Zákon velkých čísel – stanovení empirické pravděpodobnosti

- Jaké jsou důsledky zákona velkých čísel?
- Víme, že díky velkému počtu nezávislého opakování stejného náhodného experimentu můžeme získat s daným rozptylem střední hodnotu E[X] náhodné proměnné, která se váže k tomuto experimentu.
- Z velkého počtu opakování náhodných experimentů můžeme získat **pravděpodobnost**, že nastane nějaký jev: p = P(X∈C), kde C = (a,b).
- Pravděpodobnost *p* stanovíme jednoduše z počtu, jak často nastane událost $X_i \in C$ v dané sekvenci opakování.
- Hledáme relativní četnost událostí $X_i \in C$ v rámci *n* opakování.
- Definujme náhodnou proměnnou „indikátor náhodné proměnné“ Y_i indikující, jestliže nastává jev $X_i \in C$ nebo nenastává, podle předpisu:

$$Y_i = \begin{cases} 1 & \text{if } X_i \in C, \\ 0 & \text{if } X_i \notin C. \end{cases}$$

Zákon velkých čísel – stanovení empirické pravděpodobnosti

- Střední hodnota Y_i je dána předpisem:

$$\mathrm{E}[Y_i] = 1 \cdot \mathrm{P}(X_i \in C) + 0 \cdot \mathrm{P}(X_i \notin C) = \mathrm{P}(X_i \in C) = \mathrm{P}(X \in C) = p.$$

- Náhodné proměnné Y_i jsou nezávislé (X_i tvoří nezávislou sekvenci náhodných proměnných a Y_i je určeno na základě X_i a tedy ze zákona o přenosu nezávislosti mi plyne výše uvedené).
- Tedy relativní četnost jevů Y_i je dána průměrem: $(Y_1+Y_2+\cdots+Y_n)/n = \bar{Y}_n.$
- Pokud pravděpodobnost p hraje stejnou roli jako μ, pak ze zákona velkých čísel aplikovaného na $\bar{Y}_n$ plyne:

  aritmetický průměr n nezávislých náhodných proměnných se střední hodnotou p a rozptylem $p(1-p)$ a pro $\varepsilon>0$ je dán: $\lim_{n\to\infty} \mathrm{P}\left(|\bar{Y}_n - p| > \varepsilon\right) = 0$

- Vidíme, že pravděpodobnost jevu můžeme stanovit z velkého počtu realizací tohoto jevu (četnost) při n opakováních nezávislých náhodných experimentů.
- Dostali jsme tak přesnější vyjádření tzv. empirické definice pravděpodobnosti náhodného jevu.

Zákon velkých čísel – hustota pravděpodobnosti

- Můžeme stanovit hustotu pravděpodobnosti pro předchozí případ?
- Předpokládejme spojitou náhodnou proměnnou s hustotou pravděpodobnosti *f* a distribuční funkcí *F*. Mějme interval C = (a-h, a+h) pro *h* – malé kladné číslo.
- Pro velké *n* z rovnice $\lim_{n\to\infty} \mathrm{P}(|\bar{Y}_n - p| > \varepsilon) = 0$ plyne, že

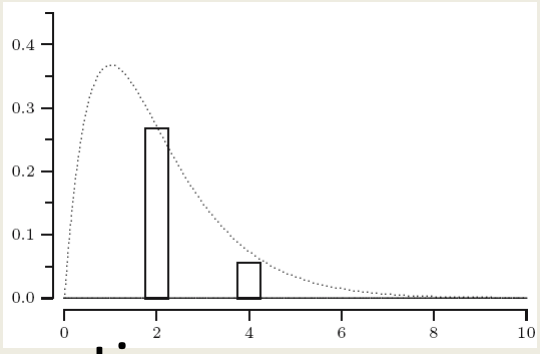
$$\bar{Y}_n \approx p = \mathrm{P}(X \in C) = \int_{a-h}^{a+h} f(x)\,\mathrm{d}x \approx 2hf(a).$$

