12 Úvod do statistiky

Břetislav Fajmon, UMAT FEKT, VUT Brno

Tato přednáška je složena z částí textu [1], a sice

- **oddíl 10.2 – empirické charakteristiky popisu dat**
- **oddíl 11.3 – základní principy statistického testu**
- **oddíl 11.4 – příklad diskrétního statistického testu**
- **oddíl 13.5 – příklad spojitého statistického testu**
- **kapitola 14 – test střední hodnoty průměru $\overline{X}$ při známém rozptylu.**

Za důležitý text na českém internetu považuji též text `http://homen.vsb.cz/~oti73/cdpast1/`, který je velmi přehledný (a obsahuje nejen partie statistiky, ale též jej doporučuji v rámci opakování části pravděpodobnost).

Já osobně jsem se statistice věnoval v první polovině textu [3] předmětu MPSO na FEKT.

12.1 Základy statistického měření

Co je to statistika?

První informace, které lze označit za statistiku, určitě souvisí se slovem status (= stav), a možná také přeneseně i se slovem stát, protože první souhrnná měření se většinou týkala měření či informací shrnujících stav na území určitého státu.

Dnes bychom možná pojem statistiky vymezili obšírněji: statistika = jakékoli zpracování informací, které souvisí s měřením určitých veličin.

Základní pojmy, které zpracovávají soubor měření jedné veličiny:

- **Průměr z naměřených hodnot $x_1, x_2, \ldots, x_n$: $\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$. Označení $\overline{x}$ je celkem standardní a používá se ve fyzice i dalších vědách k vyjádření průměrné hodnoty.**

- **Medián z hodnot $x_1, x_2, \ldots, x_n$ je prostřední z těchto hodnot vzhledem k jejich uspořádání podle velikosti.**

- **Modus z hodnot $x_1, x_2, \ldots, x_n$ je ta z hodnot, která se vyskytuje s nejvyšší četností.**

Příklad 12.1. Jsou získány výsledky kvizového skóre u 11 osob:

$$8, 5, 7, 9, 8, 1, 3, 4, 7, 7, 7.$$

Průměr těchto hodnot je $\overline{x} = \frac{1}{11} \cdot \sum x_i = \frac{66}{11} = 6$. Modus tohoto souboru je hodnota, která se vyskytuje nejčastěji, čili číslo 7. A abychom mohli určit medián, musíme hodnoty seřadit podle velikosti (například vzestupně):

$$1, 3, 4, 5, 7, 7, 7, 7, 8, 8, 9.$$

Prostřední z těchto hodnot je na šesté pozici, čili mediánem je číslo 7.

Příklad 12.2. **Mějme jiný soubor hodnot, už uspořádaný podle velikosti, například sestupně:**

$$7, 6, 5, 5, 4, 2, 1, 1.$$

Protože počet měření je sudý (budeme též říkat, že soubor měření má sudou délku), medián určíme jako průměr dvou prostředních hodnot: $\frac{1}{2}(5+4) = 4{,}5$**.**

Příklad 12.3. **Soubor měření může mít více modů (= druhý pád od slova modus). Například soubor**

$$8, 6, 6, 5, 4, 3, 3$$

je tzv. bimodální soubor, protože nejčastěji (= dvakrát) se v něm objevují hodnoty 6 (=modus1) a 3 (=modus2). Při třech modech mluvíme o trimodálním, při čtyřech o kvatromodálním souboru, atd. Některé učebnice ignorují možnost více modů a za modus označují největší nejčastěji nabývanou hodnotu, což by v našem případě bylo 6.

Z uvedených tří charakteristik je většinou nejužitečnější průměr – až na situace tzv. odkloněných hodnot (např.: měření výše platu v ČR).

Dalšími pojmy kromě pojmů již uvedených je horní a dolní kvartil – lze označit např. q_{25} a q_{75}. Dolní kvartil získáme ze souboru měření uspořádaného podle velikosti tak, že vezmeme hodnotu, která je větší než 25 procent souboru měření; horní kvartil je hodnota větší než 75 procent souboru měření (opět uspořádaného podle velikosti vzestupně).

