# Intuitivní Výklad Základních Statistických Metod

*Pavel Osecký* [1]

Když přednášíte statistické metody pro humanitní předměty, cítíte se před velkou částí studentů jako člověk předvádějící album slavných symfonií odpůrcům orchestrální hudby. Přesto byste však měli dovést posluchače k porozumění elementárním procedurám popisné i induktivní statistiky, jejich napojení na metodologii příslušného humanitního oboru a jejich vyústění směrem k inteligentnímu užívání výpočetních programů. Při tom by váš výklad neměl být pouhým vyprávěním o statistice a vůbec ne odchovem tlačítkových idiotů. — Obsah tohoto článku je založen na zkušenostech s posluchači psychologie na Masarykově univerzitě, kde jsou základnímu kursu statistiky věnovány po dva semestry tři týdenní hodiny, rozdělené na dvě hodiny přednášek prokládaných občasným cvičením u tabule, a na jednu hodinu v počítačové učebně, navazující velmi volně na přednášku.

*Něčemu z matematiky se nelze vyhnout:* Z výpočetních záležitostí je to zacházení se sumačním symbolem, zavedení pojmu matice, sloupcového a řádkového vektoru, podobně jako kontrola dovedností počítat s desetinnou čárkou. Z teoretických potřeb je to především zavedení pojmu univerza (základního prostoru) jako množiny apriori možných výsledků náhodného pokusu, jevů (slovensky udalostí) jako všech nebo některých podmnožin univerza, náhodných proměnných (náhodných veličin) jako reálných funkcí na univerzu a náhodných vektorů, které opět vyžadují, aby se student zbavil strachu z pojmu vícerozměrného prostoru. Při tom se náhodný pokus chápe v nejširším slova smyslu nejen jako klasický laboratorní experiment, ale i jako soustavné klinické či jiné pozorování, nebo opakovatelná operace, ať výrobní či jiná. Ostatní matematické problémy, jako třeba měřitelnost náhodných proměnných a tím spíše užívání nekonečných řad a integrálů, je nutno odsunout stranou: užijeme např. symbolu integrace, ale odvoláváme se na jakoby samozřejmý pojem plošného obsahu či prostorového objemu apod.

*Intuitivnost výkladu* se může založit na opětovném využívání empirického zákona velkých čísel. Skutečnost, že za podmínek stochastické stability se při mnohonásobném nezávislém opakování náhodného pokusu relativní četnost sledovaného jevu zpravidla postupně ustaluje kolem určité konstanty (zvané pravděpodobností onoho jevu), vyjádříme zkratkovitě zápisem

$$n \to \infty \Rightarrow \#_n(X \in I)/n \rightsquigarrow P(X \in I)$$

kde $n$ je počet nezávislých opakování náhodného pokusu, $X$ vyjadřuje náhodnou proměnnou, $I$ danou číselnou množinu, $\#_n(X \in I)$ absolutní četnost výskytů sledovaného jevu a $P(X \in I)$ jeho pravděpodobnost. Při tom "jevu $X \in I$" jsou příznivé právě ty z možných výsledků náhodného pokusu, při nichž se náhodná proměnná $X$ realizuje v dané číselné množině $I$. Zvlněná šipka $\rightsquigarrow$ volně symbolizuje ono ustalování, zatím co rovná $\to$ vyjadřuje neomezený růst opakování náhodného pokusu. Představa pravděpodobnosti jako zidealizované relativní četnosti by měla posluchačům nahradit axiomatickou definici; je ovšem nutné tuto nepřesnost jasně přiznat. — Obdobným způsobem se můžeme při intuitivním výkladu statistiky vyhnout obvyklému soustavnému probírání počtu pravděpodobnosti: v rámci popisné

[1] Masarykova univerzita v Brně.

statistiky nejprve zavedeme empirické charakteristiky a pak přejdeme prostřednictvím empirického zákona velkých čísel k jejich teoretickým protějškům bez důrazu na jejich logickou výstavbu.

