KATEDRA SOCIÁLNÍ PEDAGOGIKY

PEDAGOGICKÉ FAKULTY MASARYKOVY UNIVERZITY

STATISKA

Výukový text

AUTOŘI: **Jiří Němec**

Karel Pančocha

Martin Stanoev

OBSAH

Úvod

Právě v rukou držíte text, který vznikl jako výstup jedné z dílčích aktivit projektu OPVK „Zvýšení kompetencí vědecko-výzkumných pracovníků zejména v metodologické oblasti“ (CZ.1.07/2.3.00/09.0116), který probíhá/probíhal na Pedagogické fakultě MU. Projekt podporuje nabytí metodologických kompetencí nejen u akademických pracovníků, ale prostřednictvím jejich následné výuky také u studentů zmíněné Fakulty. Tímto bychom rádi poděkovali poskytovatelům grantu, bez něhož by podobný projekt byl jen těžko myslitelný.

Na následujících stránkách zjistíte, že statistika se nudná zdá být pouze těm, kteří nerozumí jejím principům. Základní znalosti statistiky jsou dnes nedílnou součástí vzdělanostního základu všech vysokoškoláků. Je tomu tak proto, že spletitost a diferenciace současného světa a naše potřeba tento svět inteligentně zvládnout nás nutí pochopit základní principy, na základě kterých jsme schopni dojít k datům relevantním pro pochopení problémové situace a jejího řešení. A to nejen v případech, kdy statistiku sami využíváme ke svému výzkumu, ale také v našich každodenních životech. Neboť je nutné, abychom dokázali odlišit skutečně významná zjištění, skutečně důležité informace od manipulativních snah všemožných reklamních agentur a agentur pro vztah s veřejností. Ty nám totiž velice často předkládají na první pohled „důležitá“, „tvrdá“ data, která se však při bližším zkoumání ukážou být nesmyslná a nerelevantní a celý jejich počin je odhalen jako další pokus nás zmást a donutit nás reagovat podle přáních zadavatelů zakázek. K odhalení takových manipulativních snah často stačí znalost základních pojmů statistiky a základních principů nejběžnějších statistických metod.

Právě předání této zdatnosti je hlavním posláním následujících kapitol.

Úvod do problematiky statistického myšlení (měření)

Kvantitativní výzkum

Při kvantitativním výzkumu se neobejdeme bez znalosti základů statistických principů a procedur. Je-li smyslem kvalitativního výzkumu daný jev co nejlépe a nejvíce poznat, důvěrně se s ním seznámit, pochopit ho a v kontextu daného kvalitativního paradigmatu ho také vysvětlit, pak cíle kvantitativního výzkumu jsou odlišné. Jde nám především o ověření dané hypotézy, tzn. objevení nebo ověření souvislostí či příčin sledovaných proměnných a pokud možno předpovědět (zobecnit) i chování těchto proměnných.

Při kvalitativním výzkumu převažují teorie **deskripční** (popisné – kvalitativně popisujeme určitý jev) a **explanační** (vysvětlujeme daný jev). Při kvantitativním výzkumu pracujeme také s teoriemi deskripčními (kvantitativní popis daného jevu) a explanačními, ale též s **predikčními** (předpovídání daného jevu prostřednictvím znalosti proměnných).

Úkol:

Zkuste se zamyslet nad možnými druhy teorií kvalitativního a kvantitativního výzkumu. Které teorie budou výzkumně hodnotné pro kvalitativní a které pro kvantitativní výzkum? Pokuste se najít výzkumný problém, na který byste vzhledem k jeho charakteru hledali odpověď prostřednictvím kvalitativního a kvantitativního výzkumu.

Při kvantitativním výzkumu budeme pracovat především s daty (číselnými charakteristikami daného jevu), která budeme prostřednictvím jasných pravidel a metod shromažďovat (1. **získávání dat**), pomocí různých statistických procedur zpracovávat (2. **analýza dat**) a podle daných pravidel i vyhodnocovat (3. **interpretace dat**).

Statistika jako vědní obor se dělí na:

A) statistiku deskriptivní (popisnou), ve které převládá třídění dat a jejich tabulkové a grafické zobrazení. Pro popis souboru používáme tzv. míry centrální tendence (aritmetický průměr, medián, modus) a rozptyl. Deskriptivní statistiky obvykle využíváme v úvodu výzkumné studie, když se snažíme o vykreslení (deskripci) výchozí situace, například při popisu výběrového souboru často používáme průměr nebo rozptyl sledované proměnné apod. Popisnou statistiku nemůžeme použít k testování hypotéz!

POZOR: Studenti se často dopouštějí závažné chyby při testování hypotéz! Např. si vytvoří hypotézu: Dívky ve sledovaném souboru budou mít lepší prospěch než chlapci. Zjistí, že 55% dívek a 45% chlapců ze souboru mělo na konci roku vyznamenání. Jejich úvaha je následující: Padesát pět je větší než čtyřicet pět, mohu potvrdit svoji hypotézu. Tento myšlenkový pochod je zcela zcestný a svědčí o základních metodologických nevědomostech studenta! Tyto hypotézy je třeba statisticky testovat (viz dále).

Úkol: K formulaci hypotézy bychom mohli mít ještě další výhrady. Poznáte jaké? Čeho bychom se měli vyvarovat při konstrukci hypotéz a jak by správně měla být sestavena? Uměli byste určit závisle a nezávisle proměnnou?

B) statistiku induktivní (zobecňující), jejíž metody využíváme především k zobecnění a predikci, při testování hypotéz apod. V praxi budeme nejčastěji využívat metodu regresní analýzy, korelaci, chí-kvadrát a další.

Jak správně vybrat aneb Základní a výběrový soubor

Statistický soubor můžeme definovat jako množinu prvků, které jsou vymezeny z hlediska věcného, prostorového (kde) a časového (kdy). O každém prvku naší množiny (souboru) bychom měli umět říct, zda do daného souboru náleží nebo nenáleží, tedy zda splňuje dané **kritérium** pro členství v množině. Jednotlivými prvky mohou být lidé, žáci, určité jevy (např. šikana, kriminalita apod.), události (počet dopravních nehod, počet brutálních napadení učitele žákem apod.). Všechny prvky tvoří **základní soubor** nebo též **universum**.

Protože by v praxi bylo velmi obtížné pracovat s celým základním souborem, například všemi dětmi, které v daném roce nastoupí do ZŠ, realizujeme výzkumy prostřednictvím **výběrových souborů** nebo též **výběrů** (**vzorků)**. Počet prvků v něm nazýváme **rozsah souboru** a obvykle ho značíme n. Podle rozsahu výběrového souboru určujeme **reprezentativnost**, která vypovídá o tom, jak dobře výběrový soubor reprezentuje soubor základní. Obvykle o větších rozsáhlých souborech (více jak tisíc respondentů) říkáme, že jsou reprezentativní. Znakem reprezentativnosti výběrového vzorku není pouze počet, ale i další charakteristiky, např. způsob, jakým jsme výběrový soubor získali. Abychom zajistili reprezentativnost, musí mít každý prvek základního souboru stejnou šanci (pravděpodobnost) účastnit se výzkumu, dostat se do výběrového souboru. Z tohoto konstatování jednoznačně plyne, že nejvýznamnější metodou, kterou bychom měli vytvářet výběrové soubory, je **metoda náhodného výběru**, též **znáhodnění**. Této metody bývá hojně využíváno např. i ve Sportce nebo podobných hrách. Do osudí jsou vložena čísla (například lístečky se jmény žáků nebo jiné označení), ze kterých náhodně losujeme patřičný počet. Tento postup si umíme představit v případě, kdy např. z každé třídy jedné školy chceme do výběrového souboru náhodným výběrem získat pět žáků.

Příklad:

Poměr mezi rozsahem výběru a velikostí dané populace (N) nazýváme výběrový poměr. Například budeme chtít vypočítat pravděpodobnost, s jakou se jeden prvek základního souboru (dané populace) dostane do našeho výběrového

souboru. Prostým náhodným výběrem budeme vybírat z 15 000 dětí, které nastoupí do první třídy ZŠ, 300 dětí.

Výběrový poměr = n/N, tj. 300/15000 = 0,02. Můžeme konstatovat, že výběrový poměr je 2%. Každý prvek základního souboru má dvouprocentní pravděpodobnost, že bude vybrán (vylosován) a zúčastní se výzkumu.

Náhrada prostého a náhodného výběru

Představa, že bychom prostým náhodným výběrem vybírali z tak velkého množství dětí, může některého z badatelů odradit již na samém počátku. Můžeme tedy použít jiné náhrady prostého výběru. Např. Hendl (2004) uvádí:

Stratifikovaný náhodný výběr

Sestává-li populace z různých subpopulací, pak je možné ji rozdělit do těchto skupin (podskupin, vrstev – podle toho stratifikovaný, například podle demografických nebo socioekonomických ukazatelů) a náhodně vybírat z každé skupiny.

Vícestupňový shlukový výběr

Například budeme chtít realizovat výzkum ve velkých sídlištních školách ve městech nad 50 000 obyvatel. Náhodně vybereme ze souboru všech krajů v ČR, náhodně ze všech měst daného kraje a náhodně také školu z daného města. V každé vrstvě shluků provádíme náhodný výběr. Tento výběr je vhodný například pro průzkum veřejného mínění.

Systematický výběr

Je podobný výběru náhodnému. Očíslujeme prvky populace a na základě rozhodnutí vybíráme například každý osmdesátý prvek. Podmínkou je, že prvky nemohou být očíslovány podle určitých charakteristik (například od nejmladších po nejstarších). V takovém případě by výběr nebyl validní.

Velmi často však budeme provádět pilotní výzkumy nebo drobné výzkumné sondy, které nebudou mít ambice na zobecnění ve vztahu k celé populaci. V mnoha případech bude překážkou i nedostatek financí, a tak se můžeme rozhodnout pro jiné způsoby výběru.