Zákon velkých čísel – hustota pravděpodobnosti

- Z poslední rovnice můžeme stanovit hustotu pravděpodobnosti v bodě *a* jako:

$$f(a) \approx \frac{\bar{Y}_n}{2h} = \frac{\text{the number of times } X_i \in C \text{ for } i \leq n}{n \cdot \text{the length of } C}.$$

- Př.: spočítáme si *f(a)* pro h = 0,25 a dvě hodnoty *a* rovné 2 a 4. Pravděpodobnostní distribuce bude Gam(2,1) a nasimulujeme 500 nezávislých opakování.
- Dostaneme sloupcový graf, kde šířka sloupce je 2*h*, jeho výška *f*(*a*) a jeho plocha $\overline{Y_n}$.
- Vidíme, že jsme se relativně přesně trefili do Gam(2,1) hustoty pravděpodobnosti.

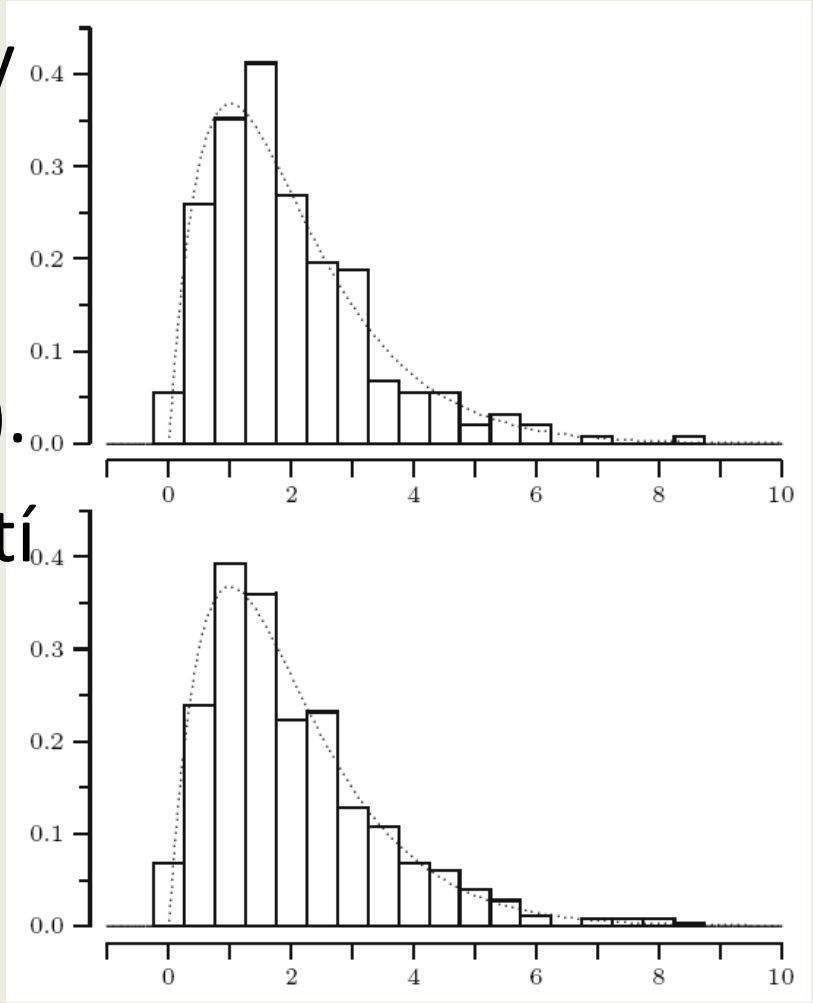


Zákon velkých čísel – hustota pravděpodobnosti

- Ve skutečnosti pokud chceme získat co nejvěrnější podobu hustoty pravděpodobnosti hledané pravděpodobnostní distribuce, tak je třeba šířku intervalu *h* mít co nejmenší a provést simulaci pro co nejvíce různých hodnot *a*. Tedy pokrýt osu *x* co nejvíce sloupci.
- Takovýto graf nazýváme jako **histogram**.
- Histogram vlastně je grafem diskrétní náhodné proměnné a zúžováním šířky sloupce limitně k nule dostáváme spojitou distribuci.

