Popisná statistika

Slovní popis problému

Naším cílem v této úloze bude stručně a přehledně charakterizovat rozsáhlý soubor dat - v našem případě počty bodů z prvního a druhého zápočtového testu z matematiky. Chceme získat představu o středním počtu získaných bodů v jednotlivých písemkách (poloha), jejich variabilitě a rozdělení mezi studenty. Dále bychom rádi posoudili, zda dobrý výsledek v prvním testu bývá doprovázen dobrým výsledkem i v druhém testu.

Data

Máme k dispozici počty získaných bodů z první a druhé paralelkové písemky z předmětu Matematika I (zimní semestr 2015/2016) a to za 762 studentů (všichni studenti, kteří psali první i druhou zápočtovou písemku). Pro spravedlivější porovnání výsledků obou písemek jsme vyloučili ze zkoumání studenty, kteří psali pouze první písemku. Následuje ukázka tabulky s body z písemek.

Student(ka)	1.PP	2.PP
1	43	65
2	37	43
3	27	7
...	...	...
762	44	28

Řešení - teorie

Než představíme jednotlivé charakteristiky, je potřeba si uvědomit, že půjde o charakteristiky **populační**, nikoliv výběrové. Vycházíme totiž z toho, že známe celý základní soubor (výsledky všech studentů) a můžeme tedy dané charakteristiky spočítat přesně. Naopak výběrové charakteristiky se používají v situacích, kdy máme k dispozici pouze malou část základního souboru (získanou náhodným výběrem), a pomocí těchto výběrových charakteristik se snažíme (co možná nejlépe) odhadnout charakteristiky celého základního souboru, které neznáme. Výběrové charakteristiky bychom tedy použili např. v situaci, kdy bychom neměli přístup k databázi dosažených bodů a tak bychom náhodně vybrali 20 studentů a zeptali se jich na jejich výsledky ze zápočtových písemek.

Míry polohy

1. **Aritmetický průměr**
   Jde o nejčastěji používanou míru polohy. Jsou-li x_i $(i = 1...n)$ jednotlivá pozorování (v našem případě počty bodů), aritmetický průměr definujeme takto:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Mezi výhody průměru patří, že je srozumitelný široké veřejnosti (i bez znalosti statistiky) a že jde o hodnotu, která v jistém smyslu nejlépe vystihuje daný soubor dat (minimalizuje součet čtverců odchylek jednotlivých pozorování od této hodnoty). Nevýhodou může být obtížnější interpretovatelnost (hodnota průměru nemusí být v oboru přípustných hodnot pro x - např. průměrný počet bodů nemusí být celočíselný) a vyšší citlivost na odlehlá pozorování (např. když se někdo při zadávání bodů do počítače splete a zapíše třímístné číslo, vychýlí to průměr a tím přestane dobře charakterizovat polohu dat).

Poznamenejme, že vedle aritmetického průměru existují i jiné průměry (např. geometrický, harmonický atd.), ty se však používají ve specifických situacích (např. při výpočtu průměrného tempa růstu, kdy nelze počítat celkový růst za několik období pomocí součtu, nýbrž pomocí součinu růstů v jednotlivých obdobích).

2. **Medián**
Další oblíbenou charakteristikou polohy je medián. Je to, zhruba řečeno, prostřední hodnota ze všech hodnot v souboru, tedy 50 % hodnot by mělo být menších než medián a 50 % hodnot větších než medián. Jedná se tedy o 50% kvantil.

Jednotlivá pozorování x_i v souboru dat uspořádejme vzestupně podle velikosti, tj.

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq ... \leq x_{(n)}. \tag{1}$$

Potom medián $\tilde{x}$ spočítáme takto

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{pro } n \text{ liché,} \\ (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})/2 & \text{pro } n \text{ sudé.} \end{cases}$$

Výhodou mediánu je, že není citlivý na odlehlá pozorování (chybně zadaná hodnota neovlivní medián, nebo jej ovlivní jen velmi nepatrně). Použití mediánu naopak není výhodné v případech, kdy sledovaný znak nabývá jen malého počtu různých hodnot (např. pouze 0 nebo 1), v takovém případě je totiž medián příliš hrubý ukazatel s nízkou vypovídací hodnotou o poloze dat a může být i dosti citlivý na malé změny v datech, jsou-li četnosti jednotlivých hodnot vyrovnané.

