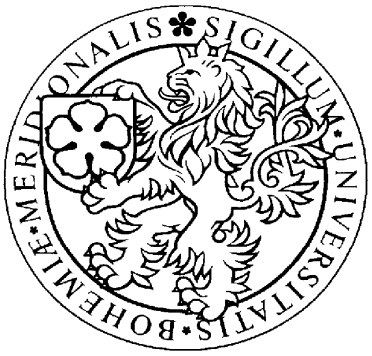
JIHOČESKÁ UNIVERZITA
V ČESKÝCH BUDĚJOVICÍCH
Zdravotně sociální fakulta



BIOSTATISTIKA

doplňkové texty pro posluchače kombinované formy studia
studijního programu „Ochrana obyvatelstva“

studijního oboru „Ochrana obyvatelstva se zaměřením na CBRNE“

Ing. Ladislav Beránek, CSc. MBA.

Mgr. Renata Havránková

ČESKÉ BUDĚJOVICE 2007

Obsah

1. Počet pravděpodobnosti

Klíčová slova: Klasické a statistické pojetí pravděpodobnosti. Náhodný jev, některé elementární poučky pro počítání s pravděpodobnostmi. Podmíněná pravděpodobnost, nezávislost náhodných jevů.

Základní pojmy

Náhodný jev

= jev, který za daných podmínek nastat může a nemusí; jeho nastání je věc náhody
= výsledek náhodného pokusu
- je to výchozí pojem počtu pravděpodobnosti a označujeme ho A, B, C, …

Jev jistý

= jev, který za daných podmínek nastane vždy

Jev nemožný

= jev, který za daných podmínek nastat nemůže

Elementární náhodný jev

= jev, který se nedá dále rozdělit na podrobnější jevy
= konečný jev

Operace s náhodnými jevy

(1) Jestliže při každé realizaci jevu A nastává i jev B, pak říkáme, že jev A má za následek jev B neboli jev A je ***částí*** jevu B.
$A \subset B$

(2) Jevy A a B jsou ***rovnocenné***, jestliže pokaždé, kdy nastal jev A, nastal také jev B a naopak.
$A = B$

(3) Jev spočívající v nastoupení jak jevu A, tak jevu B nazýváme ***průnikem*** jevů A a B.
$A \cap B$ $(A * B)$

(4) Jev spočívající v nastoupení alespoň jednoho z jevů A a B nazýváme ***sjednocení*** jevů A a B.
$A \cup B$ $(A + B)$

(5) ***Rozdílem*** jevů A a B nazýváme jev spočívající v nastoupení jevu A a současném nenastoupení jevu B.
$A - B$

(6) Jev, který spočívá v nenastoupení jevu A, je ***jevem opačným*** k jevu A.
$\overline{A}$

(7) Jevy A a B se nazývají ***neslučitelné***, jestliže výskyt jednoho z nich bude vylučovat možnost výskytu druhého jevu, tj. jejich průnik je jev nemožný.
$A \cap B = \varnothing$

Definice pravděpodobnosti

Klasická definice pravděpodobnosti

Podle klasické definice pravděpodobnosti nastání jevu A je dáno poměrem $\underline{m}$ ku $\underline{n}$, kde $\underline{m}$ je počet všech situací příznivých jevu A a $\underline{n}$ je počet všech možných situací, přičemž $\underline{n}$ musí být konečné číslo a předpokládá se, že každá z celkového počtu situací má stejnou šanci nastat.

Statistická definice pravděpodobnosti

V některých případech není splněn základní požadavek klasické definice pravděpodobnosti, tj. předpoklad stejné možnosti všech jevů.

U statistické definice je pravděpodobnost nastání jevu A přibližně rovna poměru m / n, přičemž $\underline{m}$ je počet situací, v nichž reálně nastal jev A a $\underline{n}$ je počet všech uskutečněných pokusů.

Př.: pravděpodobnost narození syna
- dle klasické definice: 50 %
- dle statistické definice: 52 % (rodí se více mužů)

Pravidla pro počítání s pravděpodobností

Náhodné jevy

neslučitelné
- $P(A \cap B) = 0$ … průnik
- $P(A \cup B) = P(A) + P(B)$ … sjednocení

slučitelné
- nezávislé … s opakováním
  - $P(A \cap B) = P(A) * P(B)$ … průnik
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ … sjednocení
- závislé … bez opakování
  - $P(A \cap B) = P(A) * P(B/A)$ … průnik
    nebo $= P(B) * P(A/B)$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ … sjednocení

Jevy neslučitelné

= nemohou nastat současně

Podmíněná pravděpodobnost

= P(B/A) … pravděpodobnost nastoupení jevu B za předpokladu, že nastal jev A
= P(A/B) … pravděpodobnost nastoupení jevu A za předpokladu, že nastal jev B

Jevy nezávislé

= jevy A a B jsou nezávislé, jestliže pravděpodobnosti nastoupení nebo nenastoupení jednoho z jevů neovlivňuje pravděpodobnost nastoupení nebo nenastoupení jevu druhého

Jevy závislé

= nastoupení jevu A ovlivňuje jevy další

OTÁZKY A PŘÍKLADY

1. Co je to náhodný jev
2. Co je to podmíněná pravděpodobnost
3. Jaká jsou základní pravidla pro počítání s pravděpodobností

2. Náhodná veličina

Klíčová slova: Náhodná veličina a její rozdělení. Pravděpodobnostní funkce, distribuční funkce, hustota pravděpodobnosti. Charakteristiky a kvantily náhodné veličiny.

Náhodná veličina

= veličina, jejíž hodnota je jednoznačně určena výsledkem náhodného pokusu

- obecně se značí X, Y, Z …, přičemž její konkrétní hodnoty jsou značeny x, y, z …
- 2 náhodné veličiny považujeme za nezávislé, jestliže zákon rozdělení jedné z nich nezávisí na hodnotě 2. náhodné veličiny

nespojitá (diskrétní)

- nabývá izolovaných celočíselných hodnot
- může nabývat pouze některých obměn vyplývajících z povahy věci
- např. hody na kostce: pouze 1, 2, 3, 4, 5, 6; nikoli: 1,7; 2,3 atp.
- Popisuje ji:

1. pravděpodobnostní funkce
2. distribuční funkce
3. tabulka, grafy

spojitá

- může nabývat všech hodnot z určitého intervalu reálných čísel
- Popisuje ji:

1. hustota pravděpodobnosti
2. distribuční funkce
3. graf, POZOR!!! ne tabulka

Rozdělení náhodné veličiny

- pravidlo, jež každé hodnotě nebo množině hodnot z každého intervalu přiřazuje pravděpodobnost, že náhodná veličina nabude této hodnoty nebo hodnoty z určitého intervalu se nazývá *zákon rozdělení náhodné veličiny*
- Zdrojem může být *funkční předpis* nebo *zobecněná zkušenost* a formou pak je *vzorec*, *tabulka* či *graf*.
- nejjednodušším způsobem zadání zákona rozdělení nespojité veličiny je tzv. řada rozdělení, což je tabulka, ve které jsou k možným hodnotám x udány pravděpodobnosti P (x)
- velmi názornou představu o rozdělení získáme tím, že ji graficky zobrazíme →polygon rozdělení pravděpodobností

NESPOJITÁ NÁHODNÁ VELIČINA

Pravděpodobnostní funkce

- k pravděpodobnostní funkci vede Bernoulliho schéma, které se týká opakovaně prováděných nezávislých pokusů
- Vzorec:

$$P(x) = p^x \ (1-p)^{n-x}$$

kde p … je podíl sledovaného jevu na celku (př. podíl žen v podniku)
n … je počet pokusů (př. velikost vzorku)

- je to fce, která každému x přiřazuje pravděpodobnost, že náhodná veličina X nabude této hodnoty a platí $P(x) = P(X = x)$, přičemž $\sum P(x) = 1$
- pomocí této P. fce můžeme stanovit i P, že náhodná veličina nabude hodnoty z intervalu <x1;x2> a platí $P(x_1 \le X \le x_2) = \sum P(x)$

Distribuční funkce (u nespojité náhodné veličiny)

- udává pravděpodobnost, že náhodná veličina X nabude hodnoty menší (nebo rovné) než právě zvolené x.
- Vzorec:

$$F(x) = P(X \le x)$$

- př.: $F(0) = P(X \le 0)$; $F(1) = P(X \le 1)$; $F(2) = P(X \le 2)$ …

- z toho vyplývá, že distribuční funkce je součtem všech pravděpodobnostních funkcí tvořených z čísel menších nebo rovných danému *x*

Graf

- pravděpodobnostní funkce = polygon
- distribuční funkce = histogram

SPOJITÁ NÁHODNÁ VELIČINA

Hustota pravděpodobnosti

- Vzorec:

$$f(x) = F'(x)$$

- Vlastnosti:
  1. $f(x) \geq 0$
  2. $\int f(x)\,dx = 1$ … definiční obor
- Platí zde:

$$P(x_1 \leq X < x_2) = \int f(x)\,dx = F(x_2) - F(x_1)$$

- je definována jako limita pravděpodobnosti, že veličina X padne do velmi malého intervalu vydělená délkou tohoto intervalu v případě, že se tato délka blíží nule

Distribuční funkce (u spojité náhodné veličiny)

- Vzorec:

$$F(x) = P(X \leq x) = \int f(t)\,dt$$

- počítám plochu pod křivkou té spojité funkce - integrál

Charakteristiky náhodné veličiny

Distribuční funkce podává o náhodné veličině úplnou informaci. Známe-li tuto funkci, víme, jakých hodnot může uvažovaná veličina nabývat a jaké jsou pravděpodobnosti, které těmto hodnotám odpovídají. V praxi však často potřebujeme nějaké koncentrovanější a přehlednější vyjádření této informace. K takovému zestručněnému popisu užíváme číselné hodnoty, které nazýváme *charakteristiky náhodných veličin*.