Další způsoby výběru

Výběr na základě dobrovolnosti

Dnes jsme svědky takového výběru především na televizních obrazovkách. Kdo chce, může poslat SMS a vyjádřit se k danému problému. O odpovědi můžeme též požádat např. návštěvníky určitých webových stránek, elektronických diskusí apod. Výběr není reprezentativní, ale může nám poskytnout hrubou představu o mínění lidí.

Výběr na základě dostupnosti

Provádíme-li například výzkum u studentů VŠ, je nejjednodušší je požádat o vyplnění dotazníku. Podobně bychom například mohli jako respondenty výzkumu využít žáky nejbližší fakultní školy, kolegy na pracovišti apod.

Kvótní výběr

Máme-li demografické údaje o základním souboru, pak se můžeme pokusit o kvótní výběr. Z každé předem definované kategorie obyvatelstva vybereme patřičný počet respondentů a dotazujeme se. Např. oslovíme skupinu studentů, nezaměstnaných, vyučených apod.

Základní statistické znaky

Předmětem našeho zkoumání jsou často **vlastnosti** prvků daného výběrového souboru (množiny). Tyto vlastnosti jsou reprezentovány tzv. **znaky** (uváděno podle Škaloudová, 1998, s. 10-11). Platí přitom, že jedna vlastnost může být reprezentovaná i více znaky. Podle charakteru znaku je dělíme na:

- **URČUJÍCÍ**, které prezentují příslušnost k danému souboru (národnost, pohlaví, věk) a
- **ZKOUMANÉ**, které jsou předmětem našeho výzkumu (inteligence, zájmy, hodnoty, postoje apod.).

Zkoumané znaky jsou označovány jako **PROMĚNNÉ.** V tomto případě by pohlaví bylo **nezávisle proměnnou** a např. zájmy nebo inteligence **závisle proměnnou**. Pro koncepci vlastního výzkumu je velice důležité, abychom si uvědomili, s jakými proměnnými pracujeme!

Druhy znaků (proměnných)

Jednotlivé znaky mohou nabývat různých charakteristik a podle toho je dělíme na:

- **KVALITATIVNÍ** – jsou vyjádřeny slovně a vyjadřují určitou kvalitu zkoumaného jevu (bydliště, profese, zájmy, hodnoty člověka, pohlaví, barva pleti, národnost apod.) POZOR – číselné hodnoty nemůžeme použít k číselným operacím.

- **KVANTITATIVNÍ** – jsou vyjádřeny numericky (věk, hmotnost, inteligenční kvocient, měsíční výdělek apod.). Vyjadřují míru zastoupení KVALITY – (tj. velký problém měření v pedagogice, viz dále)
    - Kvalitativní znaky pro snazší zpracování v PC velice často převádíme na numerické (tzv. kódujeme, krajské město = 1 atd.)

- **ALTERNATIVNÍ** – nabývají pouze dvou hodnot (pohlaví – muž nebo žena)
- **MOŽNÉ** – nabývá více hodnot (aprobace, věk apod.)
- **SPOJITÉ** (kontinuální) – v daném intervalu může nabývat jakýchkoliv hodnot – výrazným omezením je přesnost (procento). Máme-li dvě různé hodnoty, existuje mezi nimi třetí.

- **NESPOJITÉ** (diskrétní) – v daném intervalu nabývají izolovaných hodnot (ročník studia, počet narozených dětí, počet opakování ročníku atd.)

Charakteristika proměnných aneb (kategorizace)

Abychom mohli správně využít různých statistických procedur, je zcela nezbytné, abychom znali charakteristiku jednotlivých dat (proměnných). Podle toho, jaká data budeme mít k dispozici, budeme plánovat jejich analýzu, testování hypotéz apod.

NOMINÁLNÍ DATA

Poskytují málo informací, spíše je můžeme klasifikovat nebo kategorizovat. V této souvislosti rozlišujeme dvě kategorie nominálních dat.

A) Každá kategorie obsahuje jeden prvek (SPZ, rod. č.apod.)

B) Každá kategorie má více prvků (pohlaví, věk, typ školy)

Číselné označení dané kategorie nominálních dat nikdy nereprezentuje určitou naměřenou hodnotu, ale odlišuje kategorie!!!! Tuto skutečnost je třeba si dobře zapamatovat! V druhé kategorii B) můžeme vypočítat absolutní četnost, analýzu rozptylu, apod., nemůžeme však počítat PRŮMĚR.

ORDINÁLNÍ (pořadová) DATA

Měření v pedagogice bychom s trochou nadsázky mohli přirovnat k audienci, při které se král ptá svých dcer, jak ho mají rády. Myslíte, že můžeme provádět výzkum ve třídě a ptát se například, jak se dítě snaží? Jak byste například postupovali, pokud byste chtěli zjistit oblibu jednotlivých učitelů? Kam bychom oblibu, jako sledovaný znak (vlastnost učitele), zařadili?

Podívejme se na charakteristiky ordinálních (pořadových) dat. Jak sám jejich název napovídá, můžeme je seřadit do pořadí, tedy A > B > D apod. Čísla určují hodnotu pořadovou nikoliv skutečně NAMĚŘENOU. (Např.známky ve škole.)

Jak můžeme rozumět této větě? Jaký je rozdíl mezi hodnotou pořadovou a naměřenou? Jak je to se známkami ve škole? Vyjadřují skutečně naměřenou hodnotu? Už jste někdy přemýšleli o rozdělení známek? Rozkládají se podle normálního rozdělení?

- Každému stupni obvykle odpovídá více prvků. (Například stejnou známku z písemné práce ve třídě získalo více dětí.)
- Například u známek často počítáme jejich průměr, z hlediska další statistické analýzy to bývá obvyklé (např. průměrná odpověď respondentů na dané škále), ale měli bychom si uvědomit, že se jedná o zprůměrování kvalitativní odpovědi. Více než průměr můžeme např. uvádět medián a modus. S ordinálními hodnotami se také setkáváme při škálování (např. nespokojen, částečně nespokojen, částečně spokojen, spokojen). Spokojen je lepší než nespokojen, ale nemůžeme určit přesnou vzdálenost (hranici spokojenosti a nespokojenosti)!

METRICKÁ (INTERVALOVÁ) DATA

Data jsou charakteristická tím, že:

- Rozdíl mezi sousedními body (hodnotami) je stejný.
- „Měříme vzdálenosti mezi údaji“, netřídíme do skupin
- Lze použít: průměr, směrodatnou odchylku, korelační koeficient apod.
- Můžeme libovolně určit začátek 0. Když ho někdo jiný určí jinde, vzdálenost mezi daty se nezmění. (Např. vstup na VŠ = 0.)

POMĚROVÁ DATA

Velice zřídka se v pedagogice a sociálních vědách obecně setkáváme s poměrovými daty. Ty jsou především výsadou přírodovědných oborů. Výjimkou může být např. věk, výška, hmotnost apod. Poměrová data jsou charakteristická:

- Nejdokonalejší stupnicí.

- Rozšiřují intervalovou stupnici. 0 znamená nepřítomnost sledovaného znaku. (Např. pokud se někdo právě narodí, je mu 0 let.)
- Poměrová stupnice umožňuje poměřovat, kolikrát je něco větší nebo menší.

Úkol: Pokuste se roztřídit následující proměnné podle uvedených kritérií. U nominálních dat se pokuste navrhnout příslušné kategorie. Věk respondentů, počet dětí v rodině (počet sourozenců), průměrný roční plat, národnost, známky ve škole, průměrný prospěch žáka z přírodovědných předmětů, počet bodů v didaktickém testu, hmotnost, poměr hmotnosti a výšky člověka, barva očí, druh absolvovaného studia, zájmy dítěte, oblíbené předměty ve škole, růměný počet dětí ve třídách, velikost (plocha) třídy, inteligenční kvocient.

Základní statistický popis proměnných

Klíčové pojmy

Míry centrální tendence (střední hodnoty), míry rozptýlenosti, míry koncentrace, průměr, medián, modus, šikmost a špičatost, variační rozpětí, rozptyl, směrodatná odchylka.

Podívejme se nyní, jakým způsobem můžeme popisovat proměnné a jak tyto charakteristiky interpretovat. Studium proměnných, zvláště u velkých výběrových souborů, může být složité a na první pohled (např. velký počet řádků – případů), především bez použití statistických softwarů, složité. Můžeme využívat jednoduchých způsobů redukce, které nám nabízí statistický software, a které nám umožní jednoduchý popis a interpretaci našich naměřených hodnot. Než přistoupíme k složitějším analýzám, vytvoříme si základní představu o datech, se kterými budeme dále pracovat. Množiny dat našich proměnných můžeme tedy popsat (redukovat) prostřednictvím:

- Výpočtem **měr centrální tendence** (např. **průměr, medián, modus**)
- Výpočtem **měr rozptýlenosti** (např. variační **rozpětí, rozptyl, směrodatná odchylka** atd.)
- Výpočtem **měr koncentrace** (především **šikmost a špičatost**)

Míry centrální tendence

Tzv. střední hodnoty souborů nebo míry centrální tendence nám popisují jednu ze základních charakteristik číselných řad.

Průměr

Aritmetický průměr je definován jako součet všech naměřených hodnot vydělený jejich celkovým počtem (srov. Hendl, 2004). Viz vzorec.

$$M = \frac{\sum_{i=1}^{n} xi}{n}$$

Průměr můžeme popsat dle následujících charakteristik:

- Geometricky si aritmetický průměr můžeme představit jako těžiště.
- Součet odchylek od průměru je roven nule.
- Průměr je zatížen odlehlými hodnotami.
- Používáme minimálně od intervalové stupnice dále.