Viz obrázek o platech v ČR.

zmínka o α-kvantilu (netýká se statistiky, ale pravděpodobnosti, protože α-kvantil konstruujeme pomocí $F(x)$).

Příklad 12.4. **Náhodná veličina X udává počet líců při čtyřech hodech mincí. Měřením se získalo těchto dvacet hodnot veličiny:**

$$3, 1, 1, 3, 1, 2, 0, 2, 4, 4, 1, 2, 2, 1, 2, 1, 2, 3, 3, 3.$$

Určete průměr a empirický rozptyl souboru měření.

a) Klasické řešení: Vypočteme průměr, empirický rozptyl i empirickou směrodatnou odchylku:

$$\overline{x} = \frac{1}{20}\sum_{1}^{20} x_i = 2{,}05;$$

$$
\begin{aligned}
s^2 &= \frac{1}{20}\left(\sum_{1}^{20} x_i^2\right) - 2{,}05^2 = 1{,}1475;\\
s &= \sqrt{1{,}1475} = 1{,}0712.
\end{aligned}
$$

Vidíme tedy, že při čtyřech hodech mincí padalo průměrně $2,05$ *líců; odchylka* $s \doteq 1{,}07$ *znamená, že pokud měřenou veličinu lze popsat normálním rozdělením, víme, že asi 68 procent měření veličiny leží v intervalu* $(2,05 - 1,07; 2,05 + 1,07)$.

b) Řešení pomocí rozdělení četnosti: Máme li data zpracována v podobě četností – viz tabulka 1, kde ν_i jsou hodnoty, kterých veličina X nabývá (ν je písmeno řecké abecedy a čte se „ný“) –

Tabulka 1: K příkladu 4: Tabulka četností souboru měření veličiny X.

ν_i	ν_i^2	četnost $c(\nu_i)$
0	0	1
1	1	6
2	4	6
3	9	5
4	16	2

$$\overline{x} = \frac{1}{n}\sum_{\nu_i} \nu_i \cdot c(\nu_i); \qquad s^2 = \frac{1}{n}\Big(\sum_{\nu_i} \nu_i^2 \cdot c(\nu_i)\Big) - \overline{x}^2.$$

c) Řešení pomocí rozdělení pravděpodobnosti:

Úpravou

$$\frac{c(\nu_i)}{n} = p(\nu_i).$$

lze dosáhnout přepisu vzorců ve tvaru

$$\begin{aligned}
\overline{x} &= \sum_{\nu_i} \nu_i \cdot \frac{c(\nu_i)}{n} = \sum_{\nu_i} \nu_i \cdot p(\nu_i);\\
s^2 &= \left(\sum_{\nu_i} \nu_i^2 \cdot \frac{c(\nu_i)}{n}\right) - \overline{x}^2 = (\sum_{\nu_i} \nu_i^2 \cdot p(\nu_i)) - \overline{x}^2.
\end{aligned}$$

S využitím tabulky 2 empirických pravděpodobností pak dosazením do těchto vzorců dostaneme tentýž výsledek jako v případě a) a b).

Tabulka 2: K příkladu 4: Tabulka empirických pravděpodobností.

ν_i	ν_i^2	$p(\nu_i)$
0	0	0,05
1	1	0,3
2	4	0,3
3	9	0,25
4	16	0,1

I v tomto případě se stále jedná o pouhé přepsání stejných vzorců a) nebo b) s využitím označení pomocí pravděpodobnosti.