Nejčastější jsou:

- ***střední hodnota … E(X)***
  - pro nespojité náhodné veličiny
    $E(X) = \sum x\,P(x)$
  - pro spojité náhodné veličiny
    $E(X) = \int x\,f(x)dx$

1. střední hodnota konstanty je rovna konstantě
2. střední hodnota součinu konstanty a náhodné veličiny je rovna součinu této konstanty a střední hodnoty dané veličiny
3. střední hodnota součtu *s* náhodných veličin je rovna součtu jejich středních hodnot
4. střední hodnota součinu *s* nezávislých náhodných veličin je rovna součinu jejich středních hodnot

- ***rozptyl … $D(X) = E(X^2) - E^2(X)$***
  - pro nespojité náhodné veličiny
    $D(X) = \sum [x - E(X)]^2 * P(x) = \sum x^2 P(x) - [\sum xP(x)]^2$
  - pro spojité náhodné veličiny
    $D(X) = \int [x - E(X)]^2 * f(x)dx = \int x^2 f(x)dx - [\int x\,f(x)]^2$

1. rozptyl konstanty je roven nule
2. D(cX) = c2 D(X)
3. D(c+X) = D(X)
4. rozptyl součtu s nezávislých náhodných veličin se rovná součtu rozptylů těchto náhodných veličin

Kovariance je střední hodnota součinu odchylek obou náhodných veličin X,Y od jejich středních hodnot.

Koeficient korelace je poměr kovariance k součinu směrodatných odchylek obou náhodných veličin, který měří těsnost závislosti obou náhodných veličin a nabývá hodnot z intervalu <-1;1>

OTÁZKY A PŘÍKLADY

1. Co je to náhodná veličina
2. Jaké jsou základní charakteristiky náhodné veličiny
3. Co je to kovariance

3. Rozdělení náhodných veličin

Klíčová slova: Základní modely rozdělení nespojitých a spojitých náhodných veličin. Zákon velkých čísel a limitní věty.

Počet stupňů volnosti je počet nezávislých sčítanců.

Nejčastější rozdělení:

1. nespojité
   - **alternativní** – X nabývá pouze 2 hodnot, 0(nenastane-li sledovaný jev) a 1 (nastane-li sledovaný jev)
   - **binomické** – jev A může nastat s pravděpodobností p a nenastat s pravděpodobností 1-p. je obecně asymetrické. S růstem n nebo přibližováním p k hodnotě 0,5 se stává postupně symetričtějším, při p=0,5 je již symetrické
   - **Poissonovo** – pouze pokud rozsah výběru je dosti velký (n>30) a pravděpodobnost p je velmi malá, pak se poissonovo rozdělení používá jako aproximace binomického rozdělení. Střední hodnota a rozptyl jsou stejné – jsou rovny parametru rozdělení. Řídí počet jevů v prostorové jednotce nebo počet událostí v časové jednotce. Má pouze jeden parametr
   - **hypergeometrické** při výběru bez vracení. (jednotlivé pokusy jsou závislé). Jestliže rozsah N je velký a n a M/N se nemění, blíží se hypergeometrické rozdělení binomickému. To znamená, že pro velká N můžeme zanedbat rozdíl mezi výběrem bez vracení a s vracením. Prakticky: je-li tzv. výběrový poměr n/N menší než 0,05, lze hypergeometrické rozdělení nahradit rozdělením binomickým s parametry n a M/N.

2. spojité
   - **normální** – nejdůležitější, klasickým příkladem je rozdělení náhodných chyb, vzniklých při měření nějaké veličiny, má tvar zvonkovité křivky, která nabývá maxima v bodě $x = \mu$. Je vhodným pr. rozdělením tehdy, působí-li na kolísání náhodné veličiny velký počet nepatrných a vzájemně nezávislých vlivů. Pro jednodušší výpočet distribuční fce transformujeme náhodnou veličinu na normovanou veličinu $\Rightarrow$ normované normální rozdělení
   - **logaritmicko - normální** – jestliže má náhodná veličina Y = ln X normální rozdělení, pak veličina X má log normální rozdělení. Význam při modelování příjmových rozdělení
   - **exponenciální** – uplatnění v teorii spolehlivosti a v teorii hromadné obsluhy (životnost výrobků (doba mezi realizacemi 2 po sobě jdoucích náhodných jevů)
   - χ^2 – rozdělení – Jsou – li U1, U2, …,Uv nezávislé náhodné veličiny, z nichž každá má které mají normované normální rozdělení, pak součet čtverců náhodných veličin má rozdělení χ^2.Se vzrůstajícím počtem stupňů volnosti se blíží normálnímu rozdělení.

- **Studentovo** – skládá se ze 2 nezávislých náhodných veličin, jedna má normální rozdělení a druhá χ^2 s v stupni volnosti. Počet stupňů volnosti je jediným parametrem tohoto rozdělení. Pro v jdoucí k nekonečnu se t blíží normovanému normálnímu rozdělení (při v > 30 ho považujeme už za normální)
- **Fischerovo** - 2 nezávislé náhodné veličiny s rozdělením χ^2 (v1 a v2 stupňů volnosti). Toto rozdělení má dva parametry – stupně volnosti

Zákon velkých čísel – jestliže zvětšujeme počet náhodných pokusů, přibližuje se empirická charakteristika,popisující výsledky těchto pokusů, charakteristice teoretické
Bernoulliho věta: relativní četnost sledovaného jevu stochasticky konverguje k jeho pravděpodobnosti.
Limitní věty
⇨ centrální limitní věta – o náhodných veličinách, jejichž limitním zákonem je normální rozdělení, říkáme, že mají asymptoticky normální rozdělení. Konvergencí pravděpodobnostních rozdělení k normálnímu se zabývá centrální limitní věta. 2 dílčí formulace:
⇨ Věta Moivrova-Laplaceova – vyjadřuje konvergenci binomického rozdělení k rozdělení normálnímu → pro dostatečně velké n lze binomické rozdělení aproximovat normálním rozdělením
⇨ Věta Linderbergova-Lévyho – pro dost velké n má přibližně normální rozdělení i součet a průměr n nezávislých náhodných veličin, které mají stejné (libovolné) rozdělení s konečnou střední hodnotou a konečným rozptylem

OTÁZKY A PŘÍKLADY

1. Co je to počet stupňů volnosti
2. Jaké jsou základní druhy rozdělení

4. Základní pojmy statistiky

Klíčová slova: Statistický znak, statistická jednotka, statistický soubor. Druhy statistických šetření. Tabulky a grafy rozdělení četností.

Statistika se zabývá jevy, které se vyskytují opakovaně – jde o *jevy hromadné*. Ty mohou být *dvojího typu*:

1) jedná se o velký počet opakovaných pozorování jednoho a téhož prvku (např. opakované měření rozměru určité součástky...);
2) druhý typ spočívá ve zkoumání určité jedné vlastnosti velkého počtu různých prvků (např. zisk velkého počtu různých podniků, politické postoje velkého počtu různých osob a řada dalších situací…).

Pojmy:

statistický soubor, základní soubor, výběrový soubor = množina prvků, z nichž každý má celou řadu vlastností. Statistický soubor všech jednotek, který je předmětem sledování, se nazývá základní soubor. Jeho rozsah může být konečný, ale i nekonečný → proto se provádí výběrové šetření, kdy se ze zákl.souboru vyberou jen některé jednotky, čímž se získá výběrový soubor

statistické jednotky	=	jednotlivé prvky statistického souboru; jsou nositeli vlastního daného statistického souboru
rozsah souboru	=	počet jednotek statistického souboru; obvykle se označuje *n* nebo *N*. Zjišťujeme-li u stat.jednotky pouze 1 stat.znak→ jednorozměrný soubor x zjišťujeme-li 2 a více znaků → vícerozměrný soubor
statistický znak *(statistická proměnná)*	=	vyjadřuje vlastnosti statistické jednotky. Je-li konkrétně touto jednotkou např. pracovník podniku, můžeme jeho vlastnosti vyjádřit znaky věk, plat, délka praxe aj.

Druhy statistických znaků

Statistické znaky (proměnné)

kvantitativní (číselné) … počet dětí, známka, výška

- měřitelné (metrické) … počet dětí, výška
  - nespojité … počet dětí
  - spojité … výška
- pořadové (ordinální) … známka

kvalitativní (kategoriální) … vstup do EU (ano x ne), jméno, barva očí

- alternativní … vstup do EU (ano x ne) = pouze 2 alternativy
- množné … jméno, barva očí = více možných odpovědí
  - nespojité … jméno = 2 různá jména; není možný přesun mezi Petrem a Pavlem
  - spojité … barva očí = různé odstíny – modrá-modrošedá-šedá

Pozn.:

- *měřitelný znak* – umožňuje porovnávat nejenom pořadím, ale také rozdílem (poměrem), tj. o kolik, kolikrát
- *pořadové znaky* – umožňují porovnávat pouze pořadím (ten, kdo má dvojku, není 2x horší než ten co má jedničku)
- *nespojité znaky* – nabývají pouze některých hodnot, podle povahy problému (př. 1, 2, 3 děti; ne 2,5); nejčastěji přirozených čísel nebo celých nezáporných čísel; ALE
- *spojité znaky* – nabývají jakýchkoli hodnot v rámci určitého intervalu (př. výška člověka – 180,25 cm apod.)

Druhy statistických šetření

Statistický soubor všech jednotek, který je vlastním předmětem sledování, o němž chceme provádět závěry, se nazývá ***základním souborem***. Jeho rozsah může být konečný i nekonečný, zpravidla je ale velký. Proto se velmi často z úsporných i časových důvodů provádí šetření výběrové, kdy se ze základního souboru určitým způsobem vyberou pouze některé jednotky, čímž se získá ***výběrový soubor***. Z něho získané výsledky pak slouží k provádění úsudku o základním souboru.

Rozdíl mezi základním a výběrovým souborem je pouze otázkou věcného, prostorového a časového vymezení.

Většina statistických souborů, s nimiž se setkáváme v hospodářské praxi, je příznačná svou rozsáhlostí, někdy dokonce i nekonečností. Jakmile jsme tedy postaveni před úkol provést určité statistické šetření a pak analyzovat data z něho získaná, musíme nejprve rozhodnout, zda toto šetření budeme realizovat jako vyčerpávající či výběrové.

99,9 % všech souborů je výběrových.