Skutečnost, že průměr je zatížen odlehlými hodnotami bychom si měli uvědomit vždy, když budeme provádět statistické analýzy, ve kterých určitým způsobem figuruje průměr, např. rozptyl, směrodatná odchylka atd. Výrazně odlehlá hodnota (např. vložená do tabulky omylem) může zásadním způsobem zkreslit naše výsledky a potažmo i jejich následnou interpretaci.

Medián

Medián je střední hodnota, která rozděluje naše data seskládaná do pořadí na polovinu. V případě lichého souboru dat se jedná vždy o střední hodnotu, v případě sudého počtu se dvě střední hodnoty sečtou a podělí dvěma (průměr). Výpočet mediánu užíváme zejména v případech, kdy rozdělení proměnné je nesymetrické (tzn. např. levostranné nebo pravostranné). Výhodou mediánu je skutečnost, na rozdíl od průměru, že není zatížen odlehlými hodnotami. Poměrně dobře ho můžeme využít např. při interpretaci ordinálních proměnných. V uvedeném příkladu je to číslo 5.

$$\{0; 1; 2; 5; 8; 9; 10\}$$

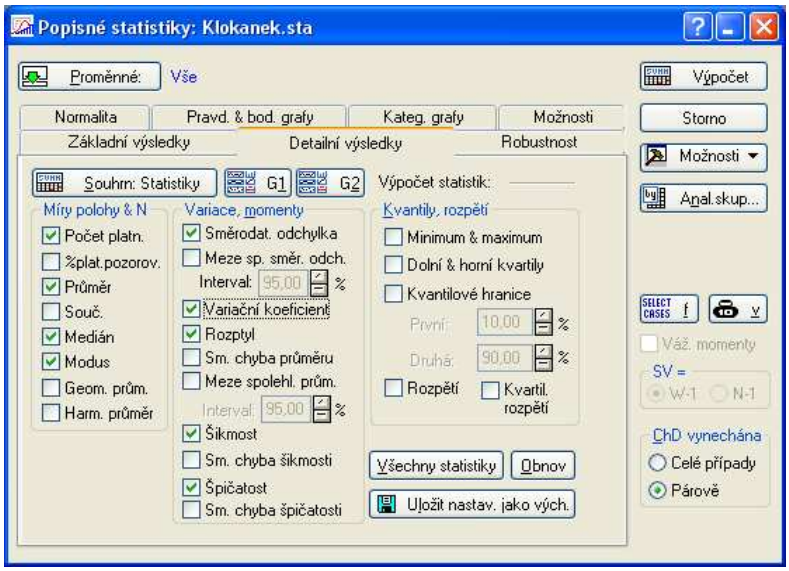
Modus

Modus je prezentován hodnotou (číslem), které je v souboru zastoupeno nejvícekrát. Modus může být i vícenásobný (např. rozdělení s dvěma nevíce vrcholy). Prostřednictvím této míry centrální tendence můžeme např. dobře interpretovat i ordinální data. Prakticky by se pak jednalo např. o tzv. typickou odpověď na dané škále apod. V našem příkladě se jedná opět o číslo 5.

{1; 2; 3; 4; 4; 5; 5; 5; 6; 7; 8; 8}

Výpočet měr centrální tendence prostřednictvím programu statistica

Všechny operace pro výpočet popisných statistik nalezneme: Statistiky – základní statistiky a tabulky – popisné statistiky. Zvolíme proměnné kliknutím na Proměnné a zobrazí se nám následující tabulka.



V záložce detailní nastavení zaklikneme vše, co požadujeme vypočítat a klikneme na Výpočet. Vytvoří se nám následující tabulka.

Proměnná	Popisné statistiky (Klokanek.sta)							
	Průměr	Medián	Modus	Četnost modu	Rozptyl	Sm.odch.	Šikmost	Špičatost
klokanek	41,36667	38,00000	Vícenás.	2	627,6885	25,05371	0,685684	-0,11446
Mat	18,00000	15,00000	10,00000	5	94,1379	9,70247	0,734895	-0,48371
Čj	18,93333	16,50000	25,00000	4	113,3747	10,64776	0,521372	-0,74866
LOC1	2,16667	2,00000	1,000000	11	1,2471	1,11675	0,445076	-1,16185
LOC2	2,30000	2,00000	Vícenás.	9	1,2517	1,11880	0,303917	-1,24367

V jednotlivých sloupcích nalezneme požadované výpočty dílčích proměnných (řádky).

Úkol: Na základě vašich znalostí se pokuste o interpretaci jednotlivých proměnných. Zjistěte, zda-li se jedná o symetrické rozdělení proměnných. Porovnejte rozptyly proměnných.

Směrodatná odchylka – statistická charakteristika vyjadřující podobně jako rozptyl rozptýlenost dat kolem průměrů. Značíme ji *s* a je definována jako druhá odmocnina rozptylu. **S. o.** je rovna nule, když všechna data nabývají stejné hodnoty, v opačném případě platí $s > 0$. **S. o.** je ovlivněna průměrem, odlehlé hodnoty ji zvětšují. **S. o**. neposkytuje věrohodné informace v případě silně zešikmeného rozdělení, v takovém případě se používají kvantilové míry.

Statistická indukce

Klíčová slova: statistická indukce, výběrový soubor, základní soubor, náhodný výběr.

Definice: Statistická indukce je formalizované vznášení soudů o základním souboru (populaci) na základě zjištění učiněných na výběrovém souboru (vzorku).

Princip: Těžiště statistické analýzy nespočívá v popisné statistice, ale v oblasti statistiky označované jako statistická indukce. V literatuře se též můžeme setkat s označeními statické rozhodování, statistické usuzování (inference). Jedná se o metody používané v případech, kdy nemáme kompletní informace o všech zkoumaných jednotkách analýzy (např. v případě dotazníkového šetření je jednotkou analýzy jeden respondent). Informace o této části - označujeme jako vzorek nebo výběrový soubor, zobecňujeme na celek, označujeme jako populaci, základní soubor.

K výběru vzorku (výzkumného souboru) používáme náhodný výběr (jednotlivé modifikace náhodného výběru viz jiná část textu). Při náhodném výběru má každý element základního souboru (populace) stejnou pravděpodobnost, že bude vybrán do výběrového souboru (vzorku). Při statistické indukci tedy využíváme náhodu. Metody statistické indukce se opírají o počet pravděpodobnosti. Existují dva základní typy statistické indukce - metody odhadování a postupy založené na statistických testech.

Statistické odhady

Klíčová slova: parametr, odhad bodový, odhad intervalový, interval spolehlivosti,

Definice: Při odhadech se snažíme na základě pozorování náhodně vybraných jednotek – výběrovém souboru odhadnout vlastnosti základního souboru. Při odhadech je potřeba rozlišit parametr a jeho odhad – odhad parametru získáváme na základě výběru z populace, můžeme ho spočítat, skutečnou hodnotu parametru můžeme znát pouze v případě, že naše šetření obsáhlo celou populaci.

Princip: Děje se tomu například v případě, kdy se na základě šetření veřejného mínění a politických preferencí snažíme usoudit na skutečný poměr preferencí politických stran v populaci. Odhad provádíme buďto jedinou hodnotou, anebo číselným intervalem. Například pokud chceme odhadnout průměr populace, můžeme využít výběrový průměr, výběrový medián, nebo průměr extrémních hodnot (minima a maxima). Místo bodového průměru však často využíváme odhad intervalový – vytvoříme tak určitý interval spolehlivosti, odvozený od určité hladiny spolehlivosti.

Interval spolehlivosti počítaný například pro populační průměr je dán výběrovým průměrem, který tvoří jeho střed, jeho hranice jsou určeny mezí chyby. Délka intervalu je určena hladinou spolehlivosti, tedy pravděpodobností, se kterou se odhadovaný parametr ocitne v tomto intervalu při opakovaném provádění výběru. Nejpoužívanější hladiny jsou 90%, 95% a 99%.

Interval spolehlivosti můžeme vypočítat i pro jakýkoliv jiný parametr – například procentuální preferenci určité politické strany. Přesnost, s jakou jsme schopni odhadnout parametr, je dán velikostí vzorku a jeho variabilitou. Nezávisí, vyjma velmi malé populace, na velikosti základního souboru. Pokud vynášíme soud o platnosti nulové hypotézy ve vztahu k určitému parametru, interval spolehlivosti obsahuje dostatečné množství informací k přijetí takového rozhodnutí.

Testování hypotéz

Klíčová slova: nulová hypotéza, alternativní hypotéza, hladina významnosti, chyba 1. druhu, chyba 2. druhu,

Definice: Při testování hypotéz formulujeme dvě vzájemně si odporující hypotézy. Nulovou hypotézu, která obvykle deklaruje "žádný rozdíl", respektive nalezený rozdíl lze přičíst přirozené variabilitě dat. Alternativní hypotézu chceme dokázat, je to naše pracovní hypotéza - znamená že nulová hypotéza neplatí, znamená existenci diference mezi skupinami, nebo existence závislosti mezi proměnnými.

Při testování hypotéz stanovujeme hladinu významnosti, což je pravděpodobnost, že zamítneme nulovou hypotézu, ačkoliv ona platí. Nejčastěji se za hraniční hodnotu bere 0,05 (nebo méně). Naměřená hodnota p, kterou získáme výpočtem testovací statistiky, nám potom odpovídá na otázku - jestliže platí nulová hypotéza, jaká je pravděpodobnost, že získáme právě vypočítanou hodnotu, nebo ještě neobvyklejší hodnotu testovací statistiky.

Princip: Pokud jsme stanovily hodnotu hladiny významnosti 0,05 potom v případě hodnoty p 0,05 a nižší zamítáme nulovou hypotézu (ne zcela přesně řečeno - pravděpodobnost že platí nulová hypotéza při hodnotě p 0,05 je 5%).