Zákon velkých čísel – hustota pravděpodobnosti

- Na obrázku vidíme dvě sady náhodných experimentů s Gam(2,1) rozdělením (dvě na sobě nezávislé simulace).
- Po každé se s jinou přesností „trefíme“ do teoretického Gam(2,1) rozdělení.



Centrální limitní věta

Centrální limitní věta

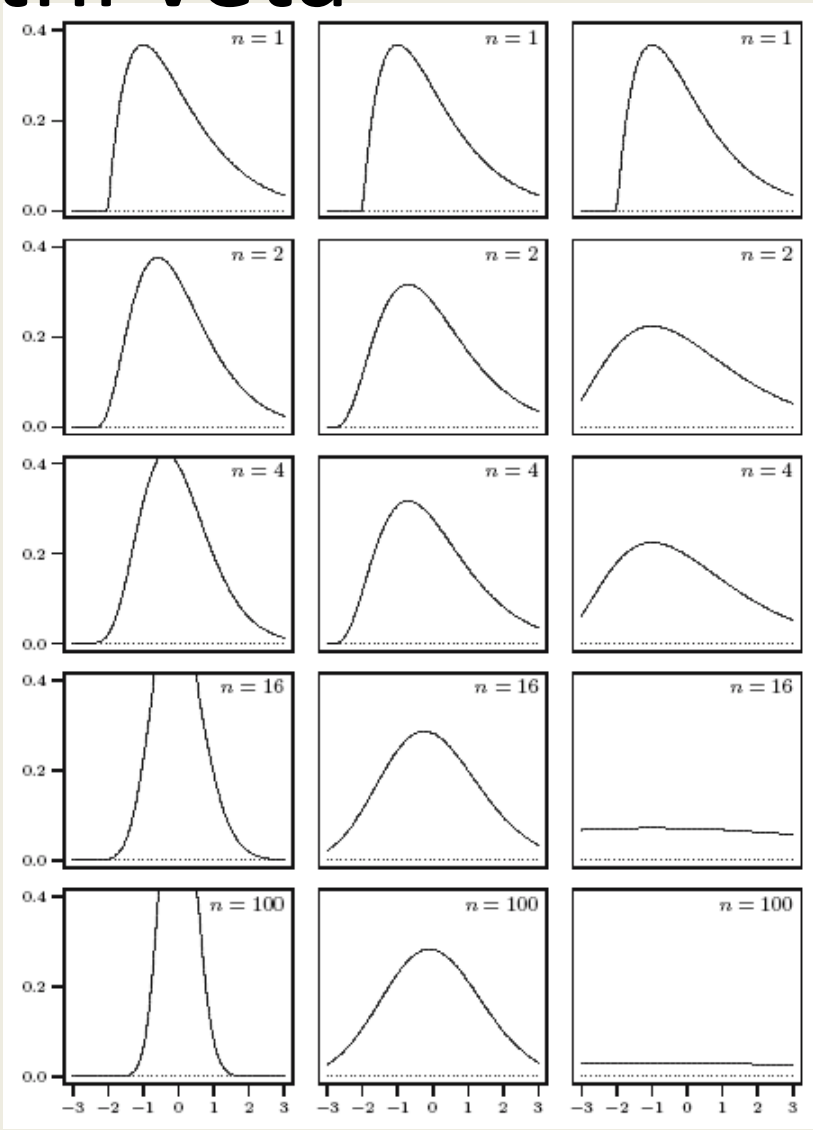
- Centrální limitní věta je zpřesnění zákona velkých čísel.
- Tedy pokud máme *n* nezávislých náhodných proměnných $X_1, X_2, X_3, \ldots, X_n$ se stejnou pravděpodobnostní distribucí a s konečných rozptylem, tak průměr $\overline{X_n}$ má přibližně normální rozdělení, aniž by záleželo na pravděpodobnostní distribuci X_i.
- Na str. 41 přednášky 4 jsme viděli, že hustota pravděpodobnosti náhodné proměnné $\overline{X_n}$ se stává stále více symetrickou a „zvonový“ tvar kolem střední hodnoty μ se stále zužuje a pro $n \to \infty$ konverguje k delta funkci s posunutím μ.
- Nicméně, když provedeme správnou normalizaci náhodné proměnné, tak lze „zvonový“ tvar udržet i pro velká *n*.