Pomocí stat. zjišťování získáváme stat. údaje, což jsou číselné nebo slovní obměny sledovaných stat. znaků. Nejprve je nutné stanovit, kdy, kdo a jak bude zjišťování

prováděno → je třeba určit zpravodajskou jednotku, která poskytne potřebné info o stat. jednotce (může i nemusí být totožná s jednotkou statistickou)

Pomocí statistického šetření (zjišťování) získáváme statistické údaje, což jsou číselné nebo slovní obměny sledovaných statistických znaků.

Zjišťované údaje mohou být dvojího typu:

(a) zjišťují se za určitý časový interval
   - musí být stanovena *rozhodná doba* (např. objem produkce za rok 1998, počet zákazníků obsloužených za jeden den…)

(b) zjišťují se k určitému okamžiku
   - musí být stanoven *rozhodný okamžik* (např. počet pracovníků k 31. 10. 1998…)

Dále rozlišujeme, zda jsme využili:

- ***primární data*** = přímé pozorování … přímo pozorujeme statistické jednotky a zjišťované hodnoty znaku získáváme sčítáním, měřením, vážením apod.;
- ***sekundární data*** = dotaz … buď expediční metodou (existují sčítací komisaři či tazatelé apod.) nebo korespondenční metodou, kdy zpravodajská jednotka požadované údaje na předem stanovených formulářích sama sdělí. Stále častěji používanou formou je i telefonické dotazování.

Jiným způsobem zjišťování je forma výkaznictví (výkaz je předem navržený a schválený formulář, který v pravidelných lhůtách předkládá zpravodajská jednotka stat. orgánům). Jejich význam klesá. Kromě výkaznictví se zjišťují údaje ve stat. zvláštními stat. šetřeními → *soupisy (cenzy),* např. soupis zásob; a *znalecký odhad* (subjektivní hodnocení jevu osobami určenými stat. úřady)

OTÁZKY A PŘÍKLADY

1. Zjišťování souvislost typu kuřáka s náchylností ke karcinomu plic

a). Jedná se o šetření nebo o experiment?

[Pokud výzkum probíhá tak, že ve skupině subjektů sledujeme, kdo co kouří a jaká je jeho náchylnost ke karcinomu plic, pak je to šetření]

b) Jak by vypadal takový výzkum, kdyby měl být sestaven jako experiment?

[Museli bychom sami náhodně subjektům určit, kdo bude co kouřit. Následně bychom u nich sledovali náchylnost ke karcinomu plic]

5. Náhodný výběr, typy výběrů

Klíčová slova: Základní soubor (konečný a nekonečný) a náhodný výběr. Varianty náhodného výběru z konečného souboru. Nejdůležitější výběrové charakteristiky a jejich rozdělení.

Většina statistických souborů, s nimiž se setkáváme v hospodářské praxi, je příznačná svou rozsáhlostí, někdy dokonce i nekonečností. Jakmile jsme tedy postaveni před úkol provést určité statistické šetření a pak analyzovat data z něho získaná, musíme nejprve rozhodnout, zda toto šetření budeme realizovat jako vyčerpávající nebo výběrové.

Při ***vyčerpávajícím zjišťování*** se prošetřují veškeré jednotky statistického souboru. Zpravidla se jedná o záležitost mimořádně nákladnou (drahé finančně i na čas, potřeba mnoha pracovníků, dlouhá doba zpracování výsledků …).Takové šetření může provádět pouze stát a jeho statistické orgány (např. sčítání lidu v desetiletých intervalech)

Proto se přistupuje k ***šetření výběrovému***, kdy se vybírají pouze některé jednotky ze základního souboru. Poté z charakteristik pořízených v rámci tohoto šetření můžeme více či méně přesně usuzovat na vlastnosti celého základního souboru.

Nejznámější vyčerpávající šetření je sčítání lidu, volby nebo HDP. Typickým výběrovým šetřením je zase inflace měřená přes CPI (= index spotřebitelských cen).

Typ výběru

reprezentativní výběr (pravděpodobnostní)
- prostý náhodný výběr
- složitější uspořádání

1. oblastní výběr
2. vícestupňový výběr

nereprezentativní výběr
- anketa
- záměrný výběr

1. typický výběr
2. kvótní výběr

Reprezentativní výběr

= výběrový soubor je pravděpodobnostní zmenšeninou základního souboru; tzn. zachovává vlastnosti pravděpodobnostního rozdělení, které platí v základním souboru

Prostý náhodný výběr

3 podmínky:
1. Každý prvek základního souboru má stejnou pravděpodobnost, že bude vybrán
2. Libovolná n-tice prvků ze základního souboru má stejnou šanci dostat se do výběru jako každá jiná n-tice
3. V základním i ve výběrovém souboru musí platit stejné rozdělení pravděpodobnosti, tj. stejný tvar i stejné parametry.

Složitější uspořádání

= mají za cíl sdružovat (homogenizovat) vlastnosti základního soubotu a tím zmenšovat riziko výběrové chyby

Oblastní výběr (stratifikovaný, proporcionální)

Má 2 kroky:
1) rozdělím základní soubor na podskupiny (= strata)
2) vlastní výběr provedu úměrně velikosti těch podskupin

- př. rozdělím ČR na 11 regionů (krajů) – např. severomoravský kraj má 13 % obyvatel

Vícestupňový výběr

Má 2 kroky:
1) rozdělím základní soubor na podskupiny
2) vyberu pouze některé podoblasti a v nich uskutečním náhodný výběr

- nejčastěji dvoustupňový

Nereprezentativní výběr

= výběr, který nezachovává vlastnosti pravděpodobnostního rozdělení, které platí v základním souboru

Anketa

= neadresné šetření s velmi nízkou návratností, nejčastěji prostřednictvím sdělovacích prostředků (tisk, televize); nezávaznost

- nejvyšší možná návratnost je 25 %

Záměrný výběr (úsudkový)

= nejčastější forma dotazování, šetření (v marketingu, ve státní statistice…)

- znalec nebo skupina odborníků na analyzovanou problematiku podle svého nejlepšího uvážení vybere jednotky, o kterých se lze domnívat, že ve svém souhrnu nejlépe umožní provést zkoumání

Typický výběr

- oslovují se jednotky typické pro daný problém na základě segmentace trhu

Kvótní výběr

- nachází uplatnění zejména tam, kde není známa velikost základního souboru (výzkumy veřejného mínění)

Má dvě fáze:

1) základní soubor se rozdělí na oblasti (podskupiny)
2) v oblastech se stanoví záměrně kvóty (psychologicky, sociologicky, demograficky apod.)

Problémy:

- špatná práce tazatelů
- nedostatečná síť agentur
- chyba je, když se vyhodnocují statistickými metodami

Prostý náhodný výběr:

při něm mají nejen všechny jednotky základního souboru stejnou pravděpodobnost, že budou vybrány, ale i všechny myslitelné n-členné kombinace mají stejnou pr. stát se výběrovým souborem

Techniky náhodného vybírání

(a) losování

(b) tabulky náhodných čísel

- obsahují náhodný výběr čísel 0, 1,…, 9 ze souboru, kde má každá číslice stejnou pravděpodobnost výskytu
- jde o sloupce číslic, které je možné libovolně spojovat do dvojmístných, trojmístných či vícemístných čísel
- ***opora výběru*** = soupis (seznam) zástupných výběrů, který jednoznačně identifikuje všechny statistické jednotky v základním výběru

(c) systematický (intervalový) výběr

- nejprve stanovím, jakou část základního souboru chci vybrat (např. 40 jednotek s tím, že to bude každá 25. … interval)

pak vylosuji první číslo mezi 1 a 25 a poté po 25 (intervalový výběr) vybírám vzorek

OTÁZKY A PŘÍKLADY

1. Jaké jsou základní typy výběrů
2. Jaké jsou techniky náhodného vybírání

6. Exploratorní analýza dat, grafy a tabulky

Klíčová slova: Základy exploratorní analýzy dat. Tabulky a grafy rozdělení četností.

Tabulky rozdělení četností

Tabulka rozdělení četností podává informaci o počtu (četnosti) výskytu jednotlivých variant znaku v souboru.

Chceme-li mezi sebou porovnávat různá rozdělení četností lišící se svým rozsahem a dospět také ke snazší interpretaci výsledků, je vhodné převést *absolutní četnosti* na *relativní četnosti*. Relativní četnosti p_i získáme jako podíl jednotlivých absolutních četností k celkovému rozsahu souboru.

U nespojitých statistických znaků se většinou využívá ***prosté rozdělení četností***.

U spojitých statistických znaků (výška v cm), případně u nespojitých znaků, jež nabývají velkého počtu obměn (plat v Kč), používáme ***intervalové rozdělení četností***, ve kterém *variační rozpětí* souboru R (tj. rozdíl mezi maximální a minimální zjištěnou hodnotou znaku) rozdělíme na určitý počet intervalů a potom zjistíme počty hodnot patřících do těchto intervalů. Intervaly se nepřekrývají a mají stejnou délku.

Při výpočtech statistických charakteristik nahrazujeme různá pozorování, která patří do jedné skupiny, jedinou zastupitelnou hodnotou. Za tuto zastupitelnou hodnotu se zpravidla volí střed intervalu.

Grafy

(a) spojnicové a sloupkové grafy

Pro grafické znázornění prostého rozdělení četností se využívá ***polygon četností***. Na ose x jsou hodnoty znaku (x_i) a na ose y jim odpovídající četnosti (n_i). Pokud bychom nahradili absolutní četnosti n_i relativními četnostmi p_i, obdržíme polygon relativních četností.

Velmi důležitá je u rozdělení četností poloha vrcholu. Tato obměna se nazývá modus x a označuje nejčastěji se vyskytující proměnnou v souboru. Podle počtu vrcholů rozlišujeme 1vrcholová (nejčastější) a vícevrcholová rozdělení četností

Pro grafické znázornění intervalového rozdělení četností se nejčastěji využívá ***histogram četností***. Je to sloupkový graf tvořený pravidelnými rovnoběžníky, jejichž základny mají délku zvolených intervalů a jejichž výšky mají velikost příslušných třídních četností (resp. relativních četností). Na ose x jsou zadány intervaly (př. počet odpracovaných hodin – od do) a na ose y je znázorněno n_i (př. počet pracovníků).