Při stanovování platnosti hypotéz se můžeme dopustit dvou typů chyb - chyby 1. druhu, kdy nulová hypotéza platí, ale my jí zamítáme, pravděpodobnost této chyby je odvislá od stanovené hladiny významnosti. Nebo chyby druhého druhu, kdy nulová hypotéza neplatí, ale my ji nezamítneme. Je dána silou testu, která závisí na zvolené testové metodě a na tom jaké je skutečné rozdělení dat.

Je důležité ještě podotknout, že nezamítnutí nulové hypotézy neznamená její důkaz. Spíše to znamená, že nemáme dosti evidence k jejímu zamítnutí.

Testy hypotéz lze rozdělit na:

1. Testy parametrů - předpokládáme, že typ rozdělení známe, pouze neznáme jeho přesné parametry. Testové hypotézy mohou nabývat rozličné podoby a mohou se

týkat jednoho nebo více parametrů - např. t-test shody úrovně ve dvou skupinách, testování průměru dvoustranným z-testem atd.

2. Testy nezávislé na typu rozdělení (tzv. neparametrické) - používají se v případě, že rozdělení veličiny v základním souboru není známo, nebo neodpovídá žádnému z obvyklých rozdělení. Výhodou je, že nemusíme mít žádné apriorní informace o základním souboru. Nevýhodou je menší síla (citlivost) oproti testům parametrickým. Patří sem např. Mann-Whitneyův test shody úrovně ve dvou nezávislých výběrech. Tyto testy jsou vhodné pro hodnocení ordinálních dat.

3.Testy typu rozdělení - ty používáme k ověření (či zamítnutí) hypotézy, že veličina má v souboru určité rozdělení. Dominantní úlohu v této skupině mají testy ověřující normalitu, protože ta je předpokladem mnoha statistických metod.

Testy normality

Klíčová slova: normální rozdělení, centrální limitní teorém, ověření normality dat

Definice: Pro normální rozdělení platí, že v intervalu ohraničeného na jedné straně průměrem sečteným se směrodatnou odchylkou a na druhé straně průměrem odečteným o směrodatnou odchylku se nachází 68,3% populace, v intervalu ohraničeném průměrem sečteným se dvěma směrodatnými odchylkami a průměrem odečteným o dvě směrodatné odchylky se nachází 95,5% populace a v intervalu ohraničeném průměrem sečteným se třemi směrodatnými odchylkami a průměrem odečteným o tři směrodatné odchylky se nachází 99,7% populace.

Princip: Pro uvedení do problémů je třeba alespoň uvést, že jsou popsána různá pravděpodobnostní rozdělení náhodných proměnných. Nejvíce používané rozdělení je normální rozdělení, zvané též Gaussovo rozdělení. Je tomu tak z těchto příčin: mnoho sledovaných proměnných můžeme aproximativně modelovat pomocí tohoto rozdělení. Některé proměnné lze převést jednoduchou transformací na proměnnou, jež má toto rozdělení a existuje mnoho statistických procedur, které jsou v důsledku předchozích dvou bodů odvozeny pro toto rozdělení. Dále je to díky platnosti centrálního limitního teorému - neformálně ho lze vyjádřit takto: proměnná, pokud vznikla jako součet velkého počtu efektů nezávisle působících příčin, má přibližně normální rozdělení. Aproximace je tím lepší, čím je počet přispívajících faktorů větší.

K ověření normality lze mimo jiné použít statistické testy. Než však přistoupíme k testům normality, je vhodné se též nad podstatou problému zamyslet - lze normální rozložení očekávat? Např. v případě platů

obyvatelstva existuje spodní hranice - minimální mzda, platy však nejsou omezené shora, nelze tedy očekávat symetrii kolem středu, lze tedy vyloučit normální rozdělení. U některých veličin je normální rozložení naopak dobře známo, lze tedy také od testování normality upustit. Pro posouzení normality je také vhodná vizualizace pomocí zobrazení grafem, kdy rozložení dat (znázorněné

histogramem nebo normálním pravděpodobnostním grafem) opticky porovnáme s normálním rozdělením. Mezi testy používané k rozhodnutí, zda data pocházejí z normálního rozdělení, můžeme použít např. Shapirův -Wilksův, Kolmogorovův - Smirnovův a Lillieforsův test.

Testování hypotéz na základě nominálních a kategoriálních (pořadových) dat

Klíčové pojmy pro daný výklad

Nominální data, kategoriální data, chí-kvadrát, test dobré shody chí-kvadrát, test nezávislosti chí-kvadrát, pozorovaná četnost, očekávaná četnost

Definice

Chí-kvadrát test dobré shody. Jedná se o jeden z testů dobré shody (další např. Kolmogorovovův test a Lillieforsův test). Slouží k tomu, abychom zjistili, zda data získána v pedagogické realitě při náhodném výběru respondentů (např. ve školní třídě, či mezi skupinou žáků se zdravotním postižením) se odlišují od teoretických četností, které odpovídají nulové hypotéze. Tedy zda hodnoty, které jsme získali, se liší od předpokládaných hodnot, které bychom získali, pokud by platila nulová hypotéza.

.

Chí-kvadrát test nezávislosti. Pomocí tohoto testu můžeme s využitím dvourozměrných kontingenčních tabulek testovat nezávislost dvou nahodilých veličin. V pedagogice jej lze využít v případech, kdy zjišťujeme, zda existuje závislost mezi dvěma jevy, které byly zachyceny měřením pomocí nominálních popřípadě ordinálních dat. (pozor, to že mezi dvěma jevy existuje závislost, neznamená, že jeden jev je příčinou druhého jevu. Testem chí-kvadrát nezávislosti nelze zjišťovat vztah příčiny a následku).

Úvod

Některé vědecké disciplíny velmi často analyzují proměnné, které mají metrický nebo poměrový charakter. V medicíně můžeme měřit teplotu těla pacienta ve stupních, ve fyzice zase rozpínání materiálů v milimetrech. Můžeme pak tvrdit, že teplota pacienta se zvýšila nebo snížila o X stupňů, nebo že materiál A se po zahřátí na určitou teplotu rozpíná 3x více než materiál B. V pedagogice však na tento typ dat narazíme jen velmi zřídka. Výjimkou mohou být výsledky různých vědomostních testů, kde může skóre dosažených výsledků nabývat hodnot např. od 0 do 100. Většinou se však v pedagogické realitě setkáváme s daty na úrovni nominální nebo ordinální. Nominální data můžeme vnímat jako data kvalitativní (pozor, nejedná o to, že by se s nimi nedalo pracovat v rámci kvantitativní metodologie), neboť si je můžeme představit na ose, na které se od sebe budou lišit pouze podle druhu. Není však možné určit, která hodnota je větší a která menší, ani o kolik je menší či větší. O sčítání či násobení proměnných mezi sebou zde nemůže být ani řeč. Typickou nominální proměnnou je etnická příslušnost nebo preference určitého předmětu ve škole. Nikoho by asi nenapadlo říci, že matematika je menší než chemie, nebo že Čech je 3x více než Francouz. Vzhledem k tomu, že se právě s nominálními a ordinálními daty setkáme v pedagogickém výzkumu nejčastěji, podíváme se v následujícím textu na možnosti použití statistických testů významnosti v této oblasti. Konkrétně se budeme zabývat dvěma užitími testu chí-kvadrát, a to testem dobré shody chí-kvadrát a testem nezávislosti chí-kvadrát.

Zajímavost

Již v předchozím textu jsme se seznámili se jménem Karl Pearson. V době svých studií na univerzitě se tento Londýnský rodák zabýval celou řadou disciplín od matematiky, přes fyziku, náboženství, historii, socialismus až po Darwinismus.

Poté co vystudoval právo na Cambridgské univerzitě a získal doktorát politických věd na univerzitě v Heidelbergu, se stal se profesorem aplikované matematiky na University College v Londýně. Ve své knize *Gramatika vědy* (1892) uvádí, že statistická analýza dat je základem všeho vědění. Z tohoto důvodu je Pearson často označován za „otce statistiky." Kromě toho, že do statistiky zavedl koncept směrodatné odchylky a zasloužil se o rozvoj korelačního koeficientu je také autorem prvního testu dobré shody chí-kvadrát. Pearson byl výjimečný člověk a jeho věhlasné přednášky a nadšení pro statistiku inspirovalo mnohé jeho pokračovatele.

Princip testu chí-kvadrát dobré shody

Nejprve se podíváme na to, jak funguje test chí-kvadrát (značíme jako χ^2) dobré schody v případech, kdy máme jen jednu proměnnou, která nabývá nejvýše dvou různých hodnot. Nejlépe si takovou situaci představíme na příkladu.

Učitel dějepisu na střední škole se rozhodne, že dá svým studentům na výběr, jakým způsobem budou na konci roku z předmětu vyzkoušeni. Studenti si mohou vybrat mezi písemným testem nebo ústním zkoušením. Učitele zajímá, zda je jeden ze způsobů zkoušení u studentů oblíbenější. Každý z padesáti studentů ze dvou tříd tedy rozhodne o tom, zda raději napíše test, nebo se nechá vyzkoušet ústně. Pokud by studenti považovali obě varianty zkoušení za rovnocenné, vybrala by si přibližně polovina z nich písemný test a přibližně druhá polovina ústní zkoušení. Určitý vliv zde ovšem hraje i náhoda a proto se může počet studentů v jedné skupině od studentů v druhé skupině lišit, přestože vnímají obě varianty zkoušení jako rovnocenné. (je to obdobné jako u házení mincí. Když ji hodíte 50x může se vám stát, že padne 21x panna a 29x orel, přestože je mince vyvážená a žádná ze stran mince není těžší. Sehrála zde roli „pouze" náhoda). Test chí-kvadrát dobré shody nám pomůže odhalit, zda jsou rozdíly mezi velikostí skupin

studentů dány pouze náhodou, nebo je jeden ze způsobů zkoušení oblíbenější než druhý.