Míry varibility

1. **Rozptyl, Směrodatná odchylka**
Nejběžněji používanou mírou variability je rozptyl, potažmo směrodatná odchylka. Rozptyl je definován takto:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Jedná se tedy o průměr druhých mocnin vzdáleností jednotlivých pozorování od průměru. Toto číslo je poněkud obtížněji interpretovatelné (je vyjádřeno v druhé mocnině jednotek, ve kterých měříme x_i), proto se dále definuje směrodatná odchylka jako odmocnina z rozptylu:

$$\sigma = \sqrt{\sigma^2}.$$

Směrodatná odchylka tedy vyjadřuje průměrnou vzdálenost jednotlivých pozorování od průměru. Čím větší je variabilita dat, těm větší je samozřejmě tato vzdálenost.

2. **Rozpětí**
   Rozpětí je vlastně rozdílem dvou kvantilů a udává šířku intervalu, ve kterém je koncentrována předem daná část celého souboru. Čím větší je rozptýlenost dat, tím větší jsou šířky těchto intervalů. $100\alpha\%$ kvantil (značíme x_α) je taková hodnota, která rozděluje daný soubor dat na dvě části - 100α % dat je menších nebo rovna dané hodnotě a $100(1-\alpha)$ % dat je větších než daná hodnota. Pokud by tato hodnota vycházela někde mezi dvěma pozorováními, používají se různé interpolace nebo zaokrouhlování (různé softwary k tomu přistupují různě) - viz. např. výpočet mediánu pro n sudé výše.

   - **Variační rozpětí**
     $$R = x_{max} - x_{min}$$
     Jde o rozdíl největší a nejmenší hodnoty, tedy šířku intervalu, ve kterém se nachází všechna pozorování.

   - **Decilové rozpětí**
     $$R_d = x_{0,9} - x_{0,1}$$
     Decilové rozpětí je rozdíl posledního a prvního decilu a odpovídá šířce intervalu, ve kterém se nachází 80 % hodnot.

   - **Kvartilové rozpětí**
     $$R_q = x_{0,75} - x_{0,25}$$
     Kvartilové rozpětí je rozdíl horního a dolního kvartilu a představuje šířku intervalu, ve kterém se nachází 50 % dat.

   Jelikož jsou rozpětí odvozená od kvantilů, nejsou citlivá na extrémní hodnoty (kromě variačního rozpětí). Nevýhodou oproti momentovým charakteristikám (rozptyl, směrodatná odchylka) je fakt, že nevyužívají veškerou číselnou informaci v datech obsaženou, ale pouze uspořádání dat.

3. **Variační koeficient**
   Na rozdíl od předchozích charakteristik variability je variační koeficient relativní (nikoliv absolutní) mírou variabilitu. Proměnlivost hodnot totiž vztahuje k velikosti těchto hodnot, přesněji směrodatnou odchylku vztahuje k průměru. Je tedy definován takto:
   $$V = \frac{\sigma}{\bar{x}}.$$
   Tento ukazatel vychází z empirického pravidla, že čím větších hodnot sledovaná proměnná nabývá, tím větší bývá její rozptýlení. Pro potřeby srovnání variability různých proměnných nebo souborů se tedy variační koeficient snaží tento jev eliminovat.

Míry (lineární) závisloti

1. **Korelační koeficient (Pearsonův)**
   Jednou z nejužívanějších (ovšem zdaleka ne jedinou) mírou závislosti (či podobnosti) dvou veličin je korelační koeficient. Uvažujme, že u každého objektu (v našem případě student) sledujeme dvě proměnné (v našem případě počty bodů v prvním a druhém testu) x, y. Sílu (a také směr) lineární závislosti mezi těmito proměnnými lze vyjádřit pomocí korelačního koeficientu, který je definován takto:

$$r(x,y) = \frac{\mathrm{cov}(x,y)}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

   Dá se ukázat, že $-1 \leq r(x,y) \leq 1$, přičemž hodnoty blízké -1 indikují silnou negativní závislost (s rostoucím x klesá y), hodnoty blízké 1 indikují silnou pozitivní závislost (s rostoucím x roste i y) a hodnoty blízké nule znamenají absenci lineární závislosti.

