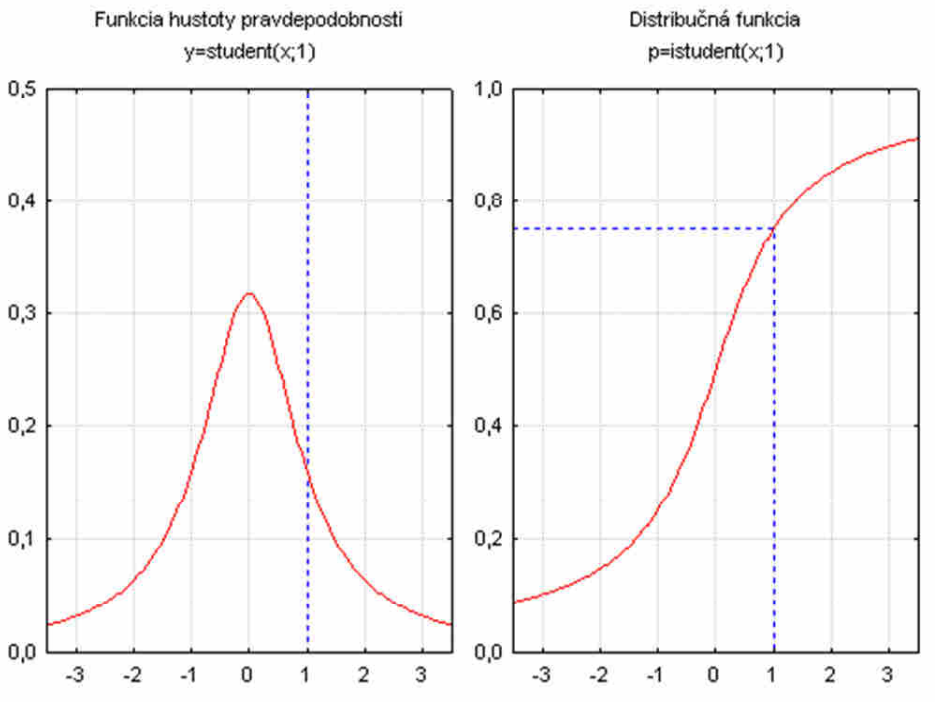


Obrázok 32 F-rozdelenie F(10, 10), p = 0,75

(Studentovo) t-rozdelenie s *k* stupňami voľnosti:

$$T_k = \frac{Z}{\sqrt{X_k^2 / k}} \sim t(k),$$

kde $X_k^2 \sim \chi^2(k)$, $Z \sim N(0, 1)$ sú nezávislé náhodné premenné (Obrázok 33).



Obrázok 33 t-rozdelenie t(1), p = 0,75

Zrejme platí, že $Z^2 \sim \chi^2(1)$, $(T_k)^2 \sim F(1,k)$.

Tabuľka 14 Prehľad rozdelení (Zvára a Štěpán, 2002)

Rozdelenie	**Označenie**	**Kritická hodnota**	**Stredná hodnota**	**Rozptyl**
Normálne	$N(\mu, \sigma^2)$	$P(Z > z(p)) = p$	μ	σ^2
Štandardizované normálne	$N(0, 1)$	$P(Z > z(p)) = p$	0	1
Chí-kvadrát	$\chi^2(k)$	$P(X_k^2 > \chi_k^2(p)) = p$	k	$2k$
Studentovo t	$T(k)$	$P(\lvert T_k \rvert > t_k(p)) = p$	0 ($k > 1$)	$\frac{k}{k-2}$ ($k > 2$)
Fisherovo F	$F(k, m)$	$P(F_{k,m} > F_{k,m}(p)) = p$	$\frac{m}{m-2}$ ($m > 2$)	$\frac{2m^2(k+m-2)}{k(m-2)^2(m-4)}$ ($m > 4$)

Nie len samotné dáta sú premenlivé, ale aj vypočítané štatistiky sú od výberu k výberu náhodne premenlivé. Premenlivosť vypočítaných charakteristík zachytávame často jedným parametrom – štandardnou chybou odhadu, ktorá je odvodená od smerodajnej odchýlky. Nemeria rozptýlenosť pôvodnej náhodnej premennej, ale rozptýlenosť vypočítanej štatistiky.

