Obohacení historiografického výzkumu o smysluplné datové analýzy

Marek Vokoun[1]

Praha 2020

[1] Ing. Marek Vokoun, Ph.D. (CEVRO Institut, z. ú.; UJEP); Marek.Vokoun@gmail.com

1 Statistika a historie

Jedna z metod práce, která se využívá pro kritiku pramenů (základní metoda zkoumání pramenů v historii). Užitím pro kritiku pramenů hovoříme o metodě s názvem **historická statistika**. Využívá se pro studium hromadných jevů a má určitá specifika. V historické statistice obvykle využíváme statistické analýzy dvou hlavních druhů pramenů: **výsledky statistických šetření** a **soubory stejnorodých pramenných údajů**.

Statistika je pomocná věda pro historické bádání podobně jako sociologie, geografie či demografie.

Hospodářské dějiny využívají historické statistiky. Na některých univerzitách (Chicago, Cambridge, London School of Economics, Štrasburk) převažuje trend propojovat hospodářské dějiny přímo s otázkami, které si klade ekonomie. Dochází k používání historické statistiky k analýze široce definovaného hospodaření v dějinách. Toto silné spojení ekonomie a historie je dlouhodobě některými historiky kritizováno.

Historická demografie jako samostatná disciplína využívá historické statistiky. Cílem je analýza demografických ukazatelů v čase a v územním srovnání – například studium populace, demografické aspekty historické skutečnosti, otázky výživy, dostupnosti lékařské péče, rodiny, manželství, úmrtnosti, porodnosti. Propojení historie s demografií a statistikou je většinou historiků považováno za relevantní pro studium hromadných jevů.

2 Tradice statistické metody pro historiky v českém prostředí

2.1 Havránek, Jan, Petráň, Josef. *Základy statistické metody pro historiky*. Praha 1963.

Anotace: *Popisuje historické procesy pomocí interpretace číselných údajů, tabulek a grafů, vypracovaných matematickou analýzou pramenných údajů. Kde? Nejčastěji v hospodářských, sociálních, ale i v kulturních dějinách aj. Použití: přednostně tam, kde jsou číselné údaje k dispozici nebo kde lze údaje z pramenů kvantifikovat. (např. koncentrace majetku, proměna poddanských dávek, daňového zatížení, šíření nových technologií – knihtisku, parního stroje, hospodářská či obchodní závislost, vědecké či umělecké kontakty podle četnosti korespondence nebo uměleckých sbírek, kultur. význam podle četnosti publikací a recenzí)*.

2.2 Kubiš, Karel. *Kvantitativní metody a historická statistika,* in: Hroch, Miroslav a kol., Úvod do studia dějepisu. Praha 1985.

„Samostatná disciplína, pracující výhradně kvantitativní metodou. Zaměření: kvantitativní výzkum hromadných jevů s cílem jejich utřídění a sledování souvislostí a příčinností v historickém procesu, příp. definování ˊzákonitostíˊ… Použití: zpravidla pro období od 18. století, kdy moderní stát začíná sbírat demografické a ekonomické údaje – první záměrné ‚statistiky' (např. sčítání lidu). V ‚předstatistickém období' jen na základě vytváření dílčích statistik (z urbářů, soupisů poddaných, městských knih, zemské desky apod.). Podmínka: Zjištění vzniku a reprezentativnosti

údajů cestou kritiky pramene, protože prameny v ‚předstatistickém' období odrážejí statistické údaje jen zprostředkovaně, a tudíž nepřesně a neúplně." (s. 216-222)

2.3 Výuka historické statistiky na Karlově univerzitě - PhDr. Karel Kubiš, CSc.

Porozumění základním statistickým pojmům, etapám a technikám historicko-statistické práce, popisu jednorozměrných statistických souborů, problematice středních hodnot (včetně používání chronologického průměru), dále mírám variace, indexům, časovým řadám a způsobům jejich vyrovnávání a rovněž i základním informacím o práci s vícerozměrnými statistickými soubory (včetně základů korelačního počtu).

Povinná literatura

Stuchlý Jaroslav, Statistika I., Praha 1999.

- *Přehled a cvičení ze statistických metod pro manažery na VŠE, obsahuje testování hypotéz o středních hodnotách a pravděpodobnostním rozdělení, teorii pravděpodobnosti a úvod do statistické regrese.*

Doporučená literatura

Stuchlý Jaroslav, Statistika II., Praha 1999.

- *Přehled a cvičení ze statistických metod pro manažery na VŠE, obsahuje metody analýzy časových řad a indexů, vícenásobnou regresi dat.*

2.4 Výzkum používání statistických metod

VOKOUN, Marek, STELLNER, František, Czech economic historians and interdisciplinary approach, in: *Économies et Sociétés. Série "Histoire Économique Quantitative*" 50, 2015, No. 6, s. 857-875.

Výsledky dotazníkového šetření ukazují, že nečekaně 90 % respondentů z oboru hospodářské dějiny (Univerzita Karlova, VŠE v Praze a Ostravská univerzita) považuje interdisciplinární přístup za užitečný. Přibližně 60 % respondentů uvedlo, že ve svých publikacích použilo některé statistické metody (průměry, analytické grafy, časové řady, regrese, korelace a další pokročilé metody).

V institucionální analýze (Univerzita Karlova, VŠE v Praze a Ostravská univerzita) došlo k text-minigovému zkoumání publikovaných nebo kolektivních děl, avšak nejsou žádné důkazy, které by naznačovaly jakoukoli formu takového interdisciplinárního přístupu a použití pokročilých statistických metod. Výsledky studie také naznačují, že čeští ekonomičtí historici, zejména v oblasti moderních dějin, jsou ochotni spolupracovat a používat některé pokročilé statistické metody, ale pokud jde o ekonomy zabývající se historickými záležitostmi, bývají velmi skeptičtí.

3 Pomůcky v anglickém jazyce

ANDERSON, Margo, Quantitative History. in: *The Sage Handbook of Social Science Methodology*, edited by William Outhwaite and Stephen Turner, London 2007, s. 246-259. Dostupné z: http://users.hist.umn.edu/~ruggles/hist5011/Margo-Outwaite-Ch-14.pdf

- Pojednání o kvantitativní historii, kterou autor považuje za množinu dovedností a technik používaných k uplatňování metody statistické analýzy údajů ke studiu historie. Tj. jde o historickou statistiku – využívání zejména statistické analýzy tabulek dat. Avšak jde také o další přístupy, které podrobuje kritice.

