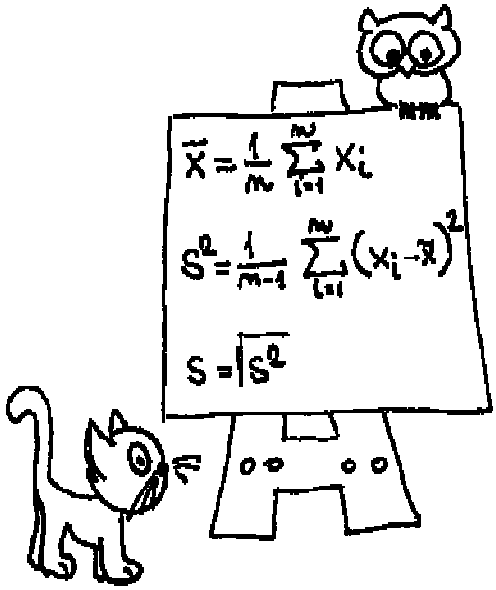
Statistika nuda je…

Jak je to s tou průměrnou mzdou, co ji průměrný pracující nikdy neuvidí? Proč může přijít stoletá voda dva roky po sobě? A proč mi při zpracování dat nestačí znát jen průměr, ale motají se k tomu ještě další pojmy?

Musíme začít pomalu a postupně. Nejprve co to je **základní soubor** a **výběrový soubor**. Vysvětlíme si to na výšce české ženy starší 18 let. Základním souborem jsou VŠECHNY ženy splňující podmínku věku a národnosti. Bude jich moc a asi se od všech najednou údaje o jejich výšce nedají zjistit (než je zjistím, tak se mi část populace obmění :o)). Proto raději pracujeme se souborem výběrovým. Ten může mít různý rozsah (údaje od 100 žen, 258 žen, 1000 žen…) a hodně záleží i na provedení výběru (data by měla určitě být nezávislá – tj. navzájem se neovlivňující a měla by postihovat co nejvěrněji celou populaci). Na základě zpracování dat z výběrového souboru můžeme vyslovit závěry o celém souboru základním. Správnost našich závěrů bude ovlivňovat jednak rozsah, jednak kvalita výběru.



Co ze získaných dat můžeme zjistit? Především takzvané charakteristiky polohy dat (aritmetický průměr, modus, medián) a charakteristiky variability dat (rozpětí, rozptyl, směrodatná odchylka). **Aritmetický průměr** umí vypočítat asi všichni. Sečtou se naměřené hodnoty a tento součet se vydělí počtem měření. Ve výsledku jsou obsažena všechna data. Je to nejpoužívanější a nejznámější charakteristika, ale sama o sobě nestačí (viz rozruch kolem každého zveřejnění aktuální průměrné mzdy…). Pro pochopení celého problému potřebuji řadu dalších charakteristik.

Modus je takzvaná módní hodnota. Je to číslo, které se nejčastěji mezi naměřenými daty opakuje. Pokud mám malý rozsah výběru (málo naměřených čísel), tak se stanovit často nedá. Prostě se tam žádná hodnota neopakuje, nebo se nám tam víc hodnot opakuje třeba dvakrát. **Medián** je takzvané prostřední měření – pokud naměřená data srovnám podle velikosti, je to hodnota ležící uprostřed. Pokud je dat lichý počet, je to snadné – prostřední číslo je jedno. Pokud je hodnot sudý počet, pak je mediánem průměr z obou prostředních čísel.

Například: z šesti naměřených hodnot výšky v centimetrech 165, 169, 174, 169, 172, 174 máme určit modus a medián. Je pohodlnější srovnat si hodnoty od nejmenší po největší: 165, 169, 169, 172, 174, 174. Modus bychom nestanovili, opakují se nám dvě hodnoty (169, 174) dvakrát. Medián leží uprostřed. Protože mám sudý počet hodnot, uprostřed leží 169 a 172. Průměr těchto dvou čísel – 170,5 – je tedy medián.

Ještě jedna data – už srovnaná podle velikosti: 158, 161, 168, 168, 168, 172, 178. Modus je 168 (třikrát se opakuje), medián je také 168, neboť je to čtvrtá - tedy prostřední hodnota z řady sedmi naměřených a srovnaných dat.