Výberové rozdelenie štatistiky je pravdepodobnostné rozdelenie hodnôt, ktoré štatistika nadobúda vo všetkých možných výberoch o danom rozsahu zo špecifikovanej populácie.

Príklady použitia rozdelení (Tabuľka 14) k popisu náhodného správania štatistík v prípade, že môžeme predpokladať normálne rozdelenie základného súboru alebo v prípade dostatočne veľkého rozsahu výberového súboru:

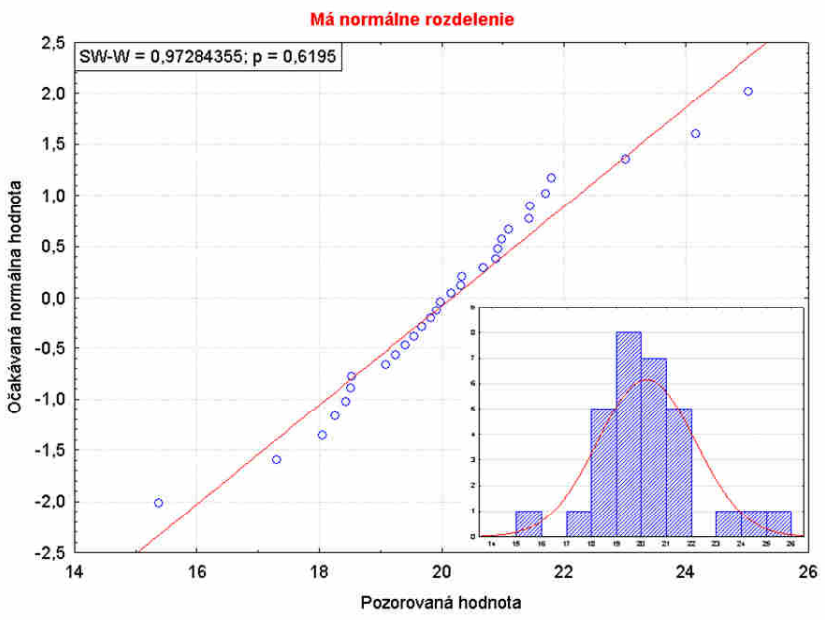
- normálne rozdelenie popisuje výberové rozdelenie aritmetického priemeru pri známom parametri σ,
- t-rozdelenie popisuje výberové rozdelenie aritmetického priemeru pri neznámom parametri σ,
- chí-kvadrát rozdelenie popisuje výberové rozdelenie rozptylu,
- F-rozdelenie popisuje výberové rozdelenie pomeru rozptylov z dvoch nezávislých náhodných výberov.

3.1.4 Overovanie predpokladu normality

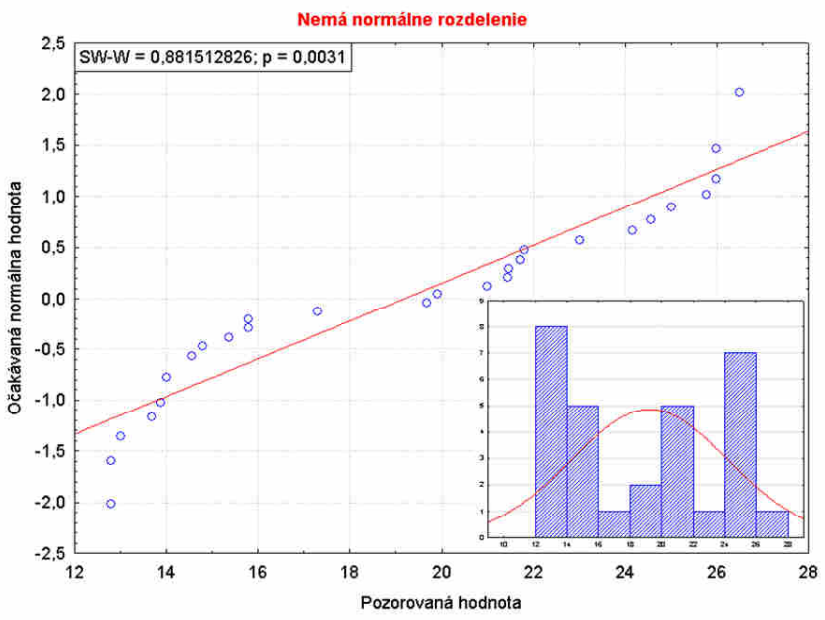
To, či premenná má normálne rozdelenie, prezrádzajú už samotné popisné charakteristiky. V prípade, že premenná pochádza z normálneho rozdelenia, priemer, medián a modus sú približne rovnaké hodnoty. V tomto prípade priemer rozdeľuje dáta na dve rovnaké polovice, t.j. rozdelenie je symetrické - koeficient šikmosti je približne rovný nule. V prípade, že aj koeficient špicatosti je približne rovný nule, rozdelenie náhodnej premennej je špicaté rovnako ako normálne rozdelenie.

