Marie Turčičová

Analýza rozptylu (ANOVA)

22. dubna 2020

Představme si situaci, kdy chceme otestovat rovnost středních hodnot nějaké spojité veličiny u více skupin. Například máme populace desetiletých chlapců z pěti států Evropy a chceme otestovat, zda je populační průměr výšky těchto chlapců stejný ve všech těchto pěti zemích. Formálně bychom to zapsali takto:

$$\begin{aligned} x_{11}, \ldots, x_{1n_1} &= \text{ výšky vybraných desetiletých chlapců z 1. země} \\ x_{21}, \ldots, x_{2n_2} &= \text{ výšky vybraných desetiletých chlapců z 2. země} \\ &\vdots \\ x_{51}, \ldots, x_{5n_5} &= \text{ výšky vybraných desetiletých chlapců z 5. země.} \end{aligned}$$

Všimněte si, že výběry nemusejí být stejného rozsahu, tedy $n_1, \ldots, n_5$ mohou být obecně různá čísla. Označíme jakožto $\mu_1, \mu_2, \ldots, \mu_5$ střední výšky (alias populační průměry výšek) chlapců v jednotlivých zemích. Chceme testovat hypotézu

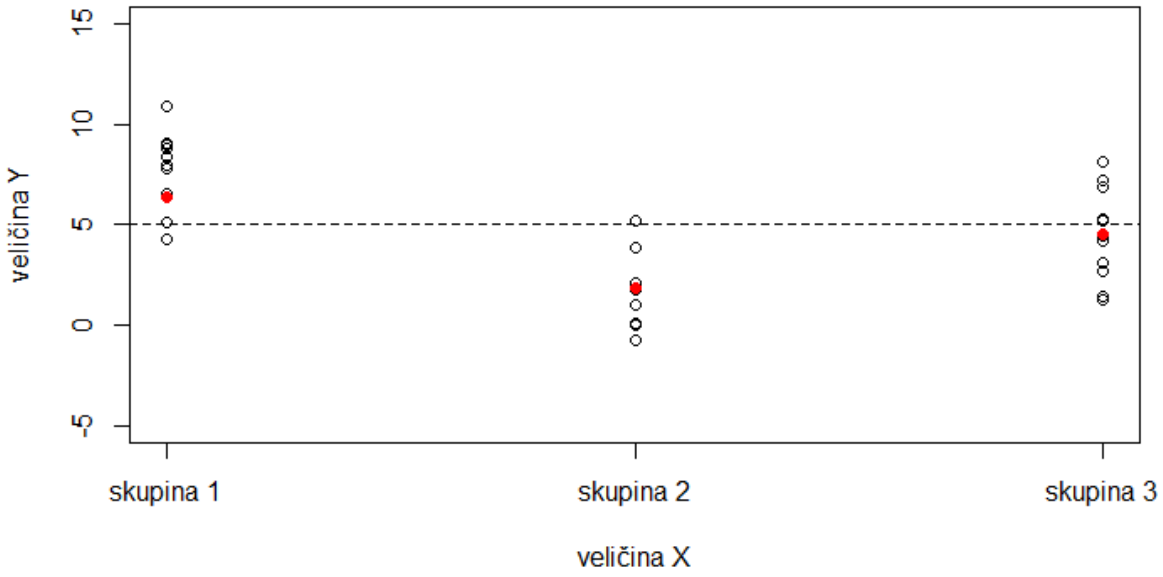
$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \tag{1}$$

proti alternativní hypotéze H_1, že H_0 neplatí, tedy existují alespoň dvě populace, které nemají stejnou střední hodnotu. Navíc předpokládejme, že daný znak (pro nás tedy výška desetiletých chlapců) má normální rozdělení, tedy

