5. Normální rozdělení. Ověřování statistických hypotéz

Adekvátní použití kvantitativních metod, které využíváme v praxi sociologických výzkumů, se opírá v určité míře o předpoklad, že se zkoumané znaky (nebo soubor znaků) podřizují určitým zákonitostem statistického rozdělení. Nejčastěji se setkáváme se zákonitostí normálního rozdělení.

Druhá skupina otázek, které v tomto oddíle budeme řešit, souvisí s ověřováním hypotéz. Můžeme vydělit dvě funkce statistických procedur. Za prvé, popis prvků sledovaného souboru, za druhé, pomoc badateli při některých výpovědích o nich. V předcházející kapitole jsme hovořili především o deskriptivní funkci statistiky. V této části krátce vysvětlíme základní pojmy i principy statistických důkazů.

NORMÁLNÍ ROZDĚLENÍ. Nejznámějším teoretickým rozdělením je normální nebo Gaussovo rozdělení. Normální rozdělení znaku pozorujeme v těch případech, kdy na množinu různých hodnot, které může určitý znak mít, působí množství nezávislých nebo slabě závislých faktorů, které souhrnně hrají stejnou, a to velmi malou úlohu (tj. neexistují nějaké dominující vlivy).

Gaussovo rozdělení popisujeme funkcí typu:

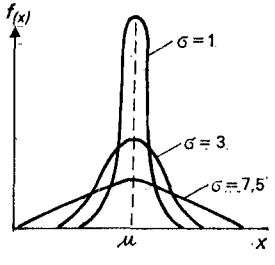
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad (18)$$

kde σ^2 — rozptyl náhodné veličiny (σ^2 je teoretický rozptyl na rozdíl od s^2, vypočítaného z výběrového souboru); μ — označení průměru (matematický předpoklad).

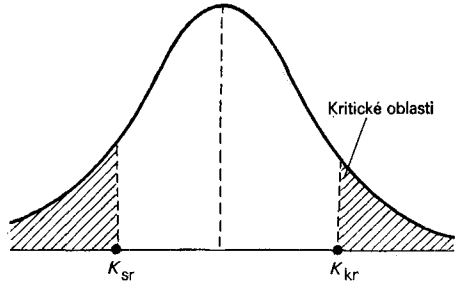
V praktických výpočtech se často používá tzv. pravidlo tří sigma, které spočívá v tom, že jen 0,27 % ze všech hodnot normálně rozděleného znaku leží vně intervalu $\mu \pm 3\sigma$, tj. téměř všechny hodnoty znaku jsou v rozmezí šesti sigma (obr. 10).

STATISTICKÉ HYPOTÉZY. Statistickou hypotézou nazýváme hypotézu o typu nám neznámého rozdělení nebo o dalších parametrech nám známého rozdělení.[1] Například statistickou hypotézou bude hypotéza o tom, že daný základní soubor má normální rozdělení. Prověřovanou hypotézu nazýváme nulovou (základní) hypotézou a označujeme ji H_0. Spolu s nulovou hypotézou se ověřuje alternativní hypotéza H_1, která má opačný charakter.

Statistická kritéria a prověrka hypotéz. K ověření nulové hypotézy se používá speciálně zvolená náhodná veličina, jejíž přesné nebo alespoň přibližné rozdělení je známé a obyčejně je máme k dispozici v tabulkách. Tato veličina se nazývá statistickým kritériem (zatím je označíme K).



Obr. 10 Křivka normálního rozdělení



Obr. 11. Rozdělení kritéria

Pro kritérium K se volí tzv. kritická oblast, tj. soubor hodnot kritéria, pro něž odmítáme nulovou hypotézu. Bod K_{kr} se nazývá kritickým bodem tehdy, když odděluje kritickou oblast od oblasti, v níž hypotézu přijímáme.

Rozlišujeme pravostrannou, levostrannou nebo dvoustrannou kritickou oblast.

Přijetí nebo odmítnutí hypotézy provádíme na základě odpovídajícího statistického kritéria s určitou pravděpodobností. Předpokládáme, že nulová hypotéza je pravdivá tehdy, jestliže pravděpodobnost toho, že kritérium K bude mít hodnotu vyšší než K_{kr}, tzn. že se bude nacházet v kritické oblasti, se rovná zvolené pravděpodobnosti, tj.

$p(K > K_{kr}) = \alpha$ (pro pravostrannou oblast);
$p(K < -K_{kr}) = \alpha$ (pro levostrannou oblast);
$p(|K| > K_{kr}) = \alpha/2$ (pro dvoustrannou oblast).