FEINSTEIN, Charles, THOMAS, Mark. *Making history count: a primer in quantitative methods for historians*. Cambridge 2002.

- Pedagogové z Cambridgeské univerzity problematiku představují bez nutné znalosti matematiky a silně „propojují" ekonometrii a historii. Vysvětleny jsou histogramy, ukazatele středních hodnot, variace, skupinových rozdílů, tvary pravděpodobnostních rozdělení, normalita dat, korelace, regrese a kontingenční tabulky. Pro porozumění je představen statistický systém testování hypotéz a možnosti interpretace odhadnutých koeficientů. Obsahuje i případové studie použití daných metod v historickém výzkumu: nezaměstnanost a dávky v nezaměstnanosti v meziválečném období ve Velké Británii, emigrace z Irska, chudoba v Anglii a další.

FLOUD, Roderick. *An introduction to quantitative methods for historians*. London 1973. eBook 2013

- Příručku z Princetonské univerzity by měl historik použít, pokud používá kvantitativní data a klasifikuje určité jevy do skupin. Jde o příručku historické statistiky, jak klasifikovat různé druhy dat do tabulek, jak interpretovat střední hodnot a histogramy, trendy a růstové koeficienty v časových řadách, jak na korelaci a co s nedokonalými daty.

Učebnice ekonometrie:

- WOOLDRIDGE, Jeffrey M. *Introductory econometrics: A modern approach*. 3rd ed. Mason, OH 2006.
- KENNEDY, Peter. *A guide to econometrics*. 5th ed., Cambridge, Mass. 2003.

4 Příklady využití statistických metod

Používání aritmetických průměrů a variačních koeficientů je poměrně běžně využívaný postup v historickém výzkumu – průměrná spotřeba, výška, věk, produkce apod. Při analýze pramenů a následném kódování dat do tabulek vznikají určité skupiny, mezi kterými můžeme sledovat rozdíly. Porovnání těchto skupin je možno provádět množinou metod, která zkoumá „rozprostření“ dat.

4.1 Jednoduché třídění na dvě skupiny

Tato analýza nám dá odpověď, zda existují rozdíly mezi průměry ve dvou nezávislých výběrech dat. Například rozdíl průměrné výroba železa v tunách na jednu železárnu v německých zemích a v českých zemích. Je třeba mít údaje o produkci za konkrétní rok o všech (nebo mít reprezentativní výběr) železárnách a kódovat je podle území.

- t-test nepárový (data mají normální rozdělení)
- Wilcoxonův test nepárový (data nemají normální rozdělení)

Také je možné sledovat určitý efekt v čase. Pro stejnou skupinu železáren sledovat, objem produkce před zavedením a po zavedení nové technologie výroby (např. zvětšení výšky pece, přechod na minerální paliva) v daném sledovaném období. Je tedy nutné mít data o produkci železáren v určitém stejném okamžiku (roce). Poté vědět, které z nich v dalším roce zavedly, a které nezavedly danou novou technologii. Tím je možné analyzovat, jaký efekt mělo zvětšení výšky pece daných železáren na produkci železa oproti starší technologii.

- t-test párový (data mají normální rozdělení)
- Wilcoxonův test párový (data nemají normální rozdělení)

4.2 Složitější třídění na více skupin a více faktorů

Počet faktorů působící na zkoumaný jev, například průměrná cena chleba (podílem na příjmu, přepočet a stejnou měnu, vyjádřeno cenou jiné komodity) ve zkoumaných zemích, může být větší než jeden a tyto faktory mohou působit v kombinaci (interakce). Například na cenu chleba bude mít vliv úroda obilí, válečný stav, průměrná velikost disponibilního důchodu domácnosti, cena ostatního pečiva, cena soli apod.

Vojenské dějiny pracují kupříkladu s databázemi příslušníků československého zahraničního odboje, což umožňuje prakticky libovolně kombinovat množství různých údajů, včetně souvislosti národnostních kategorií s některými kategoriemi sociologickými (věk, původ, povolání, konfese, zdravotní stav apod.). Příklad: MARŠÁLEK, Zdenko. *„Česká“, nebo „československá“ armáda? národnostní složení československých vojenských jednotek v zahraničí v letech 1939-1945.* Praha 2017.

- ANOVA, MANOVA, regresní model

4.3 Data sbíraná v čase

Data mají také podobu více časových okamžiků. Obvykle se setkáváme s ročními ukazateli, ale neméně zajímavé jsou denní, týdenní, měsíční ukazatele. U těchto sebraných a dále kódovaných dat je nutné velmi opatrně pracovat s finančními daty. Ty mohou být v různých měnách a mohou být nominální (bez vlivu inflace) či reálné (se započtením vlivu inflace). Příkladem dat jsou roční výdaje na zbrojení, průměrný roční plat úředníka, roční rozpočet ministerstva, čtvrtletní produkce zpracovatelského průmyslu, denní spotřeba chleba, migrace, týdenní průměrný počet odpracovaných hodin dělníka v továrně, počet kriminálních činů za jeden měsíc apod.

