

JULIUS JANÁČEK

Statistika jednoduše

PRŮVODCE SVĚTEM STATISTIKY





JULIUS JANÁČEK

Statistika

jednoduše

PRŮVODCE SVĚTEM STATISTIKY

Grada Publishing

Upozornění pro čtenáře a uživatele této knihy

*Všechna práva vyhrazena. Žádná část této tištěné či elektronické knihy nesmí být reprodukována a šířena v papírové, elektronické či jiné podobě bez předchozího písemného souhlasu nakladatele. Neoprávněné užití této knihy bude **restně stíháno**.*

ING. JULIUS JANÁČEK, Ph.D.

Statistika

jednoduše

PRŮVODCE SVĚTEM STATISTIKY

Vydala Grada Publishing, a.s.

U Průhonu 22, 170 00 Praha 7

tel.: +420 234 264 401

www.grada.cz

jako svou 8651. publikaci

Redaktorka Lenka Zajíčková

Za odborný obsah odpovídá autor

Grafická úprava a sazba Eva Hradiláková

Návrh a zpracování obálky Eva Hradiláková

Ilustrace Kamila Andresová

Počet stran 120

První vydání, Praha 2022

Vytiskla tiskárna Tribun EU s.r.o. (knihovnicka.cz)

© Grada Publishing, a.s., 2022

ISBN 978-80-271-6640-4 (pdf)

ISBN 978-80-271-1738-3 (print)

Obsah

O autorovi	8
O čem to bude?	9
Poděkování	12
Struktura	12
1. Deskriptivní statistika	13
1.1 Náhodná veličina	13
1.2 Střední hodnota a výběrový průměr	13
1.3 Medián	14
1.4 Modus	15
1.5 Rozptyl	15
1.6 Směrodatná odchylka	16
2. Náhodné veličiny	19
2.1 Diskrétní náhodná veličina	19
2.2 Spojitá náhodná veličina	19
2.3 Rozdělení pravděpodobnosti	19
2.4 Distribuční funkce náhodné veličiny	20
2.5 Hustota pravděpodobnosti spojité náhodné veličiny	21
2.6 Četnosti a histogram	23
3. Normální rozdělení a centrální limitní věta	25
3.1 Normální rozdělení	25
3.2 Pravidlo tří směrodatných odchylek	27
3.3 Centrální limitní věta a asymptotická normalita	28
4. Korelace	31
4.1 Korelace	31
4.2 Korelace není kauzalita (vliv)	36

5. Statistické testování	41
5.1 Statistický test	41
5.2 Provedení statistického testu	42
5.3 Nezamítnutí neznamená potvrzení	44
5.4 Náhodný výběr	45
5.5 P-hodnota	46
5.6 Hladina významnosti α	46
5.7 Síla testu	48
5.8 Určení hypotéz a priori (před testováním)	49
5.9 Předpoklady statistických testů	50
5.10 RCT (<i>Randomized Controlled Trial</i> – randomizovaný kontrolovaný test)	52
5.11 Intervaly spolehlivosti	56
6. Regrese: úvod	59
6.1 K čemu je regrese?	59
6.2 Jak postupovat při tvorbě regresního modelu	63
6.3 P-hodnota pro daný koeficient (proměnnou) v regresi	64
6.4 Dummy proměnné	65
6.5 Do regrese jen relevantní proměnné	65
6.6 Zobecněná regrese	66
6.7 Ceteris paribus v regresi	69
6.8 Dva typy regresní analýzy	69
6.9 Předpoklady regrese	70
7. Regrese: lineární	71
7.1 Koeficienty v lineární regresi	71
7.2 Konstanta v lineární regresi	73
7.3 Regresní rovnice v lineární regresi	73
7.4 Předpovídání pomocí regresního modelu	74
7.5 Způsob výpočtu lineárního OLS regresního modelu	75
7.6 Velikost koeficientu v regresním lineárním modelu a sklon přímky	77
7.7 Koeficient R^2 (<i>R-squared</i> , R na druhou, koeficient determinace) v lineární regresi	79
7.8 Adjustované R^2	81
7.9 Akaike kritérium (<i>Akaike criterion</i>)	81