(b) bodové grafy

Bodové grafy používají jako grafické prostředky body umísťované v souřadnicové soustavě. Slouží ke znázorňování závislosti mezi dvěma kvantitativními znaky (popř. ke znázorňování průběhu časové řady). Vodorovná osa x je přitom stupnicí pro hodnoty kvantitativního znaku x_i (nezávislá proměnná), svislá osa y je určena pro vynášení hodnot druhého kvantitativního znaku y_i (závislá proměnná).

(c) výsečové grafy

U výsečových grafů relativní četnosti obměn znaku znázorňujeme pomocí výsečí kruhu, které získáme rozdělením středového úhlu úměrně k podílu jednotlivých částí zobrazovaného jevu vyjádřených v procentech.

(d) graf STEM-and-LEAF

Tento graf je dalším znázorněním rozdělení četností obvyklým ve statistickém softwaru. První sloupec udává kumulativní absolutní četnosti od nejmenší hodnoty k mediánu (hodnota v závorce) a od největší hodnoty k mediánu. Počet číslic za čarou udává četnost příslušné obměny tarifní třídy.

(e) krabičkový graf

Tento graf se nejčastěji používá pro zobrazení kvartilů. Přehledně znázorňuje charakter analyzované proměnné pomocí kvartilů, vnitřních a vnějších hradeb a extrémů (minimum, maximum). Slouží k identifikaci odlehlých pozorování.

Základním prvkem grafu je obdélník, jehož hrany tvoří hodnoty dolního a horního kvartilu, tzn., že uvnitř obdélníku je 50 % hodnot proměnné. Uvnitř je svislou čarou vyznačen medián a popř. tečkou aritmetický průměr.

OTÁZKY A PŘÍKLADY

1. Jaké jsou základní grafy využívané ve statistice
2. Jaké používáme tabulky

7. Popisné charakteristiky, kvantily

Klíčová slova: Základní popisné charakteristiky, jejich vlastnosti, použití. Kvantily – pojem, vlastnosti, použití.

Kvantily

Kvantil je hodnota, která rozděluje soubor hodnot určitého statistického znaku na dvě části – jedna obsahuje ty hodnoty, které jsou menší (nebo stejné) než tento kvantil, druhá část naopak obsahuje hodnoty, které jsou větší (nebo stejné) než kvantil.

Nejčastěji používanými kvantily jsou medián, kvartily, decily a percentily.

Medián

- je 50 % kvantil, který člení statistický soubor na dvě stejně četné poloviny.
- je to prostřední hodnota (při lichém rozsahu souboru → je nutné ale před tím obměny statistického znaku uspořádat vzestupně podle velikosti) x u sudého rozsahu je to průměr dvou prostředních jednotek

Kvartily

- jsou hodnoty, které dělí uspořádaný statistický soubor na čtyři části, přičemž každá část obsahuje 25 % jednotek.
- Jsou celkem 3:
  - dolní kvartil x_{25} – odděluje čtvrtinu nejnižších hodnot znaku
  - prostřední kvartil = medián x_{50}
  - horní kvartil x_{75} – odděluje 75 % nejnižších hodnot znaku od zbývajících 25 % hodnot znaku

Decily

- dělí soubor na 10 stejně obsazených částí
- $x_{10}, x_{20}, \ldots, x_{90}$

Percentily

- rozdělují soubor na 100 stejně velkých částí
- $x_1, x_2, \ldots, x_{99}$

Základní popisné charakteristiky

míry úrovně (polohy)

míry variability

- absolutní variabilita

- relativní variabilita

míry asymetrie (šikmost)

míry koncentrace (špičatost)

Porovnávat několik souborů pouze pomocí tabulek rozdělení četností by bylo pracné → shrneme info ze zjištěných údajů o stat. znaku a vyjádříme je v koncentrované podobě pomocí určitých charakteristik → úrovně rozdělení četností a variability rozdělení, někdy i šikmost a špičatost.

Míry úrovně (polohy)

Za základní vlastnost rozdělení lze považovat jeho úroveň. Měří se pomocí různých druhů středních hodnot, což jsou jednoduché číselné charakteristiky, pomocí kterých můžeme nahradit a zobecnit hodnoty souboru.

Počítají-li se střední hodnoty ze všech jednotek statistického souboru, nazývají se *průměry*. Nejdůležitější jsou průměry:

- aritmetický;
- harmonický;
- geometrický (jeho aplikace je minimální – při výpočtu průměrného koeficientu růstu ČŘ a v teorii indexů),
- kvadratický.

Do druhé skupiny můžeme zařadit ty střední hodnoty, které jsou založeny pouze na některých vybraných hodnotách souboru. Jelikož nepředstavují tak kvalitní charakteristiky polohy jako průměry, jsou počítány jen jako střední hodnoty doplňkové, pokud jich je k charakteristice souboru zapotřebí. Nejdůležitější z nich jsou:

- modus = nejčastěji se vyskytující hodnota v souboru;
- medián = hodnota dělící soubor na dvě stejně velké části.

<u>*Vlastnosti aritmetického průměru:*</u>

1. Součet jednotlivých odchylek od průměru je nulový.
2. Aritmetický průměr konstanty je roven této konstantě.
3. Přičteme-li k jednotlivým hodnotám znaku konstantu, zvýší se o tuto konstantu i aritmetický průměr.
4. Násobíme-li jednotlivé hodnoty znaku konstantou, je touto konstantou násoben i průměr.
5. Násobíme-li váhy (četnost n_i ve váženém tvaru) aritmetického průměru konstantou, průměr se nezmění.
6. Je-li statistický soubor rozdělen do k dílčích podsouborů, v nichž známe jednotlivé dílčí průměry $x_1, x_2, \ldots, x_k$ a počty pozorování $n_1, n_2, \ldots, n_k$, pak průměr celkového souboru je váženým aritmetickým průměrem těchto dílčích průměrů, kde vahami jsou četnosti těchto podsouborů.

<u>Vztahy mezi jednotlivými typy průměrů:</u>

$x_H \leq x_G \leq x \leq x_K$

Rovnost nastane pouze tehdy, jsou-li všechny hodnoty statistického souboru stejné. Liší-li se alespoň jedna hodnota od ostatních, jsou nerovnosti v uvedeném vztahu ostré.

Míry variability

Měření variability má význam při posuzování vypovídací schopnosti aritmetického průměru, jeho vypovídací schopnost je tím větší, čím je variabilita sledovaného znaku menší a naopak.

Míry variability charakterizují proměnlivost hodnot statistického souboru.

Jsou rovněž důležitou informací o tom, jaká je vypovídací schopnost měr polohy – velká variabilita hodnot snižuje informační význam např. průměrů nebo mediánu.

Míry absolutní variability

- ve stejných měrových jednotkách jako sledovaný znak

- nejjednodušší, ale i nejhrubší mírou variability je variační rozpětí R - snadnost a rychlost výpočtu, jednoduchá interpretace, krajní hodnoty ovšem mohou být nahodilé a případné extrémní hodnoty se projeví především na těchto hodnotách, neříká nic o variabilitě hodnot uvnitř variačního rozpětí
- $R = x_{max} - x_{min}$
- kvantilová rozpětí:
  - kvartilové rozpětí $x_{75} - x_{25}$
    - udává, jak široký je interval obsahující prostředních 50 % hodnot souboru
  - decilové rozpětí $x_{90} - x_{10}$
  - percentilové rozpětí $x_{99} - x_1$
- Praxe dává přednost jiným mírám variability (jejichž velikost je závislá na variabilitě všech hodnot stat.souboru):
- rozptyl
  = průměrná čtvercová odchylka jednotlivých hodnot znaku od jejich aritm. průměru
  - je to míra, která je závislá na variabilitě všech hodnot statistického souboru
- směrodatná odchylka
  = kladně vzatá odmocnina z rozptylu
  - je uvedena ve stejných měrových jednotkách jako zkoumaný statistický znak
  - lze ji interpretovat
- kvantilová odchylka
  = aritmetický průměr kladných odchylek sousedních kvantilů

<u>*Vlastnosti rozptylu:*</u>

1. Rozptyl konstanty je roven nule.
2. Přičteme-li ke všem hodnotám znaku konstantu, rozptyl se nezmění.
3. Násobíme-li všechny hodnoty znaku konstantou, rozptyl je násoben čtvercem (druhou mocninou) této konstanty.
4. Rozptyl součtu (rozdílu) dvou proměnných (x +- y) je roven součtu obou znaků zvětšenému (+) nebo zmenšenému (-) o dvojnásobek tzv. kovariance, což charakterizuje vzájemnou závislost dvou proměnných *x* a *y*.
5. Násobíme-li četnosti výskytu ve váženém tvaru rozptylu konstantou, rozptyl se nezmění.

Míry relativní variability

- pokud srovnáváme variabilitu souborů lišících se svoji úrovní. Měří variabilitu v poměru k úrovni sledovaného znaku v souboru. Udává, kolik procenty se podílí směrodatná odchylka na aritmetickém průměru
- nejznámější mírou rel. variability je variační koeficient (v_x) $v_x = s_x / x$ (*100 … v %)
  = slouží k porovnání variabilit
  - $v_x > 0{,}5$ (= 50 %) $\Rightarrow$ znak značné nesourodosti statistického souboru;
  - v takovém případě se musím zamyslet, zda by nebylo vhodné roztřídit soubor na menší méně různorodé části
  - obecně se pohybuje v intervalu $(-\infty; +\infty)$, protože aritm. průměr může být i záporný

<u>Variabilita proměnné:</u>

(a) pořadové
- *poměrový koeficient diferenciace* (P_D) $P_D = 4s_x^2 / R^2 \in \langle 0;1 \rangle$
- 0 je při minimální variabilitě; 1 je při maximální variabilitě

(b) kategoriální
- základní zhušťující charakteristikou hodnot katerogiální proměnné je její modus
- pro posouzení stupně typičnosti modální varianty je však důležité i objektivní změření proměnlivosti hodnot kategoriální proměnné

- je jasné, že čím je proměnlivost hodnot této proměnné menší, tím je stupeň typičnosti modu pro daný soubor vyšší
- tato měnlivost se označuje jako *mutabilita* (M)

Míry asymetrie (šikmost)

Míry šikmosti jsou založeny na srovnání stupně koncentrace malých hodnot sledovaného statistického znaku se stupněm koncentrace velkých hodnot tohoto znaku.