Centrální limitní věta

- Je třeba „stabilizovat“ střední hodnotu μ a rozptyl σ^2.
- Ze zákona velkých čísel je zřejmé, že $E[\overline{X_n}] = \mu$ pro libovolné *n*. Na druhou stranu rozptyl je nepřímo úměrný počtu experimentů *n*.
- Tedy musíme nějak „upravit“ rozptyl, aby pro $n \to \infty$ mi hustota pravděpodobnosti náhodné proměnné $\overline{X_n}$ nekonvergovala k posunuté delta funkci.
- Na obrázcích na str. 10 jako příklad vidíme hustoty pravděpodobnosti náhodné proměnné $(\overline{X_n} - \mu)$ násobené různě umocněným *n* s pravděpodobnostní distribucí Gam(2,1) pro rostoucí *n*.

Centrální limitní věta

- Tři různě parametrizované náhodné proměnné s pravděpodobnostním rozdělením Gam(2,1):
  - první sloupec $n^{1/4}(\overline{X_n} - \mu)$
  - druhý sloupec $n^{1/2}(\overline{X_n} - \mu)$
  - třetí sloupec $n(\overline{X_n} - \mu)$



Centrální limitní věta

- Z obrázků je zřejmé, že nejlepšího výsledku dosáhneme s $n^{1/2}$. Tento faktor mi nejlépe udržuje „zvonový“ charakter hustoty pravděpodobnosti i pro velká *n*.
- Z definice rozptylu náhodné proměnné $\overline{X_n}$, viz str. 37 v přednášce 4, pro libovolnou konstantu C:

$$\mathrm{Var}\big(C(\bar{X}_n - \mu)\big) = \mathrm{Var}\big(C\bar{X}_n\big) = C^2\mathrm{Var}\big(\bar{X}_n\big) = C^2\frac{\sigma^2}{n}.$$

- Tedy, aby se rozptyl zachoval a neměnil pro velká *n*, tak je třeba zvolit $C = n^{1/2}$.
- Ve skutečnosti pokud zvolíme $C = n^{1/2}/\sigma$, tak „standardizujeme“ průměr mnoha opakovaných nezávislých náhodných experimentů s náhodnou proměnnou Z_n:

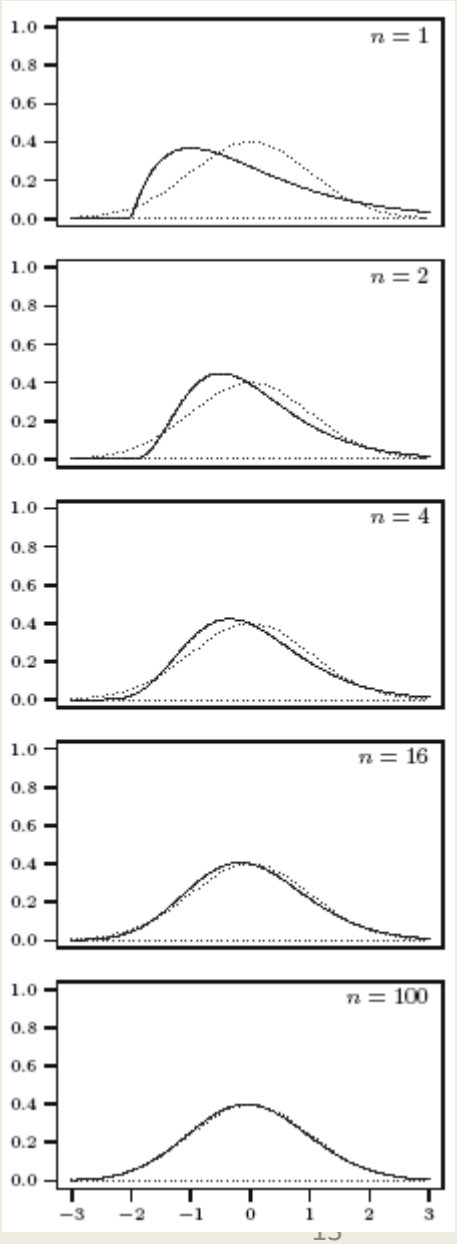
$$Z_n = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma}, \quad n = 1, 2, \ldots,$$