4.4 Kliometrie popř. nové hospodářské dějiny

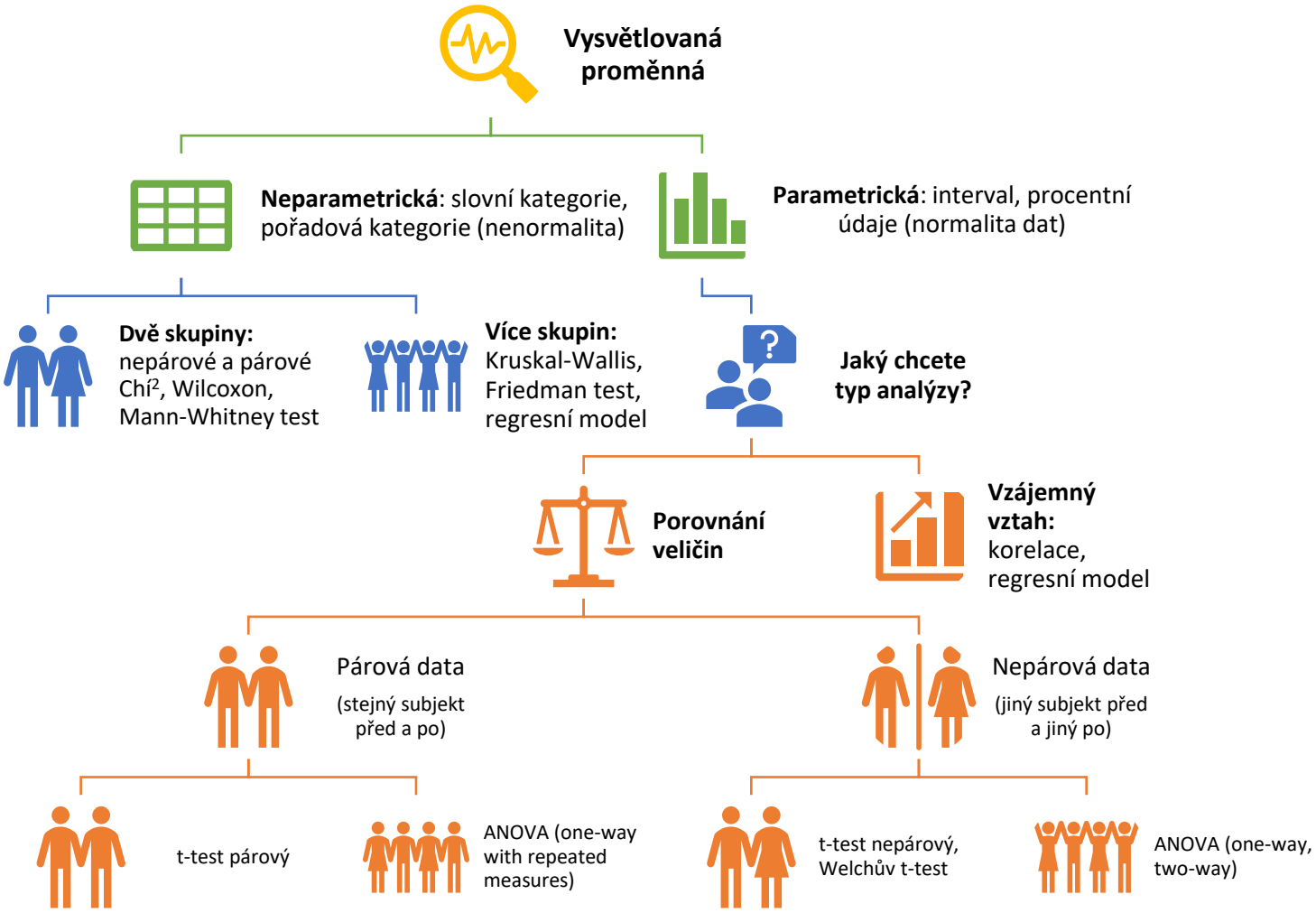
Metoda používaná v hospodářských dějinách, která hojně využívá metod ostatních věd, a tudíž statisticky zpracovává kvalitativní a kvantitativní data. Je proto charakteristická interdisciplinární prací. Vznikla v USA v 50. letech 20. století a začala být využívána v pracích Lance Davise, Alfreda Conrada a Johna Meyera, kteří se zabývali americkými hospodářskými dějinami.

Ke kliometrii tradičně patří různý stupeň použití matematických a statistických metod pro zpracování dat. Podobně jako v ekonometrii, která bývá s kliometrii ztotožňována, je třeba si dát pozor, zda naše výzkumné otázky a cíle práce využívají vhodná data, odpovídající statistické metody a zda interpretace je smysluplná.

Podmnožinou kliometrie je tedy historická statistika, která v omezené míře využívá jednoduchých statistických analýz. Avšak v rámci kliometrie je možné využít statistických metod, které se využívají v ekonomii, sociologii, psychologii, demografii, archeologii apod. Díky tomu je možné zpracovávat časové řady i panelová data (časové řady z více zemí najednou) a používat pokročilejší statistické modelování.

Kliometrie stála též u zrodu kvazi-historického a ahistorického výzkumu. Ten první, kvazi-historický, je přisuzován pracím Roberta Williama Fogela a Douglasse Cecila Northa, kteří za svoje celoživotní dílo dostali Nobelovu cenu za ekonomii. Kvazi-historický přístup využívá tzv. „hypotetickou srovnávací analýzu dějin“ (counterfactual history), která vychází z metody hojně využívané v společenských vědách, kdy srovnáváme dvě situace („Treatment effect“). Problémem této analýzy je často nemožnost najít v historii odpovídající srovnávací situaci. Například ekonomiku s otrokářstvím a podobnou ekonomiku ve stejném období bez otrokářství. Pokud vytváříme srovnávací situace, dostáváme se na tenký led. Výsledky takové analýzy jsou diskutabilní a závisí na možnostech srovnání, tj. jde o kvazi-historický přístup. Ahistorický přístup je v případech, kdy je přímo použita teorie jiné vědní disciplíny, popřípadě je statistická inference založena na společenskovědní teorii, tj. u neoklasické ekonomické teorie půjde o ekonomii, ne historii. Přímo ahistorickým oborem je kliodynamika. Kliometrii v mezích historické statistiky je možné považovat za historický, popř. kvazi-historický přístup.

4.5 Infografika



5 Výukový text

Smysluplné datové analýzy

Problémem výzkumu mnoha věd je vědomé i nevědomé zneužití statistiky, kdy dochází k chybné interpretaci nově vytvořených informací z dat. Nízká úroveň gramotnosti a nestatistická podstata lidské intuice[2] umožňuje snadno data zneužít a manipulovat současné poznání. Naším úkolem je tyto omyly odhalovat a napravovat. Jednou z cest je využívání kvantitativní historie. Jde o přístup k historickému výzkumu, který využívá kvantitativní, statistické a počítačové nástroje. Je považován za obor dějin v rámci humanitních i společenských (sociálních) věd. Výsledky výzkumu kvantitativní historie jsou publikovány například v prestižních časopisech Historical Methods (od roku 1967), Journal of Interdisciplinary History (od roku 1968), the Social Science History (od roku 1976), Cliodynamics: The Journal of Quantitative History and Cultural Evolution (od roku 2010) a další.