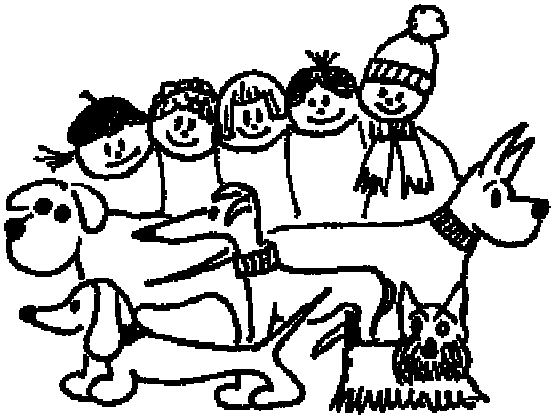
Charakteristiky variability vyjadřují, jak jsou data kolem střední hodnoty rozptýlena. Pokud je variabilita malá, znamená to, že všechna naměřená čísla leží blízko sebe. Pokud je nulová, jsou všechna naměřená čísla stejná. **Rozpětí** se často nepoužívá, ale zase se nejsnáze vypočítá. Je to rozdíl mezi největším a nejmenším naměřeným číslem. Nejběžnější charakteristikou variability je **směrodatná odchylka**, která se vypočítá odmocněním rozptylu. A jsme u nehezkého vztahu pro **rozptyl**. Pro jeho výpočet musíme mít předem spočítaný aritmetický průměr. Pak vypočítáme rozdíly mezi naměřenými hodnotami a průměrem a všechna takto získaná čísla (je jich stejně jako naměřených dat) umocníme na druhou. Potom je sečteme a výsledek vydělíme počtem měření sníženým o jedničku. Pro malý počet hodnot to jde, pro větší počet je snazší využití kalkulačky či

počítačového programu … I na běžných „školních“ kalkulačkách se dá snadno zadáním dat a zmačknutím patřičných čudlíků získat aritmetický průměr a směrodatnou odchylku.

A teď troška praxe …

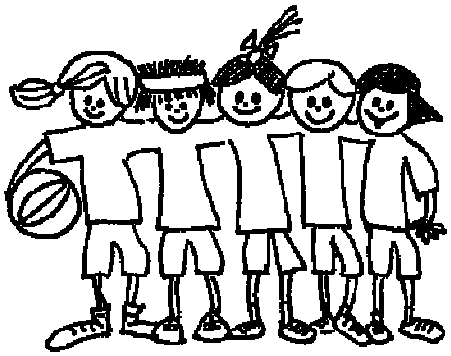
Studentky mají za domácí úkol zjistit výšku 5 žen a spočítat aritmetický průměr, medián, rozpětí, rozptyl a směrodatnou odchylku. Modus z pěti měření se většinou získat nedá.

První studentka přemýšlí: *Pokud vezmu svoji výšku a výšku maminky a babičky, budou data navzájem závislá. Geny jsou mršky. Musím tedy data získat jinde. Vezmu psa a vyrazím do parku. Tam potkám spoustu paniček spřátelených psíků a můžu se klidně zeptat, kolik která měří. A pak si vyberu výšky od různě starých žen tak, aby byly asi 10 let od sebe. Tím postihnu i různý věk.* Vezme psa, tužku a papír a vyrazí. Z tohoto „rozumného“ přístupu získá hodnoty 163, 168, 171, 173, 175 cm.

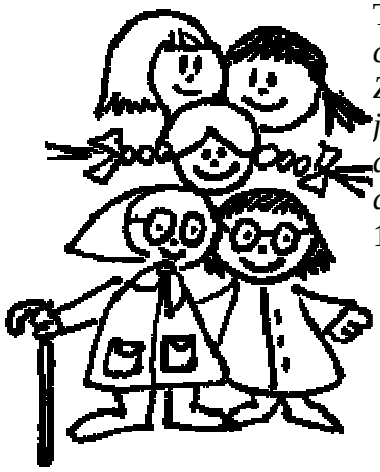


Průměr vypočítám, když součet hodnot 850 vydělím počtem měření – pěti. Průměr vyjde 170 cm. Medián je hodnota 171 cm. Rozpětí je rozdíl mezi největším a nejmenším číslem, tedy 12 cm. Pro rozptyl nejprve spočítám odchylky naměřených dat od průměru (163-170), (168-170), (171-170), (173-170), (175-170). Odchylky umocním na druhou a sečtu: 49+4+1+9+25 = 88. Nakonec tento součet vydělím čtyřmi (to je o jedničku míň než počet dat) a rozptyl tak je 22 centimetrů čtverečních. Směrodatná odchylka se získá odmocněním rozptylu a vyjde 4,7 cm.