Centrální limitní věta

- Náhodná proměnná Z_n má střední hodnotu 0 a rozptyl 1.
- Pokud X_1, X_2, X_3,… jsou nezávislé náhodné proměnné s normální distribuci $N(\mu,\sigma^2)$, tak z přednášky 3 na str. 8 víme, že náhodná proměnná Z_n má distribuci N(0,1) pro všechna n.
- Platí výše uvedené i pro X_1, X_2, X_3,… s jiným, třeba Gam(2,1) rozdělením?

Centrální limitní věta

- Pro různá *n* vidíme vývoj Gam(2,1) a normálního rozdělení.
- Tedy Gam(2,1) konverguje pro velká *n* k rozdělení N(0,1).
- Dá se ukázat, že toto chování má zcela obecný charakter a platí pro *n* nezávislých opakování náhodné proměnné s libovolnou pravděpodobnostní distribuci s definovanou střední hodnotou a rozptylem.
- Tuto vlastnost shrnuje centrální limitní věta.



Centrální limitní věta

THE CENTRAL LIMIT THEOREM. Let $X_1, X_2, \ldots$ be any sequence of independent identically distributed random variables with finite positive variance. Let μ be the expected value and σ^2 the variance of each of the X_i. For $n \geq 1$, let Z_n be defined by

$$Z_n = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma};$$

then for any number a

$$\lim_{n\to\infty} F_{Z_n}(a) = \Phi(a),$$

where Φ is the distribution function of the $N(0,1)$ distribution. In words: the distribution function of Z_n converges to the distribution function Φ of the standard normal distribution.

- Kde Z_n je transformovaná $\overline{X_n}$

$$Z_n = \frac{\bar{X}_n - \mathrm{E}\left[\bar{X}_n\right]}{\sqrt{\mathrm{Var}\left(\bar{X}_n\right)}},$$

Centrální limitní věta

- Z_n lze upravit na tvar: $Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$
- To je užitečné v případě, že známe jen hodnoty *n* nezávislých proměnných se stejnou distribucí.
- Protože platí $\bar{X}_n = \frac{\sigma}{\sqrt{n}} Z_n + \mu,$ tak $\overline{X_n}$ má přibližně pro velká *n* pravděpodobnostní distribuci $N(\mu,\sigma^2/n)$ – konverguje k delta funkci se středem μ.
- Centrální limitní věta nám poskytuje silný nástroj k aproximaci empirických pravděpodobnostních distribucí průměru nebo součtu identických nezávislých náhodných proměnných.

Aplikace centrální limitní věty

- První aplikace bude analýza získaných $\overline{X_n}$ pro různé počty opakování *n* z příkladu z přednášky 4 na str. 47.
- Tam jsme viděli, že pro *n* = 400 je $\overline{X_n}$ = 1,99, ale pro *n* = 500 je $\overline{X_n}$ = 2,06. Tedy pro větší počet opakování jsme o něco dále od střední hodnoty μ = 2, což bychom na první pohled neočekávali.
- Je tedy hodnota $\overline{X_n}$ = 2,06 pro *n* = 500 obvykle očekávatelná, nebo jsme měli během opakování experimentů smůlu na špatné měření? Odpověď získáme spočítáním P($\overline{X_n} \geq$ 2,06).

Aplikace centrální limitní věty

- Tedy chceme spočítat pravděpodobnost P($\overline{X_n} \geq 2{,}06$). Pravděpodobnost můžeme rozepsat:

$$\begin{aligned} \mathrm{P}\left(\bar{X}_n \geq 2.06\right) &= \mathrm{P}\left(\bar{X}_n - \mu \geq 2.06 - \mu\right) \\ &= \mathrm{P}\left(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \geq \sqrt{n}\frac{2.06 - \mu}{\sigma}\right) \\ &= \mathrm{P}\left(Z_n \geq \sqrt{n}\frac{2.06 - \mu}{\sigma}\right). \end{aligned}$$