Kvantitativní historie využívá při kvantitativní analýze především statistiku, vědu o získávání informací z numerických dat. Statistika se zaměřuje na (1) získávání dat – jak je sbírat a ukládat, (2) popis a analýzu dat – jakou metodu tvorby nové informace zvolit, (3) usuzování a vhodnou interpretaci – kritické zhodnocení nově vytvořených informací. Podobně jako historiografie se opírá o heuristiku, kritiku pramenů, interpretaci a syntézu. Statistická data byla využívána historiografií od nepaměti. Například kronikáři, správní či vojenský aparát využívali sběru dat pro zisk informací a sledovali změny ve vývoji či struktuře sledovaných veličin v čase. Porozumění úřední statistice je důležité pro správnou interpretaci těchto informací.

Porozumění získaných dat a tvorba vlastních nových informací z dat mohou vhodně obohatit historiografický výzkum. Smysluplné metody tvorby nových informací z dat využijí zejména jednoduchých statistik založených na jednorozměrné analýze dat, rozdílu mezi skupinami dat a zobrazení dat.[3] Nesmyslné modelují historická data a zobecňují je v historické zákony (kliodynamika), nebo nedodržují pravidla statistické interpretace dat (zaměňování kauzality a korelace).

Pokročilé analýzy využívají omezeně vícerozměrnou analýzu dat, a to zejména analýzu rozdílů mezi skupinami a analýzu tzv. kontingenčních tabulek.[4] Výzkum časových řad[5] se přímo pro historiky nabízí, avšak vyžaduje velmi kvalitní data za delší časový úsek. Časové řady se proto obvykle pouze zobrazují v absolutní hodnotě a transformované podobě meziročních změn, indexu růstu a podobně.

[2] Srov. KAHNEMAN, Daniel. *Myšlení, rychlé a pomalé*. Brno 2012; VOKOUN, Marek, STELLNER, František, Czech economic historians and interdisciplinary approach, in: *Économies et Sociétés. Série "Histoire Économique Quantitative"* 50, 2015, No. 6, s. 857-875.

[3] Průměr, variační koeficient, t-testy, ANOVA, histogramy, grafy meziročního růstu či poklesu apod.

[4] Tabulka dvou kategoriálních proměnných, které nabývají vícero možností. Například kategorie pohlaví a příjmová kategorie.

[5] Časové řady jsou chronologicky srovnaná data využívající stejný časový úsek mezi měřením pro daný objekt (či subjekt). Například počet obyvatel k 31. 12. každého roku na území České republiky nebo výška osoby k 1. 1. každého roku.

Jednorozměrné datové analýzy

V jednorozměrných datových analýzách z nasbíraných dat o určité populaci získáváme jedno číslo (jeden rozměr, souhrnné měřítko), které nám poskytuje novou informaci. Nejčastěji používáme aritmetický průměr a medián. Všechny metody tvorby nových informací o populaci mají ve statistice určitá měřítka kvality, která pomáhají správně interpretovat danou informaci. Aritmetický průměr má například jako měřítka kvality směrodatnou odchylku, šikmost a strmost. Díky této znalosti měřítek kvality je možné tvořit správné úsudky o průměrné hodnotě.

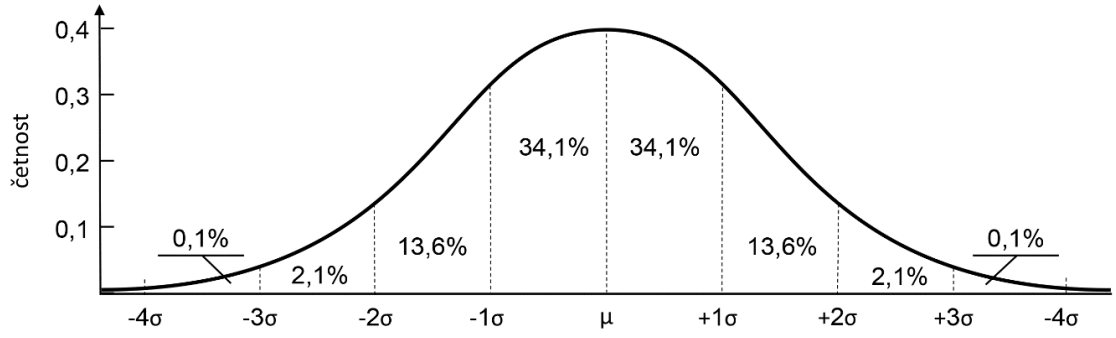
Zkoumaná populace je ve statistice označení pro množinu prvků, kterých se nově vytvořená informace týká. Máme-li data o platech úředníků v Rakousko-Uhersku na území Uherska, vyjadřujeme se interpretací pouze k této populaci. Přehnaná generalizace (platy všech úředníků z Uherska neodpovídají platům v celém Rakousko-Uhersku) nebo slabá reprezentativnost sebraných dat (údaje o 20 platech nemůže být použita pro interpretaci platů o všech úřednících v Rakousko-Uhersku).

Kolik dat je třeba pro smysluplnou intepretaci aritmetického průměru? Variabilita aritmetického průměru se od počtu okolo 40 pozorování dodatečným přidáním dalšího pozorování již tolik nemění. Avšak při takových malých počtech pozorování je nutné na toto ve vlastním výzkumu upozornit a v diskusi k výsledkům usuzovat, jak by mohla dodatečná nová pozorování ovlivnit naměřenou informaci o aritmetickém průměru.

⌖ Vhodné použití aritmetického průměru

Cílem naměřeného aritmetického průměru je podat informaci o jednom rozměru, který by charakterizoval celou populaci. Pokud chceme, aby tento jeden rozměr o populaci, např. průměrná výška 20letých až 39letých mužů v České republice byla k něčemu použitelná, musí být většina těchto mužů kolem tohoto naměřeného rozměru (okolo 68 % mužů). Většinou se kvantitativní data o populaci díky určitým zákonitostem vždy blíží tzv. „normálnímu rozdělení“. Toto rozdělení vnímejme jako umělý ideál krásy. Tento ideál má tvar Gaussovy křivky (obrázek č. 1):

Obrázek 1: Normální rozdělení, střední hodnota a směrodatná odchylka



Pokud se zabýváme výškou, tak opravdu až 68,2 % pozorování bude do jedné směrodatné odchylky a tato odchylka bude relativně malá, řekněme 15 cm. Průměrná hodnota bude řekněme 175 cm, je to pouze jeden rozměr, jedno velmi dobré číslo. Směrodatná odchylka nám poslouží jako dodatečné měřítko kvality a říká nám, že v populaci je nejvíce mužů v intervalu 175 cm plus

