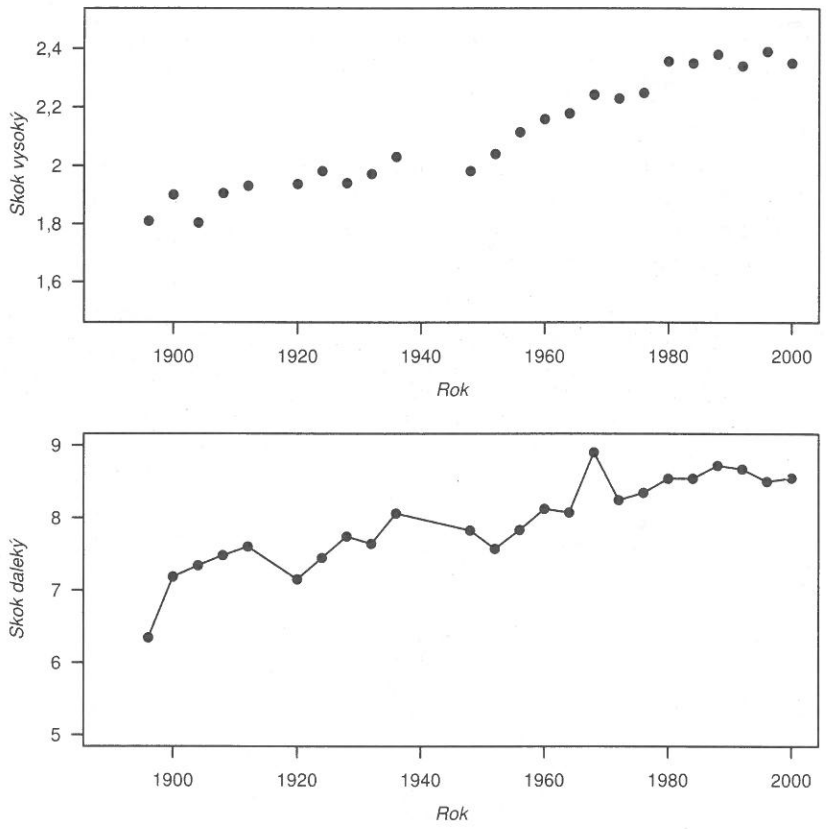
Obr. 3.6 Příklad zobrazení trendu – mistrovské výkony ve skoku vysokém a dalekém na OH



Pokud chceme znázornit trend v datech v závislosti na časovém faktoru, použijeme graf trendu. Na obrázku 3.6 jsou znázorněny výkony ve skoku vysokém a dalekém, za něž sportovec získal zlatou medaili na olympijských hrách. Data doplňujeme proloženou přímkou, jinou proloženou křivkou nebo je spojíme úsečkou.

V této knize poznáme mnoho dalších možností grafického znázornění dat.

3.2 Míry centrální tendence

Statistické zpracování dat pomocí tabulek a grafů usnadňuje jejich vizuální analýzu a celkové posouzení datové konfigurace. Pro další zpracování však potřebujeme data vhodně kondenzovat. Proto se počítají různé číselné charakteristiky – **popisné statistiky**, které zachycují různé aspekty dat. Jedná se především o charakteristiky centrální tendence a rozptýlenosti, ale i o další charakteristiky jako šikmost nebo špičatost rozdělení dat.

Míry centrální tendence se snaží charakterizovat typickou hodnotu dat. (Říká se jim také střední hodnoty, resp. míry střední hodnoty nebo míry polohy – protože určují, kde na číselné ose je vzorek rozložen.) Nejznámější z nich jsou aritmetický průměr, medián a modus.

3.2.1 Aritmetický průměr

Aritmetický průměr je definován jako součet všech naměřených údajů vydělený jejich počtem. Označujeme ho pomocí symbolu $\bar{x}$ nebo M. Výpočet má tedy podobu:

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

Poznamenejme, že stejný výpočet vyjadřují zkrácené zápisy:

$$\bar{x} = \frac{\sum_i x_i}{n} \qquad \text{nebo} \qquad \bar{x} = \frac{\sum x_i}{n}$$

kde znak $\sum$ symbolizuje součet hodnot x_i pro všechny možné hodnoty indexu i.

