Asociační tabulka

Zadání příkladu:

Zdravotní pojišťovna prováděla v jistém okresním městě průzkum očkování proti jisté nemoci. Náhodně bylo vybráno a dotázáno 60 jeho obyvatel, zdali očkování mají, nebo nemají. Výsledky jsou v tabulce 1. Otestujte hypotézu, zda pohlaví dotazované osoby a to, zdali má sledované očkování, jsou nezávislé veličiny. Hladina významnosti $\alpha = 5\ \%$

Tab. 1 - Průzkum očkování

	Očkován	Neočkován	Součet $n_{i.}$
Muž	4	26	30
Žena	21	9	30
Součet $n_{.j}$	25	35	60

Vypracování příkladu

Pokud by sledované znaky měli zařadit podle typu proměnné, jedná se o nominální proměnné, které jsou typické tím, že nemají uspořádání a jsou diskrétní, jinak řečeno kategoriální. Označme sledované znaky písmeny A (pohlaví) a B (má/nemá očkování). Pro test nezávislosti dvou kategoriálních znaků, kdy oba nabývají pouze dvou hodnot, používáme tzv. *asociační tabulku* - viz tabulka 1. V řádcích jsou hodnoty jednoho znaku (v tomto případě pohlaví dotazované osoby), ve sloupcích hodnoty druhého znaku (stav očkování). Na průnicích řádků a sloupců odpovídajících všem čtyřem kombinacím těchto dvojic hodnot jsou absolutní četnosti výskytu těchto dvojic. V posledním sloupci jsou uvedeny řádkové součty, tj. celkový počet mužů a žen, v posledním řádku sloupcové součty - celkový počet očkovaných a neočkovaných osob. V pravém dolním rohu potom celkový počet dotazovaných osob, který se musí rovnat jak součtu řádkových, tak sloupcových součtů.

Jednotlivé četnosti označíme postupně $n_{11}, n_{12}, n_{12}, n_{22}$, kde první index je řádkový, druhý sloupcový. Jejich součet označíme n. Řádkové a sloupcové součty jsou již označeny v tabulce 1. Test nezávislosti provedeme standardně v pěti krocích:

1. **Formulovat hypotézu**
   H_0: Znaky A a B jsou nezávislé.
   H_1: Znaky A a B nejsou nezávislé.

2. **Zvolit testové kritérium (TK)**
   U asociační tabulky je TK následující:

$$\chi^2 = n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}} \sim \chi^2(1)$$

   Uvedená veličina χ^2 má asymptoticky rozdělení $\chi^2(1)$.

3. **Sestrojit kritický obor** W
   Kritický obor W jsou ty hodnoty testového kritéria, které vedou k zamítnutí nulové hypotézy. V tomto případě platí:

$$W = \{\chi^2 : \chi^2 \geq \chi^2(1)_{1-\alpha}\}$$

   Konkrétně pro $\alpha = 5$ % je

$$W = \{\chi^2 : \chi^2 \geq \chi^2(1)_{0,95} = 3{,}84\}$$

4. **Spočítat hodnotu TK**

$$\chi^2 = 60.\frac{(4.9 - 21.26)^2}{30.30.25.35} = 19{,}817 \tag{1}$$

5. **Formulovat závěr**
   V tomto případě platí, že TK $\chi^2 \in W$, protože 19,817 $\geq$ 3,84. Takže nulovou hypotézu H_0 zamítáme a přijímáme H_1.

   Závěr: sledované znaky pohlaví a očkování nejsou nezávislé.

