STATISTIKA PRO EKONOMY KURZ

Úroveň Bc. stupeň studia

4. TÉMA

Aplikace statistických testů ve softwaru R

Cíle kapitoly:

1. Testování statistických hypotéz
2. Jednovýběrové t – testy
3. Dvouvýběrové parametrické a neparametrické t – testy
4. ANOVA (Analýza shodnosti rozptylu)
5. Regresní analýza

1. Testování statistických hypotéz

Statistická hypotéza je určité tvrzení o parametrech (nebo obecněji o vlastnostech) základního souboru (např. o střední hodnotě). O její přijatelnosti rozhodujeme pomocí **statistického testu.**

Hypotéza, které platnost ověřujeme, se označuje $\boldsymbol{H_0}$ a nazývá se **nulová hypotéza**. Oproti ní postavíme **alternativní hypotézu $\boldsymbol{H_A}$ (nebo $\boldsymbol{H_1}$)**, která popírá platnost H_0.

Testové kritérium je vhodná funkce výběru (např. výběrový průměr apod.). Volba testového kritéria závisí na tom, jakou hypotézu chceme otestovat (zda testujeme hypotézu o průměru, rozptylu), ale také na povaze dat ve výběrovém souboru.

U statistického testu existuje jistá pravděpodobnost, že zamítneme i tu nulovou hypotézu, která ve skutečnosti platí. Tuto pravděpodobnost nazýváme **hladina významnosti**, značíme $\boldsymbol{\alpha}$, volí se malé číslo, často $\alpha = 0{,}05$.

Je zřejmé, že jsou 4 možné případy, jak je znázorněno v tabulce:

	H_0 platí	H_0 neplatí
H_0 nezamítáme	správné rozhodnutí pravděpodobnost $1-\alpha$	chyba 2. druhu pravděpodobnost β
H_0 zamítáme	chyba 1. druhu pravděpodobnost α	správné rozhodnutí pravděpodobnost $1-\beta$

Kritický obor je množina takových hodnot testového kritéria, u kterých hypotézu H_0 zamítáme, určuje se s ohledem na zvolenou hladinu významnosti a někdy i s ohledem na počet prvků výběrového souboru.

2. Jednovýběrové t – testy

V případě testování parametrů jednoho výběrového souboru se využívají:

- ***t-test*** – pokud testujeme hypotézu o průměru a rozdělení je normální,
- ***asymptotický u-test*** – pokud testujeme hypotézu o průměru a dat je v dostatečném množství (alespoň 30),
- ***χ^2-test o rozptylu*** – pokud testuji hypotézu o rozptylu, nebo směrodatné odchylce a rozdělení je normální,
- ***asymptotický u-test pro populační poměr*** – pokud testuji hypotézu o podílu nějaké skupiny v základním souboru a dat je v dostatečném množství (alespoň 30).

Další testy používané pro jeden výběrový soubor:

- ***Wilcoxonův test*** – pokud testujeme hypotézu o mediánu, rozdělení není normální a dat je málo. V tomto případě používáme jako míru střední hodnoty medián, průměr nelze použít.
- ***Shapiro – Wilkův test*** – pokud testujeme hypotézu o tom, že data pocházejí z normálního rozdělení.

Postup ve softwaru R

V prvním kroku je důležité stanovit nulové hypotézy:

- jednovýběrový t-test
  $\mu = k$ (střední hodnota = konstanta)
- jednovýběrový Wilcoxonův test
  $\mu = k$ (střední hodnota = konstanta)
- Shapiro-Wilkův test
  rozdělení dané proměnné = normální rozdělení

Jestliže analyzujeme **jediný výběr**, jehož střední hodnotu (např. průměr, medián) srovnáváme s předem danou konstantou, volíme **jednovýběrový test** (například se ptáme, jestli průměrná výška studentů = 178 cm).

Pokud je rozdělení daného výběru normální, zvolíme **jednovýběrový t-test** (*Statistics – Means – Single-sample t-test*). Jestliže zamítneme normalitu dat (anebo víme, že data nemají normální rozdělení, volíme neparametrický test - např. **jednovýběrový Wilcoxonův** test – příkaz: wilcox.test(*dataset*$*proměnná*,mu=*konstanta*)[1]

K testování normality daného výběru použijeme např. Shapiro-Wilkův test (Statistics – Summaries – Shapiro-Wilk test of normality).