[1] Pojem hypotézy ve statistice je užší než obecný pojem vědecké hypotézy.

Příručka pro sociology. Základní pojmy a metodologické principy marxisticko-leninské sociologie. Praha. 1980. Originál Moskva 1977.

Zvolená pravděpodobnost α se nazývá hladinou významnosti. Praktické přijetí nebo odmítnutí nulové hypotézy se provádí tímto způsobem: zvolíme odpovídající kritérium (jak, to vysvětlíme později), vypočítáme pozorovanou hodnotu kritéria K_H, kdy vycházíme ze zjištěného empirického rozdělení, zvolíme hladinu statistické významnosti (obvykle 0,05 nebo 0,01).

Z tabulek rozdělení kritéria K pro danou hladinu významnosti najdeme kritický bod K_{kr}. Jestliže $K_H > K_{kr}$, nulovou hypotézu odmítáme. Jestliže $K_H < K_{kr}$, tak pro její odmítnutí důvody nemáme.

Když přijímáme nebo odmítáme nulovou hypotézu, můžeme se dopustit chyb dvou typů: odmítnout hypotézu, i když je pravdivá, a naopak, přijmout hypotézu, když není pravdivá. Když nějakou hypotézu přijmeme, je proto nesprávné se domnívat, že je zcela prokázána. Pro zvýšení její průkaznosti je nutné ji i nadále ověřovat (například zvětšením výběrového souboru).

Odmítnutí hypotézy je vždy průkaznější než její přijetí.

Typy hypotéz. Příklady statistických hypotéz: a) normální rozdělení má určitý daný průměr a rozptyl nebo má určitý daný průměr (o rozptylu nic nevíme); b) rozdělení normální nebo dvě stejná neznámá rozdělení. V prvním případě mluvíme o ověřování parametrických hypotéz (tj. prověřují se hypotézy o parametrech rozdělení), a ve druhém případě hovoříme o ověřování neparametrických hypotéz. Jako kritéria se nejčastěji uplatňují náhodné veličiny náhodně rozložené (Z — kritérium), test Fišera (F — kritérium Fišera), test Studenta (t — kritérium Studenta), test chí–kvadrát (χ^2 — kritérium) apod.

Jako konkrétní příklad si rozebereme použití chí-kvadrát testu k prověrce neparametrických hypotéz.

Chí-kvadrát kritérium. Populárnost chí-kvadrát testu je dána především tím, že jeho použití nevyžaduje předchozí znalost rozdělení sledovaného znaku. Kromě toho může znak být spojitý nebo diskrétní (přetržitý), či jej dokonce můžeme měřit na nominální škále.

Například jestliže neznáme typ rozdělení, ale můžeme předpokládat, že je typu A, tak nám χ^2 test dovoluje ověřit tuto hypotézu: zkoumaný soubor má rozdělení typu A. K ověření této hypotézy porovnáme empirické (pozorované) četnosti a teoretické četnosti (vypočítané za předpokladu určitého typu rozdělení). Tyto četnosti si vypíšeme:

Hodnoty znaku	x_1	x_2	...,	x_k
Empirické četnosti	n_1	n_2	...,	n_k
Teoretické četnosti	n_1	n_2	...,	n_k

Obvykle se empirické a teoretické četnosti liší. Je možné, že tyto rozdíly budou zcela náhodné (statisticky nevýznamné) a jsou objasnitelné malým počtem pozorování, způsobem třídění nebo jinými příčinami. Ale je také možné, že rozdíly v hodnotách (četnostech) jsou statisticky významné a objasnitelné tím, že teoretické hodnoty jsou vypočítané na základě nepravdivého předpokladu o rozdělení základního souboru. Chí-kvadrát test odpovídá na otázku, jsou-li rozdíly mezi empirickými a teoretickými četnostmi náhodné nebo ne. Jako jakýkoli jiný test ani χ^2 nedokazuje pravdivost hypotézy, ale pouze s určitou pravděpodobností α ověřuje shodu pozorovaných a teoretických četností. Chí-kvadrát test má podobu

$$\chi^2 = \sum \frac{(n_i - \tilde{n}_i)^2}{\tilde{n}_i}. \tag{19}$$

Kritický bod χ^2 rozdělení najdeme (viz tabulku B v příloze) k zvolené hladině významnosti α a počtu stupňů volnosti df. Počet stupňů volnosti najdeme na základě vzorce:

$$df = k - 1 - r,$$

kde je k — počet různých tříd (kategorií) znaku; r — počet parametrů předpokládaného rozdělení, které hodnotíme na základě výběrového souboru (například pro normální rozdělení se hodnotí dva parametry μ a s^2).

