DIVOKÝ SEN O MĚŘENÍ

Jan Klaschka, Psychiatrické centrum Praha

Danovi s omluvami: Neumím jinak a neumím líp.

Dvakrát měř

Také jste to určitě někdy četli nebo slyšeli: "Mám-li jedny hodinky, vím, kolik je hodin, mám-li dvoje, nejsem si už jist." Tento vtip je obzvláště oblíbený mezi statistiky - snad proto, že jim dává pocítit převahu nad laiky, kteří potřebují takto drastický příklad, aby si uvědomili rozdíl mezi *měřením* a *měřeným*. Statistici si jistě tyto pojmy pletou zřídka a iluze, že výsledek jediného měření je přesně tím číslem "o které jde", je u nich také vzácná. V něčem si ale přece bude leckterý statistik s laikem roven: Má-li k dispozici jediné měření nějaké veličiny, nenapadne ho na otázku, jaké hodnoty tato veličina nabývá, lepší odpověď než právě výsledek měření. (Pokud se k tomu nikdo nehlásí, týká se to alespoň autora před nějakým časem.)

Jiná odpověď by však statistika napadnout mohla. *Odhad měřeného z měření* představuje statistickou úlohu, která má smysl a řešení i v případě, že měření je jen jedno.

Uvažujme nejběžnější model měření X veličiny ξ s aditivní chybou ε

$$X = \xi + \varepsilon \,,$$
$$E\xi = \mu,\ \mathrm{var}\xi = \sigma_\xi^2,\ E\varepsilon = 0,\ \mathrm{var}\varepsilon = \sigma_\varepsilon^2,\ E\xi\varepsilon = 0. \tag{1}$$

Veličina ξ se chápe jako náhodná proto, že měřené objekty se pokládají za výběr z nějaké populace objektů. Hodnota veličiny ξ u daného objektu v daném okamžiku - tzv. *skutečná hodnota* - je pevná.

Měření X je sice nestranným odhadem veličiny ξ, tj. $E(X \mid \xi) = \xi$, lze ale sestrojit odhad $\hat{\xi} = f(X)$ s menší střední kvadratickou chybou $E(X - \hat{\xi})^2$. Jak je dobře známo ([4], str. 302), tuto střední kvadratickou chybu minimalizuje podmíněná střední hodnota $E(\xi \mid X)$.

Je-li sdružené rozdělení veličin ξ a ε normální, platí

$$E(\xi \mid X) = \mu + r(X - \xi) = (1 - r)\mu + rX \,,$$

kde

$$r = \sigma_\xi^2 \ / \ (\sigma_\xi^2 + \sigma_\varepsilon^2) \ .$$

Pro ne-normální dvojici ξ, ε může podmíněná střední hodnota $E(\xi \mid X)$ mít jiný tvar, odhad

$$\hat{\xi} = \mu + r(X - \mu) = (1 - r)\mu + rX \tag{2}$$

ovšem i tehdy minimalizuje střední kvadratickou chybu alespoň ve třídě lineárních odhadů, tj. odhadů tvaru $\hat{\xi} = aX + b$.

Odhad (2) se vcelku nepříčí intuici: Odchylku měření X od střední hodnoty μ je třeba přičíst na vrub chyby měření z tím větší části, čím větší tato odchylka je a čím větší je rozptyl chyby měření ve srovnání s rozptylem skutečných hodnot. Parametr r, pro který si z psychometrie vypůjčíme označení *reliabilita*, udává de facto, v jaké míře lze hodnotám X jakožto informaci o ξ věřit. Je-li $r = 1$, měření je bezchybné a $\hat{\xi} = X$. Je-li naopak reliabilita nulová, jsou veškeré odchylky X od μ způsobeny jen chybou měření a $\hat{\xi} = \mu$.

Odhad (2) je funkcí parametrů μ a r, jež v praktických situacích budou zpravidla neznámé. V dalším textu nicméně budeme o μ a r uvažovat jako o známých hodnotách, kdykoli budou k dispozici data, na základě kterých lze tyto parametry dostatečně přesně odhadnout. Důsledné rozlišování odhadu (2) a jeho modifikace, kde místo μ a r figurují jejich odhady, by bylo korektnější a vedlo by k bohatší a těžší teorii, ale spíše by zatemnilo než zviditelnilo jádro sdělení.