K overeniu normality môžeme použiť aj testy rozdelenia, kde na základe výsledkov zamietneme alebo nezamietneme hypotézu, ktorá tvrdí, že hodnoty premennej pochádzajú z predpokladaného rozdelenia (Obrázok 34, Obrázok 35).

Viac informácií ako testy nám ponúknu grafy. Vizualizáciou rozdelenia môžeme identifikovať odchýlky od normality, extrémne hodnoty, bimodalitu a pod. K vizualizácii môžeme použiť histogram preložený očakávanými normálnymi hodnotami - Gaussovou krivkou alebo normálny pravdepodobnostný graf (Obrázok 34, Obrázok 35).



Obrázok 34 Hodnoty premennej pochádzajú z normálneho rozdelenia



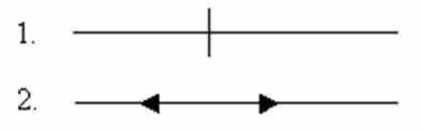
Obrázok 35 Hodnoty premennej nepochádzajú z normálneho rozdelenia

Normálny pravdepodobnostný graf porovnáva rozdelenie premenných s normálnym rozdelením pravdepodobnosti. Čím sú body bližšie k priamke, tým premenné lepšie zodpovedajú normálnemu rozdeleniu. Z histogramu zas môžeme zistiť, že ak pozorované hodnoty majú normálne rozdelenie, tak histogram má tvar zvonu.

3.2 Odhady parametrov

Parameter je číselná hodnota, ktorá platí pre celú populáciu. Odhad parametru získame z výberu z populácie.

Odhad realizujeme buď jedinou hodnotou – bodový odhad alebo číselným intervalom – intervalový odhad (Obrázok 36), ktorý pokrýva teoretickú hodnotu parametra s určitou spoľahlivosťou.



Obrázok 36 Bodový (1.) a intervalový (2.) odhad

3.2.1 Bodový odhad

Parametre a ich bodové odhady (Tabuľka 15) odlišujeme iným značením (pre parametre používame písmená gréckej abecedy). Bodové odhady nazývame aj výberové štatistiky.

Tabuľka 15 Označenie teoretických a výberových charakteristík

	parameter	**odhad**
stredná hodnota	μ	$\bar{x}$
Medián	$\tilde{\mu}$	$\tilde{x}$
Modus	$\hat{\mu}$	$\hat{x}$
smerodajná odchýlka	σ	s
Rozptyl	σ^2	s^2
pravdepodobnosť	π	p
korelačný koeficient	ρ	r
regresný koeficient	β	b

Najlepším bodovým odhadom strednej hodnoty populácie μ a rozptylu populácie σ^2 je výberový priemer $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ a výberový rozptyl $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$, kde *n* je rozsah výberu a $x_1, x_2, \ldots, x_n$ je výberový súbor.

Ak je populácia konečná s rozsahom *N*, tak $\bar{x}$ je odhadom strednej hodnoty $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ a s^2 je odhadom rozptylu $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$.

3.2.2 Intervalový odhad

Intervalový odhad predstavuje interval možných odhadov parametra populácie.