*Funkcionální charakteristiky* budeme zavádět ve shodě s bodovým a intervalovým zpracováním datového souboru

$$\begin{bmatrix} x_1 & y_1 & \dots & z_1 \\ x_2 & y_2 & \dots & z_2 \\ \vdots & \vdots & & \vdots \\ x_n & y_n & \dots & z_n \end{bmatrix}$$

— V případě bodového zpracování dat, které respektuje každou jednotlivou naměřenou hodnotu, vyjdeme z četnostní funkce $p(x)$ udávající, s jakou relativní četností se odpovídající hodnoty $x$ vyskytují v datovém souboru. Při zvětšování rozsahu $n$ se tyto relativní četnosti ustalují podle empirického zákona velkých čísel kolem hodnot $\pi(x)$ pravděpodobnostní funkce:

$$n \to \infty \Rightarrow p(x) \rightsquigarrow \pi(x)$$

Tato funkce $\pi(x)$ je podobně jako $p(x)$ nezáporná a součet jejích kladných hodnot musí dát jedničku. V analogii s relativními četnostmi položíme

$$P(X \in I) = \sum_I \pi(x)$$

čímž dostáváme formální definici pro diskrétní náhodnou proměnnou a pravděpodobnosti s ní spojené. (Sumační symbol je třeba vhodně definovat.)

— V případě intervalového zpracování dat rozložíme číselnou osu na třídicí intervaly a uvnitř téhož intervalu už mezi jednotlivými údaji nerozlišujeme. Vyjdeme z histogramu, který však na rozdíl od standardních statistických programů bude připouštět nestejně dlouhé třídicí intervaly a třídní relativní četnosti bude vyjadřovat ne výškou, ale plošným obsahem svých jednotlivých obdélníků. Samotná výška tedy bude v každém třídicím intervalu rovna podílu třídní relativní četnosti a délky intervalu a budeme ji považovat za graf funkce $f(x)$ zvané hustota četnosti. Při zvětšování rozsahu $n$ za současného zmenšování maximální délky $\delta$ třídicích intervalů si můžeme podle empirického zákona velkých čísel představit, že původní schodovitý graf ohraničující shora histogram přejde ve víceméně hladký graf funkce $\varphi(x)$ zvané hustota pravděpodobnosti:

$$n \to \infty \wedge \delta \to 0 \Rightarrow f(x) \rightsquigarrow \varphi(x)$$

Tato funkce je podobně jako $f(x)$ nezáporná a ohraničuje spolu s číselnou osou oblast o plošném obsahu rovném jedničce. V analogii s histogramem položíme

$$P(X \in I) = \int_I \varphi(x)\,dx$$

čímž dostáváme formální definici pro spojitou náhodnou proměnnou a pravděpodobnosti s ní spojené.

— Obdobně lze postupovat u diskrétních popř. spojitých náhodných vektorů a zavádět pravděpodobnosti typu

$$P(X \in I \wedge Y \in J \wedge \dots \wedge Z \in K)$$

V jednorozměrném a dvojrozměrném případě můžeme předešlé myšlenky vyjádřit snadno pochopitelnými obrázky, obdobnými situaci v popisné statistice.
— Při probírání kontingenčních tabulek je vhodná příležitost pro definici neslučitelných jevů vedoucí ke sčítání pravděpodobností a definici stochasticky nezávislých náhodných proměnných spojenou s násobením pravděpodobností. Zvláště názorné je sledování kontingenčních tabulek s podmíněnými relativními četnostmi: stochastická nezávislost se tu projevuje tendencí k rovnosti všech řádků popř. všech sloupců.

*Číselné charakteristiky* budeme zavádět ve shodě s tříděním proměnných podle stupně kvantifikace na nominální, ordinální, intervalové a poměrové, přičemž alternativním (dichotomickým) proměnným může být vyhrazeno zvláštní postavení. Např. intervalovému kvantifikačnímu stupni odpovídá aritmetický průměr $m$ jako empirická charakteristika polohy hodnot sledované proměnné, směrodatná odchylka $s$ pak vyjadřuje variabilitu těchto hodnot a koeficient korelace $r$ těsnost lineárního homoskedastického vztahu mezi dvěma proměnnými. S odvoláním na empirický zákon velkých čísel se pak zavádí např. střední hodnota $\mu$ jako teoretická charakteristika polohy pomocí už zmíněného schématu

$$n \to \infty \Rightarrow m \rightsquigarrow \mu$$

Protože by vážnější zájemce o statistiku mohl namítnout, že z této intuitivní představy není možno odvozovat žádné matematické důsledky, můžeme okrajově připojit i formální definici střední hodnoty pro diskrétní náhodné proměnné