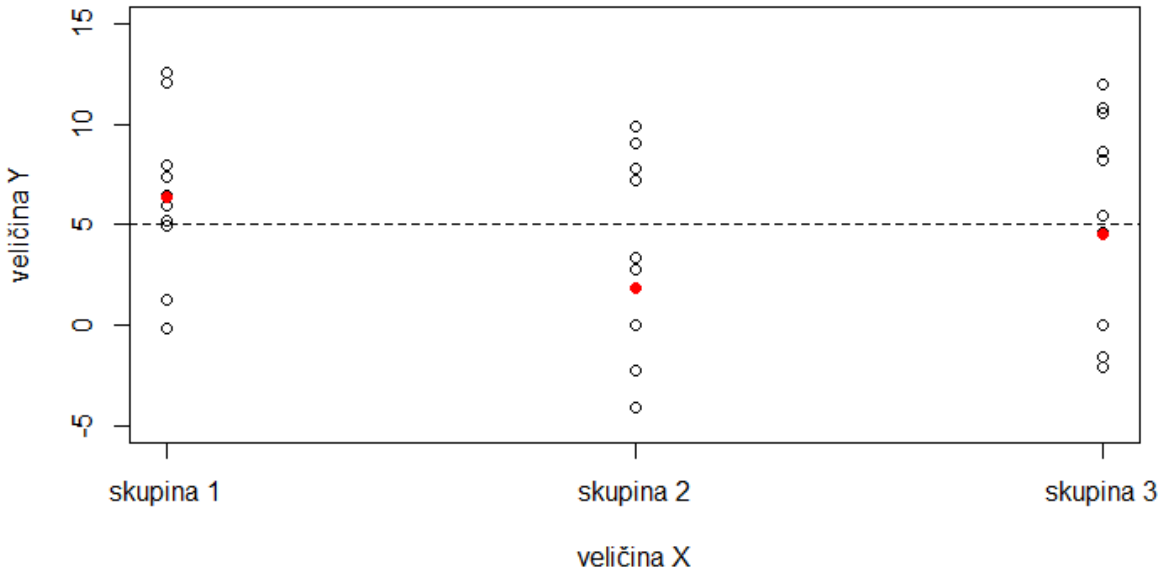
$$\begin{aligned} &X_{11}, \ldots, X_{1n_1} \text{ je výběr z } \mathcal{N}(\mu_1, \sigma^2) \\ &X_{21}, \ldots, X_{2n_2} \text{ je výběr z } \mathcal{N}(\mu_2, \sigma^2) \\ &\vdots \\ &X_{51}, \ldots, X_{5n_5} \text{ je výběr z } \mathcal{N}(\mu_5, \sigma^2). \end{aligned}$$

Někoho by mohlo napadnout použít dvouvýběrový t-test na každou dvojici (tedy celkem 10 dvojic) a naši hypotézu H_0 zamítnout, pokud alespoň pro jednu z těch 10 dvojic náš t-test zamítne rovnost středních hodnot. Problém s tímto přístupem je, že takto sestrojená metoda by nedodržela celkovou hladinu testování α. Kdyby v každém t-testu byla hladina α, tak tedy pravděpodobnost, že v každém z těchto t-testů zamítneme nulovou hypotézu, která platí, je rovna α. Avšak když máme velký počet skupin, provádíme t-testů opravdu mnoho a pravděpodobnost, že alespoň v jednom z nich zamítneme rovnost středních hodnot se s počtem skupin zvyšuje, a tedy i celková hladina naší metody, která bude ve výsledku určitě větší než α.

Test rovnosti středních hodnot je tedy potřeba provést pro všechny výběry najednou, nikoli sekvenčně. To umožňuje metoda zvaná analýza rozptylu. Na čem je založena? Podívejte se na obrázky níže, na nichž jsou vykresleny hodnoty nějakého kvantitativního znaku ve třech různých skupinách, a zkuste si odpovědět na otázku, na kterém obrázku vám skupiny připadají odlišnější a proč. Všimněte si přitom, že průměry jednotlivých skupin (označené červenými puntíky) mají na obou obrázcích stejné hodnoty.



(a)



(b)

Obrázek 1: Hodnota spojité veličiny (veličina Y) v jednotlivých skupinách definovaných kategoriální veličinou X. Červeně je znázorněn výběrový průměr v každé skupině, čárkovaná čára znázorňuje celkový průměr vypočítaný ze všech pozorování.

Pravděpodobně jste si odpověděli, že skupiny se jeví odlišnější na Obrázku 1a. Důvodem je to, že rozptýlenost červených průměrů kolem celkového průměru je na Obrázku 1b relativně malá v porovnání s rozptýleností bodů v jednotlivých skupinách. Zato na Obrázku 1a je rozptýlenost červených bodů poměrně velká vzhledem k rozptylu bodů uvnitř skupin. Při testu rovnosti středních hodnot tedy stačí porovnat rozptyl hodnot v jednotlivých skupinách s rozptylem průměrů skupin. Přesně na tomto pozorování je založena "analýza rozptylu"(angl. ANalysis Of VAriance).