Druhá studentka nemá čas: *Jdu na basket a nemám na nějakou blbou matiku čas. Ta úča bude čučet – vyberu jí ze soupisek stejně vysoké hráčky a klidně jí je dodám i se jmény, kdyby snad měla kecy. A s tím průměrem jí pěkně zacvičíme. A bude i jasný modus a s rozptylem se nemusím počítat. Ať žijí maxiženy…* A donese takto vypracovaný úkol: 181, 181, 181, 181, 181 cm.

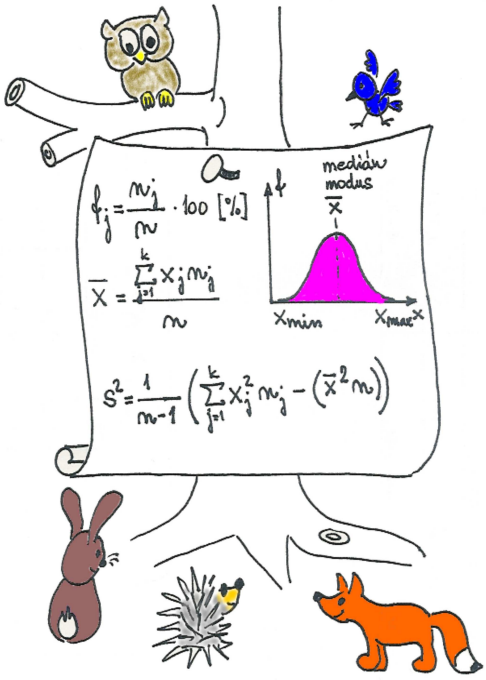


Průměr, modus i medián jsou 181 cm, rozpětí, rozptyl i směrodatná odchylka jsou nulové. Přístup je pochopitelně špatný. Data byla uměle vybrána tak, aby hodnoty byly shodné. A pak zatímco venčení psů není koníček ovlivňující výšku postavy, tak u basketbalu už to tak jednoznačné nebude. Takže i náhodně oslovená děvčata na tréninku by nedodala zrovna vhodně získaná data.



Třetí studentka na to jde zase jinak. *Asi bude zajímavé získat hodnoty co nejrůznější. Ty dvě staré dámy z přízemí jsou opravdu maličké. Zeptám se jich. Švagrová zase trošku přerostla a její nejlepší kámoška je ještě o dva centíky vyšší. No a já budu ten zlatý střed. Věk postihnu docela slušně – švagrovka s kámoškou jsou o dost starší než já a ty dvě babči už jsou dávno v důchodu.* A dodá hodnoty 154, 152, 180, 182, 172 cm.

Průměr je 168 cm, medián 172 cm, rozpětí 30 cm, rozptyl 202 cm na druhou a směrodatná odchylka 14,2 cm. Ani hon za extrémy není nejlepší nápad. V celé populaci je hodně těch „běžných“ výšek a tady z pěti hodnot jsou čtyři poměrně silně odchýlené. Průměr sice vychází sympaticky, ale ta variabilita…



Pokud bychom data získaná celou třídou (třeba 30 studentů x 5 hodnot, tj. 150 změřených žen) vyhodnotili dohromady, získali bychom už docela slušný náhled na rozložení výšek žen v populaci.

Ale to, co šlo dobře pro 5 nebo i 15 hodnot, už by nám tak snadno nefungovalo pro velká množství zjištěných dat. A my jsme si přece hned na začátku vysvětlili, že výsledky závisí nejen na kvalitě, ale i kvantitě výběrového souboru.

S tím, co jsme se naučili, by to šlo, ale trvalo by nám to dlouho. Jen ta představa, jak ta data rovnám podle velikosti, abych našla modus a medián… jak se pracně propočítávám k rozptylu… prostě to uděláme jinak.