Dĺžka **intervalu spoľahlivosti** je celkovou mierou nepresnosti vytvoreného odhadu. Krátke intervaly spoľahlivosti sú presnejšie než dlhé. Dĺžka intervalu spoľahlivosti závisí nie len od rozsahu výberu *n*, ale aj na hladine spoľahlivosti, s ktorou ho určujeme. Hladina spoľahlivosti je pravdepodobnosť, s akou sa odhadovaný populačný parameter ocitne v danom intervale pri opakovanom prevádzaní výberu. Keď napr. pracujeme s 95% hladinou spoľahlivosti, znamená to, že zo 100 vytvorených intervalov ich približne 95 prekryje hľadanú hodnotu parametra.

Hodnota **smerodajnej (štandardnej) chyby** sa zmenšuje, ak sa veľkosť vzorky zväčšuje.

Tabuľka 16 Smerodajné chyby vybraných výberových štatistík

Výberová štatistika	**Smerodajná chyba**
priemer	$\frac{s}{\sqrt{n}}$
medián	$\frac{1{,}25s}{\sqrt{n}}$
smerodajná odchýlka	$s\sqrt{2n}$
šikmosť	$\sqrt{6/n}$
špicatosť	$\sqrt{24/n}$

Všeobecný vzorec pre získanie asymptoticky platného intervalu spoľahlivosti:

bodový odhad $\pm$ *koeficient spoľahlivosti pre danú hladinu* $\times$ *smerodajná chyba odhadu.*

Napríklad:

95% interval spoľahlivosti pre parameter μ pri známom σ:

$$\bar{x} \pm 1{,}96\sigma_{\bar{x}}.$$

3.3 Testovanie hypotéz

Procedúru **testovania hypotéz** môžeme rozložiť do nasledujúcich krokov:

1. *Formulácia výskumnej otázky vo forme nulovej a alternatívnej štatistickej hypotézy.* Formulácia testovacieho problému súvisí s hypotézou výskumu a s výberom štatistickej

metódy, pomocou ktorej sa bude platnosť nulovej hypotézy overovať. Po výbere metódy je potrebné overiť jej predpoklady (validitu štatistickej metódy).

Nulová hypotéza H0 je tvrdenie, ktoré obvykle deklaruje „*žiadny rozdiel*“ alebo „*nezávislosť*“ medzi premennými.

Alternatívna hypotéza H1 znamená situáciu, kedy nulová hypotéza neplatí. Obvykle deklaruje „*existenciu rozdielu*“ alebo „*existenciu závislosti*“ medzi premennými.

Alternatívne hypotézy môžu byť jednostranné (Obrázok 37) alebo obojstranné (Obrázok 38).

Napríklad:

H1: $\mu > 100$, je jednostranne formulovaná,

H1: $\mu \neq 100$, je obojstranne formulovaná.

2. *Voľba prijateľnej úrovne chyby rozhodovania.*

Tabuľka 17 Schéma testovania hypotéz

Schéma testovania		**Záver testu**	
		H0 platí	**H0 neplatí**
Skutočnosť	**H0 platí**	správny	chyba I. druhu
	H0 neplatí	chyba II. druhu	správny

A. Chyba I. druhu – nulová hypotéza platí, ale zamietne sa.

P(chyba I. druhu) = α *= „hladina významnosti“*

P(neurobíme chybu I. druhu) = 1 - α *= „spoľahlivosť“*

Konvenčná hladina pre spoľahlivosť je 0,95 alebo 0,99.

B. Chyba II. druhu – nulová hypotéza neplatí, ale príjme sa.

P(chyba II. druhu) = β

P(neurobíme chybu II. druhu) = 1 - β *= „sila testu“*

Konvenčná hladina pre silu testu je 0,8 alebo 0,9.

3. *Vypočítanie testovacej štatistiky.*

Existuje veľa testovacích štatistík, výpočet závisí na povahe dát a hypotéz. Výberom štatistickej metódy je testovacia štatistika určená jednoznačne. Jej hodnota sa vypočíta zo získaných dát. Pri použití softvéru nepotrebujeme poznať algoritmus výpočtu, ale podstatu metódy.