Předpoklady

Prvním předpokladem je, že pozorování v jednotlivých výběrech jsou nezávislá a že i **výběry jsou navzájem nezávislé**. Tento předpoklad se nijak netestuje, pouze je potřeba si jeho rozumnost odůvodnit z designu pokusu.

Druhým předpokladem je, že **výběry pocházejí z normálního rozdělení**. Tento předpoklad se otestuje pomocí Shapiro-Wilkova testu, který se PO provedení analýzy rozptylu aplikuje na tzv. standardizovaná rezidua. Více o tom za chvíli.

Posledním předpokladem je to, že výběry pocházejí z **rozdělení se shodnými rozptyly**. Tomuto jevu se říká "homoskedasticita"a testuje se pomocí Leveneova nebo Bartlettova testu. Detailní výklad těchto testů by přesahoval rámec tohoto textu, proto se spokojíme pouze s jejich praktickým použitím v programu R, kde se vždy na základě p-hodnoty rozhodneme, zda shodu rozptylů můžeme předpokládat, nebo ne. Každý z těchto testů je konstruován zcela jinak a při použití na shodná data můžeme dostat různé výsledky (a to dokonce p-hodnoty vedoucí k opačnému závěru). Při vybírání vhodného testu je potřeba mít na paměti, že Bartlettův test je citlivý zejména na splnění předpokladu normálního rozdělení. Nicméně jak se později ukáže, předpoklad shodných rozptylů není zcela nezbytný a existuje modifikace analýzy rozptylu, která se bez shodných rozptylů obejde.

Formální zápis analýzy rozptylu

Místo zkoumání rozdílů výběrových průměrů (jak to činí dvouvýběrový t-test) se analyzuje kolísání výběrových průměrů kolem celkového průměru. Nejprve si zavedeme značení:

- K = počet skupin (v příkladu z úvodu bylo $K = 5$)
- $\bar{X}_{..}$ = celkový výběrový průměr (na Obrázku 1 znázorněn černou čárkovanou čarou)
- $\bar{X}_{i.}$ = průměr i-tého výběru (na Obrázku 1 znázorněné červenými puntíky)
- $n = \sum_{i=1}^{K} n_i$ = celkový počet pozorování (ze všech skupin dohromady).

Principem metody je rozklad celkového rozptylu, který v datech je, na variabilitu mezi skupinami a variabilitu uvnitř skupin. Formálně to můžeme zapsat takto:

$$\sum_{i=1}^{K}\sum_{t=1}^{n_i}(X_{it} - \bar{X}_{..})^2 = \sum_{i=1}^{K} n_i(\bar{X}_{i.} - \bar{X}_{..})^2 + \sum_{i=1}^{K}\sum_{t=1}^{n_i}(X_{it} - \bar{X}_{i.})^2$$

$$\text{(celková variabilita)} = \text{(variabilita mezi)} + \text{(variabilita uvnitř)}$$

Jednotlivé součty čtverců ze vzorečku výše mají následující označení a význam.

- celková variabilita $S_T = \sum_{i=1}^{K}\sum_{t=1}^{n_i}(X_{it} - \bar{X}_{..})^2$ označuje celkový rozptyl zkoumané veličiny, která v datech je (jedná se o součet čtvercových odchylek jednotlivých pozorování od celkového průměru)
- variabilita průměrů $S_A = \sum_{i=1}^{K} n_i(\bar{X}_{i.} - \bar{X}_{..})^2$ označuje tu část variability pozorování, kterou lze vysvětlit tím, že pozorování patří do různých skupin s různou střední hodnotou
- reziduální (zbytková) variabilita $S_e = \sum_{i=1}^{K}\sum_{t=1}^{n_i}(X_{it} - \bar{X}_{i.})^2$ je ta část celkové variability, která v datech zbude i když připustíme rozdílné střední hodnoty ve skupinách

Aby součty čtverců výše byly právoplatným odhadem rozptylů, je potřeba je ještě vydělit tzv. stupni volnosti. Dostáváme tak průměrné čtverce (*mean squares*):

- $MS_T = \frac{1}{n-1} S_T$ (kde označíme $f_T = n - 1$)
- $MS_A = \frac{1}{K-1} S_A$ (kde označíme $f_A = K - 1$)
- $MS_e = \frac{1}{n-K} S_e$ (kde označíme $f_e = n - K$)

Všimněte si, že také pro stupně volnosti platí vztah $f_T = f_A + f_e$. Průměrný čtverec MS_e je odhadem onoho společného rozptylu σ^2, značí se s^2 a nazývá se **reziduální rozptyl**. Tedy

$$s^2 = \frac{1}{n-K} \sum_{i=1}^{K} \sum_{t=1}^{n_i} (X_{it} - \bar{X}_{i.})^2. \tag{2}$$