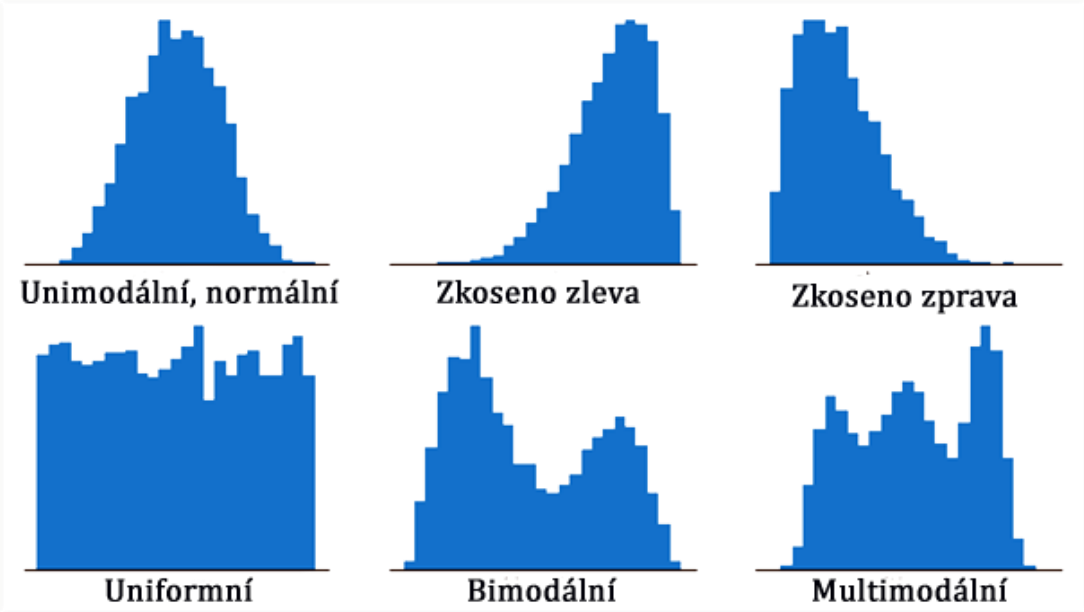
mínus 15 cm (od 160 do 190 cm). Medián výšky – hodnota uprostřed všech naměřených hodnot seřazených podle velikosti, bude také blízká tomuto průměru.

Průměrnou hodnotu vždy uvádějte společně se směrodatnou odchylkou.

Pokud se zabýváme průměrným příjmem, tak 68,2 % pozorování opět bude kolem jedné směrodatné odchylky, ale tato odchylka bude relativně veliká, řekněme 25 000 Kč. Průměrná hodnota příjmů v České republice je 34 105 Kč (2. čtvrtletí 2019 dle ČSÚ), avšak mediánová mzda[6] je nižší, střed hodnot je 29 127 Kč. Směrodatná odchylka nám říká, že máme průměrný příjem 34 105 Kč plus mínus 25 000,- Kč (reálně se ale pohybuje u 68,2 % populace v rozmezí od 9 105 až 59 105 Kč). Tato hodnota je ale jako jednorozměrný ukazatel populace na rozdíl od průměrné výšky nepoužitelná, neboť necharakterizuje dobře zkoumanou populaci. Měřítko kvality směrodatné odchylky nám říká, že toto rozpětí, je příliš široké a spodní interval je pod minimální mzdou a horní interval nereflektuje průměrné příjmy v nejlépe placených profesích.

Dalšími ukazateli kvality jsou dále i šikmost a špičatost (též strmost)[7]. V případě průměrných mezd jde o „dolů spláclé“ rozdělení (s malou špičatostí) a zešikmené směrem k mediánu (pravé zkosení). Tvar rozdělení nám vykreslí počítačový program. Histogram („statistická rozdělení“)[8] má i další označení: tvar rozdělení hodnot, distribuce hodnot, frekvence hodnot (Obr. 2).

Obrázek 2: Tvary a příklady histogramů



[6] Medián je číslo, které je uprostřed souboru hodnot, tj. dělí daný soubor na dvě stejně velké části. Např. medián souboru (1, 2, 3, 4 a 5) je číslo 3.

[7] Naměřená nenulová šikmost je odchylka od ideálu normálního rozdělení, které má šikmost nulovou. Je-li šikmost kladná, též pravostranná, tak se většina hodnot nachází pod průměrem, u záporné šikmosti se naopak většina hodnot nachází nad průměrem. Naměřená nenulová špičatost je opět odchylka od ideálu normálního rozdělení, které má špičatost nulovou. Je-li špičatost kladná (strmý kopec), tak je většina hodnot kolem průměru, u záporné špičatosti se naopak většina hodnot nachází daleko od průměru (plošší kopec).

[8] Histogram je grafické znázornění (sloupcový graf) distribuce dat. Unimodální rozdělení je podobné normálnímu, např. průměrná výška či váha ve třídě 60 studentů. Zkosené zleva je např. počet vypitých káv denně (většina nad průměrem). Zkosené zprava je např. měsíční mzda (většina pod průměrem). Uniformní může být třeba objem kyslíku v atmosféře. Bimodální je např. obrat restaurací podle hodiny (oběd a večeře). Multimodální je např. čas čekání na opravu telefonu v minutách (převládají např. 3 časově oddělené druhy oprav).

Vícerozměrné datové analýzy

Vícerozměrné datové analýzy využívají analýzu rozdílů mezi průměry dvou či více skupin. Tyto skupiny mohou mít téměř stejné histogramy a nelišit se. Nebo tyto histogramy mají sice vizuálně podobné, ale mohou se lišit pouze v naměřené hodnotě aritmetického průměru (zůstává podobná směrodatná odchylka, podobná strmost a šikmost). Nebo mají histogramy úplně odlišné.