Pokud zamítneme nulovou hypotézu jako v tomto případě, je možné ještě změřit *sílu závislosti* sledovaných znaků pomocí **koeficientu asociace** $r_{AB} \in [-1, 1]$

$$r_{AB} = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}} \tag{2}$$

Čím je r_{AB} bližší jedné, tím jsou vyšší četnosti n_{11} a n_{22} (hlavní diagonála tabulky). Naopak čím je hodnota koeficientu asociace bližší -1, převládají četnosti n_{12} a n_{21}. Síla závislosti roste s absolutní hodnotou r_{AB}. Zde je

$$r_{AB} = \frac{4.9 - 21.26}{\sqrt{30.30.25.35}} = \text{-}0{,}575 \tag{3}$$

Program STATGRAPHICS

V programu STATGRAPHICS je výpočet snadný. Předpokládejme, že četnosti z asociační tabulky jsou zadány ve dvou proměnných: Ockovan (postupně 4, 21) a Neockovan (9, 26). Výchozím bodem je funkcionalita *Describe/Categorical Data/Contingency Tables (2-way)*.

Do pole Columns zadáme proměnné Ockovan a Neockovan a klikneme na OK. V dalším formuláři zaškrtneme tabulky *Analysis Summary, Frequency Table, Tests of Independence, Summary Statistics* a opět klikneme na OK. Ve vygenerované analýze v tabulce *Tests of Independence* je uvedena hodnota testového kritéria χ^2 jako `Statistic`. Dále počet stupňů volnosti `Df` a hodnota `P-Value`. Pokud je P-Value nižší než hladina významnosti α, nulovou hypotézu zamítáme. To je uvedeno i v textu pod tabulkou.

```
The StatAdvisor
This table shows the results of a hypothesis test run to determine
whether or not to reject the idea that the row and column classifications
are independent. Since the P-value is less than 0,05, we can reject the
hypothesis that rows and columns are independent at the 95,0% confidence
level. Therefore, the observed row for a particular case is related to
its column.
```

Uveďme nyní ještě postup, jak zjistit hodnoty kvantilů rozdělení $\chi^2(n-1)$ potřebné pro konstrukci kritického oboru. Ve STATGRAPHICS je najdeme ve formuláři *Describe/Distribution Fitting/Probability Distributions*, kde jsou uvedena všechna pravděpodobnostní rozdělení, která jsou v tomto programu k dispozici. V seznamu rozdělení vybereme Chi-Square a klikneme na OK. V dalším kroku je třeba zadat počet stupňů volnosti (D. F.), v tomto případě je to 1.

V dalším formuláři *Tables and Graphs* je třeba zaškrtnout tabulku *Inverse CDF*, ve které je možné zjistit hodnotu libovolného kvantilu příslušného rozdělení. Stačí najet na tabulku myší a kliknout pravým tlačítkem. Z menu zvolit *Pane Options* a do formuláře *Inverse CDF Options* zadat příslušnou hodnotu procent číslem z intervalu $(0, 1)$, zde konkrétně 0,95. Potom klikneme na OK a z tabulky odečteme hodnotu kvantilu, která je rovna 3,84.

STATGRAPHICS počítá i hodnotu koeficientu asociace r_{AB} podle (2). Ta je v tabulce *Summary Statistics* jako `Pearson's R` a je rovna -0,5747, po zaokrouhlení -0,575, jak je uvedeno výše v (3).

Interpretace výsledků

Na základě provedeného testu byla zamítnuta hypotéza nezávislosti pohlaví a očkování respondenta na dané hladině významnosti. Vypočtená hodnota TK (1) překračuje kritickou hodnotu 3,84, leží tedy v kritickém oboru W. Koeficient asociace r_{AB} = -0,575. Sílu závislosti lze proto označit jako střední s tím, že dominují četnosti na vedlejší diagonále tabulky.

Program MS Excel

Tabulkový kalkulátor k testování nezávislosti v asociační tabulce obecně použít lze, protože uvedené vzorce jsou velmi jednoduché, nicméně přesný postup výpočtů v programu MS Excel uvádět nebudeme. Kvantily rozdělení χ^2 doporučujeme dohledat ve statistických tabulkách, které jsou nejsnáz dostupné na Internetu. Kvantilové funkce Excelu vzhledem ke špatným zkušenostem s nimi pro tyto účely nedoporučejeme.