$$\mu = \sum_{-\infty}^{\infty} x\, \pi(x)$$

představující zidealizovaný případ aritmetického průměru, který se dá vyjádřit pomocí četnostní funkce ve tvaru

$$m = \sum_{-\infty}^{\infty} x\, p(x)$$

Poněkud obtížněji je možno dospět k formální definici střední hodnoty pro spojitou náhodnou proměnnou:

$$\mu = \int_{-\infty}^{\infty} x\, \varphi(x)\, dx$$

— Vzorce pro aritmetický průměr, empirickou směrodatnou odchylku apod. by si měli posluchači pamatovat, ale to má smysl jen tehdy, když jim rozumí. Je tedy i při intuitivním způsobu výkladu zapotřebí věnovat nějaký čas jejich motivaci. Uvede-li pak student při zkoušce vzorec pro směrodatnou odchylku

$$s = \sqrt{\tfrac{1}{n-1} \textstyle\sum_{i=1}^{n} (x_i - m)^2}$$

bude zahrnut otázkami: Proč jsou ve vzorci původní data $x_i$ centrována? Co by se stalo, kdybychom z centrovaných hodnot odstranili druhou mocninu? K čemu odmocnina? Máte aspoň představu, proč bylo při průměrování užito korekce $n - 1$?
— U číselných charakteristik je třeba zdůrazňovat, že vystihují třeba jen polohu nebo jen variabilitu sledovaných hodnot, a že něco nevystihují vůbec: např. aritmetický průměr není žádnou "typickou" hodnotou, koeficient korelace sám o sobě nevypovídá nic o kauzalitě pozorovaných procesů apod.
— Každá číselná charakteristika je primárně určena pro některý z uvedených stupňů kvantifikace. Je nutno připomenout pravidlo, že každé charakteristiky je sice přípustné použít i při vyšším stupni kvantifikace (možná se ztrátou informace), ne

však na nižším (zde se dodává informace falešná). Teorie je ovšem šedá a strom života se zelená: nikdo sice nebude počítat aritmetický průměr z čísel brněnských tramvajových tratí, ale učitel počítající průměrný prospěch své třídy v matematice si možná ani neuvědomí, že nedovoleně užívá intervalové statistické charakteristiky u dat na pouhém ordinálnímu stupni kvantifikace. Není třeba studentům zatajovat, že při vší kritice statistikové nedovedou zkoušejícím učitelům nabídnout místo průměrů adekvátní a stejně praktické charakteristiky.

*Výklad induktivní statistiky* především vyžaduje zdůraznění rozdílu mezi teoretickými a empirickými charakteristikami. Teoretické představují neznámé konstanty nebo funkce, empirické je možno podle kontextu chápat dvojím způsobem: v obecných úvahách jako nové náhodné proměnné odvozené z pozorování, při numerických výpočtech jako čísla, jimiž se zmíněné náhodné proměnné realizují. I při intuitivním výkladu pak můžeme jednotlivé induktivní statistické procedury vyjadřovat schematickými zápisy:
— Při stanovení intervalu spolehlivosti je to zápis

$$P(d \leq \tau \leq h) \geq 1 - \alpha$$

v němž $\tau$ znamená neznámou hodnotu teoretické charakteristiky nebo obecněji obdobný pozorování nepřístupný parametr, $d$ a $h$ charakteristiky empirické nebo obecněji tzv. estimátory, tj. pozorování přístupné náhodné proměnné, a $\alpha$ předem volitelné riziko, tj. mezní pravděpodobnost, že při korektních pozorováních a správném výpočtu přece jen dospějeme nepravdivému rozhodnutí. Do tohoto obecného schématu se dosazují konkrétní výrazy za parametr $\tau$ (např. rozdíl dvou středních hodnot $\mu_1 - \mu_2$) a za $d$ a $h$ konkrétní vzorce podle charakteru úlohy (např. $m_1 - m_2 \pm t_{1-\alpha/2}(\nu)s_0$): podrobnější informace nalezneme v receptáři takových vzorců. — Při věcné interpretaci statistických závěrů se uživatel často setkává s větším počtem intervalů spolehlivosti

$$P(d_1 \leq \tau_1 \leq h_1) \geq 1 - \alpha \wedge \ldots \wedge P(d_v \leq \tau_v \leq h_v) \geq 1 - \alpha$$