Pro modelová data $\{2; 8; 9; 10; 1; 0; 5\}$ má průměr hodnotu

$$\bar{x} = \frac{2+8+9+10+1+0+5}{7} = 5.$$

Aritmetický průměr je optimální charakteristikou typické hodnoty množiny dat pro následující vlastnosti:

1. Součet odchylek měření od průměru se rovná nule – např. pro data z příkladu jsou odchylky $\{-3; 3; 4; 5; -4; -5; 0\}$ a jejich součet je číslo nula.
2. Fyzikálně si aritmetický průměr představujeme jako těžiště dat – součet dat pod průměrem je stejný jako součet dat nad průměrem, oba součty jsou v rovnováze. Součet vzdáleností od průměru hodnot nižších než průměr má být roven součtu vzdáleností od průměru hodnot vyšších než průměr. Každá hodnota má stejnou váhu.

3. Výraz $\sum (x_i - b)^2$ je nejmenší vzhledem k parametru b, jestliže b se rovná aritmetickému průměru. Výraz $\sum (x_i - b)^2$ jistým způsobem charakterizuje celkovou chybu, které se dopouštíme, když chceme nahradit všechny údaje jednou hodnotou b. Tvrzení vyjadřuje, že $\bar{x}$ odhaduje data s nejmenší chybou, přičemž za míru chyby považujeme kvadratickou odchylku.

Pokud máme několik průměrů spočítaných z různých podmnožin dat a známe příslušné počty měření n_i, lze vypočítat celkový průměr ze všech dat jako **vážený průměr:**

$$\bar{x} = \frac{\sum n_i \bar{x}_i}{\sum n_i}$$

Podobně se počítá průměr pro data zadaná četnostním způsobem pomocí frekvenční tabulky, v níž pro hodnoty x_i jsou ještě zadané jejich četnosti výskytu f_i:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

3.2.2 Medián a modus

Medián (označovaný *Me* nebo $\tilde{x}$) znamená hodnotu, jež dělí řadu *podle velikosti seřazených výsledků* na dvě stejně početné poloviny. Jestliže n je sudé číslo, pak *Me* je jakékoli číslo z intervalu $(x_{n/2}, x_{n/2+1})$. Jednoznačněji

$$Me = 0{,}5(x_{n/2} + x_{n/2+1}).$$

Jestliže n je liché číslo, pak

$$Me = x_{(n+1)/2}.$$

Pro modelová data seřazená podle velikosti {0; 1; 2; 5; 8; 9; 10} zjistíme hodnotu mediánu 5. Medián je na rozdíl od aritmetického průměru málo citlivý k odlehlým hodnotám. Představme si třeba jakkoli velikou změnu nejmenší hodnoty směrem dolů: medián zůstane stejný, ale jistě se změní průměr. Medián *Me* má optimální vlastnost v tom smyslu, že minimalizuje součet absolutních odchylek měření od zvoleného čísla. Tato vlastnost je analogická vlastnosti aritmetického průměru $\bar{x}$, který minimalizuje součet kvadratických odchylek.

Modus nebo modální hodnota je hodnota, jež se v datech vyskytuje nejčastěji. Tato charakteristika nalézá uplatnění především u kategoriálních dat. Symbolicky se označuje $\hat{x}$ nebo *Mo*. V případě spojitých dat se odečítá pomocí sestrojeného histogramu, kdy se spočítá jako průměr z krajních hodnot intervalu, který obsahuje nejvíc dat. Pokud existuje v histogramu více vrcholů, udáváme je všechny. Říkáme pak, že rozdělení je dvou-, tří- nebo obecně vícevrcholové.

3.2.3 Použití měr centrální tendence

Rozhodnutí, kterou charakteristiku průměru nebo polohy použijeme při popisu dat, závisí na cílech a předpokladech analýzy. První omezení představuje úroveň měřítka měření. Aritmetický průměr se jen zřídka používá pro hodnocení dat, jejichž měřítko není intervalové nebo poměrové. Pro tyto dvě měřítka často uvádíme všechny střední hodnoty a všímáme si, proč se liší. Jestliže jsou data symetricky rozdělená, všechny tyto charakteristiky jsou přibližně stejné. Uvedeme základní pokyny pro užití středních hodnot.

