Jak volit hypotézy, rozdělení atd.

(P1) Chceme otestovat hypotézu, že např. procento vysokoškolsky vzdělaných lidí je v jednotlivých krajích ČR stejné. Kvůli relativnosti (počet VŠ obyvatel vztažených na počet obyvatel kraje) bychom zde ale ztratili informaci o četnostech takovéto náhodné veličiny. Naštěstí lze daný problém ekvivalentně vyjádřit jako hypotézu, že veličiny

Y("daný člověk") := "kraj, ve kterém tento člověk žije"

Z("daný člověk") := "jestli tento člověk má VŠ nebo ne"

jsou nezávislé.

(P2) Chceme otestovat hypotézu, že např. počet lidí na Zemi roste exponenciálně. Záznamy o počtech obyvatel jsou známy z jednotlivých let, proto se zdá, že lepší bude uvažovat o diskrétní analogii exponenciálního rozdělení, tedy o geometrickém rozdělení. Data jsou dále známa od určitého roku v minulosti (rok T_0) až do jiného bodu (označme ho jako rok T_1) a navíc záznamy z některých let mohou chybět. Tomu musíme přizpůsobit i naši veličinu. Označme si I množinu všech let, ze kterých máme záznamy. Zřejmě T_1 je maximum I a T_0 je minimum množiny I. Naše veličina pak bude

X("záznam o daném člověku v daném roce, který patří do I"):= T_1 - "rok, do kterého záznam přísluší"

Obor hodnot této veličiny je $J = \{T_1 - i \mid i \text{ pochází z } I\}$. Z podstaty věci má naše veličina X nulovou pravděpodobnost pro všechny ostatní hodnoty, tedy pro hodnoty MIMO množinu J (jde o tzv. useknuté geometrické rozdělení, které ještě navíc může být "děravé"). Celkově tedy potřebujeme pracovat s veličinou X s pravděpodobnostní funkcí

$$p_X(j) = \begin{cases} c \cdot \exp(-j/T), & \text{pro } j \in J \\ 0, & \text{jinak.} \end{cases}$$

Zde c a T jsou parametry, které je potřeba odhadnout (asi nejlépe metodou momentů, případně metodou max. věrohodnosti). Parametr c snadno vyjádříme pomocí parametru T z podmínky, že součet pravděpodobností má být roven 1. K výpočtu parametru T pak musíme vyřešit polynomiální rovnici v proměnné $q := \exp(-1/T)$, která bude stupně $T_1 - T_0$. Její řešení určíme přibližně pomocí vhodného softwaru. Pro chi-kvadrát test pak budeme mít (kvůli odhadu 2 konstant) ještě o 2 stupně volnosti méně.

Podobně by se postupovalo v případě, že časové záznamy veličiny X by byly spojité (pak by šlo o useknuté exponenciální rozdělení, které ještě případně může být "děravé").

Pokud nemáme "díry"v datech, rovnice pro stanovení parametrů jsou pro obě metody stejné a sice tyto:

Mějme veličinu X s pravděpodobnostní funkcí

$$p_X(j) = \begin{cases} c \cdot q^j, & \text{pro } j = 0, \ldots, n \\ 0, & \text{jinak.} \end{cases}$$

Maximálně věrohodný odhad, stejně jako metoda momentů pak poskytují tyto rovnice pro parametry c a q:

$$(\overline{x} - n) \cdot q^{n+2} + (n + 1 - \overline{x}) \cdot q^{n+1} - (\overline{x} + 1) \cdot q + \overline{x} = 0$$

$$c = \frac{1 - q}{1 - q^{n+1}}$$

kde $\overline{x}$ je výběrový průměr veličiny X.

Podobně to bude pro spojitou veličinu X s parametry c a T a hustotou

$$f_X(t) = \begin{cases} c \cdot e^{-\frac{t}{T}}, & \text{pro } t \in \langle 0, n\rangle \\ 0, & \text{jinak.} \end{cases}$$

Metoda momentů pak dává rovnice

$$T = \frac{n}{1 - e^{-\frac{n}{T}}} + \overline{x} - n$$

$$c = \frac{1}{T \cdot (1 - e^{-\frac{n}{T}})} \quad \left(= \frac{1}{n} + \frac{1}{T} - \frac{\overline{x}}{nT} \right)$$

kde opět $\overline{x}$ je výběrový průměr veličiny X. K numerickému vyřešení první rovnice můžeme zkusit např. iterační metodu.

(P3) Mějme veličinu X s alternativním rozdělením $\mathrm{Alt}(p)$, kde $p \in (0,1)$. Chceme otestovat nulovou hypotézu

$$\mathbf{H}_0 : p \geq p_0$$

proti alternativní hypotéze:

$$\mathbf{H}_1 : p < p_0$$

kde $p_0 \in (0,1)$.

Nemáme zde sice nějakou veličinu, jejíž rozdělení by přímo bylo "tabulkové"a nezáviselo na p, ale při větším rozsahu měření (tj. větším počtu měření n) můžeme použít CLV, která nám patřičné rozdělení, v tomto případě $N(0,1)$, poskytne.