8. Regrese: nelineární logistická	83
8.1 Typy regrese: logit a probit	83
8.2 Typy logistické regrese	83
8.3 Interpretace výstupu nelineární logistické regrese	84
8.4 McFaddenovo R^2 (<i>McFadden R-squared</i>)	85
8.5 Počet (procento) správně předpovězených případů	86
9. Seznam statistických testů	89
9.1 Testy normality	89
9.2 Parametrické testy	91
9.3 Testy rozptylu	97
9.4 Neparametrické testy	100
9.5 Analogické testy	108
Závěr	115
Literatura	117
Rejstřík	119



O autorovi

Julius Janáček vystudoval obor Statistika na Matematicko-fyzikální fakultě Univerzity Karlovy. V současné době vyučuje ekonomii a statistiku na Univerzitě J. E. Purkyně v Ustí nad Labem. Ve své vědecké práci se věnuje výzkumu kvality života, štěstí a chování člověka z pohledu behaviorální ekonomie. Zaměřuje se na zkoumání, jak lidé vnímají a hodnotí změny ve veřejném prostoru (například vybudování kašny) a jaké dopady obecně mají opatření veřejných autorit na kvalitu života občanů. Za důležité považuje zapojení studentů do vědecké činnosti již od bakalářského studia včetně využití kvantitativních metod – k tomu má přispět i jeho kniha *Statistika jednoduše*.

O čem to bude?



Rád bych pomoci této knize ukázat, že statistika je potřebná, atraktivní a jednoduchá. Se statistikou je to jako s auty: rozumět celé jeho struktuře je velmi složité, ale řídit auto je jednoduché. Stejně tak statistika: rozumět statistice dopodrobna v matematické teorii tak, jak je tomu vyučováno v mnoha kurzech, je složité, ale základy nutné k jejímu používání jsou jednoduché. Tato publikace zprostředkovává základní pochopení statistických metod a návod, jak je používat – nemusíme vědět, jak funguje motor – stačí nám umět auto řídit. Pokud potřebujete porozumět statistice více do detailu, tato kniha se může dobře doplňovat s klasickou učebnicí statistiky.

JAKÉ SCHOPNOSTI VÁM PŘINESE PŘEČTENÍ TÉTO PUBLIKACE?

- Porozumět základním statistickým pojmům.
- Správně interpretovat statistické výsledky.
- Vybrat a použít adekvátní statistické metody a testy.

CO STATISTIKA MIMO JINÉ DOKÁŽE?

- Ozřejmuje mnoho věcí, které jsou součástí běžného života.
- Umožňuje posoudit pravdivost mnoha tvrzení a zpráv.
- Činí nás imunními vůči manipulaci.

JAK NÁS STATISTIKA ČINÍ IMUNNÍMI VŮČI MANIPULACI?

PŘÍKLAD 0.1

Můj lékař mi jednou doporučoval pít občas červené víno, protože je prý zdravé. Ptal jsem se ho proč. Řekl mi, že lidé, které zná a kteří pijí červené víno, jsou prý zdravější. Touto úvahou se můj lékař dopustil chyby ve statistické analýze. Opravdu je možné, že existuje pozitivní korelace (vztah) mezi pitím červeného vína a zdravím (více vína, více zdraví). To ovšem neznamená, že

Věděli jste, že...