12.2 Principy statistického textu

Příklad 12.5. Soudní proces jako příklad rozhodovacího procesu. Uvažujme jednoduchý soudní proces, ve kterém existuje pouze jediný možný trest a soud rozhodne, zda se tomuto trestu obžalovaný podrobí nebo ne. A navíc proti rozhodnutí soudu neexistuje žádné odvolání. Jedná se o jakýsi rozhodovací proces, u kterého mohou nastat čtyři možné výsledky:

1. **Obžalovaný je vinen a soud jej odsoudí.**
2. **Obžalovaný je nevinen a soud jej osvobodí.**
3. **Obžalovaný je nevinen a soud jej odsoudí. Jedná se o chybné rozhodnutí - tuto chybu budeme označovat jako chybu prvního druhu.**
4. **Obžalovaný je vinen a soud jej osvobodí. Toto rozhodnutí je rovněž chybné - budeme tuto chybu označovat chybou druhého druhu.**

V každém soudním procesu se musí hledat jistá rovnováha mezi tvrdostí a mírností. Jedním extrémem je liberální soudce, který k usvědčení obžalovaného vyžaduje velké množství důkazů.

Takový soudce jen zřídka odsoudí nevinného (zřídka se dopustí chyby prvního druhu), ale dosti často osvobodí viníka (chyba druhého druhu). Druhým extrémem je soudce, kterému k usvědčení stačí jen několik důkazů. Takový soudce posílá do vězení i jen při stínu podezření, čili častěji odsoudí nevinného (chyba prvního druhu), ale zřídka osvobodí darebáka (= zřídka se dopustí chyby druhého druhu). Slova „konzervativní“ a „liberální“ jsou termíny z politiky. V dnešní době už nikdo neví, co znamenají. Tato jejich „statistická“ definice navrhuje jejich význam, ale také upozorňuje na nebezpečí každého z těchto postojů.

Je otázkou, která z chyb je závažnější - zda chyba prvního druhu, nebo chyba druhého druhu. Všeobecně se má za to, že závažnější je uvěznit nevinného, než osvobodit darebáka. A proto se chybě odsouzení nevinného přisuzuje druh číslo 1 a věnuje se jí větší pozornost. Ale někde musí být stanovena jistá hranice, po jejímž překročení už soud přistoupí k rozhodnutí „vinen“ a bez skrupulí člověka potrestá.

Všimněme si jedné věci, která platí jako obecný princip. Pokud se soudce snaží být benevolentní a odsoudí člověka až po nahromadění velkého množství důkazů (snižuje tím možnost výskytu chyby prvního druhu), současně narůstá nebezpečí, že i když je

obžalovaný vinen, potřebné množství důkazů se nenajde a soud jej osvobodí (roste možnost výskytu chyby druhého druhu). Není to nic světoborného, ale už jsme dlouho neměli žádný rámeček, a proto jej aspoň uvnitř příkladu můžeme použít:

> **Snižováním možnosti výskytu chyby prvního druhu roste možnost výskytu chyby druhého druhu - a naopak: pokud zvyšujeme možnost výskytu chyby prvního druhu, snižuje se možnost výskytu chyby druhého druhu.**

Z uvedeného je vidět, že žádnou z chyb není možné naprosto vyrušit: pokud totiž snižujeme možnost výskytu chyby prvního druhu až téměř na nulu,

roste tím možnost výskytu chyby druhého druhu do obludných rozměrů a rozhodnutí učiněná tímto stylem jsou nerozumná, až nemoudrá. Strategií v rozhodovacích procesech tohoto typu je tedy zvolit pravděpodobnost výskytu chyby prvního druhu malou, ale ne příliš malou.

Shrňme předchozí úvahy do pěti kroků, které popisují celý soudní proces:

1. *Stojí proti sobě dvě možná rozhodnutí soudu:*

$$H_0 \ldots \textit{obžalovaný je nevinen}$$
$$H_1 \ldots \textit{obžalovaný je vinen}$$

Soud musí rozhodnout právě jednu z těchto variant a toto rozhodnutí je nezvratné, neexistuje proti němu odvolání.