Pokud tato podobnost je blízká i normálnímu rozdělení, pak lze provádět tzv. „t-test“, který vezme v potaz strmost, šikmost, průměr, odchylku a počet pozorování a zhodnotí je pro obě skupiny pozorování (např. muže a ženy) a dá verdikt v podobě testové statistiky. Pokud je skupin více než dvě, např. v kontingenční tabulce, používá se F-test. Označení kontingenční tabulka se ve statistice užívá k přehledné vizualizaci vzájemného vztahu dvou statistických znaků (např. pohlaví, kategorie příjmů), které nabývají více proměn (u pohlaví: muž, žena, ostatní; u příjmů: nulové, do minimální mzdy, do mediánové mzdy).

Uvedené testy t-test a F-test nám umožňují získat nové poznatky. Tj. zjišťujeme, zda potvrdit či zamítnout předpoklad onoho námi použitého testového kritéria. Tento testovaný předpoklad pro t-test či F-test je, že obě skupiny (nebo více skupin) mají podobné rozdělení (histogram) a tudíž i stejný průměr a variabilitu hodnot (mají podobné histogramy). Tomuto předpokladu testu říkáme nulová hypotéza. Každý pojmenovaný statistický test má nějakou pevně danou a neměnnou hypotézu. Tu je možné díky výpočtu testové statistiky vždy potvrdit, či vyvrátit.

Nulová hypotéza t-testu: Testované průměry obou skupin jsou stejné.

Nulová hypotéza F-testu: Všechny průměry všech skupin jsou stejné.

Hranice toho, kdy ještě nulovou hypotézu přijmeme je na nás. Volíme si procentně možnost, že „nevinného pošleme do vězení“, statisticky jde o možnost udělat chybu I. druhu. Tuto možnost chceme mít ideálně co nemenší, obvykle je to méně jak 5% pravděpodobnost ve společenských (sociálních) a humanitních vědách.

Alfa = hranice přijatelnosti chyby, obvykle by měla být ve společenskovědním výzkumu méně jak 5 %. Tato hranice je arbitrární, např. výsledek p-value$_1$= 4,98 % a p-value$_2$=5,04 % jsou si velmi blízké, jedno je důvod k zamítnutí hypotézy, druhé ne, toto by mělo být reflektováno v diskusi k výsledkům, případně pokračovat v analýze.

Testové kritérium (t-test, F-test, z-test, chí kvadrát test apod.) má podobu jak absolutního bezrozměrného čísla (např. t = 3,2) tak percentilu (např. p-value=0,07 tj.7 %), který nám udává, „kolik lidí posíláme do vězení“, pokud zamítneme tvrzení, které je napsané v nulové hypotéze. Pokud je to vysoké číslo (např. vyšší jak pět procent) nulová hypotéza platí.

Nulová hypotéza t-testu: Testované průměry obou skupin jsou stejné.

Při testové statistice nám vyjde p-value = 0,079, to znamená, že nulovou hypotézu lze zamítnout na úrovni 7,9 %, tj. že pošleme „7,9 % nevinných do vězení“. My jsme si dali hranici Alfa = 5 %, proto je to pro nás nepřijatelné a nulovou hypotézu necháme platit. Proto platí tvrzení, že testované průměry obou skupin jsou stejné.

Nulová hypotéza F-testu: Všechny průměry všech skupin jsou stejné.

Při testové statistice nám vyjde p-value = 0,016, to znamená, že nulovou hypotézu lze zamítnout na úrovni 1,6 %, tj. že pošleme „1,6 % nevinných do vězení“. My jsme si dali hranici Alfa = 5 %, proto je to pro nás akceptovatelné a nulovou hypotézu nenecháme platit. Proto neplatí tvrzení, že testované průměry obou skupin jsou stejné. Můžeme tedy říci, že průměry skupin se liší.

⌖ Vhodné použití t-testu pro dva histogramy „normálních“ dat

Porovnání dvou histogramů, které mají normální rozdělení (lze opět otestovat, popř. otestovat pohledem na histogram) je možné pomocí párového a nepárového přístupu ke dvěma skupinám dat. Při nepárovém srovnání je to například rozdíl mezi muži a ženami ve výšce, lékaři a dělníky v příjmech, tj. dvěma zcela nezávislými skupinami. Je to nejčastější varianta t-testu. Párové t-testování porovnává dva histogramy, ale u stejných subjektů, a to v režimu před a po. Obvykle je to např. stejní muži před léčbou a stejní muži po léčbě, stejná auta s olejem od Mogulu a stejná auta s vyměněným olejem od Shell.

Tabulka 1: Příklad zápisu skupin

Věk žoldáků Sparta	Věk žoldáků Řím
17	20
18	21
19	22
20	23
22	25
25	28
15	18
19	22
21	24
21	24
22	25
…	…
21	24

Oba testy lze dělat v Excelu. Postup je takový, že data o skupinách musíte mít ve dvou sloupcích.

1. Vybereme prázdnou buňku, zvolíme: „Vložit funkci“
2. V seznamu funkcí vybereme kategorii: Statistické
3. V seznamu kategorie vybereme „T.TEST“
4. Matice 1: Zde vybereme všechny buňky v sloupečku pro první skupinu
5. Matice 2: Zde vybereme všechny buňky v sloupečku pro druhou skupinu
6. Chvosty: Zde napíšeme 2, Typ: Zde napíšeme 3