Dříve jste mohli dostat láhev červeného vína na předpis? Mnoho lékařů si dříve myslelo, že pití červeného vína je velmi zdravé. Ještě v padesátých letech 20. století bylo v ČR červené víno předepisováno jako lék. Pokud jste si chtěli udělat pěkný večer s vašim partnerem, stačilo (pokud se to dá) předstírat vysoký tlak. Stejným způsobem mohla být dříve předepisována i čokoláda. Dříve se tedy velmi vyplatilo mít dobrý vztah s praktickým lékařem.



platí kauzalita (vliv) „víno přináší zdraví“. Daná korelace (vztah) může platit například díky tomu, že bohatší lidé si mohou dovolit zdravější životní styl a možná i kvalitnější lékařskou péči, a proto jsou

zdravější. A mimo to pijí i červené víno. Naproti tomu méně bohatí lidé občas pijí tvrdý alkohol, mohou žít méně zdravě a možná si někdy nemohou dovolit kvalitní zdravotní péči. Příčinou zdraví a dlouhověkosti tedy nemusí být přímo víno, ale zdravější životní styl a kvalitní lékařská péče. Finanční bohatství tedy může způsobovat jak zdraví, tak pití červeného vína – proto statistický pozitivní vztah „červené víno – zdraví“. Víno jako takové je možná zdravé a možná nikoliv. Jeho vliv ovšem nelze určit na základě znalosti, že konzumenti vína jsou zdravější. Bližší vysvětlení rozdílu mezi korelací (vztahem) a kauzalitou (vlivem) se nachází v kapitole 4 Korelace.

PŘÍKLAD 0.2

Podobným způsobem může manipulovat firma snažící se prodat své výrobky. Říká: „Lidé, kteří užívají náš lék, jsou zdravější.“ Tím se snaží přesvědčit zákazníky, aby kupovali její výrobek. Ovšem to, že lidé, kteří daný lék užívají, jsou zdravější, neznamená, že tento lék zlepšuje zdraví. Korelace mezi užíváním tohoto léku a zdravím může platit ze stejného důvodu, jako v případě červeného vína a zdraví v příkladu 0.1: bohatší lidé více dbají o své zdraví, a proto jsou zdravější. Také mohou mít více peněz na nějaký lék. Tyto léky tedy nemusí přinášet zdraví. Korelace mezi lékem a zdravím může platit jen díky třetímu faktoru peníze. Ten přináší jak zdraví, tak užívání léku. Umět rozlišovat korelaci a kauzalitu nám umožňuje nebýt cílem manipulace.

PŘÍKLAD 0.3

Premiéři různých zemí často mluví takto: „Naše vláda hospodařila v roce 2014 odpovědně a dosáhla nejnižšího schodku státního rozpočtu od roku 2003.“ Takto dotyčný premiér chválí svou vládu. Toto může být manipulativní tvrzení, které hraje na neschopnost lidí správně interpretovat informace. Premiér vlastně řekl, že existuje negativní korelace mezi vládnutím jeho vlády a schodkem rozpočtu (když jeho strana vládne, tak je nižší schodek). Tímto vychloubáním dal najevo, že korelaci chybně ztotožňuje s kauzalitou (vlivem). Situace může být odlišná. Ano, v roce 2014 byl nízký schodek rozpočtu. Ale je jeho vláda skutečnou příčinou, nebo je zde jiný faktor, který schodek snížil? Nižší schodek rozpočtu mohl například být důsledkem úsporných opatření minulé vlády, která snížila státní výdaje. Možností je více: možná, že schodek byl nižší díky nové vládě, možná díky staré vládě a možná z úplně jiného důvodu. Ale skutečnost, že v roce vlády nového premiéra je schodek nižší, neříká nic o tom, či je to zásluha. Daný premiér by si tedy neměl automaticky přisvojovat zásluhu za úspornost. Toto je konkrétní situace, ve které schopnost statisticky a logicky uvažovat funguje jako ochrana proti manipulaci – statisticky znalý člověk zde nepodléhá lži.