Reliabilita r se (na rozdíl od střední hodnoty μ) nedá odhadnout na základě jediného měření náhodně vybraných objektů. V psychometrii je běžná technika odhadu reliability pomocí opakovaného měření, tj. dvojice měření $X = \xi + \varepsilon$, $X' = \xi + \varepsilon'$, jež se vzájemně liší pouze chybou. (Ponechme stranou, že právě v psychologii je dosti obtížné takové opakování uskutečnit.) Pokud se obě trojice (X, ξ, ε) i (X', ξ, ε') řídí modelem (1) a chyby ε, ε' mají stejné rozptyly a nulovou korelaci, je reliabilita měření X (stejně jako X') rovna korelačnímu koeficientu mezi X a X'. Reliabilitu r lze tedy odhadnout výběrovým korelačním koeficientem vypočteným z dostatečného počtu dvojic měření X_i, X'_i, $i = 1, 2, \ldots, N$ získaných u N náhodně vybraných objektů.

Jak se ukazuje, úloha odhadu skutečné hodnoty z jediného měření je v podstatě triviální a není divu, že řešení je v literatuře dávno popsáno (viz např. [3]). Dále se touto věcí zabývat by bylo nudné a uspávající...

Sen

Zeměměřič K. sestrojil přístroj, kterým mohl změřit hodnotu každé věci na světě. Provedl množství měření, která zaznamenal. Když zestárl, nechtělo se mu už běhat po světě a měřit, i pustil se do teoretické práce, jejíž výsledky shrnul ve spisu "Závěť zeměměřiče". Doporučil tu budoucímu zeměměřiči změřit dříve, než se pustí do díla, určité množství věcí (snad 1000) dvakrát, vypočítat z těchto měření aritmetický průměr μ a výběrový korelační koeficient r, a hodnotu věcí ξ pak odhadovat z měření X podle vzorce (2).

Zeměměřič K. odkázal svým synům dva shodné exempláře svého přístroje. V rodině prvního syna se pak přístroj dále dědil z otce na syna a po celé generace se prováděla a zaznamenávala měření. Z těchto záznamů dosti jasně vyplývalo, že hodnota věcí na světě si zachovává dlouhodobě stále stejné rozložení.

Druhému synovi - označme jej K_1 - odkázal starý zeměměřič kromě přístroje také své zápisky a především spis "Závěť zeměměřiče". Pamětliv otcova odkazu, vypočítal K_1 z posloupnosti opakovaných měření průměr μ a korelaci r. Otcův přístroj P pak zabudoval do přístroje P_1, který výstupní hodnoty X přístroje P používal jako vstup a na výstupu poskytoval transformované hodnoty $\hat{\xi} = \mu + r(X - \mu) = (1 - r)\mu + rX$. Celý život pak K_1 měřil a zaznamenával hodnoty $\hat{\xi}$. Nakonec přístroj P_1, své a otcovy záznamy a také "Závěť zeměměřiče" odkázal synovi K_2.

Také K_2 byl pamětliv dědova teoretického odkazu, i pustil se nejdříve do opakovaných měření věcí přístrojem P_1. Co zjistil?

1.) Výstup $\hat{\xi}$ z přístroje je zřejmě roven součtu skutečné hodnoty ξ a chyby měření ε.

2.) Aritmetický průměr z naměřených hodnot $\hat{\xi}$ je roven μ ($= E\hat{\xi} = EX$).

3.) Korelační koeficient mezi opakovanými měřeními $\hat{\xi}$ a $\hat{\xi}'$ je roven r (neboť $\hat{\xi} = aX + b$, $\hat{\xi}' = aX' + b$, kde $a \neq 0$, takže $\mathrm{cor}(\hat{\xi}, \hat{\xi}') = \mathrm{cor}(X, X')$).

Vzhledem k těmto poznatkům vložil K_2 přístroj P_1 do nitra přístroje P_2, který výstupní hodnoty $\hat{\xi} = \hat{\xi}^{(1)}$ přístroje P_1 používal jako vstup a na výstupu poskytoval transformované hodnoty $\hat{\xi}^{(2)} = \mu + r(\hat{\xi}^{(1)} - \mu) = (1 - r)\mu + r\hat{\xi}^{(1)} = \mu + r^2(X - \mu) = (1 - r^2)\mu + r^2X$. Celý život pak K_2 měřil a zaznamenával hodnoty $\hat{\xi}^{(2)}$. Nakonec přístroj P_2, své, otcovy a dědovy záznamy a také "Závěť zeměměřiče" odkázal synovi K_3.