Rozdělení může být:

- symetrické = stejný stupeň hustoty malých a velkých hodnot
- sešikmené
  - kladně = větší stupeň koncentrace malých hodnot ve srovnání s hustotou velkých hodnot (příslušné míry šikmosti vycházejí jako kladná čísla)
  - záporně = větší stupeň koncentrace velkých hodnot ve srovnání s hustotou malých hodnot (příslušné míry šikmosti vycházejí záporně)

Míry koncentrace (špičatost)

Míry špičatosti jsou založeny na srovnání stupně koncentrace hodnot prostřední velikosti se stupněm nahuštěnosti ostatních hodnot, respektive všech hodnot proměnné.

Rozdělení může být:

- relativně ploché = podíl četností prostředních hodnot je srovnatelný s četnostmi ostatních hodnot
- relativně špičaté = větší stupeň koncentrace prostředních hodnot ve srovnání s četnostmi ostatních hodnot

OTÁZKY A PŘÍKLADY

1. Při dopravním průzkumu byla sledována vytíženost vjezdu do určité křižovatky. Student, provádějící průzkum, si vždy při naskočení zeleného světla zapsal počet aut, čekajících ve frontě u semaforu. Jeho zapsané výsledky jsou:

3 1 5 3 2 3 5 7 1 2 8 8 1 6 1 8 5 5 8 5 4 7 2 5 6 3 4 2 8 4 4 5 5 4 3 3
4 9 6 2 1 5 2 3 5 3 5 7 2 5 8 2 4 2 4 3 5 6 4 6 9 3 2 1 2 6 3 5 3 5 3 7
6 3 7 5 6

Nakreslete krabicový graf, empirickou distribuční funkci a vypočtěte následující výběrové statistiky: průměr, výběrová směrodatná odchylka, medián, modus a interkvartilové rozpětí.

2. Data byla získána měřením a představují napětí baterií ve voltech:

1,2 1,5 1,4 0,6 1,1 1,6 1,4 1,7.

Určete:

a) maximum, minimum, dolní a horní kvartil, medián, interkvartilové rozpětí; zakreslete krabicový graf;
b) shorth, modus, MAD;
c) odlehlá pozorování použitím mediánové souřadnice.

3. Maximální teploty naměřené v průběhu jednoho týdne na různých místech ČR

byly: 16, 14, 18, 13, 20, 19, 19 (° C).

Určete:

a) maximum, minimum, dolní a horní kvartil, medián, interkvartilové rozpětí; zakreslete krabicový graf;
b) shorth, modus, průměr, směrodatnou odchylku;
c) odlehlá pozorování použitím z-souřadnice.

8. Bodový odhad

Klíčová slova: Bodový odhad – pojem, požadované vlastnosti, směrodatná chyba. Bodové odhady některých důležitých parametrů a charakteristik.

Bodový odhad

= charakteristiku ze základního souboru odhaduji jediným číslem (jedinou statistikou) → to prohlásíme za odhad odpovídající charakteristiky základního souboru
- např. modus, medián, aritmetický průměr…

Požadované vlastnosti bodového odhadu

1. nestrannost odhadu
   - odhad není zkreslený (nepodhodnocuje ani nenadhodnocuje odhadovanou charakteristiku, tj. nevede k systematickým chybám). Chceme, aby střední hodnota výběrové statistiky E(g) byla rovna odhadované charakteristice G, potom g je nezkresleným odhadem G. Rozdíl E(g) – G se nazývá zkreslení (vychýlení) odhadu
2. konzistence
   - se zvyšováním rozsahu výběru se odhadnutá statistika stále více přibližuje odhadované charakteristice základního souboru
3. vydatnost
   - v některých případech lze nalézt více statistik, které splňují podmínku nezkreslenosti a konzistence
   - v takovém případě použijeme k odhadu charakteristiky základního souboru tu, která má nejmenší rozptyl … ta je vydatným, nezkresleným odhadem charakteristiky základního souboru
4. aby byl postačující
   - tzn., že mimo této výběrové statistiky neexistuje žádná jiná statistika, která by poskytovala další doplňující informace o odhadované charakteristice základního souboru

I když výběrová statistika bude splňovat všechny výše uvedené požadavky, je zřejmé, že její hodnota vypočtená na základě údajů získaných jedním náhodným výběrem se bude prakticky vždy určitým způsobem lišit od odhadované charakteristiky základního souboru.

Důsledkem této odlišnosti je vznik tzv. ***výběrové chyby***, tj. rozdílu mezi výběrovou statistikou (g) a odhadovanou charakteristikou základního souboru (G).. Je-li g nezkresleným odhadem G, pak je přesnost bodového odhadu možno měřit směrodatnou odchylkou g ($\sqrt{D(g)}$) = střední chyba

Bodové odhady některých charakteristik

Odhadovaná charakteristika		*Bodový odhad = statistika*	
střední hodnota	E(X)	výběrový aritmetický průměr	x
úhrn	N * E(X)	součin	N * x
rozptyl	D(X)	výběrový rozptyl	$s^{2´}$
relativní četnost	π	výběrová relativní četnost	p
absolutní četnost	N * π	součin	N * p

Statistika = funkce napozorovaných hodnot ve výběru

OTÁZKY A PŘÍKLADY

1. Co je to bodový odhad?
2. Jaké jsou jeho základní vlastnosti?
3. Co je to směrodatná chyba?
4. Určete bodové odhady některých důležitých parametrů a charakteristik.

9. Intervaly spolehlivosti

Klíčová slova: Interval spolehlivosti – pojem, význam, jednostranný a dvoustranný interval. Intervalové odhady některých důležitých parametrů a charakteristik.

Intervalový odhad

= k odhadu charakteristiky základního souboru používáme interval

Interval spolehlivosti (= konfidenční interval)

- odhad je reprezentován intervalem (G_d;G_h), který s danou vysokou pravděpodobností bude obsahovat skutečnou hodnotu odhadované charakteristiky základního souboru
- tato pravděpodobnost se nazývá spolehlivostí odhadu a značí se 1-α
- interval, jehož dolní a horní meze jsou G_d, resp. G_h (G_d<G_h), pak nazýváme 100(1-α)% interval spolehlivosti pro charakteristiku G a platí pro něj $P(G_d < G < G_h) = 1-\alpha$

- Spolehlivost odhadu je dána zvolenou pravděpodobností. Čím je tato pravděpodobnost větší, tím je i daný odhad spolehlivější. Čím více roste spolehlivost odhadu (tj. roste 1-α), tím se zvětšuje i příslušný interval, který udává přesnost odhadu. Čím bude daný interval širší, tím bude odhad spolehlivější → hodnota odhadované charakteristiky bude ležet s vysokou pravděpodobností uvnitř intervalu, ale odhad bude méně přesný (→ mezi přesností a spolehlivostí existuje nepřímá úměrnost). Uspokojivé výsledky dostáváme, když volíme spolehlivost odhadu 1-α = 0,95 (=5%ní hladina významnosti)

Intervaly

- jednostranné
  - je dána buď dolní nebo horní mez
    - pravostranné = je dána horní mez G_h
    - levostranné = je dána dolní mez G_d
- dvoustranné

Šířka intervalu spolehlivosti závisí na:

1. rozsahu souboru (n)
    - čím větší je rozsah souboru, tím menší je interval spolehlivosti
2. rozptylu (s^2)
    - čím větší je rozptyl souboru, tím větší je interval spolehlivosti
3. spolehlivosti odhadu ($1-\alpha$)
    - čím větší je spolehlivost odhadu, tím větší je i šířka intervalu spolehlivosti

Intervalové odhady

- ✓ střední hodnoty
    - při výběru z normálního rozdělení
    - při velkém výběru z libovolného rozdělení – limitní věty
- ✓ úhrnu
- ✓ rozptylu výběru z normálního rozdělení
- ✓ relativní četnosti (velký výběr) – limitní věty
- ✓ absolutní četnosti (velký výběr) – limitní věty

OTÁZKY A PŘÍKLADY

1.. Pracovníci obchodní inspekce kontrolují váhu porce masa v určitém výrobku konzervárenského průmyslu. Technologická norma konzervy a tomu odpovídající cenová
kalkulace udávají váhu masa v konzervě 90 g. Inspekce vyhodnotila 15 výrobků s těmito výsledky:

87 88 90 90 85 88 86 90 89 89 88 92 87 90 89 g.

Najděte 95% interval spolehlivosti pro střední hodnotu hmotnosti porce masa.

2. Ze základního souboru 10.000 automaticky balených sáčků piškotů bylo vybráno 1% sáčků a zjištěna průměrná váha 15,8g a směrodatná odchylka 4,8g. Určete se spolehlivosti
0,99, v jakých mezích lze očekávat průměrnou váhu balíčků piškotů.

3. Při kontrole data spotřeby určitého druhu masové konzervy ve skladech produktů masného
průmyslu bylo náhodně vybráno 320 konzerv a zjištěno, že 59 z nich má prošlou záruční
lhůtu. Stanovte 95% interval spolehlivosti pro odhad procenta konzerv s prošlou záruční lhůtou.

10. Testování hypotéz o rozdělení v základním souboru

Klíčová slova: Vybrané neparametrické testy. Testy hypotéz o rozdělení v základním souboru.

Chí-kvadrát test dobré shody

V určitých případech hledáme rozdělení, které by odpovídalo provedenému náhodnému výběru a sloužilo tak jako teoretický model. Přitom vycházíme z výběrového rozdělení, které se přirozeně od rozdělení teoretického více či méně liší. Volba teoretického rozdělení nemusí být vždy správná, a proto je aktuální ověřit dobrou shodu, čímž rozumíme právě shodu empirického rozdělení s teoretickým, vhodným testem.

Nejznámější z nich je χ^2-test dobré shody.