2. *Vystoupí žalobce, který předloží nashromážděné důkazy pro platnost H_1.*

3. *Vystoupí obhájce a vysvětlí všechny souvislosti za předpokladu, že platí H_0. Snaží se vidět a vysvětlit všechny argumenty obžaloby ve světle toho, že obžalovaný je nevinen.*

4. *Porota soudu se odebere k rokování. Bere v úvahu jak množství důkazů a jejich závažnost, tak i argumenty obhajoby a možnost, že tyto důkazy neznamenají nutně vinu obžalovaného, ale jen shodu náhod.*

5. *Porota se vrací a vyslovuje svůj verdikt: pokud byla překročena míra závažnosti důkazů pro platnost H_1, obžalovaný je vinen. pokud ne, obžalovaný je osvobozen. Toto rozhodnutí soudu je nezvratné.*

Tabulka 3: Čtyři možné výsledky statistického testu.

	skutečnost: H_0 platí	skutečnost: H_1 platí
rozhodnutí: H_0 nezamítáme	O.K.	chyba 2.druhu
rozhodnutí: H_0 zamítáme	chyba 1.druhu	O.K.

Právě uvedených pět kroků v příkladu 5 se vyskytuje v mnoha rozhodovacích procesech, které nazý-

váme statistické testy. Tyto principy platí obecně, vyslovme je tedy obecně, už oproštěni od příkladu soudce a obžalovaného (ovšem analogie se soudním procesem zde existuje velice přímá):

(K1) Statistický test obyčejně rozhoduje o tom, zda platí hypotéza H_0 (tzv. nulová hypotéza) nebo H_1 (tzv. alternativní hypotéza). Tyto dvě hypotézy přitom stojí ve vzájemném rozporu. Ve většině testů H_0 tvrdí, že jistá veličina *nezávisí* **na hodnotách určité další veličiny, kdežto H_1 tvrdí, že naopak** *závisí***.**

(K2) Stanovíme kritérium (zpravidla určitou funkci), které ukazuje na míru platnosti alternativní hy-

potézy H_1 (určuje „závažnost důkazů“ pro H_1). Pak provedeme experiment, změříme data a dosadíme do našeho kritéria.

(K3) Kritériem bývá jistá funkce, která při různých měřeních nabývá různých hodnot, je to tedy náhodná veličina. Určíme teoretické rozdělení kritéria za předpokladu, že platí hypotéza H_0. Jinými slovy, popíšeme vlastnosti kriterijní veličiny ve světle toho, že platí H_0.

(K4) Na základě teoretického rozdělení kriterijní veličiny stanovíme určitý interval hodnot, kam když padne empirická hodnota kritéria, tak nezviklá naše přesvědčení o platnosti H_0, ale

eventuelní dopad hodnoty kritéria mimo tento interval nás povede k názoru, že byla překročena jistá kritická míra, takže usoudíme, že H_0 neplatí. Kritickou míru zpravidla určujeme tak, aby pravděpodobnost výskytu chyby prvního druhu (tj. že rozhodneme, že H_0 neplatí, když ve skutečnosti H_0 platí) byla dostatečně malá, např rovna 0.05 (to se chyby prvního druhu dopustíme nejvýše v pěti procentech případů), ale ne příliš malá, aby nerostla možnost výskytu chyby druhého druhu (tj. že rozhodneme, že H_0 platí, když ve skutečnosti H_0 neplatí) do nerozumných rozměrů.

(K5) Porovnáme empirickou hodnotu kritéria s kritickou mírou. Pokud je kritická míra překročena (hodnota kritéria leží mimo interval nalezený v bodě 4), zamítáme hypotézu H_0 ve prospěch alternativní hypotézy H_1. Pokud není kritická míra překročena, hypotézu H_0 nezamítáme.

Další standardní označení (viz tabulka 3) se používá pro pravděpodobnost výskytu chyby 1.druhu (značí se α) a pravděpodobnost výskytu chyby 2.druhu (značíme β).