Zbytek historie je už nasnadě. Každý další potomek K_i vždy nejdříve vypočítal z měření zděděným přístrojem P_{i-1} střední hodnotu μ a reliabilitu r, načež sestrojil přístroj P_i, kterým pak celý život měřil hodnoty $\hat{\xi}^{(i)} = \mu + r(\hat{\xi}^{(i-1)} - \mu) = (1 - r)\mu + r\hat{\xi}^{(i-1)} =$

$\mu + r^i(X - \mu) = (1 - r^i)\mu + r^iX$.

Srovnání záznamů různých generací čím dál tím jasněji prozrazovalo nezadržitelný trend "glajchšaltování" světa. Rozptyl naměřených hodnot věcí byl stále menší, až se posléze přestaly vyskytovat jakékoli pozorovatelné odchylky od hodnoty μ.

Zeměměřič K. měl ještě třetího syna. Ten však neměl o měření valný zájem, a tak mu otec odkázal jen dřevěný model přístroje P. Starobylá maketa byla v rodu tohoto syna po generace hrdě oprašována. Nestálo by to za zmínku, nebýt zvláštního kontrastu: Zatímco ručička dřevěného modelu, ukazující jednou provždy hodnotu μ nikdy nikoho nevyvedla z duševní rovnováhy, byli zeměměřiči K_i obdobnými projevy přístrojů P_i s rostoucím i stále více deptáni. Spíše než měření se pak věnovali návštěvě pražských hostinců, kde posléze jejich empiricky podložená filosofie vyznačující se paušálními soudy typu "Teď už stojí všechno za μ" měla velký ohlas a zapustila pevné kořeny.

Výklad snu

Z celého pokolení K_i, $i = 1, 2, 3, \ldots$ jen K_1 opravou podle vzorce (2) skutečně minimalizoval střední kvadratickou chybu odhadu veličiny ξ. Všichni ostatní vzorec (2) aplikovali v situaci, kdy byly porušeny předpoklady modelu (1). Měření pořízená přístrojem P_i pro $i \geq 1$ lze sice vyjádřit jako součet skutečné hodnoty ξ a chyby měření, ale chyba měření není nekorelovaná se skutečnou hodnotou.

Porušení předpokladů (1), s nímž se zeměměřiči K_i, $i \geq 2$ setkali, bychom nicméně mohli nazvat *porušením skrytým*. Výstupy přístroje P_{i-1}, $i \geq 2$ byly jen lineární transformací výstupů vevnitř ukrytého původního přístroje P, takže nesoulad s modelem (1) nemohl být odhalen, pokud by jej neindikoval také samotný přístroj P. (Býval by pomohl např. etalon, tj. objekt u nějž je známa skutečná hodnota, nebo znalost rozptylu σ^2_ξ.)

Počínání zeměměřičů K_i, $i \geq 2$ tedy můžeme interpretovat více způsoby. Možná byli prostě slepě poslušní instrukcí, o jejichž smyslu nepřemýšleli. Spíše ale si byli teoretického pozadí odhadu (2) vědomi, jenže dostupná měření jim neumožňovala platnost předpokladů (1) ověřit. Jednali tedy dále, jako by model (1) platil - obdobně, jako nám (prosím, některým z nás) mnohdy v aplikacích k přijetí předpokladu normality stačí, že není porušen příliš zjevně.

Konec konců se dá úsilí potomků K_1 zhodnotit i tak, že opravdu úspěšně minimalizovali střední kvadratickou chybu lineárního odhadu skutečné hodnoty, pouze se jednalo o skutečnou hodnotu jiné veličiny než ξ, totiž $\mu + r^{1-1}(\xi - \mu)$. Z tohoto pohledu se pak jako omyl nejeví samotná měření, ale jen snaha srovnávat mezi sebou záznamy různých generací.

V praxi naštěstí snad spuštění série oprav, jaké jsme byli svědky, nehrozí. Při fyzikálním měření lze zpravidla pomocí etalonů porušení modelu (1) odhalit. A v psychologii, kde je přeci jen obtížnější uchovávat neměnný vzorek v Sèvres při teplotě 0°C, se odhad (2) (pokud jsem si všiml) příliš nepoužívá a není ani v literatuře často doporučován. Důvod bude spočívat asi v tom, že $\hat{\xi}$ je jen lineární transformací X, takže volba mezi X a $\hat{\xi}$ je v podstatě jen volbou jednotky a jako taková je do značné míry lhostejná. Nasvědčuje tomu fakt, že se o vzorci (2) v literatuře mluví převážně jen v souvislosti s psychometrickým hodnocením změny v čase: Rozdíl Y - X dvou postupných měření má různé nepříjemné vlastnosti, které lze částečně oslabit nahrazením počátečního měření X odhadem $\hat{\xi}$ (viz [2]). Na rozdíl od "jednoduchého" měření ovšem korigovaná hodnota $Y - \hat{\xi}$ není funkcí (tím méně funkcí lineární) nekorigované diference Y - X.