Rigorózní porovnání rozptylu v rámci skupin a rozptylu mezi skupinami se provádí pomocí F-statistiky, která je podílem těchto dvou rozptylů:

$$F_A = \frac{S_A/(K-1)}{S_e/(n-K)} = \frac{MS_A}{MS_e}. \tag{3}$$

Je-li variabilita mezi skupinami menší nebo srovnatelná s variabilitou uvnitř skupin (tj. F_A je malá), není důvod H_0 zamítat. Je-li však meziskupinová variabilita větší než vnitroskupinová (tj. F_A je velká, jak je tomu např. na Obrázku 1a), tak střední hodnoty skupin prohlásíme za různé a zamítneme nulovou hypotézu. Kvantitativní hranice pro to, kdy už je statistika F_A moc velká, je dána kvantilem Fisherova-Snedecorova rozdělení (zkráceně jen F-rozdělení) se stupni volnosti f_A a f_e. Tedy H_0 zamítnu pokud

$$F_A = \frac{S_A/(K-1)}{S_e/(n-K)} \geq F_{f_A, f_e}(1-\alpha). \tag{4}$$

Poznámka pro experty:. (možno přeskočit) Za platnosti H_0 je taktéž rozptyl MS_A odhadem σ^2. Pokud H_0 neplatí, má MS_A tendenci být větší než σ^2. Tedy, pokud H_0 platí, je testová statistika F_A podílem dvou výrazů odhadujících tutéž věc (σ^2) a měla by být tedy přibližně rovna jedné.

Výstupem analýzy rozptylu je tabulka, která má v prvním sloupci uvedeny jednotlivé složky variability (v prvním řádku je variabilita, která je vysvětlená různými středními hodnotami ve skupinách definovaných kategoriální proměnnou) a v dalších sloupcích jsou jednotlivé čtverce a testová statistika. Pořadí sloupců se může v různých statistických programech lišit. Stejně tak poslední řádek mnohdy chybí.

variabilita	součty čtverců	stupně volnosti	průměrné čtverce	testová statistika	p-hodnota
skupiny	S_A	$f_A = K - 1$	$M_A = S_A/f_A$	F_A	p
reziduální	S_e	$f_e = n - K$	$M_e = S_e/f_e$		
celková	S_T	$f_T = n - 1$			

Test předpokladu o normálním rozdělení

Po provedení analýzy rozptylu je potřeba se vrátit k předpokladu, že všechny výběry pocházejí z normálního rozdělení. Až nyní jsme totiž schopni tento předpoklad rigorózně otestovat. K testování se použije klasický Shapiro-Wilkův test, avšak aplikovaný na takzvaná "standardizovaná rezidua". Podívejme se nyní podrobněji, o co se jedná. Předpokládáme-li, že skupiny mají různé střední hodnoty, předpokládáme vlastně, že pozorování X_{it} (t-té pozorování z i-tého výběru) splňuje model

$$X_{it} = \mu + \gamma_i + e_{it}, \tag{5}$$

kde μ je jakýsi základ pro střední hodnotu, γ_i je korekce pro střední hodnotu i-tého výběru a e_{it} je nějaký zbytek (reziduum). Platí tedy, že $\mathrm{E}\, X_{it} = \mu + \gamma_i$. Když všechna pozorování X_{it} očistíme od rozdílných středních hodnot, zbydou nám pouze rezidua e_{it}. Tato rezidua se ještě dále standardizují (aby měla nulovou střední hodnotu a jednotkový rozptyl) a pak teprve se na ně může provést Shapiro-Wilkův test. Důvodem, proč se toto dělá je to, že Shapiro-Wilkův test potřebuje, aby vstupní data byla stejně rozdělená (tj. měla mimo jiné stejnou střední hodnotu). Kdybychom do Shapiro-Wilkova testu nasypali přímo hodnoty X_{it}, nebylo by toto splněno, protože připouštíme, že výběry mohou mít různé střední hodnoty. Standardizovaná rezidua v R-ku získáme pomocí příkazu `rstandard(nazev_modelu)`.