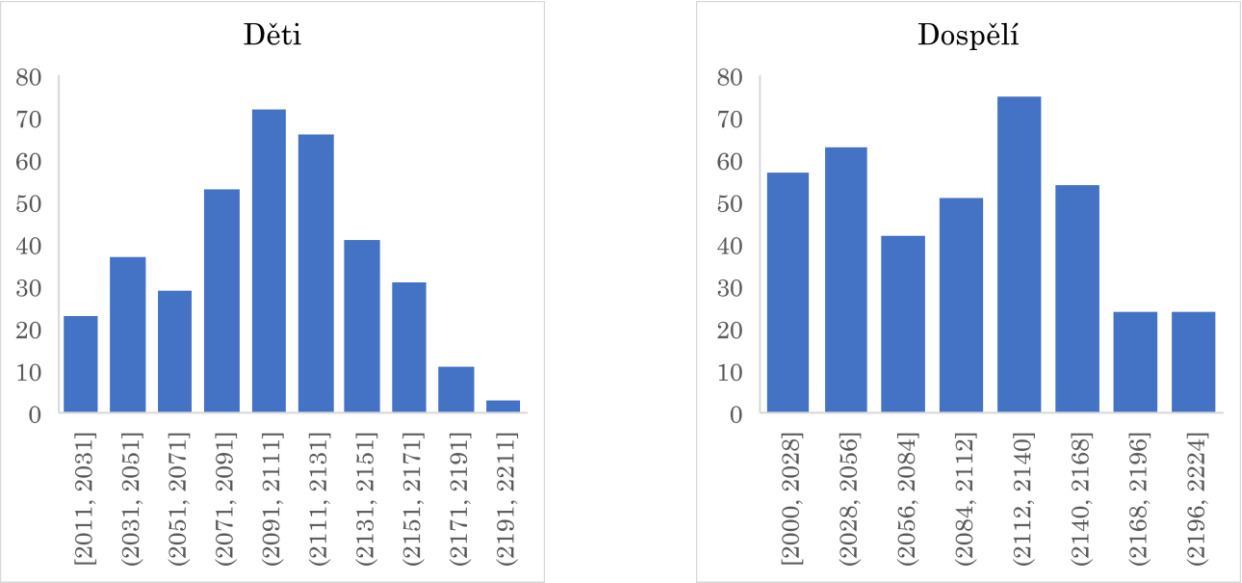
V našem případě nám vyjde v dané buňce např. číslo 0,012. Nulová hypotéza je v t-testu dána obecně: nejsou rozdíly mezi průměry obou skupin. P-value 0,012 nám napovídá, že když „pošleme do vězení 1,2 % nevinných“ lze tuto hypotézu zamítnout a tvrdit, že mezi žoldáky Sparty a Říma jsou rozdíly v průměrném věku a tyto průměry interpretovat. My akceptujeme chybu (poslat

nevinného do vězení) až na úrovni 5 %, proto můžeme tvrdit, že mezi žoldáky jsou statistiky významné rozdíly v aritmetickém průměru.

⌖ Vhodné použití Wilcoxonova testu pro dva histogramy „nenormálních“ dat

Porovnání dvou histogramů, které **nemají** normální rozdělení (lze opět otestovat, popř. otestovat pohledem na histogram) je možné opět otestovat pomocí párového a nepárového přístupu ke dvěma skupinám dat. Tyto testy jsou obsaženy v novějších verzích Excelu či statistických programech. Jde o testy jako je Wilcoxonův test, nebo Mann-Whiteův test apod. Tím, že oba histogramy jsou „nenormální“ je obvyklé porovnat např. mediány, nebo je zobrazit (Obr. 3).

Obrázek 3: Tvary a příklady „nenormálních“ histogramů



⌖ Vhodné použití Fisherova kombinatorického testu pro více rozdělení najednou v kontingenčních tabulkách

Tabulka 2: Kontingenční tabulka

Možnosti	**Výzkumné objekty – kategorie**			
	Muži	**Ženy**	**Děti**	**Řádky celkem**
Podvýživa	2200	5200	7200	Celkem podvýživa 14600
Normální	1250	2250	3250	Celkem normální 6750
Obezita	4400	7400	8400	Celkem obezita 20200
Sloupce celkem	Celkem muži 7850	Celkem ženy 14850	Celkem děti 18850	**Celkem pozorování 41550**

Kontingenční tabulky (Tab. 2) lze testovat F-testem, protože je zde vícero skupin. Klasický test nezávislosti je založen na tzv. testu dobré shody, tedy porovnání očekávaných četností a skutečných četností v jednotlivých políčcích tabulky za předpokladu, že hodnoty sledovaných znaků (objekty a možnosti) na sobě nezávisí.

Tabulka 3: Kontingenční tabulka s procenty pro sloupce

Možnosti	**Výzkumné objekty – kategorie**			
	Muži	**Ženy**	**Děti**	**Řádky celkem**
Podvýživa	2200 **28** %	5200 **35** %	7200 **38** %	Celkem podvýživa 14600
Normální	1250 **16** %	2250 **15** %	3250 **17** %	Celkem normální 6750
Obezita	4400 **56** %	7400 **50** %	8400 **45** %	Celkem obezita 20200
Sloupce celkem	Celkem muži 7850 **100** %	Celkem ženy 14850 **100** %	Celkem děti 18850 **100** %	**Celkem pozorování 41550**

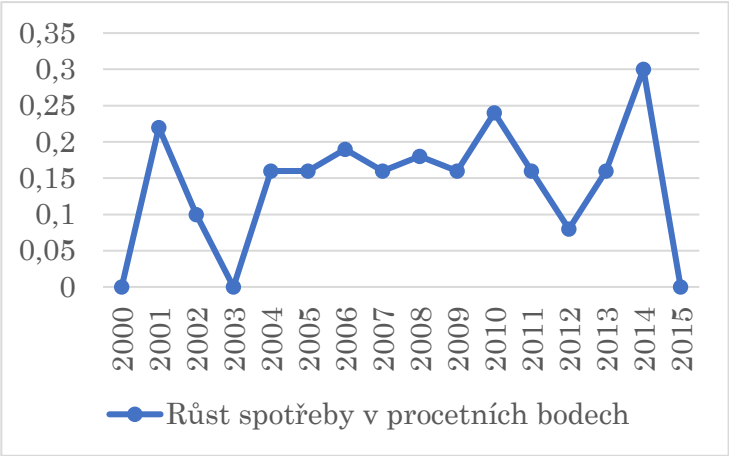
Zde vidíme procentní srovnání našich subjektů v rámci sledovaných kategorií. Z procent plyne, že F-test vyjde pro naše pozorování s p-value menší jak 5 % a budeme zamítat nulovou hypotézu o rovnosti. Můžeme potvrdit, že mezi skupinami jsou statisticky významné rozdíly. Toto se provádí ve statistickém software[9]. Ne vždy jsou rozdíly na první pohled patrné jako v našem příkladě (Tab. 3), proto F-test je pro nás určitou jistotou, že náš úsudek je správný.

Časové řady

Časové řady představují pozorování jednoho jevu (objem produkce, příjmy skupiny obyvatel, počty vojáků, velikost území, vybrané daně, počty sňatků apod.) ve více časových (t) okamžicích. Pohledem na jejich průběh v čase rozeznáváme časové řady tzv. stochastické, které se vyznačují odchylkou od rovnovážné hodnoty. Tyto jsou tedy stacionární, jinak řečeno jsou náhodně ustálené kolem určitého trendu či konstanty (Obr 4.).