Vizualizace dat

Velmi užitečným nástrojem pro pochopení a prezentaci dat jsou grafy. V literatuře existuje nepřeberné množství různých typů grafů zaměřených na zobrazení různých aspektů či vlastností datového souboru. My se zaměříme na zobrazení polohy, variability a rozdělení počtu bodů v jednotlivých testech a na zobrazení souvislosti mezi počtem bodů v prvním a v druhém testu.

1. **Krabicový graf (Box plot)**
   Hlavním cílem krabicového grafu je znázornit polohu a variabilitu datového souboru a to vykreslením vybraných kvantilů (mediánu, kvartilů). V omezené míře také vypovídá něco o rozdělení dat a napomáhá identifikovat odlehlé hodnoty. Krabicový diagram budeme kreslit na výšku a to tímto postupem (pro ilustraci viz. obr. níže):

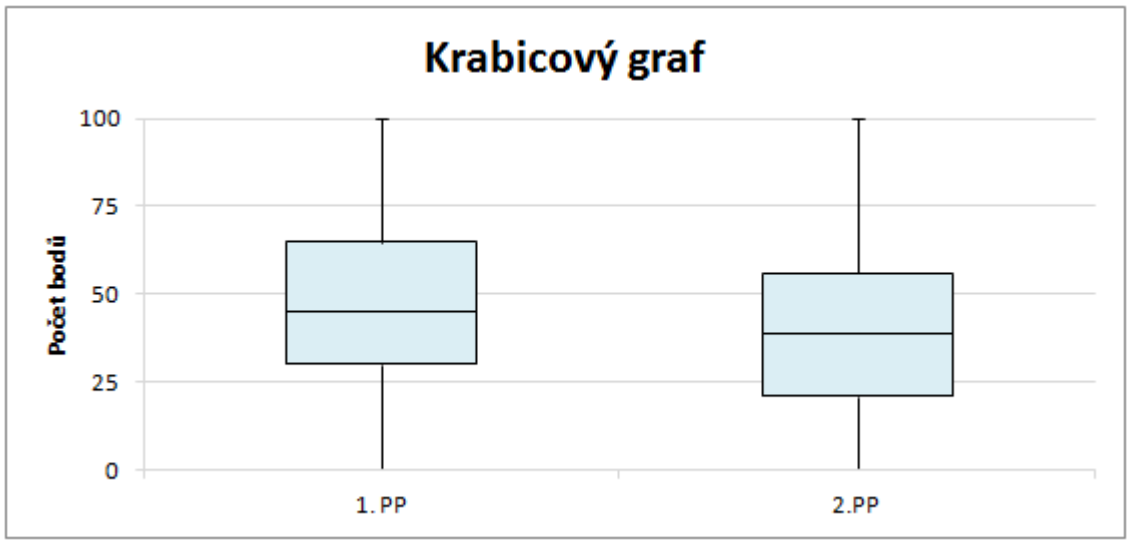
   (a) Vykreslí se základní obdélník (krabice), jehož dolní a horní strana odpovídají dolnímu a hornímu kvartilu $(x_{0,25}, x_{0,75})$. Výška krabice tedy odpovídá kvartilovému rozpětí.

   (b) Uvnitř krabice se udělá vodorovná čára ve výšce mediánu ($\tilde{x} = x_{0,5}$). Tím jsme rozdělili soubor dat na čtyři stejně velké (co do četnosti) části.

   (c) Ze spodní části krabice se pak vede svislá úsečka směrem dolů (dolní fous), přičemž její dolní mez odpovídá takové nejmenší hodnotě ze souboru, která leží pod krabicí ve vzdálenosti nejvýše 1,5 násobku výšky krabice. Podobným způsobem se sestrojí horní fous krabice.

   (d) Hodnoty ze základního souboru, které leží mimo tyto fousy (buď nad nebo pod nimi) se zobrazí v grafu individuálně (každá zvlášť) a odpovídají odlehlým pozorováním.