3. Dvouvýběrové parametrické a neparametrické t – testy

Dvouvýběrovými testy zkoumáme shodu parametrů (obecněji vlastností) u dvou nezávislých nebo závislých souborů. U dvouvýběrových testů je důležité v první fázi identifikovat, zda se jedná o závislé či nezávislé výběry. Poté je možné postupovat dále ve statistické analýze a aplikovat vhodný statistický test podle povahy a vlastností případové studie. U závislých výběrů není nutné analyzovat rozptyly výběrů. U nezávislých výběrů je nutné aplikovat dodatečné statistické testy, které slouží k analýze předpokladů shodnosti rozptylů dvou výběrů.

Používané **parametrické testy** jsou:

- ***párový t-test*** – používá se pro porovnání průměrů pro závislé soubory (hodnota v jednom souboru závisí na hodnotě ve druhém – např. spotřeba auta u paliva s aditivy a bez nich). Data musejí být jasně ve dvojicích a rozdíly hodnot mají normální rozdělení;

- ***dvouvýběrový t-test*** – používá se pro nezávislé soubory (dvě skupiny dat, které netvoří páry). Oba soubory musejí mít normální rozdělení. Je zde dále důležité, zda jsou rozptyly u obou souborů shodné, nebo ne – tento test má dvě varianty;
- ***dvouvýběrový F-test*** – *používá* se pro porovnání rozptylů pro nezávislé soubory. Oba soubory musejí mít normální rozdělení;
- ***asymptotický dvouvýběrový u-test*** – používá se pro testování průměrů pro nezávislé soubory (dvě skupiny dat, které netvoří páry). Oba soubory musejí mít dostatek dat (alespoň 30);
- ***asymptotický dvouvýběrový u-test pro populační poměr*** – používá se pro porovnání podílů v populaci. Oba soubory musejí mít dostatek dat (alespoň 30).

Některé **neparametrické dvouvýběrové** testy jsou:

- ***párový Wilcoxonův test, dvouvýběrový Wilcoxonův test, dvouvýběrový Mannův-Whitneyův test*** – nahrazují podobné t-testy, pokud rozdělení některé skupiny (nebo obou) není normální a dat je málo.

Postup ve softwaru R

V prvním kroku je důležité stanovit nulové hypotézy:

- F-test

  $s^2_1 = s^2_2$ (var(1) = var(2) neboli rozptyly se sobě rovnají)
- Shapiro-Wilkův test

  rozdělení dané proměnné = normální rozdělení
- párový Wilcoxonův test

  $\mu_1 - \mu_2 = k$ (rozdíl středních hodnot obou výběrů = konstanta (často 0))
- dvouvýběrový Wilcoxonův test

  $\mu_1 = \mu_2$ (střední hodnoty obou výběrů se sobě rovnají)
- párový t-test

  $\mu_1 - \mu_2 = k$ (rozdíl středních hodnot obou výběrů = konstanta (často 0))
- dvouvýběrový t-test (Studentův t-test, t-test nezávislých výběrů), $\mu_1 = \mu_2$ (střední hodnoty obou výběrů se sobě rovnají)

Pokud srovnáváme **dva výběry** (A, B) musíme rozhodnout, zda jsou výběry na sobě závislé či nikoli (**závislé** – hodnota v A ovlivňuje hodnotu v B – např. délka levé a pravé ruky u jednoho člověka; výška bratra a sestry; **nezávislé** – hodnota v A a B se navzájem neovlivňují (např. náhodně vybrané délky levé a pravé ruky u různých lidí, výška náhodně vybraných mužů a žen).

Pokud jsou výběry **závislé**, volíme **párové** testy, pokud jsou **nezávislé**, volíme testy **dvouvýběrové**.

Závislé i nezávislé výběry mohou a nemusí mít normální rozdělení. Pokud zamítneme normalitu výběrů (Shapiro – Wilkův test), volíme **neparametrické testy** – např. **párový Wilcoxonův test** (*Statistics – Nonparametric tests – Paired samples Wilcoxon test*), resp. **dvouvýběrový Wilcoxonův test** (*Statistics – Nonparametric tests – Two sample Wilcoxon test*).

Pokud výběry mají normální rozdělení, volíme t-testy (**párový t-test** (*Statistics – Means – paired t-test*), resp. **dvouvýběrový t-test** (*Statistics – Means – Independent samples t – test*)).

Použití dvouvýběrového t-testu však předpokládá homogenitu (shodnost) rozptylů (variancí). Podle toho, jestli se rozptyly obou výběrů rovnají nebo nerovnají, zvolíme odpovídající variantu t-testu. Shodnost rozptylů testujeme pomocí **F-testu** (*Statistics – Variances – Two-sample F-test*). Pokud zamítneme shodnost rozptylů, zvolíme v dialogu dvouvýběrového t-testu možnost *Assume equal variance? **No***). Pokud nezamítneme shodnost rozptylů, zvolíme *Assume equal variance? **Yes***.