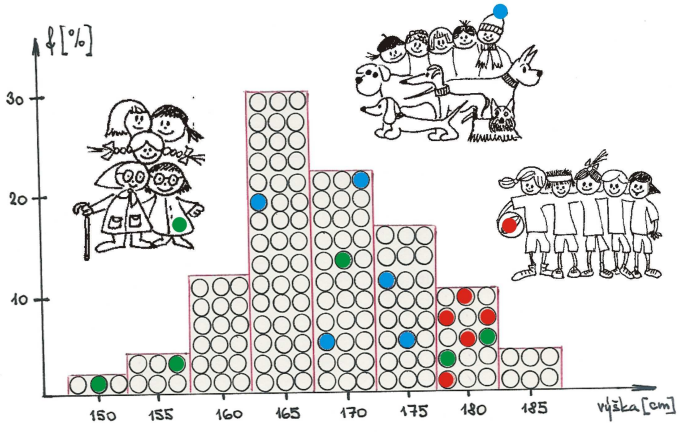
Vytvoříme takzvaná **sdružená data**. Prostě si řeknu, že ty ženy podle výšky rozdělím do skupinek = intervalů, kterým říkáme třídy. Tříd by nemělo být moc (pak je s tím zbytečně moc práce), ani málo (pak už je výsledek hóódně zaokrouhlený). V praxi se říká – ne méně než 5 (raději 7), ne víc jak 20.

j	x_d - x_h	x_j	n_j	f_j [%]
1	148 – 152	150	3	2
2	153 – 157	155	6	4
3	158 – 162	160	18	12
4	163 – 167	165	45	30
5	168 – 172	170	33	22
6	173 – 177	175	24	16
7	178 – 182	180	15	10
8	182 - 187	185	6	4
		Součet :	150	100

A tak se nám naše bádání o výšce ženy smrskne do docela přehledné tabulky. První sloupeček je pořadí třídy (počet tříd se značí *k*, pro nás *k* = 8), druhý jsou intervaly, do kterých jsem roztřídila získané hodnoty výšek. Třetí sloupeček *xj* je velmi důležitý – **leží uprostřed třídy** a je to významný prvek, který ve výpočtech nahradí všechna měření v intervalu. Počet žen, které „spadly“ do té které třídy se značí *nj*. Třeba všechny basketbalistky jsou ve třídě sedmé. A že jich tam není 5, ale patnáct? Nezapomeňte, že najednou už zpracováváme data z celé třídy a i jiní znali někoho vysokého. Dole pod sloupečkem máme součet *n* = 150.

Co to je *fj*? Je to **relativní četnost**. Říká, kolik procent žen patří do té které třídy. Vzoreček je na tabuli, ale není to nic jiného, než výpočet procent z celku (tedy součet *f* nám musí dát 100%). Jestliže v první třídě těch nejmenších (jejich výška kolísá kolem 150 cm) máme tři ženy, jejich počet (3) vydělím *n* (150) a výsledek vynásobím 100. Vyjdou mi 2%.

Relativní četnost se vynáší do často používaného grafu – **histogramu**. Najednou se nám 150 žen (koleček) pěkně uspořádá do přehledného obrázku. Do grafu jsem zakreslila i našich 15 „známých“ postaviček z minula (pochopitelně, že normálně se do grafu ta kolečka nekreslí, to jen na poprvé pro lepší představu :o))

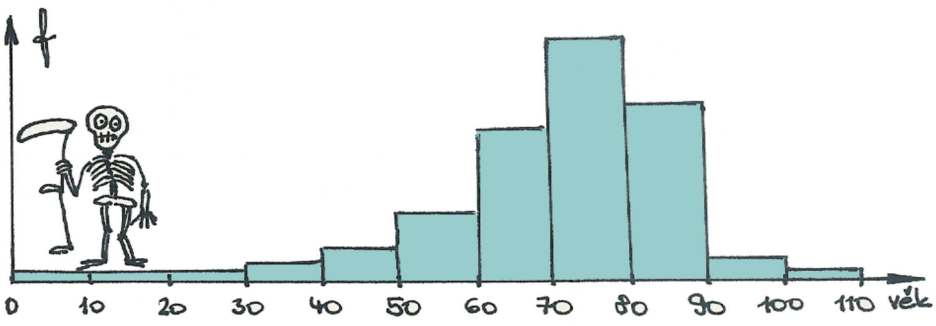


Aritmetický průměr se zjistí tak, že se roznásobí prostředek třídy *xj* počtem prvků

ve třídě *nj*. Tyto násobky se sečtou (bude jich stejně jako tříd) a součet se vydělí počtem *n*. Pozor – dělí se to *n* a nikoli *k* (to je častá chyba). Modus a medián budeme jen odhadovat. Tam, kde je nejvyšší sloupeček v histogramu, tam někde bude modus. A my ho stanovíme jako střed (*xj*) té nejčetnější třídy. Modus nám vyjde 165 cm. Medián je přece hodnota uprostřed setříděného souboru, a tak budu sčítat *fj*, až se dopočítám nad 50%. Ve čtvrté třídě je to 2+4+12+30, tj. jen 48%. Takže medián leží až ve třídě páté – tam jsou data mezi 48% a 70%, pro jednoduchost ho určíme jako střed mediánové třídy – 170 cm.