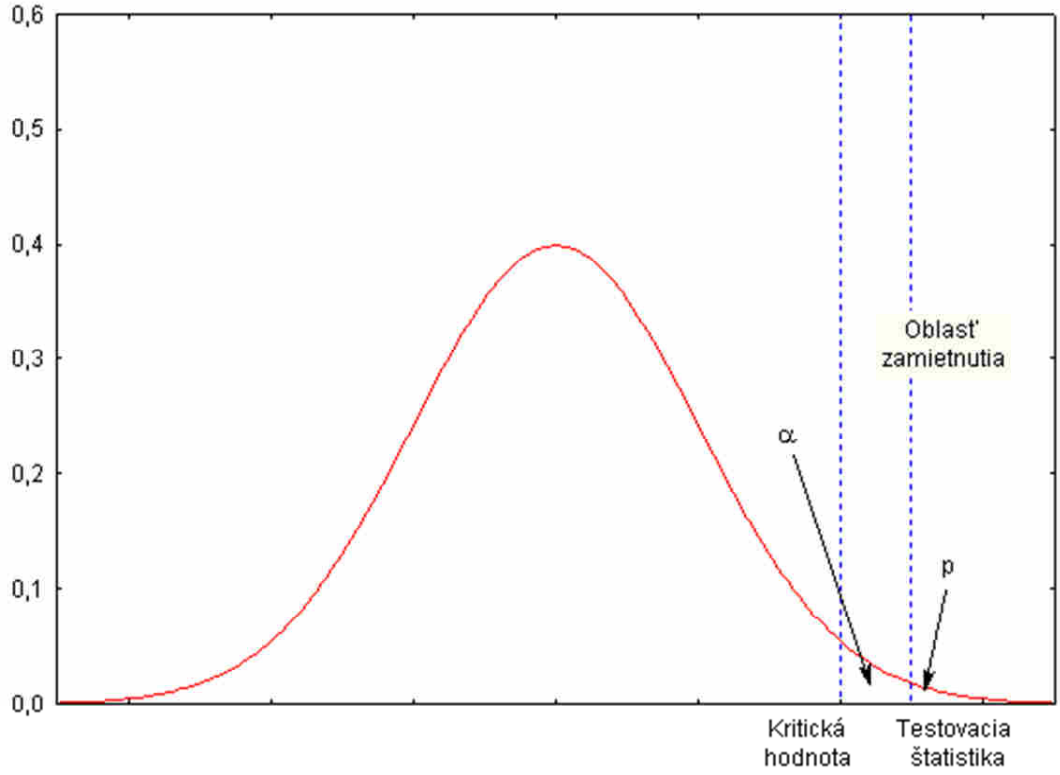
V mnohých prípadoch sa používa ako testovacia štatistika štandardizovaná vzdialenosť odhadu od nulovej hypotézy. Testovacia štatistika má v tomto prípade všeobecný tvar:

Testovacia štatistika = (bodový odhad – hypotetická hodnota)/smerodajná chyba odhadu.

