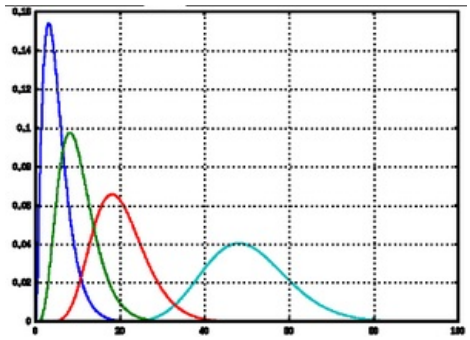
Čtyřpolní a kontingenční tabulka

Kontingenční tabulka je nástrojem používaným při studiu kategoriálních dat. Umožňuje především testovat *hypotézy o struktuře, hypotézy o nezávislosti* a *hypotézy o symetrii.* Kontingenční tabulku lze použít nejen pro kvalitativní a diskrétní kvantitativní data, po rozdělení do kategorií lze kontingenční tabulku použít i pro studium spojitých kvantitativních dat.

Čtyřpolní tabulka je speciálním případem kontingenční tabulky, kdy měřená data mohou nabývat právě jedné ze dvou kategorií, např. ano/ne, muž/žena nebo remise/progrese.

Kontingenční tabulky jsou přímou aplikací vlastností χ^2 (čti *„chí kvadrát“*) rozdělení, které popisuje chování nezávislých veličin s normalizovaným normálním rozdělením. U rozdělení χ^2 se udává **počet stupňů volnosti**; vlastně nejde o nic jiného, než o počet nezávislých náhodných veličin.



Příklad χ^2 rozdělení pro několik stupňů volnosti (modrá křivka pro 5 stupňů volnosti, zelená 10, červená 20 a světle modrá pro 50)

Test nezávislosti

Kontingenční tabulka představuje rozložení výskytů jednotlivých kombinací sledovaných znaků. Příkladem může být studium rozložení Rh faktoru a barvy očí. Výsledky (smyšleného) stanovení Rh faktoru u 100 modrookých a 300 hnědookých lze přehledně zapsat do tabulky:

barva očí	Rh^+	Rh^-	součet
modrá	35	65	100
hnědá	94	206	300
součet	129	271	400
zjištěné četnosti kombinací			

Nyní formulujeme **nulovou hypotézu**: *Rozložení znaků se navzájem neovlivňuje*. Za předpokladu **nezávislosti znaků** by mělo platit, že například podíl modrookých a hnědookých bude stejný jak v celkovém souboru, tak i ve skupině Rh^+ a Rh^-, analogicky i podíl Rh^+ a Rh^- ve skupině modrookých nebo hnědookých by měl být stejný jako v celém souboru. Pak lze sestavit tabulku odhadnutých hodnot:

barva očí	Rh^+	Rh^-	součet
modrá	32,25	67,75	100
hnědá	96,75	203,25	300
součet	129	271	400
odhadnuté četnosti kombinací			

Výpočet odhadu hodnoty v *i*-tém sloupci a *j*-tém řádku se určí tak, že se vynásobí součet hodnot v *i*-tém sloupci se součtem hodnot v *j*-tém řádku a podělí se počtem všech prvků v tabulce. Zjištěné četnosti výskytu se obvykle označují n_{ij}, kde *i* je příslušný řádek a *j* příslušný sloupec kontingenční tabulky. Četnosti odhadnuté za předpokladu nezávislosti znaků se obvykle označují m_{ij}. Pro testovou statistiku χ^2 platí:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}.$$

Pro data z příkladu bude platit:

$$\chi^2 = \frac{(35-32.25)^2}{32.25} + \frac{(65-67.75)^2}{67.75} + \frac{(94-96.75)^2}{96.75} + \frac{(206-203.25)^2}{203.25} \doteq 0.2289$$

Tuto hodnotu je třeba porovnat s kritickou hodnotou $\chi^2_\nu(\alpha)$. Parametr ν (někdy též *df)* se označuje jako **počet stupňů volnosti** a má hlubší matematický význam. V případě kontingenční tabulky se počet stupňů volnosti určí jako **součin počtu řádků zmenšeného o jedna s počtem sloupců zmenšeným o jedna**. Počet stupňů volnosti čtyřpolní tabulky je tedy roven jedné. Parametr α je požadovanou nejistotou prvního druhu, v biomedicínské statistice se značí jako hladina významnosti p.

Obvykle si předem stanovíme hladinu významnosti p a podle ní vyhodnocujeme výsledek. Stanovíme-li si tedy v modelovém příkladu ν rovno 0,05 (hladina významnosti je 5 %) a počet stupňů volnosti bude u čtyřpolní tabulky 1, vyjde nám:

$$\chi^2_1(0.05) = 3.84$$

Protože testová statistika nám vyšla nižší než kritická hodnota, nelze zamítnout nulovou hypotézu, tedy hypotézu, že krevní skupina není závislá na barvě očí.

Důležitým omezením použití kontingenčních tabulek při testování nezávislosti je požadavek, aby všechny odhadnuté četnosti byly větší než 5. Pokud by tato podmínka nebyla splněna, nebyly by splněny teoretické předpoklady, a tudíž by nebyla zaručena správnost výsledků. Této situaci lze v praxi čelit **slučováním několika skupin** do jedné. Pokud by se například ve výše uvedeném příkladu vyskytl jeden zelenooký člověk, vycházela by jeho odhadovaná četnost nízká. Problém by se vyřešil sloučením se skupinou modrookých, ovšem za cenu zvýšení rizika chyby druhého druhu.

Odkazy

Použitá literatura

- ANDĚL, Jiří. *Základy matematické statistiky.* 2. vydání. Praha : MATFYZPRESS, 2007. ISBN 80-7378-001-1.
- ZVÁROVÁ, Jana. *Základy statistiky pro biomedicínské obory* [online] . 1. vydání. Praha : Karolinum, 1998. Dostupné také z <http://new.euromise.org/czech/tajne/ucebnice/html/html/statist.html>. ISBN 80-7184-786-0.