Second helping

Můžeme poznatky o odhadu (2) nějak zužitkovat v situaci, kdy se měření X veličiny ξ řídí modelem (1), ale reliabilitu r (ani její dostatečně přesný odhad) neznáme? Svým způsobem ano, pokud víme alespoň, že $r \leq r^*$ pro nějaké $r^* < 1$. Odhad

$$\hat{\xi}(\rho) = \mu + \rho(X - \mu) \tag{3}$$

má totiž menší střední kvadratickou chybu než X pro $2r - 1 < \rho < 1$. Lepší než spokojit se s neupravenou hodnotou X je tedy, zdá se, použít odhad (3) a zvolit kteroukoli hodnotu ρ mezi $2r^* - 1$ a 1.

Podívejme se, jak bude tato metoda fungovat v životě. Představte si, že se ucházíte v konkursu o místo vedoucího pracoviště. Jediným konkurentem je váš velmi neoblíbený kolega. Výsledky jsou zatím vyrovnané a o všem má rozhodnout test schopnosti rychlého učení. Na rozdíl od svého soupeře jste pochopil(-a), že důležitým předpokladem úspěchu v tomto testu je nízká naměřená úroveň vstupních znalostí, proto veškeré vstupní znalosti zatajíte. A ejhle - vychází vám, že jste se naučil(-a) 150 jednotek, zatímco váš kolega, sláva, jen 50. Už se těšíte, jakými úkoly ho v nejbližší době pověříte... Reliabilita rozdílu mezi výsledky dvou testů (vstupního a výstupního) je bohužel obecně dosti malá (viz [1]), v případě daného testu učení není větší než 0,4. Lze tedy volit třeba $\rho = -0,1$ ($> 2 \ . \ 0,4 - 1 = -0,2$). Střední výsledek uchazečů v obdobných konkursech je 100 bodů. I přepočítá konkursní komise výsledky testu takto: váš ze 150 na $100 - 0,1 \ . \ (150 - 100) = 95$, kolegův z 50 na $100 - 0,1 \ . \ (50 - 100) = 105$. Je to sice těsné, ale komisi to k určení vítěze stačí. Když se zajímáte, kdo si takový způsob hodnocení vymyslel, dozvíte se, že metodika byla vyvinuta v Psychiatrickém centru v Bohnicích a že tamní statistik dokázal, že to je tak lepší než výsledky testu vůbec nekorigovat. Chcete si na celou záležitost ještě pořádně posvítit, ale nějak to stále odkládáte - dostáváte teď v práci moc úkolů...

Závěr

nechť si utvoří každý sám. Osobně jsem si z překvapení vlastnostmi odhadů (2) a (3) odnesl hlavně trochu nedůvěry - možná iracionální - k odhadům s nejmenší střední kvadratickou chybou (obecně, ne jen v modelu (1)).

Nezbytnou vlastností kvalitní metody měření je bezpochyby značná shoda pořadí naměřených hodnot podle velikosti s pořadím hodnot skutečných. Odhad $\hat{\xi}(\rho)$ dle (3) pro ρ záporné je z tohoto hlediska katastrofa. Nemuseli bychom tyto patologické případy brát vážně, vždyť optimální hodnota ρ (rovná reliabilitě měření) záporná nikdy není, takže v praxi není důvod záporné ρ volit. A odhad $\hat{\xi}$ podle (2) není podle uvedeného kritéria horší (i když není ani lepší) než nekorigované měření X.

Ale: Je-li *nejlepší* usilovat o *nejmenší* čtverce chyb, očekával bych také, že bude *dobré* je jakkoli *zmenšovat*. Toto očekávání se nesplnilo. Reflexy praktického statistika velí zamítnout hypotézu.

Literatura

[1] Bereiter, C.: Some persisting dilemmas in the measurement of change. In: Harris, Ch. W. (ed.): Problems in measuring change. Univ. of Wisconsin Press, Madison 1963.

[2] Cronbach, L. J. - Furby, L.: How we should measure "change" - or should we? Psychological Bulletin 74 (1970), 68 - 80.

[3] Kelley, T. L.: Fundamentals of statistics. Harvard Univ. Press, Cambridge, MA 1948.

[4] Rao, R. C.: Lineární metody statistické indukce a jejich aplikace. Academia, Praha 1978.