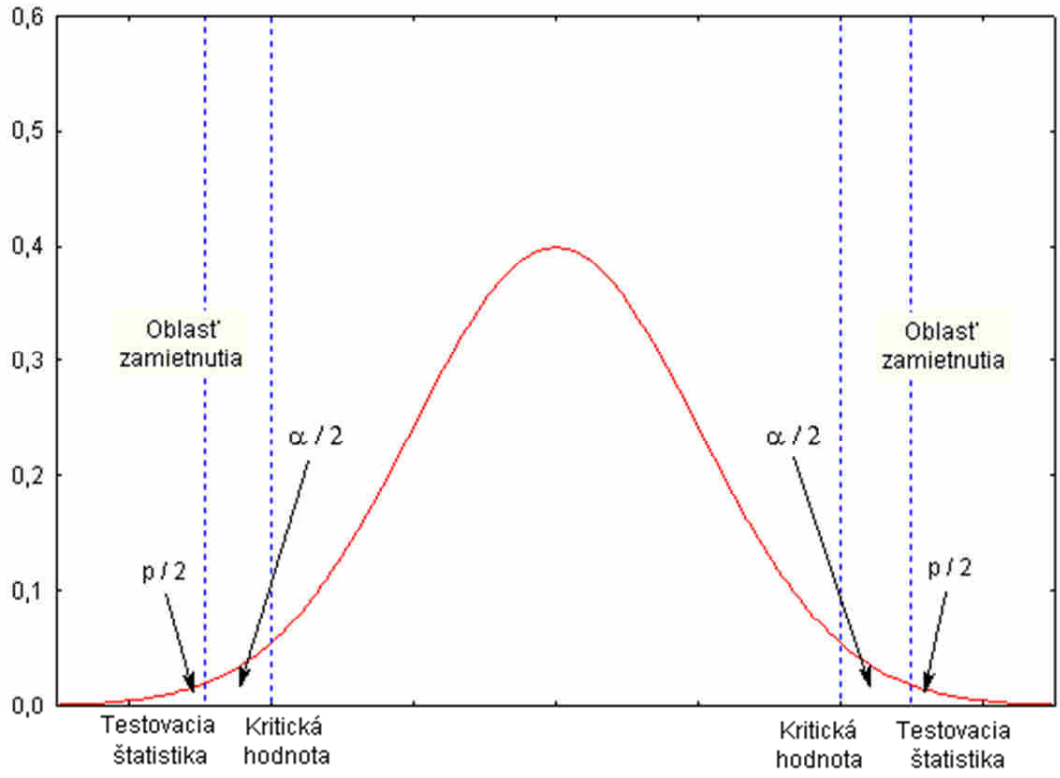
4. *Odporučenie.*

   Testovaciu štatistiku porovnáme s kritickou hodnotou alebo ju prevedieme do pravdepodobnostnej škály na tzv. hodnotu významnosti *p* (*p*-hodnota, hodnota pravdepodobnosti). Hodnota významnosti *p* predstavuje najmenšiu hladinu významnosti, na ktorej môžeme zamietnuť nulovú hypotézu. Pri využívaní štatistického softvéru formulujeme záver testovania pomocou *p*. Ak je *p* rovné alebo menšie ako zvolená hladina významnosti α, tak nulovú hypotézu zamietame. Inak nulovú hypotézu nezamietame.

   Druhý spôsob spočíva v priamom porovnaní testovacej štatistiky s kritickou hodnotou, ktorá sa určuje v závislosti na zvolenej hladine významnosti. Kritická hodnota určuje kritickú oblasť, resp. oblasť zamietnutia (Obrázok 37, Obrázok 38). Ak sa hodnota testovacej štatistiky nachádza vo vnútri kritickej oblasti, zamietame nulovú hypotézu. Samozrejme, ak testovacia štatistika je vo vnútri kritickej oblasti, potom hodnota významnosti *p* je menšia ako príslušná hladina významnosti.



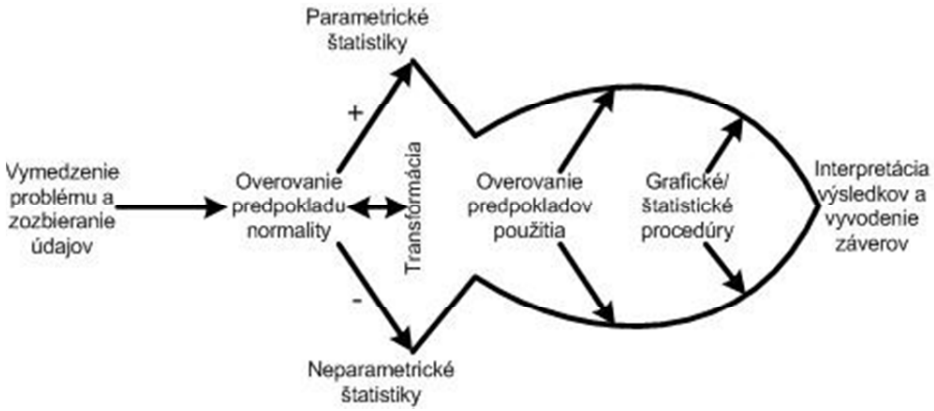
Obrázok 37 Rozdelenie testovacej štatistiky a plochy odpovedajúcej štatistickej významnosti jednostranného testu



Obrázok 38 Rozdelenie testovacej štatistiky obojstranného testu

Aby sme vypočítali hodnotu významnosti *p* obojstranného testu, testujeme nulovú hypotézu proti obojstrannej alternatívnej hypotéze, musíme uvažovať obidva konce rozdelenia štatistiky.