Postup:

1. Nejprve si musíme stanovit hladinu významnosti α, jinými slovy pravděpodobnost že zamítneme nulovou hypotézu i v případě, že je ve skutečnosti pravdivá. Jak již možná víte, ve statistice není nic na 100% a tak mají všechny naše závěry pouze pravděpodobnostní charakter. Ve společenských vědách se většinou používá hladina významnosti α = 0,05, což znamená, že počítáme s 5% možností, že náš závěr bude chybný.

2. Dále si musíme stanovit nulovou (H0) a alternativní (H1) hypotézu.

   H0: Popularita písemného testu je stejná jako popularita ústního zkoušení.
   H1: Popularita písemného testu není stejná jako popularita ústního zkoušení.

   Dané hypotézy si můžeme zapsat i numerickou formou s předpokládanou proporcí studentů, kteří si vyberou zkoušení pomocí písemného testu, tedy:

   H0: Popularita písemného testu = 0,5
   (Tento zápis značí, že v případě že popularita obou forem zkoušení je stejná, písemný test si vybere právě polovina studentů)

   H1: Popularita písemného testu ≠ 0,5
   (Tento zápis značí, že v případě že popularita obou forem zkoušení není stejná, písemný test si vybere více nebo méně než polovina studentů)

3. Nyní jdeme s učitelem do tříd a zeptáme se 50 studentů, kterou formu zkoušení si vyberou. Pokud by studenti nepreferovali žádnou z forem, bylo by jich asi 25 pro test a 25 pro ústní zkoušení. Těmto hodnotám říkáme očekávané četnosti. My jsme však zjistili, že 17 studentů chce psát test a 33 chce být zkoušeno ústně. To jsou tzv. pozorované četnosti.
4. V této fázi nám přichází na pomoc Karl Pearson se svým vzorcem pro chí-kvadrát test, do kterého doplníme hodnoty, které jsme naměřili.

$$\chi^2 = \sum \left[\frac{(f - f)^2}{f} \right]$$

Fo = pozorované četnosti

Fe = očekávané četnosti

Po dosazení našich dat do vzorce získáme hodnotu $\chi^2 = 5{,}12$.

Co nám hodnota chí-kvadrát 5,12 říká? Obecně lze říci, že čím vyšší hodnota chí-kvadrát, tím větší rozdíl mezi očekávanými a pozorovanými četnostmi. Zdali je tento rozdíl dostatečně vysoký, abychom mohli zamítnout hypotézu H0 a přijmout hypotézu alternativní, záleží na tom, zdali hodnota chí-kvadrát překračuje tzv. kritickou hodnotu pro daný stupeň volnosti (SV) a zvolenou hladinu významnosti α. Požadovanou kritickou hodnotu je možné najít ve statistických tabulkách, v současnosti však spíše využijeme některý se softwarů pro statistickou analýzu dat, který nejenže nám po zadání výše uvedených informací hodnotu chí-kvadrát vypočítá, ale také nám vyhodnotí, zdali překračuje kritickou hodnotu.

Než se podíváme na to, jak lze tuto operaci provést v softwaru Statistica, můžeme prozradit, že v našem případě je pro hladinu významnosti α 0,05 a stupeň volnosti 1 kritickou hodnotou chí-kvadrát 3,841. Jak je z výše uvedeného patrné,

námi vypočtený chí-kvadrát 5,12 tuto hodnotu překračuje. Je tedy možné říci, že rozdíly mezi očekávanými a pozorovanými četnostmi je statisticky významný (tento rozdíl není pouze náhodný). Je tedy možné přijmout alternativní hypotézu, popularita písemného testu není stejná jako popularita ústního zkoušení. Z výsledků je patrné, že popularita písemného testu je nižší.

V našem příkladě pracujeme s proměnnou, která nabývá pouze dvou hodnot (písemný test nebo ústní zkoušení), chí-kvadrát test dobré shody však lze analogicky využít i v případech, kdy proměnná nabývá většího množství hodnot (studenti by si mohli volit například ze čtyřech typů zkoušení).

Princip testu chí-kvadrát nezávislosti

Ve výše uvedeném příkladu jsme se věnovali použití chí-kvadrátu na případ s jednou proměnnou. Chí-kvadrát však může být použit i pro bivariační analýzu. V tomto případě budeme používat test chí-kvadrát nezávislosti. Pokusíme se zjistit, zdali je jedna proměnná (jeden jev) souvisí s druhou proměnnou (druhým jevem). Tato situace je častá např. při zpracovávání dotazníků nebo hodnocení účinnosti pedagogické intervence.

Princip si opět nejlépe ilustrujeme na jednoduchém příkladě:

Speciální pedagog pracuje se žáky s dyslexií v pátých třídách ŽŠ. Rozhodne se vyzkoušet novou metodu reedukace specifických poruch učení a za tímto účelem si žáky náhodně rozdělí na dvě skupiny. S první skupinou bude pokračovat v práci za použití standardních metod reedukace, u druhé skupiny vyzkouší nový přístup. Na konci školního roku jsou všichni žáci ohodnoceni v rámci svých čtenářských dovedností a rozděleni do dvou skupin podle toho zda jejich dyslektické problémy přetrvávají nebo se zlepšili. Šetření bylo provedeno u 120 žáků, z nichž 61vyzkoušelo novou metodu reedukace a 59 jich využilo metod

původních. V obou skupinách byli na konci roku žáci, kteří se ve čtenářských schopnostech zlepšili i ti, kteří se nezlepšili. Z výsledků šetření můžeme sestavit jednoduchou tabulku:

Zlepšení čtenářských dovedností

	ANO	**NE**	**Celkem žáků**
Nová metoda	A 40	B 21	61
Původní metoda	C 27	D 32	59
N			120

Z tabulky vyplývá, že při použití nové metody se z 61 žáků zlepšilo na konci roku 40 a 21 zlepšení nevykazovalo. Na druhé straně, při použití původní metody ke zlepšení došlo u 27 žáků, zatímco u 32 nikoliv. Otázkou je, zdali mezi těmito dvěma proměnnými, tedy čtenářskými dovednostmi a použitou metodou reedukace existuje souvislost nebo zda je rozdíl v počtu žáků, kteří vykazují zlepšení a kteří ne čistě náhodný.

V tomto případě použijeme test chí-kvadrát nezávislosti pro tabulku 2x2:

$$\chi^2 = \frac{n(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)}$$

Pokud do vzorce doplníme výše uvedené hodnoty, získáme výsledek chí-kvadrát testu v hodnotě 4,77. Nyní již postupujeme analogicky jako v prvním případě.

Srovnáváme tedy, zda naměřená hodnota chí-kvadrát překračuje kritickou hodnotu pro zvolenou hladinu významnosti a stupeň volnosti. Pro vyhodnocení bude opět pohodlnější využít software, který nám hodnotu chí-kvadrát nejen vypočítá, ale automaticky také srovná s kritickou hodnotou a vyhodnotí, zda se od sebe skupiny studentů na základě užité metody reedukace liší.

Výsledek testu bude v našem případě následující:
Je tedy možné uvažovat o tom, že skupiny nejsou stejné a rozdíl mezi nimi není pouze náhodný. Jak je patrné z tabulky, skupina, u které byla využita nová metoda reedukace, zahrnuje více studentů, kteří se ve čtení zlepšili. Můžeme sice usuzovat, že faktorem, který ovlivnil toto zlepšení je právě nová reedukační metoda.
Zde je však na místě velká opatrnost a prozíravost výzkumníka, neboť k větší míře zlepšení mohlo dojít i z jiných důvodů, které jsou mimo naši kontrolu. Ve skupině se např. mohla vytvořit lepší atmosféra, která vedla k větší motivaci studentů.

Způsob řešení za použití softwaru Statisitca:

Vzhledem k tomu, že se jedná o tzv. 2 x 2 tabulku, tedy každá ze dvou proměnných nabývá pouze dvou možných hodnot bude náš postup následující:

1. V horní liště klikněte na Statistiky
2. Vyberte Neparametrické statistiky
3. V nově otevřeném okně zvolte 2 x 2 tabulky
4. Do tabulky zadejte naměřené hodnoty. V našem případě 40 a 21 do prvního řádku a 27 a 32 do druhého.
5. Stiskněte Výpočet

6. V nově vytvořené tabulce najdete na šestém řádku Chí-kvadrát 4,77 a p hodnotu 0,289.

Nyní máme dostatek informací k tomu, abychom mohli říci, že skupiny se od sebe statisticky významně liší. Nemusíme ani srovnávat námi naměřenou hodnotu chí-kvadrát s kritickou hodnotou uvedenou ve statistických tabulkách. K našemu závěru nám postačí p hodnota, která je nižší než 0,05. Pokud si ještě vzpomínáte, tak právě tuto hladinu významnosti jsme si na začátku stanovili jako rozhodující. Pokud je naměřená hodnota p menší než 0,05 pak je možné usuzovat, že skupiny žáků se od sebe opravdu liší a nejedná se pouze o náhodu.

Vztahy, relace a korelace

Klíčové pojmy pro daný výklad

Vztah, relace, korelace, parciální korelace, korelační koeficient, kauzalita, zdánlivá korelace, intenzita asociace (korelace), pozitivní a negativní korelace, nekorelované proměnné, statisticky významná korelace.

Definice

Korelace - z lat. correlatio – souvislost, vzájemný vztah. Korelace, též korelační analýza, zkoumá vzájemný vztah kvantitativních proměnných. (U nominálních popřípadě u obecně kategoriálních dat mluvíme o asociaci.) Dvě proměnné jsou korelované, když určité hodnoty jedné proměnné mají tendenci vyskytovat se společně s hodnotami druhé proměnné. Míra tohoto vztahu může nabývat intervalu od neexistence vztahu mezi proměnnými až po absolutní vztah – korelaci. Míru korelačního vztahu vyjadřují korelační koeficienty.