Použití ve 2 typických situacích:

1. Nulová hypotéza předpokládá, že v konečném základním souboru roztříděném podle nějakého znaku (kvantitativního nebo kvalitativního) do *k* skupin jsou podíly variant v základním souboru rovny číslům $\pi_{0,1}$; $\pi_{0,2}$; …, $\pi_{0,k}$.
2. Nulová hypotéza předpokládá, že nekonečný základní soubor má rozdělení určitého typu (např. normální).

   2 typy modelu:
   - úplně specifikovaný model
     - nulová hypotéza udává nejen typ rozdělení, ale i jeho parametry
   - neúplně specifikovaný model
     - nulová hypotéza udává pouze typ rozdělení; tzn., že parametry musím odhadnout ⇒ odlišnost v počtu stupňů volnosti
     - častější

Chí-kvadrát test se používá k ověřování normality.

Podmínky použití:

- je potřeba dostatečně velké *n*
- a zároveň dostatečné zastoupení všech skupin … očekávané četnosti v jednotlivých skupinách by neměly být menší než 5

<u>*Kolmogorovův-Smirnovův test pro jeden výběr*</u>

- dáváme přednost před chí-testem v případě malého rozsahu (má totiž větší sílu testu)
- vychází z původních napozorovaných hodnot, nikoli z údajů setříděných do tříd, tím nedochází ke ztrátě informace, obsažené ve výběru
- ověření hypotézy, že pořízený výběr pochází z rozdělení se spojitou distribuční funkcí, která musí být plně specifikovaná
- distribuční funkce rozdělení, z něhož náhodný výběr pochází se nazývá teoretická distribuční fce

<u>*Test chí-kvadrát o nezávislosti dvou znaků(v kombinační tabulce)*</u>

- kombinační tabulka vzniká, třídíme-li jednotky souboru podle variant dvou kvalitativních znaků
- test, který používáme k ověření nezávislosti v kombinační tabulce porovnává získané četnosti s teoretickými

OTÁZKY A PŘÍKLADY

1. Ověřte, zda má streptomycin vliv na léčbu plicní tuberkulózy, pokud skupina léčená streptomycinem není závislá na skupině léčené placebem.

Hodnocení	Streptomycin	Placebo	Celkem

Významné zlepšení	28 16.45	4 15.55	32
Střední/malé zlepšení	10 11.82	13 11.18	23
Beze změn	2 2.57	3 2.43	5
Střední/malé zhoršení	5 8.74	12 8.26	17
Významné zhoršení	6 6.17	6 5.83	12
Smrt	4 9.25	14 8.75	18
Celkem	55	52	107

11. Statistické testování hypotéz

Klíčová slova: Princip statistického testování hypotéz a základní pojmy. Testy hypotéz o některých důležitých parametrech a charakteristikách.

Statistickou hypotézou se rozumí určitý předpoklad o parametrech či tvaru rozdělení zkoumaného znaku.

Na základě vyčerpávajícího šetření celého základního souboru by bylo možné bezpečně rozhodnout o správnosti či nesprávnosti hypotézy. Takovéto vyčerpávající šetření je však většinou neekonomické nebo dokonce technicky neproveditelné, a proto šetření podrobujeme pouze část základního souboru – výběrový soubor.

Testování hypotéz

= proces ověřování správnosti nebo nesprávnosti hypotézy pomocí výsledků získaných náhodným výběrem

Nulová hypotéza H_0

= předpoklad, který vyslovíme o určité charakteristice či tvaru rozdělení v základním souboru

Alternativní hypotéza H_1

= tato hypotéza nějakým způsobem popírá konstatování formulované nulovou hypotézou

Protože při testování hypotézy jde o úsudek prováděný z údajů získaných náhodným výběrem, můžeme se ve svých úsudcích dopustit i chybných závěrů.

Může se nám stát, že:

zamítneme testovanou hypotézu H_0, ačkoliv ve skutečnosti platí;
- pak se dopouštíme tzv. ***chyby I. druhu***
- pravděpodobnost této chyby značíme α.

přijmeme testovanou hypotézu H_0, i když ve skutečnosti platí alternativní hypotéza H_1
- pak se dopouštíme tzv. ***chyby II. druhu***
- pravděpodobnost této chyby značíme β.

Pravděpodobnost 1-β se nazývá *síla testu*. Síla testu vlastně tedy vyjadřuje, s jakou pravděpodobností zamítneme nulovou hypotézu, platí-li alternativní hypotéza. Jinak řečeno, udává pravděpodobnost, že se nedopustíme chyby II. druhu.
Klasický způsob spočívá v tom, že předem zvolíme pevnou pravděpodobnost chyby I. druhu, tzv. hladinu významnosti v přijatelné výši (nejčastěji 5%). Testovací postup je odvozen tak, aby při dané hladině významnosti zajišťoval minimální pravděpodobnost chyby II.druhu.
Popis standardního testu zejména uvádí, jaké použít v dané situaci testové kritérium (T). Množinu hodnot, jichž může T nabýt, nazýváme výběrový prostor (S). Dříve, než testování provedeme, musíme mít připraveno pravidlo, umožňující rozhodnou ve prospěch H0 nebo H1. To bude nejjednodušeji vyjádřeno, když rozdělíme S na 2 podprostory:

1. podprostor V obsahující hodnoty svědčící ve prospěch H0 (obor přijetí)
2. podprostor W obsahující hodnoty svědčící ve prospěch H1 (kritický obor)
3.

Postup testování

1) zformulujeme nulovou a alternativní hypotézu
2) zvolíme a vypočteme testové kritérium
3) vymezíme kritický obor W
4) zformulujeme závěr, tedy výsledek testu
   (a) hodnota testového kritéria je v kritickém oboru W
      - testem byla prokázána alternativní hypotéza
      - prohlásíme-li, že platí H_1, neseme $100\alpha\%$ riziko nesprávnosti tohoto výroku
   (b) hodnota testového kritéria je v oboru přijetí V
      - testem jsme alternativní hypotézu neprokázali
      - kdybychom prohlásili, že platí H_0, neseme riziko omylu, že tomu tak ve skutečnosti není, o velikosti β

Testy hypotéz

parametrické

- o 1 parametru
  - o parametrech μ, π, o střední hodnotě E(X) při velkém i malém rozsahu výběru, o rozptylu aj.
  - test hypotézy o průměru testujeme, že $\mu = \mu_0$
  - test hypotézy o relativní četnosti – testujeme, že rel. četnost určité varianty znaku v zákl. souboru se rovná určitému číslu
  - test hypotézy o rozptylu - testujeme, že rozptyl zákl. souboru je roven určité hodnotě σ_0^2
  - test hypotézy o parametru δ (tj. o střední hodnotě) exponenciálního rozdělení (pokud potřebujeme provést test, který probíhá v podmínkách výběru menšího rozsahu, kdy navíc testovým kritériem je parametr některého jiného něž normálního rozdělení.

- o shodě
  - test hypotézy o shodě 2 průměrů
    - známe rozptyly
    - neznáme rozptyly, ale předpokládáme, že jsou shodné
    - rozptyly neznáme a jsou různé
  - test hypotézy o shodě 2 rozptylů, HO: $\sigma_1^2 = \sigma_2^2$

neparametrické

- chí-kvadrát test dobré shody
- Kolmogorovův-Smirnovův test pro 1 výběr; 2 výběry aj.

OTÁZKY A PŘÍKLADY

Jednovýběrové testy

testujeme střední hodnotu při známém rozptylu σ^2

1. Odběratel s dodavatelem uzavřeli smlouvu o dodávce pytlů obilí. Při známém rozptylu $\sigma^2 = 0{,}1$ plnícího stroje má být střední hodnota hmotnosti pytlů 10 kg. Pro ověření skutečnosti, že plnící stroj pracuje dobře, bylo náhodně vybráno 40 pytlů a získán průměr jejich hmotnosti $x = 9{,}6$ kg. Rozhodněte, zda dodavatel dodržuje stanovenou střední hodnotu hmotnosti.

testujeme střední hodnotu při neznámém rozptylu

2. Balíčky soli mají mít hmotnost 1 kg. Bylo zváženo 10 balíčků a zjištěny odchylky od váhy 1 kg:

-1,2 0,5 -0,6 -0,3 0,2 -1,0 0,4 -0,8 0,5 -0,4.

Zjistěte, zda lze na základě zjištěných hodnot konstatovat, že průměrná hmotnost jednoho balíčku nedosahuje 1 kg.

testujeme podíl (procentuální vyjádření)

3. V náhodném výběru čipů vyráběných velkou světovou společnosti 10% čipů nevyhovuje novým požadavkům na kvalitu. Sestrojte 95% interval spolehlivosti pro podíl p čipů (v celé populaci), které nevyhovují dané normě, jestliže rozsah výběru je:

a) n = 10
b) n = 100

Dvouvýběrové testy

testujeme rozdíl podílů

4. TV stanice zjišťuje sledovanost určitého pořadu a zajímá ji, zda u dospělých osob do 25 let („mladší osoby“) je tato sledovanost jiná, než u věkově starších osob. Daný pořad sledovalo 80 z 500 náhodně vybraných mladších osob a 100 z 1000 náhodně vybraných starších osob.

a) Najděte 99% interval spolehlivosti pro rozdíl podílů sledovanosti

uvedeného pořadu u těchto dvou věkových skupin .

b) Otestujte danou hypotézu.

testujeme rozdíl středních hodnot

5. U 12-ti náhodně vybraných rodin se 2-mi dětmi byly zjištěny roční výdaje na průmyslové
zboží (v tisících Kč):
41,2 39,4 36,3 38,7 39,9 38,3 40,6 41,5 37,4 43,1 35,7 35,8. Obdobně u šesti náhodně vybraných rodin se 4-mi dětmi byly údaje následující: 39,2 43,8 38,9 44,3 41,2 44,1.
Zjistěte, zda se střední hodnota ročních výdajů na průmyslové zboží liší u rodin se 2-mi a 4-mi dětmi.

12. Zkoumání závislosti

Klíčová slova: Některé elementární metody zkoumání závislosti – dvourozměrné tabulky, bodový diagram, podmíněné průměry a rozptyly. Čára podmíněných průměrů, korelační poměr. Kovariance a empirický korelační koeficient. Jednofaktorová analýza rozptylu.