   Poznamenejme, že chceme-li porovnat polohu a variabilitu více různých datových souborů (popř. více proměnných v jednom souboru), bývá výhodné nakreslit všechny krabicové grafy do jednoho obrázku.

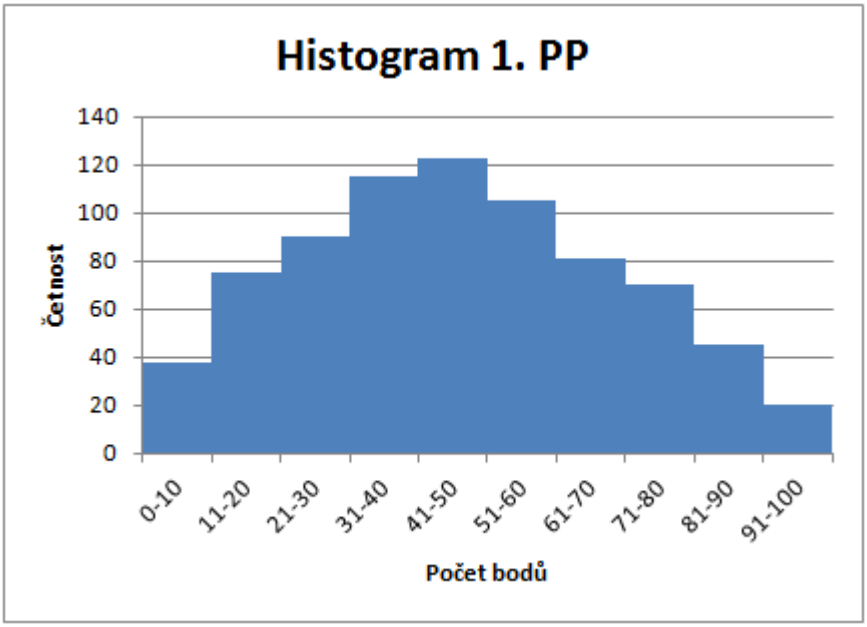


2. **Histogram**
   Podobně jako krabicový graf i histogram dokáže do určité míry znázornit polohu a variabilitu dat. Jeho hlavním cílem je však přehledně zobrazit rozdělení dat v souboru.

   Prvním (a nejdůležitějším) krokem je rozdělit jednotlivá pozorování do intervalů dle zjištěných hodnot. Intervaly jsou zpravidla ekvidistantní (mají stejnou šířku), musí pokrývat celý obor hodnot sledované proměnné a v každém intervalu by měl být dostatečný počet pozorování. Finální volba intervalů je více méně arbitrární, protože neexistuje žádné univerzální pravidlo pro jejich konstruování. V literatuře se takových (zpravidla empirických) pravidel dá najít více a žádné z nich není to jediné nejlepší. Vzhledem k tomu, že vzhled výsledného grafu může být dosti citlivý na volbě intervalů, bývá někdy dobré s hranicemi intervalů trochu experimentovat a nakonec vybrat tu možnost, která nejlépe zobrazí rozdělení dat.

   Jakmile máme zvolené intervaly, zjistíme, kolik hodnot z datového souboru leží v jednotlivých intervalech. Tím získáme tabulku četností.