JAK PROVÁDĚT VÝPOČTY

Pro základní výpočty a využití metod popsanych v této publikaci (např. průměr a směrodatná odchylka) stačí tužka a papír, případně kalkulačka. Pro použití pokročilejších metod (např. korelace či některé statistické testy) je dostatečný program Microsoft Excel. Pro složité operace a metody (např. regresní analýza a složitější statistické testy) je víceméně nutné použít statistický software – např. SPSS (placený) či Gretl (volně dostupný). Pro některé statistické výpočty lze také využít mnoho online statistických softwarů. Počítat regresní modely samozřejmě můžete i na papíře, ale tvorba a výpočet středně složitého logistického modelu vám může zabrat přibližně sedm týdnů čistého času. Doporučuji spíše počítač.

FUNKCE

Pro názornost a vyšší porozumění jsou v této knize často používány klasické matematické grafy. Matematický graf ukazuje, jaká je hodnota jedné proměnné (osa y), při určité hodnotě jiné proměnné (osa x). Graf tedy ilustruje vztah mezi dvěma proměnnými.

SUMY

V knize je používán matematický koncept „suma“, jehož symbolem je $\sum_{k=0}^n$. Konkrétním příkladem sumy je $\sum_{k=0}^n x_k$. Tento symbol říká, že máme sečíst všechna x přes k od nuly do n . Tedy $\sum_{k=0}^n x_k = x_0 + x_1 + x_2 + \dots + x_n$.

DESETINNÁ ČÍSLA JSOU LEPŠÍ NEŽ PROCENTA

Pro pravděpodobnost je lepší používat desetinná čísla než procenta. Například říkáme, že jev A nastane s pravděpodobností 0,7. Neříkáme 70 procent. Proč? Protože s desetinnými čísly se lépe pracuje. Například pokud jev A nastane s pravděpodobností 0,7 (70 %) a jev B nezávisle s pravděpodobností 0,3 (30 %), tak pravděpodobnost, že nastane A i B je $0,7 \times 0,3 = 0,21$ (21 %). S procenty to nefunguje: $70 \times 30 = 2100$. V textu se tedy pracuje s pravděpodobnostmi vyjádřenými desetinnými čísly.

Poděkování

Příčinou vzniku této publikace jsou konzultace s vysokoškolskými studenty, kterým jsem měl tu čest pomáhat. Všem patří můj dík. Obzvláště těm, kteří se nestyděli a kladli mi otázky, díky kterým jsem poznával, co je dobře pochopitelné a co je třeba vysvětlit lépe. Dále děkuji Michaele Ulrichové, Janu Popelkovi a Marku Vokounovi, kteří mi pomohli s finálními úpravami textu. Velké poděkování patří Kamile Andresové, která svými ilustracemi vdechla knize více života. Na závěr děkuji pracovníkům vietnamského původu z blízkého obchodu s potravinami: kdykoliv jsem měl hlad, byli připraveni.

Struktura

Kapitoly 1–4 pojednávají o základních statistických principech. Ty by měly být dobře stravitelné a použitelné pro všechny. Běžnému člověku, který chce získat základní orientaci ve statistice, by měly stačit.

Kapitola 5 popisuje systém statistického testování. V kapitole 9 pak najdete seznam statistických testů a jejich použití. Pro běžného vysokoškoláka tedy stačí kapitoly 1–5 a 9.

Kapitoly 6–8 mluví o regresní analýze. Tato statistická metoda je již relativně složitější. Pro porozumění je třeba více času.

1. Deskriptivní statistika

Abychom dosáhli co nejvyššího porozumění statistice, začneme od nejjednoduššího – od základů. Stejně jako při stavbě domu.

Termínem deskriptivní statistika označujeme oblast základních statistických ukazatelů (průměr, směrodatná odchylka atd.). Při zkoumání určitého datového souboru nám tyto informace poskytují základní orientaci.

1.1 Náhodná veličina

Náhodná veličina je veličina, která může při opakované realizaci (měření) nabývat různých hodnot s určitou pravděpodobností.