Cílem zkoumání je hlubší vniknutí do podstaty sledovaných jevů a procesů určité oblasti a tím i přiblížení k tzv. příčinným souvislostem (např. když existence určitého jevu vyvolá existenci jiného jevu)

Z hlediska metody zkoumání je vhodné rozlišení tzv. pevných a volných závislostí.

Závislostí pevnou se označuje případ, kdy výskytu např. jednoho jevu nutně odpovídá výskyt druhého jevu. Z pravděpodobnostního hlediska jde o vztah, který se projeví s jistotou, neboli s pravděpodobností rovnou jedné. Závislost pevná se objevuje třeba. ve fyzice, kdy zkoumáme např. závislost mezi časem a ujetou dráhou. Ve skutečnosti však takový průběh funkce není obvyklý.

Závislostí volnou je pak případ, kdy výskyt jednoho jevu ovlivňuje výskyt druhého jevu v tom smyslu, že se zvýšila pravděpodobnost nastoupení druhého jevu při nastoupení jevu prvního. Pokud se volná závislost týká kvantitativních statistických znaků, bývá zvykem ji označovat za ***závislost statistickou***.

Statistická závislost může být:

a) silná
   - lze vypozorovat určitou tendenci – př. kousek paraboly, S křivky apod.
   - nelze však prostoupit všemi body jako tomu je u závislosti pevné

b) slabá … v podstatě to už je nezávislost
   - o průběhu závislosti proměnných nelze říct vůbec nic

V reálných empirických situacích se setkáváme prakticky výhradně s volnými závislostmi, ale s tím, že za obecnými tendencemi projevujícími se v souboru statistických údajů se mohou skrývat hlubší zákonitosti vztahů mezi veličinami.

K poznání těchto statistických závislostí, jakož i k ověřování deduktivně učiněných teorií, slouží metody regresní a korelační analýzy.

Regresní analýza

= zabývá se jednostrannými závislostmi

= jedná se o situaci, kdy proti sobě stojí vysvětlující (nezávisle) proměnná *x* v úloze „příčin" a vysvětlovaná (závisle) proměnná *y* v úloze „následků"

= úlohou regresního počtu je zkoumat obecné tendence ve změnách vysvětlovaných proměnných vzhledem ke změnám vysvětlujících proměnných

=

Korelační analýza

= zkoumá vzájemné závislosti mezi vysvětlovanými a vysvětlujícími proměnnými

= klade se zde důraz více na sílu vzájemného vztahu než na zkoumání veličin ve směru příčina – následek

Dvourozměrné tabulky

Při zkoumání závislosti mezi dvěmi proměnnými (*x* a *y*) je možné jednotlivá pozorování uspořádat do tabulky, která se někdy označuje jako *korelační*. Tato tabulka má ve sloupci (legendě) jednotlivé varianty znaku *x* a v hlavičce jednotlivé varianty znaku *y*. Do jednotlivých políček uvnitř tabulky zapisujeme tzv. simultánní (sdružené) četnosti, které vyjadřují, kolikrát se v souboru vyskytují kombinace variant obou znaků.

S takovouto tabulkou souvisí dvě podmíněná rozdělení četností. V prvním případě se jedná o podmíněné rozdělení četností znaku *y* za podmínky, že $x_i = ?$ (př. 2 … tzn. koukám na druhý řádek). Ve druhém případě jde o rozdělení četností znaku *x* za podmínky, že $y_j = ?$ (př. 4 … tzn. koukám na čtvrtý sloupec).

Jako každé jednorozměrné rozdělení četností je možné i podmíněné rozdělení četností popsat soustavou statistických charakteristik.

Podmíněný průměr:

$$y_i = \sum y_j \, n_{ij} / n_i$$

Podmíněný rozptyl:

$$s_i^2 = \sum (y_j - y_i)^2 \, n_{ij} / n_j$$

V regresní analýze se především zajímáme o změny podmíněných průměrů vysvětlované (závisle) proměnné při změnách vysvětlujících (nezávisle) proměnných.

Pokud pracujeme jen s jednou vysvětlující proměnnou a jednou vysvětlovanou proměnnou, můžeme údaje získané pozorováním *n* statistických jednotek znázornit i graficky. graficky je snadné znázornit i průběh podmíněných průměrů *y* vzhledem k různým hodnotám *x*. Spojením jednotlivých bodů dostaneme tzv. ***čáru podmíněných průměrů***.

Bodový diagram (graf)

Grafické vyjádření 2rozměrného rozdělení četností. Bodové grafy používají jako grafické prostředky body umísťované v souřadnicové soustavě. Slouží ke znázorňování závislosti mezi dvěma kvantitativními znaky (popř. ke znázorňování průběhu časové řady). Vodorovná osa *x* je přitom stupnicí pro hodnoty kvantitativního znaku x_i (nezávislá proměnná), svislá osa *y* je určena pro vynášení hodnot druhého kvantitativního znaku y_i (závislá proměnná). Pokud ovšem jako analyzovaná proměnná vystupuje spojitý stat. znak nebo statistický znak diskrétní, který nabývá mnoha variant, je vhodnějším graf. znázorněním tzv. *3dimenzionální histogram.*

Kovariance … C(X,Y)

= střední hodnota součinu odchylek obou náhodných veličin X, Y od jejich středních hodnot

- Vzorec:

$$C(X,Y) = E(XY) - E(X)E(Y)$$

Index determinace a korelace:

Rozptyl empirických hodnot lze rozložit na rozptyl vyrovnaných hodnot a rozptyl reziduálních hodnot. Kdyby mezi závisle a nezávisle proměnnou existovala funkční závislost, pak všechny empirické hodnoty byly zároveň hodnotami vyrovnanými a rovnaly by se i jejich rozptyly. Tím pádem by reziduální rozptyl byl nulový.

Naopak kdyby byla úplná nezávislost mezi oběma proměnnými, pak by všechny vyrovnané hodnoty byly stejné a jejich rozptyl by byl nulový. Rozptyl skutečných hodnot by se pak rovnal rozptylu reziduálních hodnot.

Intenzitu závislosti a kvalitu regresní funkce můžeme měřit podle toho, jak se podílí na rozptylu skutečně zjištěných hodnot rozptyl vyrovnaných hodnot. Závislost bude tím silnější, čím větší bude podíl rozptylu vyrovnaných hodnot na celkovém rozptylu a naopak. V případě funkční závislosti roven 1, v případě nezávislosti roven 0 → čím více se bude blížit 1, tím považujeme závislost za silnější, a tedy dobře vystiženou regresní funkcí a naopak. (Index determinace vynásobený 100 udává tu část rozptylu y, kterou se podařilo vysvětlit danou regresní funkcí). Odmocninou z indexu determinace je index korelace, poskytuje stejné info o těsnosti, má však menší vypovídací schopnost. Oba indexy se používají u jiných regresních funkcí než u přímky.

Koeficient korelace: je zvláštním případem indexu korelace a měří těsnost závislosti popsané lineární regresní funkcí (přímkou). Oproti indexu korelace může nabývat i záporných hodnot. Jeho definiční obor je od -1 do +1. Jestliže je roven +1, existuje mezi proměnnými přímá lineární závislost x pokud je roven -1, je mezi proměnnými nepřímá funkční lineární závislost. Když je nulový, značí to lineární nezávislost.

Koeficient determinace: udává, kolik procent variability závisle proměnné lze vysvětlit zvoleným regresním modelem

= udává, jaké procento proměnlivosti $\underline{y}$ lze vysvětlit variabilitou modelových hodnot $\underline{Y}$

= použití pouze u přímky; u ostatních funkcí se používá *index determinace* (I_{yx}^2)

= Vzorec:

$$r_{yx}^2 = s_Y^2 / s_y^2 \qquad \in (0;1) \qquad \text{neboli} \qquad \in (0;100\%)$$

= odmocnina z koeficientu determinace je ***korelační koeficient*** (r_{yx}) … u přímky, příp. *index korelace* (I_{yx}) … u ostatních funkcí

OTÁZKY A PŘÍKLADY

1. Popište metody zjišťování závislostí
2. Co je to index determinace a korelace?

13. Regresní a korelační analýza

Klíčová slova: Regresní a korelační analýza – regresní modely, regresní funkce. Regresní koeficienty, reziduální součet čtverců, reziduální rozptyl, determinační index. Sdružené regresní přímky a roviny.

Regresní analýza

= zabývá se jednostrannými závislostmi

= jedná se o situaci, kdy proti sobě stojí vysvětlující (nezávisle) proměnná x v úloze „příčin“ a vysvětlovaná (závisle) proměnná y v úloze „následků“

= úlohou regresního počtu je zkoumat obecné tendence ve změnách vysvětlovaných proměnných vzhledem ke změnám vysvětlujících proměnných

= tzn. hlavním úkolem je vystihnout pomocí regresní funkce na základě znalosti dvojic empirických (výběrových) hodnot x_i a y_i průběh závislosti mezi oběma proměnnými, což nám umožní provádět odhad hodnot závisle proměnné y na základě zvolených hodnot nezávisle proměnné x.