4. ANOVA (Analýza shodnosti rozptylu)

Analýza rozptylu (ANOVA) se používá ke zkoumání, zda průměrné hodnoty v několika souborech jsou stejné, tj. ke zkoumání závislosti číselného znaku na slovním (slovním znakem je zařazení do skupiny). ANOVA funguje na principu F-test dvou rozptylů, její závěry se však týkají průměrů ve skupinách.

Předpokladem ANOVY je normalita dat v rámci každé skupiny a shoda rozptylů u všech skupin.

Postup ve softwaru R

V prvním kroku je důležité stanovit nulové hypotézy:

- Bartlettův a Levenův test

  $s^2_1 = s^2_2 = s^2_3 = s^2_4$ .....= s^2_k (var(1) = var(2) = var(3) = var(4)..... = var(k)
- neboli všechny rozptyly se sobě rovnají) • jednofaktorová ANOVA
- $\mu_1 = \mu_2 = \mu_3 = \mu_4 =$..... = μ_k (střední hodnoty všech výběrů se sobě rovnají)

  Kruskal – Wallisův test
- $\mu_1 = \mu_2 = \mu_3 = \mu_4 =$..... = μ_k (střední hodnoty všech výběrů se sobě rovnají)

Pokud srovnáváme mezi sebou více než dva výběry, lišící se v jediném faktoru (například výše příjmů lidí s ukončeným vzděláním ZŠ, SŠ a VŠ – liší se ve vzdělanosti; nebo výše denního obratu firmy v zimě, na jaře, v létě a na podzim – liší se v ročním období)), můžeme je porovnat jednofaktorovou ANOVOU (Analysis of Variance, analýza rozptylu). Jinak řečeno – tímto testem sledujeme závislost kvantitativní proměnné na kategoriální (příjem na vzdělání, obrat na ročním období). V R potřebujeme data ve dvou sloupcích – jedna proměnná je kvantitativní (závislá, odpověď), druhá je kategoriální (nezávislá, grupovací). Test provedeme *Statistics – Means – One-way ANOVA*.

Předpokladem ANOVY je normalita dat a shodnost rozptylů. Testy předpokladů lze provést v R. Homogenitu variancí (shodnost rozptylů, homoskedascitu) lze testovat Bartlettovým testem (*Statistics – Variance – Bartlett's test*) anebo Levenovým testem (*Statistics – Variance – Levene's test).* Test normality se aplikuje na reziduály. Proměnnou reziduály je nejprve třeba vytvořit (*Data – Manager variables in active data set – Compute new variable*; zde vyplnit *New variable name*: residuals; *Expression to compute*: residuals(AnovaModel.1). Na proměnnou residuals potom aplikujeme test normality (Shapiro – Wilk). ANOVU lze provést, pokud jsou výsledky testů předpokladu neprůkazné (větší než 0.05). Podle výsledků ANOVY nezjistíme, která ze skupin se liší od jiných a jak moc (viz H0).

Pokud víme, že data předpoklady pro ANOVU (především normalitu dat), lze několik výběrů srovnat neparametrickou obdobou ANOVY – např. Kruskal – Wallisův test. V R zadáme *Statistics – Nonparametric tests – Kruskal – Wallis test*.

5. Regresní analýza

Pomocí **lineárních regresních modelů** popisujeme **statistickou závislost** mezi **dvěma číselnými veličinami.** Popis této závislosti má podobu rovnice, jak veličina ***Y* závisí na *X***.
Z těchto dvou veličin je jedna ***vysvětlovaná* (*regresand*)** a druhá je ***vysvětlující* (*regresor*).** Často je (už i intuitivně) vhodnější jedno pořadí – například vysvětlovat výkon motoru pomocí jeho objemu, ale teoreticky jsou vždy možné obě pořadí.

Nejjednodušším případem je **regresní přímka**, rovnice regresní přímky má tvar:

$$\hat{y} = b_0 + b_1 \cdot x, \text{ kde}$$

x je hodnota vysvětlující veličiny X,
$\hat{y}$ je předpovídána hodnota vysvětlované veličiny Y pro danou hodnotu x,
$b_0 + b_1$ jsou **regresní parametry** nebo **regresní koeficienty**.