Aritmetický průměr se má používat:

- jestliže data jsou získána minimálně v intervalovém měřítku (tzn. průměr neužíváme pro údaje kategoriální);
- jestliže je rozdělení symetrické;
- jestliže chceme použít statistické testy.

Medián se má použít:

- jestliže data jsou získána minimálně v ordinálním měřítku;
- jestliže chceme znát střed rozdělení dat;
- jestliže data mohou obsahovat odlehlé hodnoty;
- jestliže rozdělení dat je silně zešikmené.

Modus se má použít:

- jestliže rozdělení má více vrcholů;
- jestliže chceme získat o rozdělení jenom základní přehled;
- jestliže se slovem „průměrně" míní nejčastější hodnota.

3.3 Míry rozptýlenosti

Náhodně proměnlivé údaje nestačí charakterizovat jenom střední hodnotou. Omezenost středních hodnot spočívá v tom, že udávají pouze to, kolem jaké hodnoty se data „centrují", resp. které hodnoty jsou nejčastější. Data se stejnou střední hodnotou mohu mít různou rozptýlenost. Velikost proměnlivosti dat zachycujeme vhodně vybranou mírou rozptýlenosti dat. Existuje mnoho měr rozptýlenosti a záleží na okolnostech, kterou nebo které použijeme. Numerické charakteristiky tvaru rozdělení dat mají důležitý význam při kondenzaci dat do několika málo popisných údajů; pamatujme však, že nejlepší představu o datech nám poskytuje graf.

3.3.1 Variační rozpětí

Přestože se maximální a minimální hodnota uvádějí pravidelně při popisu dat, **variační rozpětí** R se počítá zřídka, ačkoli je jeho zjištění jednoduché:

$$R = x_{max} - x_{min}$$

Nevýhodou variačního rozpětí je velká citlivost vůči odlehlým hodnotám.

Pro modelová data {2; 8; 9; 10; 1; 0; 5} má R hodnotu 10.

3.3.2 Rozptyl a směrodatná odchylka

Obě tyto charakteristiky popíšeme v jednom odstavci, protože spolu úzce souvisejí. Oběma je společná vlastnost, že na rozdíl od variačního rozpětí R využívají při výpočtu všechny údaje a obě se vztahují k aritmetickému průměru – měří rozptýlenost dat kolem aritmetického průměru dat. Dávají větší váhu extrémnějším hodnotám než průměrná absolutní odchylka.

Rozptyl je definován jako průměrná kvadratická odchylka měření od aritmetického průměru, přičemž při průměrování této odchylky dělíme číslem $(n - 1)$:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Při větších rozsazích není rozdíl mezi dělením číslem n nebo $n - 1$ významný. Dělení číslem n se použije, jestliže rozptyl počítáme pro všechny prvky populace. Při výpočtech někdy vycházíme ze vzorce

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n - 1},$$

který vede ke stejné hodnotě.

Pro modelová data {2; 8; 9; 10; 1; 0; 5} má rozptyl hodnotu

$$s^2 = \frac{(2-5)^2 + (8-5)^2 + (9-5)^2 + (10-5)^2 + (1-5)^2 + (0-5)^2 + (5-5)^2}{6}$$

$$= 16{,}66.$$

Rozptyl se především používá v inferenční statistice při výpočtu různých testovacích statistik. Počítá se pomocí čtverců odchylek dat od průměru, proto má jiný rozměr než původní data.