Mnohonásobné porovnávání

Je-li zamítnuta $H_0 : \mu_1 = \ldots = \mu_K$, tak nás zajímá, u kterých dvou skupin jsou ty střední hodnoty odlišné. K tomu slouží **Tukeyův test**. Ten umožňuje porovnání výběrových průměrů od všech dvojic současně, a to při dodržení celkové hladiny α. i-tá a j-tá skupina jsou prohlášeny za rozdílné, pokud

$$|\bar{X}_{i.} - \bar{X}_{j.}| > q_{K,n-K}(1-\alpha)\sqrt{\frac{s^2}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}, \tag{6}$$

kde s^2 je reziduální rozptyl (vzorec (2)) a číslo $q_{K,n-K}(1-\alpha)$ by se našlo v tabulkách (jedná se o kvantil rozdělení tzv. studentizovaného rozpětí). V R-ku se tento test provádí příkazem `TukeyHSD` a jeho výstupem jsou konfidenční intervaly pro jednotlivé rozdíly $\mu_i - \mu_j$ a dále také p-hodnoty pro test dílčích hypotéz $H_0 : \mu_i = \mu_j$. Všechny tyto intervaly spolehlivosti a p-hodnoty berou v úvahu to, kolik dvojic je porovnáváno a zachovávají celkovou hladinu testování α. Odtud také pochází zkratka HSD v názvu metody, která znamená „honest significant difference“. Rozdílné střední hodnoty jsme prokázali u těch dvojic, u nichž interval spolehlivosti neobsahuje nulu, nebo příslušná p-hodnota je menší než 0.05. V R je možné tyto intervaly spolehlivosti přehledně graficky znázornit pomocí příkazu `plot(TukeyHSD(nazev_modelu))`.

Co dělat při nesplnění předpokladů

Nesplnění předpokladu shodných rozptylů

Pokud nelze předpokládat shodné rozptyly (tj. Bartlettův či Leveneův test zamítá homoskedasticitu), nic se neděje, protože existuje jednoduchá modifikace analýzy rozptylu pro nestejné rozptyly. V R-ku k jeho provedení stačí použít příkaz `oneway.test`. Jeho výstupem však bohužel není klasická tabulka analýzy rozptylu, ale pouze hodnota testové statistiky a p-hodnota, podle které je potřeba se orientovat. Pro mnohonásobné porovnání v případě nestejných rozptylů již nelze použít Tukeyho metodu, ale je potřeba zvolit nějakou jinou metodu.

Nesplnění předpokladu normality

Pokud Shapiro-Wilkův test aplikovaný na standardizovaná rezidua zamítá předpoklad normálního rozdělení, můžeme celý výsledek analýzy rozptylu zahodit, protože použití této metody nebylo oprávněné. Místo ní je potřeba se uchýlit k pořadovému testu, který se nazývá **Kruskalův-Wallisův test**.

Kruskalův-Wallisův test je zobecněním dvouvýběrového Wilcoxonova testu a jeho jedinými předpoklady je, aby naše výběry byly vzájemně nezávislé a aby pocházely ze spojitých rozdělení. Podobně jako Wilcoxonův test, ani Kruskalův-Wallisův test netestuje rovnost středních hodnot jednotlivých výběrů, nýbrž shodu celých jejich rozdělení (což speciálně znamená i shodu středních hodnot). Tedy:

H_0 : rozdělení zkoumaného znaku je ve všech skupinách stejné

H_1 : neplatí H_0

Představme si, že shrneme všechna data dohromady (bez ohledu na příslušnost ke skupině) a přidělíme jednotlivým hodnotám pořadí (podle velikosti) v rámci tohoto sdruženého výběru. Pak jako T_i označíme součet pořadí od hodnot, které původně příslušely i-tému výběru. Testová statistika Kruskalova-Wallisova testu má tvar

$$Q = \frac{12}{n(n+1)}\sum_{i=1}^{K}\frac{T_i^2}{n_i} - 3(n+1), \tag{7}$$

kde $n = \sum_{i=1}^{K} n_i$. Nulovou hypotézu zamítneme, bude-li

$$Q \geq \chi^2_{K-1}(1-\alpha), \tag{8}$$

kde $\chi^2_{K-1}(1-\alpha)$ je kvantil χ^2 rozdělení s $K-1$ stupni volnosti. V R-ku se tento test provádí pomocí příkazu `kruskal.test`.

Možná vás napadne, zda-li lze i v tomto případě provést mnohonásobné porovnávání a v případě zamítnutí H_0 identifikovat rozdílné skupiny. Možné to je, ale nepoužívá se k tomu Tukeyho test, nýbrž test jiný, který však již přesahuje rámec tohoto textu.