- Protože X_i jsou náhodné proměnné s Gam(2,1), tak $E[X_i] = 2$ a $Var(X_i) = 2$. Pro n = 500 dostaneme:

$$\begin{aligned} \mathrm{P}\left(\bar{X}_{500} \geq 2.06\right) &= \mathrm{P}\left(Z_{500} \geq \sqrt{500}\frac{2.06 - 2}{\sqrt{2}}\right) \\ &= \mathrm{P}(Z_{500} \geq 0.95) \\ &= 1 - \mathrm{P}(Z_{500} < 0.95). \end{aligned}$$

Aplikace centrální limitní věty

- Podle centrální limitní věty platí:

$$\mathrm{P}\left(\bar{X}_{500} \geq 2.06\right) \approx 1 - \Phi(0.95) = 0.1711.$$

- Dostali jsem vyčíslenou pravděpodobnost, která je velmi blízko k P = 0,1710881 z výpočtu pravděpodobnosti z hustoty pravděpodobnosti náhodné proměnné $\overline{X_n}$ - viz str. 40 přednáška 4.
- Tedy máme stále 17% pravděpodobnost, že po 500 opakováních náhodného experimentu bude spočtený průměr všech 500 náhodných hodnot o 0,06 větší, jak očekávaná střední hodnota $E[X_i] = 2$.
- Tedy hodnota $\overline{X_n}$ = 2,06 není neobvyklá pro *n* = 500.
- Pozn.: pokud *n* = 5000, tak $P(\overline{X_n} \geq 2{,}06)$ = 0,13%.

Aplikace centrální limitní věty

- Uvažujme situaci, že máme test s 10 otázkami. Pro úspěšné splnění testu musíme mít správně aspoň 6 otázek. Pro každou otázka máme na výběr ze 4 možností. Jaká je pravděpodobnost, že uděláme test, pokud odpovědi budeme volit náhodně?
- Lehce nahlédneme, že se jedná o diskrétní náhodné rozdělení Bin(10,1/4) – viz str. 3 přednášky 2. Z distribuční funkce binomického rozdělení pro k = 6 dostaneme, že $P(X \geq 6) = 0{,}0197$.
- Ačkoli je *n* = 10 malé, zkusíme k nalezení $P(X \geq 6)$ použít centrální limitní větu.

Aplikace centrální limitní věty

- Víme, že náhodná proměnná s binomickým rozdělení Bin(n,p) je součtem *n* náhodných proměnných s rozdělením Ber(p).
- Tedy $X = R_1 + R_2 + R_3 + \ldots + R_n$. Mějme $n = 10$, $\mu = p = ¼$ a $\sigma^2 = p(1-p) = 3/16$. Pak z centrální limitní věty plyne:

$$\begin{aligned}
\mathrm{P}(X \geq 6) &= \mathrm{P}(R_1 + \cdots + R_n \geq 6) \\
&= \mathrm{P}\left(\frac{R_1 + \cdots + R_n - n\mu}{\sigma\sqrt{n}} \geq \frac{6 - n\mu}{\sigma\sqrt{n}} \right) \\
&= \mathrm{P}\left(Z_{10} \geq \frac{6 - 2\frac{1}{2}}{\sqrt{\frac{3}{16}}\sqrt{10}} \right) \\
&\approx 1 - \Phi(2.56) = 0.0052.
\end{aligned}$$

Aplikace centrální limitní věty

- Jak vidíme pravděpodobnost 0,0052 je velmi špatná aproximace ke správné hodnotě 0,0197.
- Na druhou stranu můžeme psát:

$$\begin{aligned} \mathrm{P}(X \geq 6) &= \mathrm{P}(X > 5) \\ &= \mathrm{P}(R_1 + \cdots + R_n > 5) \\ &= \mathrm{P}\left(Z_{10} \geq \frac{5 - 2\frac{1}{2}}{\sqrt{\frac{3}{16}}\sqrt{10}} \right) \\ &\approx 1 - \Phi(1.83) = 0.0336, \end{aligned}$$

- Spočtená pravděpodobnost, že X > 5 je zase moc velká.
- Tedy nejlepší hodnotu bychom získali pro $k \in < 5, 6 >$.
- Pokud bychom hledali podle centrální limitní věty pravděpodobnost P(X > 5,5), dojdeme k výsledku 0,0143. Což je blíže k teoretické hodnotě a lépe to aproximuje P(X ≥ 6).
- Tedy při použití centrální limitní věty musíme být obezřetní.