Korelační koeficient vyjadřuje lineární vztah dvou proměnných. Míru intenzity vztahu dvou náhodných spojitých proměnných X a Y můžeme např. vyjádřit prostřednictvím Pearsonova korelačního koeficientu *r*, který se nejčastěji používá k měření lineární závislosti mezi kvantitativními proměnnými. Pearsonův koeficient korelace (někdy též koeficient pořadové korelace) nabývá hodnot z intervalu (-1, 1). Záporné hodnoty vyjadřují nepřímou závislost (jedna veličina roste a druhá klesá, nebo naopak), kladné hodnoty značí přímou závislost (obě veličiny proměnné se pohybují stejným směrem). Jestliže je $|r| = 1$ můžeme vztah proměnných interpretovat jako silnou závislost, při vynesení této korelace na graf by všechny body ležely na jedné přímce. Naopak koeficient $|r| = 0$ vyjadřuje úplnou lineární nezávislost proměnných, které také nazýváme nekorelované proměnné.

Úvod

Smyslem mnoha výzkumných kvantitativních šetření je hledání (popis) nebo ověření vztahů mezi proměnnými. Abychom toto konstatování mohli správně pochopit, je třeba zdůraznit, že ve vědeckém světě dbáme na exaktnost, a nemůžeme se tudíž například při pohledu na teploměr smířit s výrokem: „Dnes je opravdu teplo.“ Toto konstatování sice vychází z přesného odečtení výšky rtuťového sloupce na exaktně kalibrované stupnici, ale postrádáme zcela prostou informaci o ročním období, případně zeměpisné poloze. Zastaví-li se hladina rtuťového sloupce na pěti stupních Celsia v období ledna, budeme tuto informaci vnímat jinak, než když stejné zjištění učiníme v červenci. V druhém případě pak můžeme daný výrok označit za ironický. Z příkladu je zřejmé, že se ani v obyčejném reálném a „nevědeckém“ světě neobejdeme bez uvedení a definování přesnějších kontextů, které se k informaci o teplotě vztahují. Kerlinger (1972, s. 89) v této souvislosti hovoří o „vztahové podstatě lidského vědění.“ Ve vědeckém světě neusilujeme pouze o deskripci, explanaci, predikci, definici, nebo třídění jevů, ale též o jejich uvádění do vztahů s jevy jinými. Abychom s určitou přesností a jistou mírou přesvědčivosti mohli o někom nebo o něčem říci, že je velké, mále, tvrdé, měkké, rychlé nebo pomalé, musíme předmět našeho zajmu uvést do vztahu. O problematice určitého vztahu, intenzity (míry) vztahu a polaritě (zda-li je pozitivní, nebo negativní) se můžeme dovědět právě prostřednictvím tzv. korelačního koeficientu, ke kterému ve výkladu postupně dospějeme. Nejdříve se budeme věnovat jednoduchým lineárním vztahům dvou proměnných, později se zmíníme též o analýze tzv. vícerozměrných dat.

Zajímavost

Jak uvádí Eileen Magnello ve své populárně naučné publikaci (2010, s. 115), „pojem korelace se používal téměř o sto let dříve, než se našel způsob, jak ji měřit. O korelaci se poprvé zmiňuje biolog Comte de Buffo (1707-1788) a dále paleontolog baron Georges Cuvier (1769-1832), který v roce 1801 poprvé užívá výraz „korelace částí.“ Jak vlastně o korelaci uvažovali? Domnívali se totiž, že

organismy existují jako specificky organizované celky, a tudíž, např. při rekonstrukci určitého živočicha, můžeme vycházet ze specifického druhu kostí apod., které se váží (mají vztah) ke konkrétnímu druhu. Korelace se však jednoznačně pojí se jménem Carla Pearsona (1857-1936), který je považován za zakladatele matematické statistiky. Prvním tvůrcem metody korelace se však podle Magnello (s. 117) stal Francis Galon (1822 - 1911), „když vytvořil graf zjišťující vztah mezi mateřskou a dceřinou rostlinou hrachoru vonného. Do doby, než Galon přišel s představou korelace, byla kauzalita základním způsobem, jímž se vysvětloval vztah mezi dvěma vzájemně souvisejícími jevy, zejména v oblasti přírodních věd. Než se Pearson setkal s Galonem, domníval se, že formální matematika se může uplatnit jen v oblasti přírodních jevů způsobených kauzalitou. Později u něj tuto původní představu kauzality nahradily Galtonovy myšlenky. (…) Stal se zapáleným antikauzalistou, přesvědčeným o tom, že vesmír není ovládán zákonem kauzality v jeho nejužším smyslu, ale větší roli při vysvětlování jevů hraje variabilita.“ Jak uvedeme dále, právě variabilita se stala jednou z proměnných v matematickém vztahu pro výpočet korelace.

Jak je zřejmé z uvedeného, nemůžeme zaměňovat korelaci s kauzalitou, tedy skutečností, že když dvě proměnné jsou ve vzájemném vztahu, jedna má tendenci se vyskytovat společně s druhou, (např. absolventi školy s vyšším vzděláním mají vyšší plat), nemusí to nutně znamenat, že příčinou vyšších platů ve společnosti je vyšší vzdělání. Vyšší příjem může být způsoben druhem absolvovaného oboru na dané vysoké škole (např. absolventi fakult informatiky jednoznačně vedou nad absolventy fakult připravujících učitele) nebo též prostou pracovitostí, motivací apod. O takových korelacích říkáme, že jsou **zdánlivé**. Např.: Může existovat silný korelační vztah mezi velikostí nohy a inteligencí?

Princip

Abychom se zodpovědně mohli vyjádřit o existenci či neexistenci, popřípadě určité míře vztahu, např. mezi. sociálním statusem rodičů a dosaženým

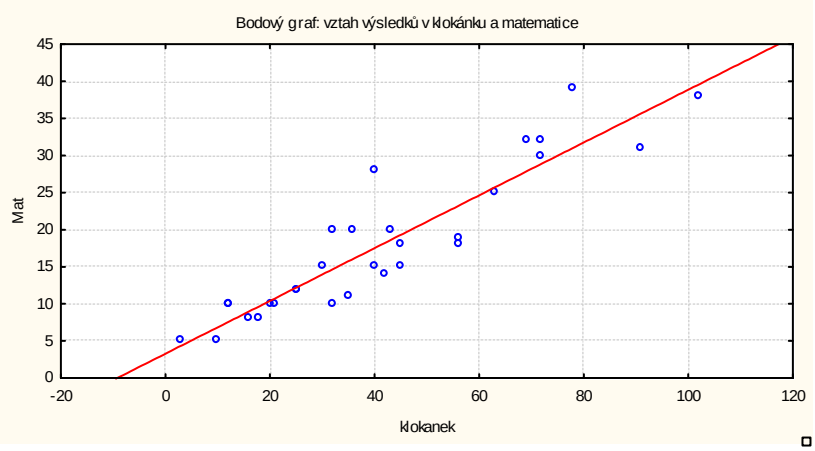
vzděláním jejich dětí, nemůžeme tak učinit na základě případové studie jedné rodiny, ale potřebujeme větší počet měření (obsáhlejší výběrový soubor) rodin (dosažené vzdělání rodičů) a jejich dětí (dosažený stupeň vzdělání). Předmětem porovnání pak bude např. úroveň vzdělání Novákových s úrovní vzdělání jejich dětí – mladých Novákových. Porovnáváme tedy určité uspořádané dvojice nebo-li svazky, které můžeme označit za relace (česky vztahy).

Uveďme i jiné příklady relace, které mohou být ještě názornější. Podívejme se na následující tabulku (z třiceti žáků ve třídě jsme vybrali jen prvních deset).

	Klok.	**Mat.**	**Čj.**
1	18	8	16
2	102	38	12
3	43	20	10
4	45	18	12
5	12	10	8
6	72	32	20
7	30	15	38
8	69	32	38
9	40	28	35
10	25	12	20

Děti v jedné z pátých tříd se zúčastnili mezinárodní soutěže Klokan v kategorii Klokánek (výsledky jsou v prvním sloupci – Klok.), zároveň napsali pololetní práci z matematiky a českého jazyka (druhý a třetí sloupec). Z tabulky je patrné, že žákovi č. 1 je přiřazen právě jeden výsledek z testu (Klokánku) a právě jeden z výsledků matematiky a českého jazyka. Také bychom mohli říct, že každému prvku množiny prvků Klok. odpovídá právě jeden prvek množiny Mat. a Čj. Relaci (vztah) tedy můžeme definovat jako množinu uspořádaných dvojic,

tedy prvků s přesně určeným pořadím. Pokusme se tento vztah zobrazit prostřednictvím jednoduchého grafu.



Z grafu je patrné, že každý bod je definován v souřadnicovém systému na ose x a y. Přímka napovídá, že se jedná o lineární vztah proměnných (jedna proměnná má tendenci se měnit společně s proměnnou druhou). Ve vlastní výzkumné praxi se může jednat o různé druhy relací. Např. mezi výší platu a sportovní aktivitou provozovanou ve volném čase, oblibou a množstvím kamarádů (kamarádek) ve skupině, náboženstvím a hodnotovou orientací, výškou a váhou apod. Jedno je však zřejmé. Vytváření relací by nám vždy mělo dávat určitý věcný smysl. Graficky můžeme například znázornit relace mezi délkou vlasů a inteligencí, nabízí se však otázka, zda-li je to smysluplný a pro naši věcnou interpretaci použitelný vztah.

Abychom si představili různé druhy relací, využijme vysvětlení, které nám podává Kerlinger (s. 96 – 97). Pokuste se sestrojit jednoduché grafy (náčrtky) podle následujících tabulek uspořádaných dvojic. Načrtněte osy x a y a vyneste na ně body, a ty se pokuste propojit přímkou.