Směrodatná odchylka s je odmocnina z rozptylu a vrací míru rozptýlenosti do měřítka původních dat:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Pro modelová data {2; 8; 9; 10; 1; 0; 5} má směrodatná odchylka hodnotu

$$s = \sqrt{\frac{(2-5)^2+(8-5)^2+(9-5)^2+(10-5)^2+(1-5)^2+(0-5)^2+(5-5)^2}{6}} = 4{,}08.$$

Při pokusu porozumět výpočtu směrodatné odchylky si všímáme jednotlivých operací:

1. Nejdříve vypočteme jednotlivé odchylky od průměru $(x_i - \bar{x})$, které pro daný údaj vyjadřují, jak se liší od typické hodnoty.
2. Čtverec odchylky (umocnění na druhou) převádí záporné odchylky na kladná čísla a zároveň větším odchylkám dává větší váhu. Například odchylce -2 dává váhu 4, ale odchylce 3 dává váhu 9.
3. Součet (suma) čtverců odchylek zachycuje všechny odchylky jedním číslem.
4. Dělením číslem $(n - 1)$ počítáme průměr s kvadratických odchylek.
5. Odmocnina převádí druhou mocninu do původního měřítka dat.

Základní vlastnosti směrodatné odchylky:

- směrodatná odchylka měří rozptýlenost kolem průměrů a má se používat jenom tehdy, když průměr je vhodný jako míra střední hodnoty (viz s. 95);
- $s = 0$ pouze tehdy, když se všechna data rovnají stejné hodnotě, jinak $s > 0$;
- stejně jako průměr $\bar{x}$ je i směrodatná odchylka s silně ovlivněna extrémními hodnotami – jedna nebo dvě odlehlé hodnoty zvětšují silně s;
- jestliže je rozdělení dat silně zešikmené, směrodatná odchylka neposkytuje dobrou informaci o rozptýlenosti dat – v takovém případě používáme kvantilové míry, které vysvětlíme v další části odstavce.

Jestliže chceme posoudit relativní velikost rozptýlenosti dat vzhledem k průměru, použijeme **koeficient variace** neboli **variační koeficient** VK. Počítáme ho, když chceme porovnat rozptýlenost dat skupin měření stejné proměnné s různým průměrem, nebo v těch případech, kdy se mění velikost směrodatné odchylky tak, že je přímo závislá na úrovni měřené proměnné ($s = k\bar{x}$, kde k je konstanta):

$$VK = \frac{s}{\bar{x}}$$

Pro modelová data {2; 8; 9; 10; 1; 0; 5} má koeficient variace hodnotu

$$VK = \frac{4{,}05}{5} = 0{,}85$$

Někdy se uvádí v procentech:

$$VK = \frac{s}{\bar{x}} \times 100\%$$

Pro naše modelová data má VK hodnotu 85 %.

3.3.3 Míry rozptýlenosti založené na empirických kvantilech

Empirický kvantil je hodnota, pod níž leží definovaná část údajů. U empirického kvantilu udáváme jeho hladinu q a označujeme ho symbolem x_q. Parametr q je z intervalu hodnot $0 < q < 1$. Hladina q určuje relativní podíl údajů, které se nacházejí pod empirickým kvantilem x_q.

Pro data můžeme vypočítat mnoho různých empirických kvantilů. Některé z nich se však používají pravidelně. Slouží k popisu jednotlivých částí rozdělení dat a vypočítávají se z nich také míry rozptýlenosti.

Výpočet empirického kvantilu s hladinou q se děje tímto způsobem: Nechť $j = [qn]$, kde operace [.] znamená zaokrouhlování na nejbližší menší celé číslo. Jestliže $qn = [qn]$, pak $x_q = (x_j + x_{j+1})/2$, jinak $x_q = x_{j+1}$, kde x_j ($j = 1, 2, \ldots, n$) jsou data výběru seřazená podle velikosti.

Hladiny q někdy uvádíme v procentech. V tomto případě nalezené hodnoty označujeme jako **percentily** nebo přesněji empirické percentily na dané úrovni. Je tedy 25% percentil rovný kvantilu o hladině 0,25.

Percentily s hladinou 25 %, 50 % a 75 % nazýváme kvartily a označujeme je takto:

- Q_I je *první* neboli *dolní kvartil* ($q = 25\,\%$);
- Q_{II} je *druhý* neboli *medián* ($q = 50\,\%$) – ten již známe z výkladu o mírách centrální tendence;
- Q_{III} je *třetí* neboli *horní kvartil* ($q = 75\,\%$).

