Teorie potřebná ke zpracování datových souborů o zoonózách

a) Interval spolehlivosti pro pravděpodobnost úspěchu

b) Testování hypotézy o nezávislosti dvou nominálních veličin a Cramérův koeficient

c) Podíl šancí ve čtyřpolní kontingenční tabulce

d) Interval spolehlivosti pro podíl šancí

a) Interval spolehlivosti pro pravděpodobnost úspěchu

Asymptotické rozložení statistiky odvozené z výběrového průměru

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení $A(\vartheta)$ a nechť je splněna podmínka $n\vartheta(1-\vartheta) > 9$.

Pak statistika $U = \frac{M - \vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}}$ konverguje v distribuci k náhodné veličině se standardizovaným normálním rozložením.

Vysvětlení:

Protože $X_1, ..., X_n$ je náhodný výběr z rozložení $A(\vartheta)$, bude mít statistika $Y_n = \sum_{i=1}^{n} X_i$ (výběrový úhrn) rozložení $Bi(n, \vartheta)$. Y_n má střední hodnotu $E(Y_n) = n\vartheta$ a rozptyl $D(Y_n) = n\vartheta(1-\vartheta)$. Podle centrální limitní věty se standardizovaná statistika $U = \frac{Y_n - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}$ asymptoticky řídí standardizovaným normálním rozložením N(0,1). Pokud čitatele i jmenovatele podělíme n, dostaneme vyjádření:

$$U = \frac{\frac{Y_n}{n} - \vartheta}{\sqrt{\frac{n\vartheta(1-\vartheta)}{n^2}}} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}} = \frac{M - \vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}} \approx N(0,1)$$

Vzorec pro meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro parametr ϑ:

Meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro parametr ϑ jsou:

$$d = m - \sqrt{\frac{m(1-m)}{n}} u_{1-\alpha/2}, h = m + \sqrt{\frac{m(1-m)}{n}} u_{1-\alpha/2}.$$

Vysvětlení:

Pokud rozptyl $D(M) = \frac{\vartheta(1-\vartheta)}{n}$ nahradíme odhadem $\frac{M(1-M)}{n}$, konvergence náhodné veličiny U k veličině s rozložením N(0,1) se neporuší. Tedy

$$\forall \vartheta \in \Xi : 1-\alpha \le P\left(-u_{1-\alpha/2} < \frac{M-\vartheta}{\sqrt{\frac{M(1-M)}{n}}} < u_{1-\alpha/2} \right) =$$

$$= P\left(M - \sqrt{\frac{M(1-M)}{n}} u_{1-\alpha/2} < \vartheta < M + \sqrt{\frac{M(1-M)}{n}} u_{1-\alpha/2} \right)$$

(Tyto meze lze vypočítat ve STATISTICE pomocí modulu Analýza síly testu.)

b) Testování hypotézy o nezávislosti dvou nominálních veličin a Cramérův koeficient

Testujeme nulovou hypotézu H_0: X, Y jsou stochasticky nezávislé náhodné veličiny proti alternativě H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny. Přitom X má r variant a Y s variant.
Kdyby náhodné veličiny X, Y byly stochasticky nezávislé, pak by platil multiplikativní vztah

$\forall j = 1,\dots,r\ ,\ \forall k = 1,\dots,s : \pi_{jk} = \pi_{j.}\,\pi_{.k}$ neboli $\frac{n_{jk}}{n} = \frac{n_{j.}}{n}\cdot\frac{n_{.k}}{n}$, tj. $n_{jk} = \frac{n_{j.}n_{.k}}{n}$.

Číslo $\frac{n_{j.}n_{.k}}{n}$ se nazývá teoretická četnost dvojice variant $(x_{[j]}, y_{[k]})$.

Testová statistika: $K = \sum_{j=1}^{r}\sum_{k=1}^{s}\frac{\left(n_{jk} - \frac{n_{j.}n_{.k}}{n}\right)^2}{\frac{n_{j.}n_{.k}}{n}}$. Platí-li H_0, pak K se asymptoticky řídí rozložením $\chi^2((r-1)(s-1))$.

Kritický obor: $W = \left\langle \chi^2_{1-\alpha}((r-1)(s-1)), \infty \right)$.

Hypotézu o nezávislosti veličin X, Y tedy zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}((r-1)(s-1))$.

Podmínky dobré aproximace

Rozložení statistiky K lze aproximovat rozložením $\chi^2((r-1)(s-1))$, pokud teoretické četnosti $\frac{n_{j.}n_{.k}}{n}$ aspoň v 80 % případů nabývají hodnoty větší nebo rovné 5 a ve zbylých 20 % neklesnou pod 2. Není-li splněna podmínka dobré aproximace, doporučuje se slučování některých variant.

Měření síly závislosti

Cramérův koeficient: $V = \sqrt{\frac{K}{n(m-1)}}$, kde m = min{r,s}. Tento koeficient nabývá hodnot mezi 0 a 1. Čím blíže je k 1, tím je závislost mezi X a Y těsnější, čím blíže je k 0, tím je tato závislost volnější.

Význam hodnot Cramérova koeficientu:
mezi 0 až 0,1 … zanedbatelná závislost,
mezi 0,1 až 0,3 … slabá závislost,
mezi 0,3 až 0,7 … střední závislost,
mezi 0,7 až 1 … silná závislost.

c) Podíl šancí ve čtyřpolní kontingenční tabulce

Výsledek pokusu	okolnosti		$n_{j.}$
	I	II	
úspěch	a	b	a+b
neúspěch	c	d	c+d
$n_{.k}$	a+c	b+d	n

Poměr počtu úspěchů ku počtu neúspěchů za okolností I je $\frac{a}{c}$ (šance na úspěch za okolností I).

Poměr počtu úspěchů ku počtu neúspěchů za okolností II je $\frac{b}{d}$ (šance na úspěch za okolností II).

Podíl těchto dvou poměrů je podíl šancí: $OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}$.

Pokud OR =1, pak okolnosti nemají vliv na výskyt jevu.
Pokud OR > 1, pak za okolností I je vyšší šance na výskyt jevu než za okolností II.
Pokud OR < 1, pak za okolností I je nižší šance na výskyt jevu než za okolností II.

Podíl šancí považujeme za odhad neznámého teoretického podílu šancí $o\rho = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$.

d) Interval spolehlivosti pro podíl šancí

Logaritmus teoretického podílu šancí oρ má přibližně normální rozložení a směrodatná odchylka jeho odhadu, tj. logaritmu podílu šancí OR, je $\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$.

Meze 100(1-α)% asymptotického intervalu spolehlivosti pro ln oρ jsou

$$\ln OR - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}u_{1-\alpha/2}, \ln OR + \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}u_{1-\alpha/2}.$$

Odlogaritmováním dostaneme meze 100(1-α)% asymptotického intervalu spolehlivosti pro oρ:

$$d = \exp\left(\ln OR - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}u_{1-\alpha/2}\right), h = \exp\left(\ln OR + \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}u_{1-\alpha/2}\right)$$

(Tyto meze lze ve STATISTICE vypočítat pomocí modulu Pokročilé lineární/nelineární modely – Zobecněné lineární/nelineární modely – Logitový model.)