Jestliže provádí $v$ separátních interpretací, z nichž každá je založena na jediném intervalu spolehlivosti, může pokaždé právem uvádět riziko $\alpha$. Jestliže však uživatel chce uvést jedinou simultánní interpretaci, k níž je zapotřebí současné platnosti všech nerovností, je jeho situace popsána vztahem

$$P(d_1 \leq \tau_1 \leq h_1 \wedge \ldots \wedge d_v \leq \tau_v \leq h_v) \geq 1 - \gamma$$

kde se ovšem riziko $\gamma$ za jinak stejných podmínek zvětšuje (v nejhorším případě $\gamma = v\alpha$). Proto není dobře poukaz na simultánní intervaly spolehlivosti vynechávat ani při velmi zjednodušeném výkladu.
— Při testování statistické hypotézy ověřujeme nejčastěji tvrzení ve tvaru

$$\tau = c$$

(čti "neznámý parametr $\tau$ se rovná danému číslu $c$") konkurující s jedním z tvrzení

$$\tau < c, \quad \tau \neq c, \quad \tau > c$$

Nechce-li se statistik podobat kouzelníku, vytahujícímu s cylindru králíka, kterého tam před tím sám vložil, musí při výkladu zdůraznit, že jak testovaná hypotéza, tak volba jedné z hypotéz alternativních musí být provedena bez přihlédnutí k vlastním datům, tedy např. na základě literárních údajů nebo jako výraz metodické skepse. Hlavní myšlenku testování můžeme vyjádřit implikací

$$h = c \;\Rightarrow\; P(t \in W_\alpha) \leq \alpha$$

kde $t$ znamená náhodnou proměnnou zvanou testová statistika, $W_\alpha$ kritický obor testu a $\alpha$ předem volitelné riziko neoprávněného zamítnutí testované hypotézy v případě, že je ve skutečnosti pravdivá. I při intuitivním výkladu by měli posluchači porozumět, že pravdivost testované hypotézy se až na riziko $\alpha$ neslučuje s nastoupením jevu $t \in W_\alpha$ a že je nutno v takovém případě testovanou hypotézu zamítnout ve prospěch alternativní. Naopak nastoupení opačného jevu $t \neq W_\alpha$ vede k závěru o slučitelnosti testované hypotézy s analyzovanými daty, a to je pravý význam přijetí testované hypotézy: nejde tedy o důkaz její pravdivosti. Potíže vyvolává nešťastný termín statistické významnosti rozdílu mezi skutečnou a hypotetickou hodnotou parametru, protože statisticky významným se při velkém počtu pozorování stane i sebemenší a věcně zcela bezvýznamný rozdíl mezi parametrem $\tau$ a číslem $c$. Doporučoval bych spíše termín "statistická průkaznost". K výkladu asi může posloužit analogie s trestním soudem, kde testovaná hypotéza odpovídá presumpci neviny, alternativní hypotéza obvinění, zamítnutí testované hypotézy odsouzení a její přijetí osvobození pro nedostatek důkazů. — V souvislosti s výstupy standardních statistických programů musí být součástí výkladu i zmínka o tzv. dosažené hladině významnosti (raději "průkaznosti") $p$, udávající riziko, při němž by se vypočtená realizace testové statistiky $t$ ocitla právě na hranici kritického oboru $W_p$. — Nověji se při analýze dat objevují snahy přihlížet k riziku neoprávněného přijetí testované hypotézy $\beta$ v případě, že je ve skutečnosti nepravdivá. Tato myšlenka je proveditelná i na intuitivní úrovni statistického výkladu, ale dosud se mi ji nepodařilo realizovat, víceméně z časových důvodů.

*Třídění metod statistické indukce.* Zdánlivou džungli různých statistických postupů je možno i nepříliš matematicky zaměřenému posluchači zpřístupnit zavedením klasifikace podle těchto čtyř hledisek:

| | *Situace při porovnávání dat:* | | *Stupeň kvantifikace:* |
|---|---|---|---|
| I | = jednovýběrové vyšetřování | a | = alternativní proměnná |
| II | = dvouvýběrové porovnávání | n | = nominální proměnná |
| III | = vícevýběrové porovnávání | o | = ordinální proměnná |
| IV | = párové porovnávání | i | = intervalová proměnná |
| V | = blokové porovnávání | p | = poměrová proměnná |

Pracujeme-li s více proměnnými, uvedeme kód pro každou z nich.