PŘÍKLAD 1.1

To, co nám padne na šestistranné hrací kostce, je náhodná veličina, protože může při opakovaném měření (zde házení) nabývat různých hodnot s určitou pravděpodobností.

1.2 Střední hodnota a výběrový průměr

Střední hodnota je charakteristika náhodné veličiny. Značí se μ . Je to číslo, které je průměrem při mnoha pozorováních. Odhadem střední hodnoty je výběrový průměr – značí se písmenem s pruhem nad ním (např. \bar{x} – „x s pruhem“). Vypočítá se jako **součet daných čísel podělený jejich počtem**:

$$\bar{x} \{x_1, x_2, \dots, x_n\} = \frac{\sum_{i=1}^n x_i}{n},$$

x_1, x_2, \dots, x_n jsou naměřené hodnoty
 n je počet pozorování (naměřených hodnot)

PŘÍKLAD 1.2

$$\text{Průměr } \bar{x} \{3, 4, 9, 20, 22\} = \frac{3 + 4 + 9 + 20 + 22}{5} = \frac{58}{5} = 11,6$$

Věděli jste, že...

Průměrný počet nohou u člověka je 1,99? Tedy (pravděpodobně) máte nadprůměrný počet nohou. To je důvod k radosti!

Průměrný lidský mozek obsahuje 78 % vody? Teď už nás tedy nemusí překvapovat, když vidíme, co všechno děláme.

Člověk v průměru stráví během svého života ve spánku 25 let?



1.3 Medián

Medián je prostřední hodnota číselné řady, která vznikla seřazením všech čísel od nejmenšího po největší. Pokud je počet čísel sudý, tak se za medián považuje buď jedno ze dvou čísel uprostřed, anebo průměr dvou čísel uprostřed.

PŘÍKLAD 1.3

Medián {3, 4, 9, 20, 22} = 9

Někdy je vhodné použít medián místo průměru. Proč? Protože mediánová hodnota může lépe charakterizovat statistický vzorek.

PŘÍKLAD 1.4

Představme si zemi, ve které tři lidé vydělávají 20 000 Kč měsíčně a jeden člověk 100 000 Kč měsíčně. Průměrná mzda je pak 40 000 Kč měsíčně. Na základě této hodnoty by si někdo mohl myslet, že většina obyvatel této země je velmi bohatá. Ovšem ztracenou informací by bylo, že tři čtvrtiny obyvatel této země vydělávají vysoce podprůměrnou mzdu (20 000 Kč). Mediánová mzda je v této zemi 20 000 Kč / měsíc. Tato hodnota zde lépe popisuje platovou situaci.



Věděli jste, že...

Průměrná měsíční nominální mzda v České republice ve druhém čtvrtletí roku 2019 byla 34 105 Kč? Naproti tomu mediánová měsíční mzda byla 29 127 Kč (ČSÚ, 2020). Proč takový rozdíl? Vysvětlení je podobné jako v příkladu 1.4. V ČR žije určité malé procento lidí, jejichž plat je velmi vysoký (statisíce). Tito lidé výrazně zvyšují českou průměrnou mzdu, ovšem více než 60 procent Čechů pobírá podprůměrnou mzdu.

1.4 Modus

Modus je hodnota, která se v dané skupině čísel vyskytuje nejčastěji.

PŘÍKLAD 1.5

Modus $\{1, 2, 8, 29, 29\} = 29$

Pokud neexistuje v souboru čísel hodnota, která se vyskytuje častěji než ostatní, tak modus neurčujeme.

1.5 Rozptyl

Rozptyl je charakteristika variability náhodné veličiny. Čím vyšší je rozptyl, tím budou hodnoty dále od sebe. Rozptyl se značí σ^2 . Odhad rozptylu na základě naměřených dat se nazývá výběrový rozptyl, značí se s^2 a vypočítá se takto:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

x_1, x_2, \dots, x_n jsou naměřené hodnoty

\bar{x} je výběrový průměr

n je počet pozorování