Vezmeme si proto nezávislé náhodné veličiny X_i (kopie veličiny X) a jejich výběrový průměr

$$\overline{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

pro který je $E(\overline{\boldsymbol{X}}) = p$ a $D(\overline{\boldsymbol{X}}) = \frac{p(1-p)}{n}$. Teď si zvolíme tuto statistiku

$$T = \frac{\overline{\boldsymbol{X}} - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n} \ .$$

Z CLV víme, že T má pro $p = p_0$ přibližně (normované) normální rozdělení $N(0,1)$ (pro patřičně velká n).

Poznámka: Tato statistika je analogií statistiky

$$T' = \frac{\overline{\boldsymbol{X}'} - \mu_0}{\sigma} \sqrt{n},$$

pro případ veličiny X' s normálním rozdělením $N(\mu, \sigma)$ a pro nulovou hypotézu

$$\mathbf{H}_0' : \mu \geq \mu_0 \ .$$

Pozor! Nenaznačujeme tím, že by naše původní veličina X s alternativním rozdělením snad měla vlastnosti nějaké jiné veličiny X' s normálním rozdělením! Jde tu o to, že při hledání kritického oboru pro X (při dané hladině významnosti α) je postup principiálně stejný jako pro případ, kdy X' má normální rozdělení - viz dále.

Kritérium pro **ZAMÍTNUTÍ** teď bude stejné jako pro test u normálního rozdělení, a sice

$$t < -\Phi^{-1}(1-\alpha) \quad \Rightarrow \quad \text{zamítáme } \mathbf{H}_0 \text{ (na dané hladině } \alpha) \ .$$

Zdůvodnění tvaru zamítacího kritéria: Abychom si více uvědomili závislost veličiny X na parametru p, budeme ji vyznačovat jako $X_{(p)}$ a podobně pro statistiku $T_{(p)} = \frac{\overline{\boldsymbol{X}}_{(p)} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$. Protože předpokládáme $\mathbf{H}_0 :\ p \geq p_0$, a tedy $E(\overline{\boldsymbol{X}}_{(p)}) = p$, dostáváme, že $E(T_{(p)}) = \frac{p - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \geq 0$. Takže očekávané hodnoty statistiky $T_{(p)}$ budou především v intervalu $\langle 0, \infty)$ a jeho blízkém okolí. Jako kritický obor si proto zvolíme

$$W : (-\infty, u_1) \ ,$$

kde požadujeme, aby $u_1 \in \mathbb{R}$ bylo největší takové, aby chyba 1. druhu byla nejvýše α, tj.

$$\Big(\forall\ p \geq p_0\Big) \quad P\big(T_{(p)} \in W\big) = P\big(T_{(p)} < u_1\big) \leq \alpha \ .$$

Je nutné zdůraznit, že za předpokladu nulové hypotézy (tj. že $p \geq p_0$) statistika $T_{(p)}$ obecně **NEMÁ** rozdělení $N(0,1)$ (toto rozdělení se díky CLV objeví právě jen pokud $p = p_0$).

Přesto ho ale nakonec použijeme, protože případ $p = p_0$ je za předpokladu $\mathbf{H}_0$ ten "nejhorší" možný, jak bude vidět z následujícího:

Protože platí

$$p \geq p_0 \quad \Rightarrow \quad \Big(\forall\ t \in \mathbb{R}\Big) \ \ F_{\mathrm{Bi}(n,p)}(t) \quad \geq \quad F_{\mathrm{Bi}(n,p_0)}(t)$$

(což lze snadno ukázat derivováním funkce $\varphi(p) = F_{\mathrm{Bi}(n,p)}(t)$ podle p pro pevně zvolená n a t) tak dostáváme, že také

$$p \geq p_0 \quad \Rightarrow \quad \Big(\forall\ t \in \mathbb{R}\Big) \ \ F_{T_{(p)}}(t) \quad \geq \quad F_{T_{(p_0)}}(t)$$

protože $T_{(p)} = \frac{\widetilde{\boldsymbol{X}}_{(p)} - np_0}{\sqrt{np_0(1-p_0)}}$ je jen lineární transformací veličiny $\widetilde{\boldsymbol{X}}_{(p)} := \sum_{i=1}^{n}(X_i)_{(p)}$ s rozdělením $\mathrm{Bi}(n,p)$. Máme tak, že

$$p \geq p_0 \quad \Rightarrow \quad P\Big(T_{(p)} < u_1\Big) = F_{T_{(p)}}(u_1\ -) \leq F_{T_{(p_0)}}(u_1\ -) = P\Big(T_{(p_0)} < u_1\Big)$$

Vidíme tedy, že $P\Big(T_{(p)} < u_1\Big) \leq P\Big(T_{(p_0)} < u_1\Big)$ a hledané u_1 tak musí splňovat, že

$$P\Big(T_{(p_0)} < u_1\Big) = \alpha$$

tedy aproximací podle CLV to je

$$\Phi(u_1) = \alpha$$

neboli

$$u_1 = \Phi^{-1}(\alpha) = -\Phi^{-1}(1-\alpha)$$

a kritérium pro **ZAMÍTNUTÍ** je tak skutečně tvaru

$$t < -\Phi^{-1}(1-\alpha) \quad \Rightarrow \quad \text{zamítáme } \mathbf{H}_0 \text{ (na dané hladině } \alpha) \ .$$