Zpracovávaná charakteristika:
Např. $\pi(x), \mu, \mu_1 - \mu_2, \sigma, \sigma_1/\sigma_2, \rho, \ldots$ Typ statistické procedury:
S = stanovení intervalu spolehlivosti
P = stanovení predikčního intervalu
T = stanovení intervalu tolerance
H = testování statistické hypotézy
N = určení potřebného rozsahu výběru

Jednoduché úlohy statistické indukce je možno sevřít sítí těchto čtyř hledisek, student však současně pochopí, že tato síť je poněkud děravá a zdaleka neobsahuje odpovědi

na všechny jednoduché otázky, které by mohl uživatel induktivní statistice klást — na něco prostě odpovědět neumíme.

*Receptář jednoduchých úloh statistické indukce* vychází z předešlého třídění: např. stanovení intervalu spolehlivosti pro rozdíl středních hodnot intervalových náhodných proměnných má kód $II\mu_1 - \mu_2 S$ a hledá se v oddílu IIi:

IIi. Dvouvýběrové porovnávání intervalové proměnné

| | | |
|---|---|---|
| $\mu_1,\ \mu_2$ | = | porovnávané střední hodnoty |
| $\sigma_1,\ \sigma_2$ | = | porovnávané teoretické směrodatné odchylky |
| $n_1,\ n_2$ | = | rozsahy porovnávaných výběrů |
| $m_1,\ m_2$ | = | porovnávané výběrové průměry |
| $s_1,\ s_2$ | = | porovnávané výběrové směrodatné odchylky |

$II\ \mu_1 - \mu_2\ S$ Interval spolehlivosti pro rozdíl středních hodnot

- Najdi označení IIi a ověř předpoklady receptu:
  *normální rozložení* nebo $n_1 > 30$ u prvního výběru a
  *normální rozložení* nebo $n_2 > 30$ u druhého výběru
- Zvol jednu ze tří forem intervalu spolehlivosti a riziko $\alpha$.
- Je-li předem známo, že variabilita u obou výběrů je stejná, tj. že neznámé teoretické směrodatné odchylky $\sigma_1 = \sigma_2$ se rovnají, vypočti pomocné hodnoty:

$$s_0 = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$\nu = n_1 + n_2 - 2 \qquad t_{1-\alpha/2}(\nu) \qquad t_{1-\alpha}(\nu)$$

  Není-li takové předběžné informace, vypočti pomocnou hodnotu:

$$s_0 = \sqrt{s_1^2/n_1 + s_2^2/n_2}$$

  Pak vyhledej kvantily:

$$\nu \approx \left(\frac{1}{n_1-1}\left(\frac{s_1^2/n_1}{s_1^2/n_1+s_2^2/n_2}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2/n_2}{s_1^2/n_1+s_2^2/n_2}\right)^2\right)^{-1}$$

$$t_{1-\alpha/2}(\nu) \qquad t_{1-\alpha}(\nu)$$

- Dosaď do zvolené formy intervalu spolehlivosti:

$$\begin{aligned} m_1 - m_2 - t_{1-\alpha}(\nu)s_0 \le\ & \mu_1 - \mu_2 \\ m_1 - m_2 - t_{1-\alpha/2}(\nu)s_0 \le\ & \mu_1 - \mu_2 \le m_1 - m_2 + t_{1-\alpha/2}(\nu)s_0 \\ & \mu_1 - \mu_2 \le m_1 - m_2 + t_{1-\alpha}(\nu)s_0 \end{aligned}$$

*Literatura:*


Z.Roth, M.Josífko, V.Malý, V.Trčka: Statistické metody v experimentální medicíně. SZN, Praha 1962.

V.Fabian: Základní statistické metody. NČSAV, Praha 1963.

H.Swoboda: Moderní statistika. Svoboda, Praha 1977.

T.Havránek: Statistika pro biologické a lékařské vědy. Academia, Praha 1993.

S.Komenda: Biometrie. Univerzita Palackého, Olomouc 1994.

L.Osecká, P.Osecký: Receptář jednoduchých metod statistické indukce. Psychologický ústav AV, Brno 1996.

K.Zvára: Biostatistika. Karolinum, Praha 1998.