3.4 Parametrické verzus neparametrické testy



Obrázok 39 Parametrické verzus neparametrické testy

Vymedzenie problému a zozbieranie údajov

Pre používanie parametrických alebo neparametrických štatistík sa nevyžaduje odlišná príprava dát. Ku každému parametrickému testu existuje niekoľko neparametrických alternatív.

Overenie predpokladu normality a transformácia

V prípade, že rozdelenie skúmanej premennej nie je normálne, využívajú sa na testovanie neparametrické postupy. Neparametrické testy môžeme použiť prakticky na ľubovoľnú situáciu. Na druhej strane cenou za túto univerzálnosť sú slabšie výsledky testu (nižšia sila testu) ako pri parametrickom teste, ktorý využíva aj informácie o rozdelení. Preto sa snažíme, ak je to možné vždy použiť parametrické testy. To znamená, že ak skúmaná premenná nemá normálne rozdelenie, tak ju transformujeme a následne znovu overujeme predpoklad normality (Obrázok 39).

Parametrické štatistiky

Pri použití parametrických štatistík sa predpokladá, že typ rozdelenia základného súboru je známy. Zvyčajne sa predpokladá, že rozdelenie je normálne. Ak sú vzorky dostatočne veľké, odchýlka od normality nie je veľmi podstatná, pretože na základe centrálnej limitnej vety má

aritmetický priemer približne normálne rozdelenie. Niektorí autori uvádzajú, že normalitu je nutné overovať iba pri malých vzorkách s rozsahom menším ako 30 (Chajdiak, Rublíková a Gudába, 1994; Chajdiak, Komorník a Komorníková, 1999; Hill a Lewicki, 2007).

Neparametrické štatistiky

Neparametrické štatistiky sa nespoliehajú na odhad parametrov (priemeru alebo smerodajnej odchýlky) popisujúcich rozdelenie premennej v základnom súbore, ale pracujú s početnosťami (napr. Chí-kvadrát test) alebo s poradovými číslami, ktoré boli pridelené pôvodným údajom (napr. Kruskal-Wallisov test).

Overovanie predpokladov použitia

Zo schémy (Obrázok 39) vidíme, že od predpokladu normality sa nám línia postupu rozvetvila, to znamená, že iné predpoklady použitia platia pre parametrické a iné pre neparametrické testy. Tu si treba uvedomiť, že určité predpoklady musia spĺňať aj neparametrické štatistiky. Napríklad neparametrický chí-kvadrát test sa dá použiť iba za predpokladu, že očakávané početnosti sú väčšie alebo rovné ako 5.

Grafické/štatistické procedúry

V tomto prípade tiež dve odlišné línie naznačujú, že ide o rôzne algoritmy výpočtu.

Interpretácia výsledkov a vyvodenie záverov

Interpretácia v obidvoch prípadoch je totožná (za predpokladu, že sme použili ekvivalentné testy) napriek tomu, že ide o dva kategoricky odlišné algoritmy výpočtu.

3.5 Závislé verzus nezávislé vzorky

V štatistickej analýze musíme rozlišovať závislé a nezávislé vzorky, napr. iné testy sa používajú na testovanie rozdielov medzi závislými a iné medzi nezávislými vzorkami.

Závislé vzorky (Tabuľka 18) predstavujú merania na tých istých objektoch, napríklad by sme chceli porovnať vedomosti študentov z matematiky na začiatku a na konci semestra. Nezávislé vzorky (Tabuľka 18) predstavujú merania na rôznych objektoch, napríklad by sme chceli porovnať vedomosti študentov z matematiky u mužov a žien.