X	Y
1	1
2	2
3	3
4	4
5	5

X	Y
1	2
2	5
3	3
4	1
5	4

X	Y
1	5
2	4
3	3
4	2
5	1

A B C

Jaký je náš výsledek? Jak vypadají relace u jednotlivých grafů?

Kerlinger celou problematiku shrnuje následovně (s. 97). „Součinový moment a patřičné korelační koeficienty jsou založeny na souběžné variabilitě prvků v množinách uspořádaných dvojic.“ Jestliže se sdružené hodnoty mění ve vzájemné závislosti – prvky jedné proměnné se zvyšují společně s druhou proměnnou, pak tento vztah (korelaci) popisujeme jako pozitivní (korelační koeficient nabývá hodnoty od 0 do 1). Jestliže hodnoty jedné proměnné se snižují a zároveň hodnoty druhé proměnné se zvyšují, pak hovoříme o negativním vztahu (korelační koeficient nabývá hodnoty od 0 do -1 (případ B). A konečně (případ C), pokud hodnoty jedné proměnné nemají žádný vztah (společně se nezvyšují ani

nesnižují) s druhou proměnnou, pak říkáme, že se jedná o proměnné nekorelované (koeficient se pohybuje kolem 0).

Výpočet

Jestliže jsme pochopili princip korelace, podívejme se nyní na vztah, který nám popisuje výpočet korelačního koeficientu. Výše jsme zmínili jméno slavného zakladatele matematické statistiky Pearsona, a tudíž princip výpočtu budeme demonstrovat prostřednictvím jeho koeficientu. Dobře a přehledně celý postup popisuje Chráska (2007, s. 114 – 115). Vychází nejdříve z obecné definice Pearsonova koeficientu (r_p), který je charakterizován jako poměr kovariance (v podstatě variance jedné a druhé proměnné) a součinu směrodatných odchylek obou proměnných.

Výpočet kovariance

$$Kovariace = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Výpočet korelace

$$r_p = \frac{Kovariace}{s_x \cdot s_y}$$

Celkový výpočet korelace po úpravě (x, y – jednotlivé dvojice hodnot proměnných; n – počet srovnávaných dvojic hodnot)

$$r_p = \frac{n \cdot \sum xy - \sum x \cdot \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2] \cdot [n\sum y^2 - (\sum y)^2]}}$$

Výpočet prostřednictvím programu Statistica

Nyní se podívejme, jak můžeme provést výpočet korelačního koeficientu dvou kvantitativních proměnných prostřednictvím programu Statistica. Budeme pracovat se souborem kolokanek.sta. (Otevření souboru: Soubor – otevřít – klokanek.sta). Vidíme, že výběrový soubor obsahuje 30 žáků (n=30). Řádky tedy prezentují jednotlivé případy (jména žáků jsou nahrazena číslem) a sloupce proměnné. U této třídy žáků byly prostřednictvím testu změřeny tři proměnné, dosažené body v testu soutěže Klokan, dále z matematiky a českého jazyka. Jedná se o kvantitativní – metrické proměnné. Naším úkolem je zjistit, zda-li spolu tyto proměnné korelují.

Postup: Statistiky – základní statistiky/tabulky – korelační matice – ok. Otevře se nám dialog, vybereme všechny tři proměnné (Crtl + levé tlačítko myši) – ok - výpočet. Vygeneruje se korelační matice (viz obr.), ze které je patné, jaký je korelační vztah vždy mezi konkrétními dvěmi proměnnými.

	Korelace (Klokanek.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=30 (Celé případy vynechány u ChD)				
Proměnná	Průměry	Sm.odch.	klokanek	Mat	Čj
klokanek	41,36667	25,05371	1,000000	0,918517	0,459105
Mat	18,00000	9,70247	0,918517	1,000000	0,535385
Čj	18,93333	10,64776	0,459105	0,535385	1,000000

Nyní se můžeme pokusit o interpretaci. Červené hodnoty program zobrazuje tehdy, jsou-li statisticky významné. Můžeme tedy říci, že existuje statisticky významný vztah (korelace) mezi výsledky v matematice a testu klokánka, ale též mezi matematikou a českým jazykem atd. Dále můžeme interpretovat intenzitu vztahu (asociaci) a směr. Je zřejmé, že hodnoty jsou kladné, a tudíž je patrné, že zvyšuje-li se jedna hodnota proměnné, např. výsledky testu klokánek, zvyšuje se i hodnota druhé proměnné (výsledky v matematice). Abychom mohli interpretovat intenzitu vztahu, podívejme se např. na tabulku,

kterou uvádí Hendl (2004, s. 246). V jiných publikacích ale nalezneme i další druhy členění míry asociace (vztahu).

Síla asociace	r
malá	0,1 - 0,3
střední	0,3 - 0,7
velká	0,7 - 1

Následně bychom korelaci proměnné Klokánek a Mat. mohli interpretovat jako velkou, mezi klokánkem a Čj. jako střední apod. Důležitá je však věcná (obsahová) interpretace. Co jsme vlastně prostřednictvím výpočtu zjistili? Na základě výsledků bychom mohli říci, že existuje statisticky významný vztah mezi výsledky testu klokánek a výsledky testu v matematice, ale též významný vztah mezi testem klokánek a výsledky v českém jazyce, přestože intenzita je zde menší (dle tabulky střední).

Nabízí se však obecná otázka, zda-li výsledky nereprezentují obecnou inteligenci žáků a zda-li opravdu existuje souvislost mezi výsledky žáků např. v matematice a českém jazyce? Tento předpoklad si můžeme jednoduše ověřit, vyloučíme-li z našeho výpočtu výsledky testu klokánek a zkusíme korelovat

pouze proměnné Mat. a Čj. Tento postup označujeme jako výpočet **parciální korelace**.

Postup: Statistiky – základní statistiky/tabulky – korelační matice – ok – parciální korelace. V levém sloupci vybereme pouze proměnné, které chceme korelovat (Mat. a ČJ.) a v druhém vyloučíme proměnnou klokánek. Statistika vypočte korelace výhradně z těchto dvou proměnných.

	Parciální korelace (Klokanek.sta) Controlling pro: klokanek Označ. korelace jsou významné na hlad. p < ,05000 N=30 (Celé případy vynechány u ChD)			
Proměnná	Průměry	Sm.odch.	Mat	Čj
Mat	18,00000	9,70247	1,000000	0,323670
Čj	18,93333	10,64776	0,323670	1,000000

Vidíme, že vztah mezi proměnnými není statisticky významný a jejich intenzita je menší. Tuto skutečnost můžeme interpretovat tak, že obecná inteligence je předpokladem vztahu mezi proměnnými a že žáci, kteří dosahují dobrých výsledků v klokánku, dosahují i lepších výsledků v matematice a českém jazyce. Existuje však specifické nadání pro matematiku a český jazyk. Někteří žáci, kteří jsou úspěšní v matematice, mohou (ale nemusí) být úspěšní i v českém jazyce. Neprokázali jsme tedy vztah mezi těmito proměnnými.

Korelace (asociace) pořadových proměnných

Ve výzkumných designech nepracujeme pouze s daty metrickými, ale též s proměnnými, které jsou definovány pouze svým pořadím (tzv. ordinální nebo též pořadová data). Takových proměnných je v pedagogickém a obecně sociálně vědním výzkumu celá řada. Jsou to např. odpovědi respondentů na škálách (např. na pořadovém kontinuu od absolutního souhlasu po nesouhlas), hodnocení výsledků (dobré, průměrné a podprůměrné), hodnocení subjektivních pocitů (cítím se dobře, jako obvykle, špatně), řazení učitelů dle oblíbenosti (velice

oblíbený, oblíbený, neutrální, neoblíbený, velice neoblíbený) apod. Škály přitom mohou být třístupňové, ale též vícebodové. Může se též jednat např. o uspořádání metrických dat. Např. počet bodů v testu z matematiky můžeme uspořádat do pořadí od nejlepší po nejhorší výsledek. Z dat metrických vzniknou data pořadová.

Celý postup si vysvětlíme na příkladu dvou položek dotazníku locus of kontrol (více k nástroji např. Němec 2010). Zajímá nás vztah mezi odpověďmi žáků na dvě položky dotazníku. Žáci ke každé položce odpovídali na pořadovém kontinuu: rozhodně souhlasím (4), spíše souhlasím (3), spíše nesouhlasím (2), rozhodně nesouhlasím (1).

1) Myslíš si, že když se budeš učit, můžeš dosáhnout lepších výsledků?
(4 – 3 – 2 - 1)
2) Myslíš si, že můžeš předejít nachlazení? (4 – 3 – 2 - 1)

Výpočet prostřednictvím programu Statistica

Postup: I nadále budeme pracovat se souborem klokanek.sta. Nebudeme však používat klasický výpočet korelace, ale použijeme tzv. Spearmanův test pořadové korelace, který v daném programu nalezneme v tzv. neparametrické statistice. Statistiky – neparametrické statistiky – korelace (Spearman, Kendaullovo tau, gama). Postup je analogický jako v předchozím případě. Vybereme proměnné a stiskneme tlačítko Spearmanův koef. R. Opět se nám vygeneruje korelační matice. Významnost a intenzitu vztahu interpretujeme podobně jako v předchozím zadání. V našem případě jsme zjistili, že vztah je významný a jeho intenzita je 0,74. Můžeme tedy říct, že mezi hodnocením dvou položek dotazníku existuje silný vztah. Jinými slovy, že žáci, kteří odpověděli na první otázku určitým způsobem, odpověděli podobně i na druhou otázku. Reálná interpretace v rámci celého nástroje by však byla složitější.