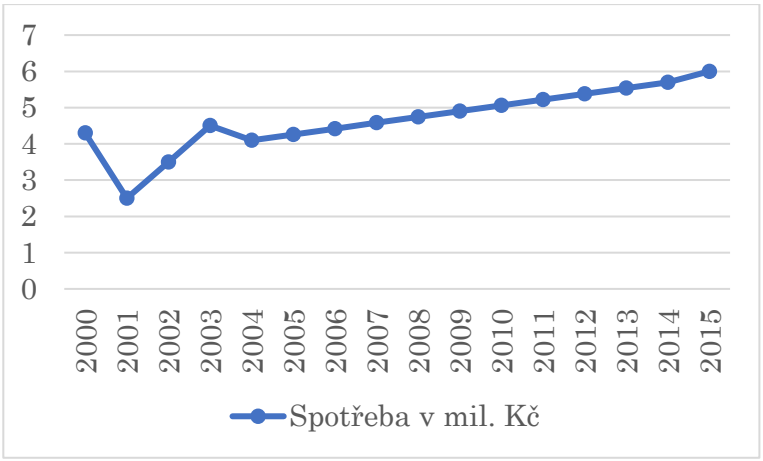
[9] Počítačový program, který usnadňuje práci s daty, například SPSS, STATA, EXCEL.

Obrázek 4: Časové řady stacionární (stochastické)



Na druhou stranu rozlišujeme časové řady deterministické, které se od počátku vyznačují předurčeným trendem. Jinak řečeno, mají v sobě jakoby zakomponovánu počáteční podmínku a díky ní „jdou dál“. Odchylka od předurčeného trendu je nahodilá, ve statistice říkáme, že má povahu „náhodné procházky“. To znamená, že hodnoty se „nevrací“ k nějaké konstantě či trendu jako u stochastické časové řady, proto jsou nestacionární (Obr. 5).

Obrázek 5: Tvary a příklady „nenormálních“ histogramů



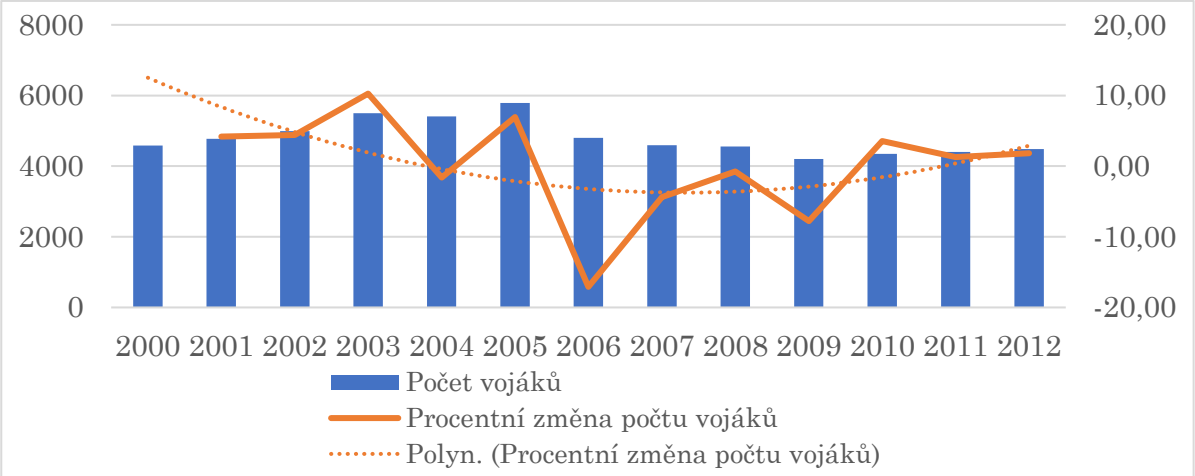
U časových řad je nejlepší jejich zobrazení a kombinace řad v jednom grafu. Pro historika by mělo být minimum vytvořit řadu absolutních hodnot, např. počty vojáků, a do stejného grafu umístit procentuální růst této veličiny. Tak je možné komentovat vývoj dané časové řady a případně i usuzovat slovně na vývoj této časové řady.

Rovnice meziočního výpočtu růstu v procentních bodech

$$Růst_{(t)} = [(Hodnota_{(t)} \div Hodnota_{(t-1)}) - 1] \times 100$$

Pro každý řádek vydělíme současnou hodnotu minulou hodnotou, odečteme jedničku a vynásobíme stem.

Obrázek 6: Kombinovaný graf absolutní a procentní změny



Pokud bychom chtěli usuzovat na trend, pak ve statistické analýze je toto označení používáno obvykle jen pro časové řady, které jsou stacionární a kde lze trend odhadnout. U deterministických časových řad je odhad trendu problematický, může jít o zdánlivý trend. Proto dochází k transformaci deterministických časových řad a interpretaci například změnové veličiny (růst, index, nebo jiná transformace). V našem posledním příkladu v obrázku č. 6 u řady absolutního počtu vojáků nelze na trend dobře usuzovat, jde na první pohled spíše o řadu stochastickou a možnost interpretovat trend bychom museli ověřit statistickým testem (Dickey–Fullerův test).

Růstové řady jsou zpravidla vždy řady stacionární. Časová řada růstu se vždy pohybuje kolem nějaké rovnováhy. V obrázku č. 6 sledujeme nejprve klesající trend procentní změny počtu vojáků a poté růstový trend procentní změny počtu vojáků od roku 2008. Ve statistice je možné se odvolávat kromě přímkového trendu i na další trendy, jako je polynomický, exponenciální, logaritmický a další. V Excelu přidáváme spojnici trendu pouze k růstovým časovým řadám nikoliv absolutním, které zpravidla bývají deterministické a trend zde může být problematický pro interpretaci.

Vhodné statistické nástroje pro výpočty na osobním počítači:

Volně ke stažení:

GRETL: http://gretl.sourceforge.net/win32/

R: https://www.r-project.org/

Placené programy:

STATA: https://www.stata.com/

SPSS: https://www.ibm.com/analytics/spss-statistics-software

Jde o studijní podklad určený studentům výběrového, bakalářského a magisterského semináře doc. PhDr. Františka Stellnera, Ph.D. na FF UK.

Ing. Marek Vokoun, Ph.D.