12.3 Znaménkový test – příklad testu parametru diskrétní veličiny

Příklad 12.6. **Podle expertního předpokladu má mít o nový výrobek zájem dvacet procent zákazníků. Ze 80 dotázaných jich projevilo zájem 25. Testujte hypotézu, že předpoklad expertů se naplnil.**

Projdeme při řešení příkladu pět kroků statistického testu. Kritický interval: pro $\alpha = 0{,}05$ **zamítáme mimo interval 10 až 22, protože** $p(0)+p(1)+\cdots+p(9) = 0{,}0287 > 0{,}025$ **(tím je určena mez 10, protože je to první číslo, pro které součet všech předchozích pstí přesáhne hodnotu** $0{,}025 = \frac{\alpha}{2}$**), a dále** $p(23)+p(24)+\cdots+p(80) \doteq 0{,}0388 > 0{,}025$ **(tím je určena mez 22, protože to je první číslo zprava, pro které součet všech následujících pstí přesáhne hodnotu** $0{,}025 = \frac{\alpha}{2}$**). Jednotlivé psti se vypočtou pomocí binomického rozdělení** $Bi(n = 80; p = 0{,}2)$**.**

12.4 U-test – příklad testu parametru spojité veličiny

Příklad 12.7. **Podle expertního předpokladu má mít o nový výrobek zájem dvacet procent zákazníků. Ze 80 dotázaných jich projevilo zájem 25. Testujte hypotézu, že předpoklad expertů se naplnil.**

Tentýž příklad nyní vyřešíme testem na normální rozdělení: budeme potřebovat znát EX a DX, které vypočteme způsobem „binomickým“ (= nahrazujeme binomické rozdělení normálním se stejnou střední hodnotou a rozptylem):

$$EX = Np = 80{\cdot}0{,}2 = 16 \quad DX = Np(1-p) = 80{\cdot}0{,}2{\cdot}0{,}8 = 12{,}8;$$

využijeme spíše hodnotu $\sigma = \sqrt{DX} \doteq 3{,}5777$**. Nyní převedeme naměřenou hodnotu** $X = 25$ **na** U**-hodnotu**

$$U = \frac{25 - 16}{3{,}5777} \doteq 2{,}516,$$

a to překračuje sice interval $\langle -1{,}96; 1{,}96 \rangle$ **standardizované normálně rozdělené veličiny pro** $\alpha = 0{,}05$**, ale spadá do intervalu** $\langle -2{,}58; 2{,}58 \rangle$ **pro** $\alpha = 0{,}01$ **– čili hodnota** $X = 25$ **je slabě významně vychýlená od předpokládané střední hodnoty** 16**.**

Uvedené intervaly lze transformací $X = U \cdot \sigma + \mu$ do „řečí“ hodnot veličiny X: pro $\alpha = 0{,}05$ je interval pro nezamítnutí H_0 roven

$$\langle -1{,}96; 1{,}96\rangle \cdot 3{,}577 + 16 = \langle 8{,}99; 23{,}01\rangle;$$

pro $\alpha = 0{,}01$ je interval pro nezamítnutí H_0 roven

$$\langle -2{,}58; 2{,}58\rangle \cdot 3{,}577 + 16 = \langle 6{,}77; 25{,}23\rangle;$$

12.5 Test střední hodnoty průměru $\overline{X}$ při známém rozptylu

Příklad 12.8. Výrobce pružin prohlašuje, že jeho výrobek má střední hodnotu zlomu $\mu = 58$ kg a směrodatnou odchylku $\sigma = 3{,}5$ kg. Testujte, zda odpovídá tvrzení skutečnosti, je-li pro 49 testovaných pružin průměrný bod zlomu 55 kg – použijte hladinu významnosti $\alpha = 0{,}01$.

(K1) Hypotéza H_0: $\mu = 58$ kg (= střední hodnota je skutečně rovna 58 kg);
H_1: $\mu \neq 58$.

(K2) Kritériem je průměr 49 měření: $\overline{X} = \frac{1}{49} \sum X_i$ . . .

průměr normálně rozdělených veličin.