Určitě nás nepřekvapí, že průměrná výška, modus i medián jsou zhruba uprostřed naměřených dat. Protože v populaci je málo procent těch „mrňavých“ i těch „přerostlých“. Kdybychom žen změřili opravdu velký počet a udělali roztřídění do mnoha tříd, histogram by se nápadně podobal tomu malému grafu na tabuli. Je to **Gaussova křivka**, která popisuje normální rozdělení hodnot.

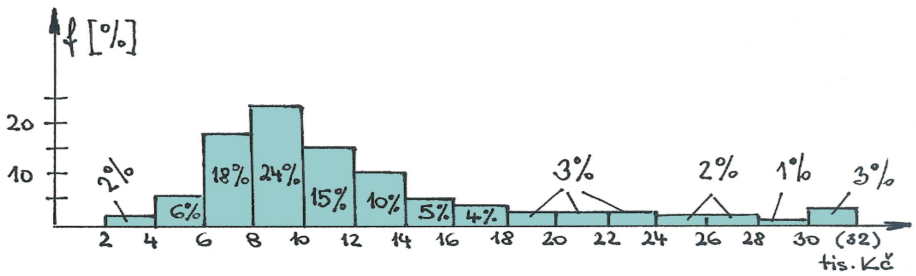
Normální rozdělení je symetrické, průměr, modus i medián jsou stejně veliké a leží uprostřed naměřených hodnot. Ne nadarmo se tomuto rozdělení říká normální – opravdu je nejrozšířenější, dobře popisuje velké množství jevů v přírodních i společenských vědách. A protože je tak běžné, nějak automaticky ho očekáváme i tam, kde nefunguje. Třeba u té průměrné mzdy… On totiž ten průměr je uprostřed jen u symetrických rozdělení.



Nyní trošku morbidní příklad. U dat dosaženého věku jsme docela rádi, že nemá normální rozdělení, to by nám ta průměrná délka života vycházela kolem 55 let. Průměr, medián i modus jsou zde totiž posunuty k vyšším hodnotám.

No a na závěr data zešikmená obráceně. Data (*nejsou přesně dle skutečnosti – sice vychází z reálu, ale mám je upravené pro snadné výpočty)* na posledním grafu vyjadřují průměrný hrubý příjem na hlavu v rodině. Začíná u existenčního minima – to je těsně nad 2000 Kč. Z dat vyplývá, že nejvíce – 24% rodin – má příjem na osobu mezi 8 a 10 tisíci korunami. Do příjmu 30 tisíc na osobu je 97 % všech domácností. Jen 3% mají příjem vyšší. Abych mohla vypočítat průměr, shrnula jsem ta 3 % nejbohatších rodin do skupiny mezi 30 a 32 tisíci. To je pochopitelně zjednodušené, jejich příjem je možná i několikanásobně vyšší. I když se jedná o malé procento rodin, toto zjednodušení nám průměr ještě trochu sníží.

Modus určíme jako prostředek nejčetnější třídy – tedy 9.000 Kč. To je tedy částka, kterou má nejvíce rodin k dispozici. Medián – tedy tu částku, kterou má k dispozici „prostřední“ rodina umíme najít také. 50 % je v prvních čtyřech třídách (2+6+18+24), 50 % v následujících 11 třídách. Medián je tedy mez mezi čtvrtou a pátou třídou - 10.000 Kč. A průměrnou hodnotu jsem vypočítala na 12.000 Kč…



Tak tedy vyšlo, že 65% rodin má na člena menší příjem, než je průměrný příjem na osobu. Obdobně to platí i u průměrné mzdy. Na průměrnou mzdu dosáhne dokonce jen asi 25 % pracujících. Ale nás, co už víme, že ne všechna data jsou symetricky rozdělena (nejsou „normální“), to nepřekvapí. A už také víme, že u nesymetricky rozložených veličin je podstatně rozumnější udávat vedle průměru i medián – tedy toho „středního pracujícího“, či modus – tedy toho „nejběžnějšího“.