Koeficienty v rovnici regresního modelu se určují například pomocí **metody nejmenších čtverců**, statistické programy mají potřebné metody naprogramovány.

Regresní model musí splňovat tři základní předpoklady:

- předpoklad normality dat,
- předpoklad homoskedasticity,
- předpoklad sériové nezávislosti.

Regresní model se skládá ze dvou částí. Ze systematické a nesystematické složky. Právě nesystematická složka, aby se stala součástí regresního modelu, musí splňovat prostřednictvím reziduí výše zmíněné předpoklady regresní analýzy.

Index determinace slouží k identifikaci, na kolik procent aplikovaný statistický model dokáže vysvětlit danou případovou studii. Čím vyšší je index determinace, tím vhodnější je aplikovat statistický model.

Korelační koeficient r umožňuje stanovit míru (sílu) závislosti mezi uvažovanými veličinami. Veličiny jsou tím těsněji propojené, čím je r bližší k 1 nebo -1. Pokud je r blízké k 0, veličiny jsou nezávislé (neovlivňují se), nebo závislost nemá tvar přímky. Záporný korelační koeficient vyjadřuje nepřímou úměru a kladný korelační vyjadřuje koeficient přímou úměru.

Další informace korelační analýza nepodává. Pouze informuje analytika o možných dalších postupech.

Postup ve softwaru R

Nulová hypotéza testu významnosti lineární regrese zní následovně:

vysvětlovaná proměnná nezávisí na vysvětlující proměnné (Y nezávisí na X). Pokud je hodnota *p* menší než hladina významnosti, zamítáme nezávislost.

Pokud potřebujeme srovnat závislost jedné kvantitativní proměnné na jiné (případně na několika) kvantitativní, použijeme lineární regresi. Jedna z proměnných je přitom vysvětlující proměnná (X, nezávisle proměnná, prediktor), druhá je vysvětlovaná (Y, závisle proměnná, odpověď). Mezi proměnnými je tedy zřejmý kauzální vztah (co je příčina a co je důsledek). Např. závisí (a jak moc) výše útraty na příjmu? Závisí příjem člověka na jeho IQ? Data máme zadaná ve dvou sloupcích (každá proměnná jeden sloupec).

Reziduály jsou užitečným pro zjištění, zda je lineární regrese vhodně zvoleným modelem.

V programu R klikneme na *Statistics – Fit models – Linear regression.* Vybereme vysvětlovanou proměnnou (*Response variable*) a vysvětlující proměnnou (*Explanatory variable*).

Výsledkem dostaneme:

- koeficient determinace = *Multiple R-squared*
- test významnosti = *F-statistic, p-value*
- koeficienty regresní rovnice = sloupec *Estimate* v části *Coefficients.*

Shrnutí:

1. Testování statistických hypotéz
2. Jednovýběrové t – testy
3. Dvouvýběrové parametrické a neparametrické t – testy
4. ANOVA (Analýza shodnosti rozptylu)
5. Regresní analýza

Pokud zkoumáme závislost **kvantitativní** proměnné na **kategoriální**, přičemž kategoriální proměnná **nabývá jen dvou hodnot**, použijeme nějaký **dvouvýběrový** test (porovnáváme dva výběry).

Pokud zkoumáme závislost **kvantitativní** proměnné na **kategoriální**, přičemž kategoriální proměnná nabývá více než dvou hodnot, použijeme model **ANOVA** (porovnáváme více výběrů).

Pokud zkoumáme závislost **kvantitativní** proměnné na jiné **kvantitativní**, použijeme buď **regresi** (jedna proměnná je vysvětlovaná a druhá vysvětlující), anebo **korelaci** (mezi proměnnými není zřejmý vztah příčina – důsledek).

Zdroje:

ADAMEC, V., L. STŘELEC, a D., HAMPEL, 2017. Ekonometrie I: učební text. Druhé nezměněné vydání. Brno: Mendelova univerzita v Brně. ISBN 978-80-7509-480-3. (s. 67-216)

HINDLS, R., 2018. Statistika v ekonomii. [Průhonice]: Professional Publishing. ISBN 978-80-88260-09-7. (s. 120-216)

MOŠNA, F., 2017. Základní statistické metody. Praha: Univerzita Karlova v Praze. ISBN 978-80-7290-972-8. (s. 14-39)

STUCHLÝ, J., 2015. Statistické analýzy dat: vysokoškolská učebnice. České Budějovice: Vysoká škola technická a ekonomická v Českých Budějovicích. ISBN 978-80-7468-087-8. (s. 49-122)