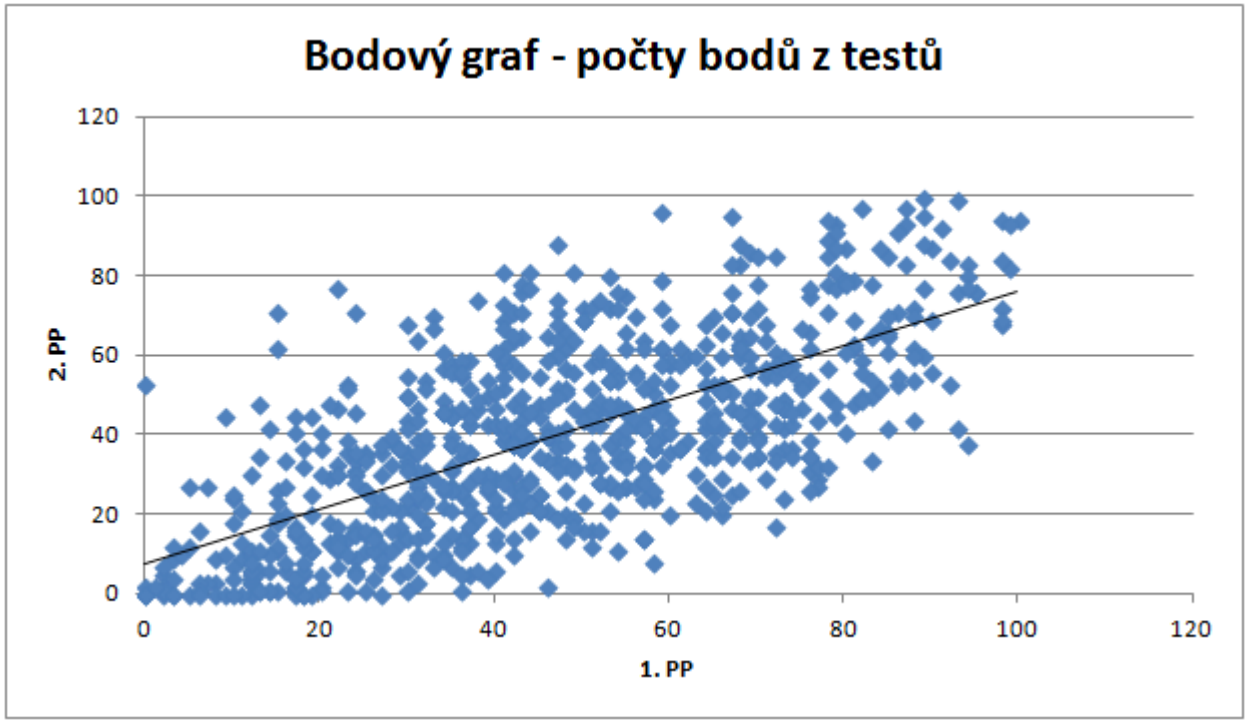
   Z tabulky četností pak vytvoříme sloupcový graf a to tak, aby šířka jednotlivých sloupců odpovídala šířce intervalu a aby se sloupce svými stranami dotýkaly. V případě ekvidistantních intervalů výška sloupce odpovídá četnosti hodnot v daném intervalu. Pozor, pokud však intervaly nejsou ekvidistantní, je třeba výšky sloupců volit tak, aby obsahy (nikoliv výšky) jednotlivých sloupců odpovídaly příslušným četnostem. Pro ilustraci uvádíme příklad histogramu níže.



3. **Bodový graf (Scatter plot)**
   Bodový graf (též nazývaný korelační diagram) je velmi jednoduchý grafický nástroj pro znázornění vzájemné závislosti (či podobnosti) dvou proměnných. Uvažujme, že u každého objektu (v našem případě student) sledujeme dvě vlastnosti, které kvantifikujeme pomocí dvou proměnných (v našem případě počet bodů z prvního testu a počet bodů z druhého testu). Každému objektu tak odpovídá dvojice čísel a můžeme jej tedy znázornit jako bod v rovině (x-ová souřadnice daného bodu bude opovídat hodnotě první proměnné - např. počet bodů v prvním testu, zatímco y-ová souřadnice bude odpovídat hodnotě druhé proměnné - počet bodů z druhého testu). Podle rozložení jednotlivých bodů v rovině poznáme, zda vyšším hodnotám jedné proměnné odpovídají i vyšší hodnoty druhé proměnné, či nižší hodnoty druhé proměnné, popř. zda je závislost složitější (nemonotónní). Může se také ukázat, že žádná zjevná souvislost mezi proměnnými není.

   Pro větší názornost se často do bodového diagramu přidává regresní přímka (přímka, která na základě kritéria nejmenších čtverců nejlépe aproximuje zakreslené body).

   Ukázka bodového grafu s regresní přímkou viz. níže.


Bodový graf - počty bodů z testů
120
100
80
60
40
20
0
2. PP
0
20
40
60
80
100
120
1. PP