Vysvětlíme si to na příkladu, v němž daný znak nabývá tří hodnot „velmi nízký", „střední", „vysoký" a kdy jsme získali následující rozdělení četností v těchto třech kategoriích (třídách):

velmi nízký	střední	vysoký	
5	10	9	$\sum n_i = n = 24.$

Teoretické rozdělení četností v těchto třech skupinách získáme na základě předpokladu nezávislosti volby jednotlivých kategorií, tj. u každého respondenta je stejná pravděpodobnost jejich volby. Je zřejmé, že očekávané (teoretické) četnosti budou rovny $24 : 3 = 8$.

Potom porovnáme empirické a teoretické četnosti:

5	10	9	n_i
8	8	8	$\tilde{n}_i$

Ověřujeme hypotézu, že počet respondentů je ve všech kategoriích přibližně stejný, tj. rozdíly empirických a teoretických četností jsou statisticky nevýznamné, náhodné.

Vypočítáme veličinu χ^2 ($df = k$) podle vzorce (19);

$$\chi^2 = \frac{(5-8)^2}{8} + \frac{(10-8)^2}{8} + \frac{(9-8)^2}{8} = 1{,}74.$$

V tabulce rozdělení χ^2, například pro hladinu statistické významnosti 0,05 a stupně volnosti $df = 3 - 1 = 2$, najdeme kritický bod $\chi^2_{kr} = 5{,}991$. Vypočítaná hodnota χ^2 z pozorovaných četností je menší než χ^2_{kr}, a z toho vyplývá, že rozložení četností je v souhlase s nulovou hypotézou, jinak řečeno, nedovoluje ji odmítnout.

Chí-kvadrát test používáme především v těch případech, kdy empirické údaje třídíme podle dvou nebo více znaků. Například máme výběrový soubor o 190 osobách a u něho rozložení názorů na nějakou otázku (tab. 8). Názory tohoto výběrového souboru roztřídíme do tří nezávislých věkových skupin (kategorií). Budeme testovat hypotézy: H_0 — neexistuje rozdíl v názorech vyjádřených v jednotlivých věkových skupinách; H_1 — existují rozdíly v názorech podle věku osob.

Tuto hypotézu budeme testovat pro hladinu významnosti $\alpha = 0{,}05$.

Pro výpočet očekávaných (teoretických) četností v každém políčku tabulky je nutné vynásobit odpovídající marginální četnosti (okrajové četnosti pod „celkem") a vydělit násobek celkovou sumou všech četností v tabulce. Například očekávaná četnost pro políčko (*a*) se rovná

$$\frac{60 \cdot 41}{190} = 12{,}9.$$

Tabulka 8

Odpověď	Věk dotazovaného (počet let)			Celkem
	více než 40	25—40	méně než 25	
Zcela nesouhlasím	(*a*) 18	(*b*) 13	(*c*) 10	41
Nesouhlasím	(*d*) 23	(*e*) 13	(*f*) 12	48
Souhlasím	(*g*) 11	(*h*) 14	(*ch*) 23	48
Zcela souhlasím	(*i*) 8	(*j*) 16	(*k*) 29	53
Celkem	60	56	74	190

Postup výpočtu najdeme v tabulce 9. Počet stupňů volnosti určíme podle vzorce

$$df = (r-1)(s-1),$$

kde r — počet řádků tabulky; s — počet sloupců tabulky.

Pro náš příklad (tab. 8) $df = (4-1) \cdot (3-1) = 6$. V tabulce B (viz příloha) najdeme hodnotu $\chi^2_{kr} = 16{,}812$. Z výpočtu provedeného v tabulce 9 vyplývá, že je nutné zamítnout nulovou hypotézu o tom, že nejsou rozdíly v názorech v různých věkových skupinách, tj. můžeme předpokládat, že existuje statisticky významná závislost mezi věkem respondenta a názorem, který zastává. Velikost χ^2 kritéria však nevypovídá nic o síle závislosti mezi proměnnými, pouze ukazuje na pravděpodobnost existence této závislosti. K určení intenzity závislosti je nutné vypočítat určitou míru této závislosti.