Testy rozdielov medzi nezávislými vzorkami sa často označujú ako testy medzi skupinami a medzi závislými vzorkami ako testy medzi premennými. Tieto názvy vyplývajú zo zobrazenia vstupnej dátovej tabuľky.

Tabuľka 18 Premenné verzus skupiny

Závislé vzorky

premenné

1.meranie	2.meranie
25	40
30	35
15	40
35	40
⋮	⋮
20	30
35	40
30	40
15	25

Nezávislé vzorky

meranie	pohlavie	
25	m	Skupina muži
30	m	
15	m	
35	m	
⋮	⋮	
20	ž	Skupina ženy
35	ž	
30	ž	
15	ž	

Treba však podotknúť, že tento zápis nebýva vždy dodržaný, napr. skupiny často bývajú zapísané v stĺpcoch. Buď sami vytvoríme neštandardnú vstupnú maticu, respektíve niektoré programy pre vybrané analýzy vyžadujú z nepochopiteľných dôvodov mať každú nezávislú vzorku v jednom stĺpci tabuľky.

3.6 Jednorozmerné verzus viacnásobné a viacrozmerné analýzy

Najskôr si stručne vysvetlíme túto terminológiu na miere závislosti medzi premennými.

Ak zisťujeme mieru závislosti medzi dvoma premennými, tak sa jedná o **jednorozmernú/univariačnú analýzu (univariate analysis)**.

V prípade, že zisťujeme mieru závislosti medzi premennou *Y* a premennými *X*, *Z* už hovoríme o **viacnásobnej analýze (multiple analysis)**.

Ak však zisťujeme mieru závislosti medzi dvoma skupinami premenných, tak hovoríme o **viacrozmernej/multivariačnej analýze (multivariate analysis)**.

V prvých dvoch prípadoch by sme použili korelačnú analýzu, konkrétne v prvom korelačný koeficient a v druhom viacnásobný korelačný koeficient. V treťom prípade by sme použili kanonickú analýzu, ktorá patrí medzi viacrozmerné prieskumné techniky.

Ak koeficient korelácie premenných *Y* a *X* je štatisticky významný, potom nás zaujíma odhad lineárneho regresného modelu zobrazujúci lineárnu závislosť medzi premennými *Y* a *X*. Určenie regresného modelu je dôležité z hľadiska predvídania veľkosti hodnoty závislej (vysvetľovanej) premennej *Y* pri určitej známej hodnote nezávislej (vysvetľujúcej) premennej *X*. V tomto prípade by sme hovorili o jednorozmernej (párovej) lineárnej regresii. Ak však by nás zaujímal odhad lineárneho regresného modelu zobrazujúci závislosť medzi závislou premennou *Y* a nezávislými premennými *X* a *Z*, hovorili by sme o viacnásobnej lineárnej regresii a nie o viacrozmernej.

Dosť často sa tieto dva termíny zamieňajú v lokalizovaných štatistických programoch, konkrétne multiple regression sa často prekladá ako viacrozmerná a nie viacnásobná regresia.

Rozdiely medzi jednorozmernou, viacnásobnou a viacrozmernou analýzou ilustrujeme na analýze rozptylu:

- pri jednorozmernej analýze vystupuje jedna závislá a jedna nezávislá premenná (napr. analýza rozptylu jednoduchého triedenia (one-way analysis of variance/ANOVA) – skúma závislosť jednej závislej kvantitatívnej premennej od jednej nezávislej nominálnej premennej),
- pri viacnásobnej analýze vystupuje jedna závislá a viac nezávislých premenných (napr. viacfaktorová analýza rozptylu (multi-way analysis of variance/ANOVA) – skúma závislosť jednej závislej kvantitatívnej premennej od viacerých nezávislých nominálnych premenných),
- pri viacrozmernej analýze vystupuje viac závislých premenných (napr. viacrozmerná analýza rozptylu (multivariate analysis of variance/MANOVA) – skúma závislosť viacerých závislých kvantitatívnych premenných od jednej, resp. viacerých nezávislých nominálnych premenných).