Korelační analýza

= zkoumá vzájemné závislosti mezi vysvětlovanými a vysvětlujícími proměnnými

= klade se zde důraz více na sílu vzájemného vztahu než na zkoumání veličin ve směru příčina – následek

Modely

1. deterministický
   - jde o případ pevné závislosti, kdy teoretická regresní funkce platí s pravděpodobností rovnou jedné
   - neexistuje chyba ε, a tudíž funkce Y představuje předpis, který s jistotou přiřazuje hodnotě proměnné x hodnotu proměnné y
2. stochastický
   - případ, kdy odchylky ε jsou nenulové následkem působení všech neuvažovaných činitelů, které ovlivňují změny proměnné y
   - vzhledem k tomu, že počet těchto činitelů je prakticky nekonečně velký, toto nezměřitelné ε považujeme za náhodnou veličinu

Regresní funkce

υ lineární v parametrech…model (ty parametry) lze odhadnou metodou nejmenších čtverců

- přímka
- parabola
- hyperbola
- logaritmické

Až po transformaci:

- exponenciální
- mocninná

υ nelineární v parametrech

regresní funkce:

- přímková regrese – nejjednodušší a nejčastěji používaný typ regrese
  - $\eta = \beta_0 + \beta_1 x$
- parabolická regrese
  - $\eta = \beta_0 + \beta_1 x + \beta_2 x^2$
- polynomická regrese p-tého stupně
  - $\eta = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p$
- hyperbolická regrese
  - $\eta = \beta_0 + \beta_1 \frac{1}{x}$
- hyperbolická regrese p-tého stupně
- logaritmická regrese
  - $\eta = \beta_0 + \beta_1 \log x$
- exponenciální regrese – nelineární z hlediska parametrů
  - $\eta = \alpha\beta^x \rightarrow \ln \eta = \ln \alpha + x \ln \beta$ (odhad parametrů regresních funkcí, které nejsou lineární z hlediska parametrů, neprovádíme metodou nejmenších čtverců přímo, protože to vede k soustavě nelineárních rovnic, ale najde se počáteční vhodný odhad a postupným zlepšováním řešení nalezneme odhad s požadovanou přesností →linearizující transformace→spočívá v tom, že pomocí logaritmů, převrácením hodnot, apod. dojdeme k takovému tvaru regr. fce, že její parametry bude už možné odhadovat metodou nejmenších čtverců)

Další pojmy:

Regresní koeficienty (b1) – směrnice regresní přímky, udává průměrnou změnu závislé proměnné y při jednotkové změně (u exponenciální funkce značí b1 koeficient růstu)

Regresní konstanta (b0) – odpovídá průsečíku regresní přímky s osou Y

- odhadnutý parametr b_{yx};
- tento parametr značíme b_{yx} proto, abychom vyjádřili, že závisle proměnnou je *y* a nezávisle proměnnou *x*
- dílčí regresní koeficienty – odhad toho, jak by se změnila v průměru vysvětlovaná proměnná y při jednotkové změně vysvětlují proměnné před tečkou

Reziduální součet čtverců

$$S_R = \sum_{i=1}^{n} \left(y_i - Y_i\right)^2 = S_y - S_T$$

Reziduální rozptyl – rozptyl skutečně zjištěných hodnot kolem regresní čáry (tj. rozptyl empirických hodnot vyrovnaných)

Sdružené regresní přímky – pokud v roli proměnné vystupuje i y, oboustranná závislost, vedle regresní přímky $Y = \beta_0 + \beta_1 x$ používáme i $X = a_0 + a_1 y$

Vícenásobná regrese a korelace:

Jestliže je závisle proměnná lineárně závislá na každé z vysvětlujících proměnných a jsou-li zároveň tyto vysvětlující proměnné vzájemně nezávislé, používáme mnohonásobnou lineární funkci.

Dílčí regresní koeficienty udávají odhad toho, jak by se změnila v průměru závisle proměnná při jednotkové změně nezávisle proměnné za předpokladu neměnnosti ostatních nezávisle proměnných.

B-koeficienty jsou normalizované regresní koeficienty. Tyto koeficienty jsou bezrozměrné – tzn. jsou nezávislé na měrových jednotkách, v nichž jsou jednotlivé proměnné uvažovány. Díky tomu je lze vzájemně srovnávat.

***Vícenásobná lineární závislost*:**

Pokud vystihujeme závislost proměnné y na větším počtu vysvětlujících proměnných lineární funkcí, používáme k měření těsnosti závislosti koeficienty dílčí korelace nebo koeficienty vícenásobné korelace. Koeficient dílčí korelace měří intenzitu lineární závislosti závisle proměnné na nezávisle proměnné x1 za předpokladu, že všechny ostatní proměnné jsou konstantní. Vyjadřuje těsnost lineárního vztahu mezi dvěma proměnnými při vyloučení vlivu jedné nebo více dalších proměnných. Nabývají hodnot od –1 do 1.
Koeficient vícenásobné korelace měří těsnost závisle proměnné na všech vysvětlujících proměnných dohromady. Vyjadřuje těsnost závislosti jedné proměnné na lineární kombinaci jiných proměnných.
Druhou mocninou tohoto koeficientu je vícenásobný koeficient determinace.
Párové korelační koeficienty: určují se pro každou dvojici proměnných a sestavují se pro ně korelační matice symetrické podle hlavní diagonály.

OTÁZKY A PŘÍKLADY

1. Byla zjišťována závislost počtu druhů na velikosti plochy: při každé velikosti plochy byla 4 nezávislá stanovení. Byly získány následující výsledky:

velikost plochy (m^2):	počet druhů
0.01	4, 6, 5, 8
0.25	9, 5, 8, 11
1.00	12, 14, 18, 11
4.00	20, 12, 25, 28
9.00	22, 28, 31, 18
16.00	25, 34, 19, 30
64.00	36, 39, 43, 22

Spočtěte regresi, proveďte vhodnou transformaci. Předpokládáme, že platí závislost počtu druhů (S) na ploše (A): $S = c\,A^z$, kde c a z jsou regresní analýzou odhadnuté koeficienty.

2. Předpokládáme exponenciálně rostoucí populaci. Velikost populace byla zjišťována v jednotlivých časech. Odhadněte růstovou rychlost populace.

čas	velikost populace
0	5
1	7
2	10
3	16
4	19
5	28
6	35
7	49
8	59
9	71
10	101

Můžeme rozumně předpokládat, že variabilita (vyjádřená jako směrodatná odchylka velikosti populace) roste s velikostí populace zhruba lineárně.

14. Analýza přežívání

ÚČEL

Analýza přežívání - sleduje určitý jev v čase a odhaduje pravděpodobnost jeho výskytu v závislosti na čase. Podobný účel má i analýza spolehlivosti (využívá také podobné metody).

· Analýza přežívání je jedna z nejčastěji používaných metod v lékařství:
· Pravděpodobnost přežití pacienta určitou dobu po jisté operaci
· Pravděpodobnost obnoveného výskytu tumoru
· Pravděpodobnost, že osoba v jistém věku již dosáhla určitého vývojového stadia
· Pravděpodobnost, že dojde k poruše kardiostimulátoru
· …

Mimo lékařství například:

· Jak dlouho vydrží manželství („přežití“ manželství před rozvodem)
· Jak dlouho vydrží student na VŠ (než bude „vyhozen“)
· Jak dlouho bude existovat firma (než bude vtlačena z trhu)
· Jak dlouho vydrží zaměstnanec ve firmě (než bude vyhozen nebo sám odejde)
· Jak dlouho vydrží auto jezdit bez opravy
· …

CENZOROVÁNÍ

V průběhu studie je pro každý subjekt sledovaný jev buď zaznamenán (například osoba zemřela) nebo zaznamenán není.

Subjekty, u nichž jsme jev nezaznamenali, se označují jako cenzurované:

· U nich na konci studie sledovaný jev ještě nenastal
· Nebo jsme s nimi ztratili kontakt a v době poslední kontroly sledovaný jev ještě nenastal

Jednotlivé subjekty se zúčastní studie různou dobu, jsou v tomto ohledu neporovnatelné, a proto by se klasické metody uplatňovaly jen obtížně

METODY

Neparametrické

„Tabulky života“ (life tables)

Kaplan-Meierovy empirické kumulativní distribuční funkce

Semiparametrické

Vosova regrese umožňuje posoudit vliv vnějších faktorů na hazardní funkci; pokud je regresní koeficient kladný, daný faktor zvyšuje riziko smrti

Parametrické

OTÁZKY A PŘÍKLADY

1. Vytvořte Kaplan-Meierův odhad pravděpodobnosti přežití v závislosti na čase pro data:

Týdny	Počet lidí	δ S(1), C(0)	Proporce přeživších	Kumulativní proporce S(t)
10	18	S		
13*	17	C		
18*	16	C		
19	15	S		

23*	14	C		
30	13	S		
36	12	S		
38*	11	C		
54*	10	C		
56*	9	C		
59	8	S		
75	7	S		
93	6	S		
97	5	S		
104*	4	C		
107	3	S		
109*	2	C		

15. Doporučená literatura

Anděl, J.: Statistické metody. 2. vydání, Matfyzpress, Praha, 1998

Havránek, T.: Statistika pro biologické a lékařské vědy. Academia, Praha, 1993

Havránek, J., Havránková, R., Vurm V., Záškodný P.: Základy zdravotnické statistiky, skripta JCU, České Budějovice, 2004

Kubánková, V., Hendl, J.: Statistika pro zdravotníky. Avicenum, Praha, 1986

Lepš, J.: Biostatistika. Jihočeská univerzita České Budějovice, České Budějovice, 1996

Likeš, J., Machek, J.: Matematická statistika. SNTL, Praha, 1983

Skalská, H., Stránský, P.: Základy biostatistiky. Univerzita Karlova, Praha, 1994

Zvára, K.,: Biostatistika, Universit Karlova, Praha, 1997

Zvárová, J.: Základy statistiky pro biomedicíncké obory, http://ucebnice.euromise.cz

Altman, D. G.: Practical Statistics for Medical Research. 2nd edition, Chapman & Hall, London, 1994

Bowers, D.: Statistics from Scratch (An Introduction for Health Care Professionals), J.Wiley & Sons, 1997

Jordan, K., Ong, B.N., Croft, P.: Mastering Statistics (A Guide for Health Service Professionals and Researchers)

Statsoft, Inc. (2004) Electronic Statistic Textbook. Tulsa, Oklahoma, USA, http://www.statsoft.com/textbook/statníme.html

Briš R., Litschmannová M. : Statistika I. Pro kombinované a distanční studium, VŠB-TU Ostrava, 2004,

Cyhelský L., Kalounová J., Hindls R. : Elementární statistická analýza, Management Press Praha, 1996,

Friedrich V. : Statistika 1., Vysokoškolská učebnice pro distanční studium, Západočeská Univerzita, Plzeň 2002,

Hindls R., Hronová S., Seger J. : Statistika pro ekonomy, Professional Publishing Praha, 2004

Likeš J., Cyhelský L., Hindls R. : Úvod do statistiky a pravděpodobnosti, VŠE Praha, 1994

Likeš J., Machek J. : Matematická statistika, SNTL Praha, 1988,

Novovičová J. : Pravděpodobnost a základy matematické statistiky, ČVUT Praha, 2002