Pro korektní použití testů založených na χ^2 kritériu předpokládáme tyto podmínky. Výběrový soubor by měl být složen z nezávislých pozorování. Proměnné mohou být měřeny na jakékoli úrovni měření, ale žádná z očekávaných četností by neměla být menší než 5. Jestliže jsou v tabulce očekávané četnosti menší než 5, potom je nutné znak překategorizovat do menšího počtu kategorií (tříd), zmenšit počet stupňů volnosti sloučením některých řádků nebo sloupců nebo použít jiného, vhodnějšího testovacího kritéria.[1]

[1] Viz *J. V. Urbach*, Biometričeskije metody, Moskva 1964.

Tabulka 9

Políčko (z tab. 8)	Četnost	Očekávaná četnost $\tilde{n}_i$	$n_i - \tilde{n}_i$	$(n_i - \tilde{n}_i)^2$	$\frac{(n_i - \tilde{n}_i)^2}{\tilde{n}_i}$
a	18	12,9	5,1	26,00	2,015
b	13	12,1	0,9	0,81	0,056
c	10	16,0	6,0	36,00	2,250
d	23	15,2	7,8	60,84	4,002
e	13	14,1	1,1	1,21	0,085
f	12	18,7	6,7	44,89	2,400
g	11	15,2	4,2	17,64	1,160
h	14	14,1	0,1	0,01	0,000
ch	23	18,7	4,3	18,49	0,988
i	8	16,7	8,7	76,69	4,532
j	16	15,6	0,4	0,16	0,010
k	29	20,6	8,4	70,56	3,425
					$\chi^2 = 20{,}923$

6. Statistické souvislosti a jejich analýza

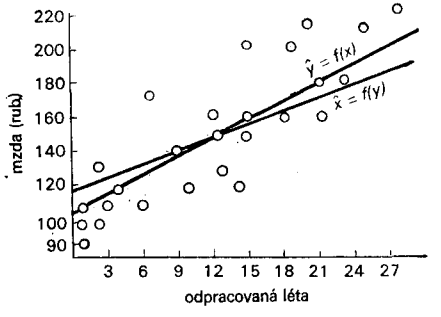
POJEM STATISTICKÉ ZÁVISLOSTI. Vyjděme ze známého tvrzení historického materialismu o všeobecné souvislosti a vzájemné podmíněnosti jevů společenského života, z něhož vyplývá, že sociolog se nemůže omezit na zkoumání jednotlivého a izolovaného jevu nebo procesu, ale musí se pokusit o zachycení celého komplexu jevů, které mají vztah k určitému společenskému procesu, a určit jejich vzájemnou závislost a podmíněnost.

Rozlišujeme dva typy závislostí: funkcionální (jako příklad můžeme uvést Newtonovy zákony v klasické fyzice) a statistické.

Zákonitosti hromadných společenských jevů se vytvářejí pod vlivem množství příčin vzájemně propojených v čase. K objasnění zákonitostí tohoto typu ve statistice přispívá výpočet statistických závislostí. Při řešení úkolu je vhodné vyčlenit dva aspekty: stanovení vzájemné závislosti mezi několika veličinami a stanovení souvislosti jedné nebo více veličin s ostatními. V principu spočívá první aspekt na teorii výpočtu korelací (korelační analýza), druhý princip souvisí s teorií regrese (regresní analýza). V tomto oddílu věnujeme pozornost nalezení vzájemné souvislosti několika znaků a základní principy regresní analýzy jen načrtneme.

Základem korelační analýzy statistické závislosti několika znaků je předpoklad o formě, směru i těsnosti (zhuštěnosti) vzájemné závislosti.

V tabulce 10 je ukázáno empirické roztřídění velikosti mzdy pracujících podle množství odpracovaných let (závislá proměnná) ve výběrovém souboru 25 osob a na obrázku 12 jsou tyto empirické údaje uvedeny ve formě diagramu. Obecně řečeno, vizuálně není vždy možné určit, zda existuje mezi proměnnými nějaká statisticky významná závislost, i když zpravidla je zřetelná nějaká tendence o vzájemné závislosti znaků i o směru této závislosti.



Obr. 12 Diagram korelační závislosti a regresní přímky pro vztah velikost platu a počet odpracovaných let

Regresní rovnice. Statistická závislost jednoho nebo většího počtu znaků na ostatních se vyjadřuje pomocí regresních rovnic. Předpokládejme dvě veličiny x a y, obdobně jako na obrázku 12. Jestliže budeme považovat x za stabilní, pak y, jak je vidět z obrázku, může nabývat různých hodnot. Označme y jako průměr z veličin y při stabilním x. Rovnice, která popisuje závislost průměru veličiny $\bar{y}_x$ na x, se nazývá regresní rovnicí y na x:

$$\hat{y}_x = F(x).$$