(K3) Za předpokladu platnosti H_0 má veličina X rozdělení

$$No(\mu = 58; \sigma_{\overline{X}} = \frac{3{,}5}{\sqrt{49}} = 0{,}5)$$

(oba parametry jsou v kilogramech).

(K4) Pro $\alpha = 0{,}01$ kritické hodnoty oboustranného testu vytvářejí interval pro nezamítnutí H_0 roven $\langle -2{,}58; 2{,}58 \rangle$, což po transformaci ne veličinu X dělá

$$\langle -2{,}58; 2{,}58 \rangle \cdot 0{,}5 + 58 = \langle 56{,}71; 59{,}29 \rangle;$$

a to jsme ještě konzervativní ($\alpha = 0{,}01$!!!).

(K5) Průměr zlomu 55 kg pro 49 měření neleží v intervalu pro nezamítnutí H_0, je tedy statisticky významným potvrzením, že výrobce lže (bod zlomu je nižší, nové pružiny tolik nevydrží).

Příklad 12.9. Určete sílu testu z předchozího příkladu, je-li ve skutečnosti střední hodnota průměru pružin bodu zlomu rovna $\mu = 56$ kg (síla testu = pst správného zamítnutí hypotézy H_0, když ve skutečnosti platí H_1).

Řešení: hodnota $\mu = 56$ je tedy hodnotou, ze které budeme v tomto určení síly testu vycházet: Máme spočítat pst zamítnutí H_0, tj. pst, že X naměříme mimo interval $\langle 56{,}71; 59{,}29\rangle$ (viz předchozí příklad). To tedy znamená:

$$
\begin{aligned}
p &= P(X < 56{,}71) + P(X > 59{,}29) = F(56{,}71) + (1 - F(59{,}29)) = \\
&= \Phi\left(\frac{56{,}71 - 56}{0{,}5}\right) + 1 - \Phi\left(\frac{59{,}29 - 56}{0{,}5}\right) = \\
&= \Phi(1{,}42) + 1 - \Phi(6{,}58) \doteq 0{,}9222 + 1 - 1 = 0{,}9222.
\end{aligned}
$$

Jedna drobná poznámka ke statistickým testům: někdy pokud to teoretické předpoklady veličiny dovolují, provádíme tzv. jednostranný statistický test, tj. hypotéza H_1 není tvaru $\mu \neq const$, ale je tvaru $\mu > const$ (respektive tvaru $\mu > const$) – viz příklad v oddílku 13.5.1, nebo příklad 14.4 ve skriptech [1].

(v takových příkladech při určení intervalu pro nezamítnutí H_0 „neusekneme“ $\frac{\alpha}{2}$ obsahu na obou stranách reálné osy, ale usekneme celý obsah α na jedné straně – na té straně, na kterou je od uvažované konstanty testu vychýlena střední hodnota μ v hypotéze H_1).

Do základního kursu statistiky patří též povídání o intervalech spolehlivosti, které často obsahují více informací než statistické testy – ovšem to už není možné stihnout v úvodní přednášce na toto téma (ve skriptech [1] též už intervaly spolehlivosti nejsou zmíněny).

Literatura

[1] **Fajmon, B., Růžičková, I.: Matematika 3. Skriptum FEKT 2003 (identifikační číslo v informačním systému VUT: MAT103).** `http://www.rozhovor.cz/souvislosti/matematika3.pdf.`

[2] **Petr Otipka, Vladislav Šmajstrla: Pravděpodobnost a statistika. Studijní text v rámci distanční formy studijních textů. VŠB - Technická unjiverzita Ostrava. Přístupné na adrese** `http://homen.vsb.cz/~oti73/cdpast1/`.

[3] **Fajmon, B., Koláček, J.: Pravděpodobnost, statistika a operační výzkum. Elektronický text VUT (MAT503), Brno 2005. K dispozici na adrese** `http://www.rozhovor.cz/ma+fy/index.php`.