Regresní analýza

Klíčové pojmy pro daný výklad

Regresní přímka, metoda nejmenších čtverců, regresní koeficient, regresní model, regresor, vysvětlovaná a vysvětlující proměnná.

Definice

Regrese – (z lat. regredere – jít zpět), též regresní analýza. Analýza statistických dat založená na matematicko-statistických modelech regresních rovnic. Tradiční název této analýzy je nepřesný, protože se využívá spíše k predikci. Regresní analýzu používáme tehdy, chceme-li zjistit závislost určité kvantitativní proměnné na jedné nebo více kvantitativních proměnných (regresorech). Jestliže jsme předem definovali nezávisle proměnnou (vysvětlující – např. inteligenci) a závisle proměnnou (např. průměrný prospěch ve škole), pak u **r. a.** předpokládáme pouze jednostrannou závislost, tzn. že vysvětlovaná proměnná zpětně neovlivňuje nezávisle proměnnou (na rozdíl např. od korelační analýzy). Pomocí **r. a.** se snažíme objasnit vztah mezi proměnnými X a Y a s určitou mírou pravděpodobnosti i předpověď, zda-li se změní závisle proměnná při změně nezávisle proměnné.

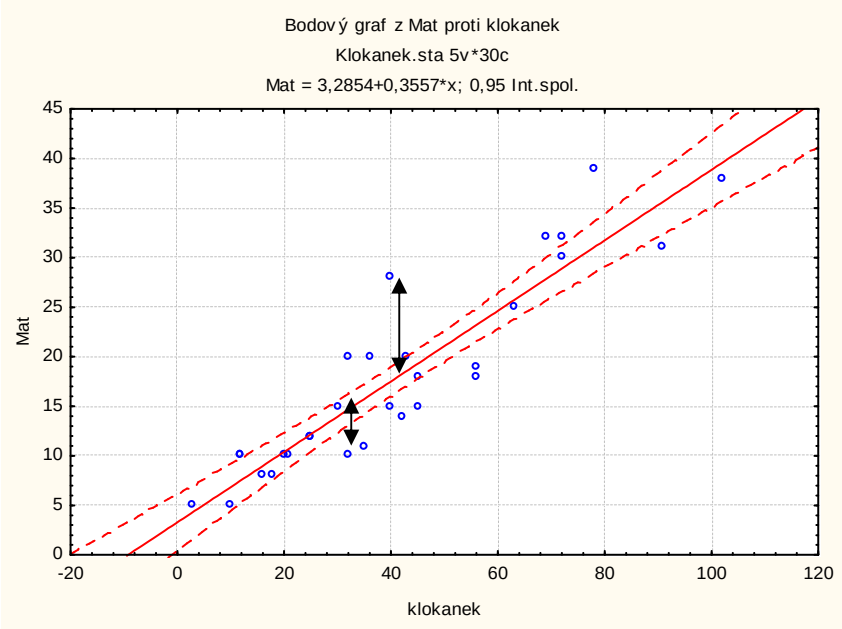
V předchozím případě (korelací) jsme si kladli otázku, zda-li existuje vztah mezi dvěma proměnnými. V této situace si budeme klást otázku, zda-li se změnou hodnot jedné proměnné (tzv. nezávisle proměnné) se budou měnit také hodnoty proměnné druhé (závislé nebo též vysvětlované proměnné). Mluvíme tedy o kauzálních (příčinných souvislostech). Např. nás může zajímat, zda-li obecné inteligenční předpoklady (nezávisle proměnná) mají vliv na výsledky testů v matematice nebo českém jazyce (závisle proměnná)? Nebo, zda-li z výšky člověka můžeme odhadovat jeho hmotnost? Z uvedených příkladů je patrné, že k řešení podobných situaci budeme potřebovat vždy dvě nebo i více

kvantitativních (metrických) proměnných (v případě vícerozměrné regrese), přičemž je dáno, která z proměnných je nezávislá (x), tzv. vysvětlující (prostřednictvím ní vysvětlujeme chování závisle proměnné), a závisle proměnná (y), tzv. vysvětlovaná (její chování je determinováno vlivem jedné nebo více nezávisle proměnných).

Chráska (2007, s. 113) popisuje smysl regresní analýzy následovně: „Úkolem regresní analýzy je nalézt regresní funkci, pomocí níž lze ze známých hodnot nezávisle proměnné určit příslušné hodnoty závisle proměnné. Nejjednodušší situace nastává v případě, že mezi proměnnými je lineární statistická závislost. Příkladem je např. vztah mezi věkem učitelů na středních školách (x) a jejich průměrnou mzdou (y). V tomto případě je mzda učitelů funkcí věku učitelů. Čím vyšší je věk učitelů, tím vyšší je jejich průměrná mzda. Platí, že: $y = f(x) = a + bx$, kde výrazy a, b jsou konstanty, jejich skutečnou velikost neznáme, ale můžeme ji odhadnout ze zjištěných empirických dat. Konstanta b se nazývá koeficient regrese a udává, o kolik se změní (průměrně) hodnota závisle proměnné y, jestliže hodnota nezávisle proměnné x se zvětší o jednotku. Známe-li obě konstanty (a,b), můžeme pro kteroukoliv hodnotu nezávisle proměnné x (např. věk učitelů) vypočítat hodnotu závisle proměnné.“

Podobně jako v předchozím případě i nyní budeme vycházet z datového souboru (tabulky), který je typický tím, že každý respondent (prvek) je charakterizován vždy dvěma nebo více měřeními, která se k němu vztahují. Podívejme se nyní znovu na vztah výsledku testů v klokánkovi (nezávisle proměnná) a matematiky (závisle proměnná). Z předchozího grafu je patrné, že mezi proměnnými existuje lineární vztah. O zachycení tohoto vztahu se pokusíme prostřednictvím tzv. regresní přímky, tj. přímky, která by se přiblížila (aproximovala) všem bodů (v ideálním případě by je všechny protnula). Jedná se o přímku, která by co nejlépe predikovala hodnoty proměnné y prostřednictvím x. Jak uvádí Henld (2004, 268) „Rozdílu mezi naměřenou a predikovanou hodnotou říkáme reziduální hodnota predikce nebo chyba predikace a značíme ji symbolem

e. Dobře proložená přímka (regresní přímka) minimalizuje velikost tzv. reziduálních hodnot pro hodnoty (x_i; y_i), kterými přímku prokládáme. Nejčastěji se používá tzv. metoda nejmenších čtverců (viz šipky v grafu). Hodnoty paramentrů a, b (viz výše) přímky y = bx získáme metodou nejmenších čtverců tak, aby byl minimální součet druhých mocnin reziduálních hodnot e.“



Oprostíme-li náš výklad o interpretaci dalších matematických postupů, pak můžeme konstatovat, že na základě sestrojené regresní přímky (funkce) pak můžeme predikovat hodnoty proměnné y z hodnot x. Prakticky si to nyní ukážeme v programu statistica.

Postup: Statistiky – vícerozměrná regrese – proměnné. Zaklikneme jako závisle proměnnou Mat a nezávisle proměnnou klokanek. Stiskneme OK a dále výpočet regrese. Vygeneruje se následující tabulka.

N=30	Výsledky regrese se závislou proměnnou : Mat (Klokanek.sta) R= ,91851735 R2= ,84367412 Upravené R2= ,83809105 F(1,28)=151,11 p<,00000 Směrod. chyba odhadu : 3,9041					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(28)	p-hodn.
Abs.člen			3,285411	1,393158	2,35825	0,025577
klokanek	0,918517	0,074720	0,355711	0,028937	12,29280	0,000000

Potvrdila se nám významnost absolutního členu, ale i odhad koeficientu závislosti matematiky na výsledku testu klokánek. Dobrým ukazatelem pro interpretaci regresního modelu je koeficient determinace $\mathbf{R^2}$. V našem případě je velký 0,84. Tuto skutečnost můžeme interpretovat tak, že model, který jsme vytvořili, vysvětluje přibližně 84% variability proměnné. Jinými slovy, výsledky testu v matematice můžeme vysvětlit prostřednictvím dosažených výsledků v testu klokánek. O silném vztahu jsme se již přesvědčili v rámci korelací, nyní obě proměnné můžeme dát do příčinných souvislostí. V praxi se často nesetkáváme s tak velkým vysvětleným procentem variability. Pokud jsme uvedli, že regresní model nám umožňuje předpovídat chování závisle proměnné podle nezávisle proměnné, podívejme se nyní na záložku residua/předpoklady/předpovědi a stiskneme předpověď závisle proměnné. Pokud vložíme jakoukoliv hodnotu výsledků testu v klokánku, dostaneme pravděpodobnou hodnotu výsledku z matematiky.

Úkol: Podobně zkuste analyzovat vztah mezi proměnnou kolokanek a Čj. Stejnou analýzu pak proveďte i na souboru vahy.sta. Zkuste předpovědět na základě výšky (nezávisle proměnná) váhu a posléze do daného modelu ještě přidejte jako druhou nezávisle proměnnou věk. Jak můžeme interpretovat daný regresní model?

Použitá a doporučená literatura:

DISMAN, M. *Jak se vyrábí sociologická znalost*. Praha : Karolinum, 2000. ISBN 80-246-0139-7.

HENDL, J. *Přehled statistických meto zpracování dat*. Praha : Portál, 2004. ISBN 80-7178-820-1.

CHRÁSKA, M. *Metody pedagogického výzkumu*. Praha : Grada, 2007. ISBN 978-80247-1369-4.

KERLINGER, F. N. *Základy výzkumu chování*. Praha : Academia, 1972.

MAGNELLO, E., Van LOON B. *Statistika*. Praha Portál, 2010. ISBN 978-80-7367-753-4.

ŠKALOUDOVÁ, A. *Statistika v pedagogickém a psychologickém výzkumu*. Praha: PdF UK, 1998. ISBN 80-86039-56-0.