Pro data o výkonech v skoku dalekém z tabulky 2.9 (s. 77) mají kvartily hodnotu $Q_I = 3{,}35$, $Q_{II} = Me = 3{,}55$, $Q_{III} = 3{,}75$.

Při popisu krajních hodnot rozdělení udáváme percentily s hladinami buď 2,5 % a 97,5 %, anebo 5 % a 95 %. Tyto extrémní percentily se často používají při určování referenčních hodnot laboratorních údajů v biomedicíně.

Interkvartilové rozpětí $Q = Q_{III} - Q_I$ je charakteristikou rozptýlenosti, jež se používá spolu s kvartily k popisu tvaru dat, když se z nějakého důvodu nechceme opřít o průměrové charakteristiky, jako je aritmetický průměr nebo směrodatná odchylka. Z definice vyplývá, že v intervalu (Q_I, Q_{III}) se nachází 50 % údajů. Interkvartilové rozpětí má intuitivnější obsah než směrodatná odchylka a není na rozdíl od směrodatné odchylky tak citlivé vůči odlehlým hodnotám.

Pro data o výkonech v skoku dalekém z tabulky 2.9 má kvartilové rozpětí hodnotu $Q = Q_{III} - Q_I = 3{,}74 - 3{,}34 = 0{,}40$.

Mediánová absolutní odchylka je mírou rozptýlenosti vycházející z dvojnásobného použití výpočtu mediánu. Jedná se o míru rozptýlenosti, která – podobně jako interkvartilové rozpětí – není citlivá k odlehlým hodnotám. Spočítá se jako

medián z absolutních hodnot odchylek jednotlivých měření od mediánu. Označuje se někdy zkráceně *MAD – median absolute deviation*. Zkráceně vyjádříme výpočet této míry vzorcem:

$$MAD = Me\{|x_i - Me|\}$$

U údajů {0; 1; 2; 5; 8; 9; 10} jsme zjistili, že medián je 5. Absolutní diference mají hodnoty {5; 4; 3; 0; 3; 4; 5}. Seřadíme je podle velikosti a zjistíme z uspořádané sekvence {0; 3; 3; 4; 4; 5; 5} medián. *MAD* má tedy hodnotu 4.

3.4 Míry špičatosti a šikmosti

Tyto charakteristiky se používají méně často, ale obvykle společně. Slouží k jemnějšímu popisu specifických stránek dat. Hodnotíme pomocí nich také to, jak se rozdělení dat podobá normální (Gaussově) křivce. K výpočtu těchto charakteristik se přistupuje různě. Nejčastěji se využívají tzv. centrální momenty třetího a čtvrtého stupně. Centrální moment k-tého stupně m_k je obecně definován vzorcem

$$m_k = \frac{\sum (x_i - \bar{x})^k}{n}.$$

Šikmost S_1 měří zešikmenost, resp. nesymetrii dat a vypočítá se pomocí druhého a třetího momentu podle vzorce

$$S_1 = \frac{m_3}{m_2^{3/2}}.$$

$S_1 = 0$ platí přibližně pro rozdělení přibližně symetrické, $S_1 > 0$ pro rozdělení s prodlouženým pravým koncem, naopak $S_1 < 0$ pro rozdělení s prodlouženým levým koncem (obr. 3.7).

Koeficient špičatosti S_2 měří odchylku špičatosti zkoumaného rozdělení od normálního rozdělení:

$$S_2 = \frac{m_4}{m_2^2} - 3$$

Takto vypočtená špičatost má pro normální rozdělení hodnotu 0. Symetrická rozdělení mohou mít stejný rozptyl, ale odlišnou špičatost. Plošší křivky ($S_2 > 0$) nazýváme platykurtické, špičatější křivky ($S_2 < 0$) leptokurtické.

Zešikmenost se také měří pomocí dalších koeficientů. U symetrických dat medián dělí na polovinu interkvartilové rozpětí. Tento poznatek je možné využít k definování koeficientu šikmosti KS pomocí kvartilů

$$KS = \frac{Q_{III} + Q_I - 2\tilde{x}}{Q},$$