Aplikace centrální limitní věty

- Praktická otázka zní: jak velké by mělo být *n*, aby šla použít centrální limitní věta? Jinými slovy, jak rychle konverguje pravděpodobnostní distribuce náhodné proměnné $\overline{X_n}$ k normálnímu rozdělení s rostoucím *n*?
- Odpověď není univerzální.
- Záleží na typu distribuce X_i, jestli je asymetrická, bimodální, diskrétní. Záleží na tom, jestli číslo *a* v P($\overline{X_n} > a$) leží příliš daleko od středu distribuce X_i nebo jestli je *n* příliš malé.
- Na druhou stranu, pokud aproximujeme diskrétní rozdělení spojitým rozdělením, tak můžeme pomocí centrální limitní věty získat relativně přesné výsledky – viz předchozí aplikace.
- Před aplikací centrální limitní věty je třeba dobře zvážit na jakou náhodnou proměnnou ji aplikujeme a mít pokud možno co největší *n*.

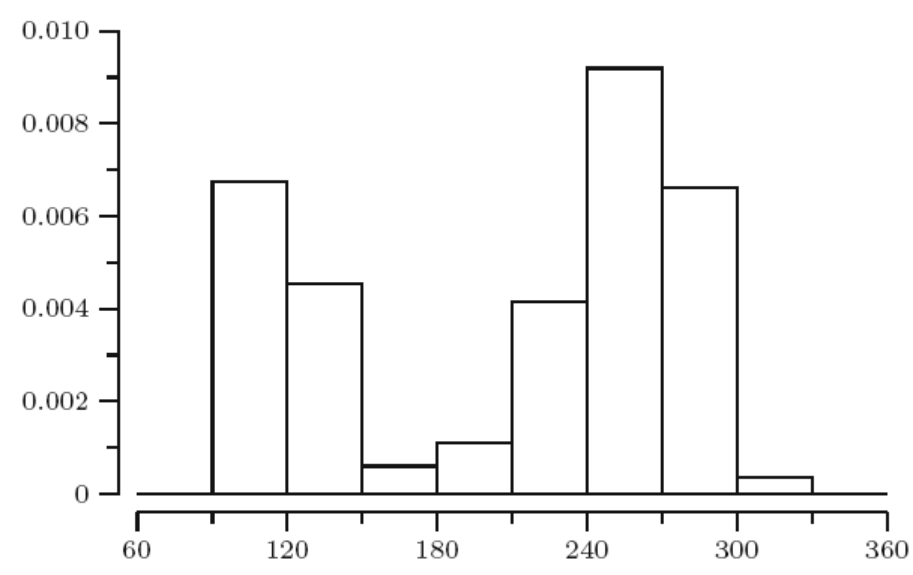
Grafické zobrazení náhodných dat

Grafické zobrazení náhodných dat

- V praxi většinou studujeme náhodné jevy, které dostaneme z nějakého náhodného experimentu.
- Záznam pozorování nebo měření dostaneme ve formě souboru dat – **statistický (výběrový) soubor**.
- První základní informace o získaných datech vidíme z grafického zobrazení statistického souboru.
- Ze souboru dat můžeme např. hned určit střední hodnotu, maximální nebo minimální hodnotu, rozptyl atp.
- Ale např. grafické zobrazení, nám může hned ozřejmit, kde se nachází maximální četnost, existuje-li více maxim v distribuci, jestli je distribuce asymetrická atp.
- Příklad grafického zobrazení dat si ukážeme na měření doby délky trvání jednotlivých erupcí gejzíru v Yellowstonském parku, jak byly pozorovány během 15 dnů. Celkem se naměřilo 272 erupcí.

Histogram

- V tabulce je 272 záznamů délky trvání jednotlivých erupcí v sekundách.
- Pojem **histogram** byl poprvé použit K. Pearsonem.



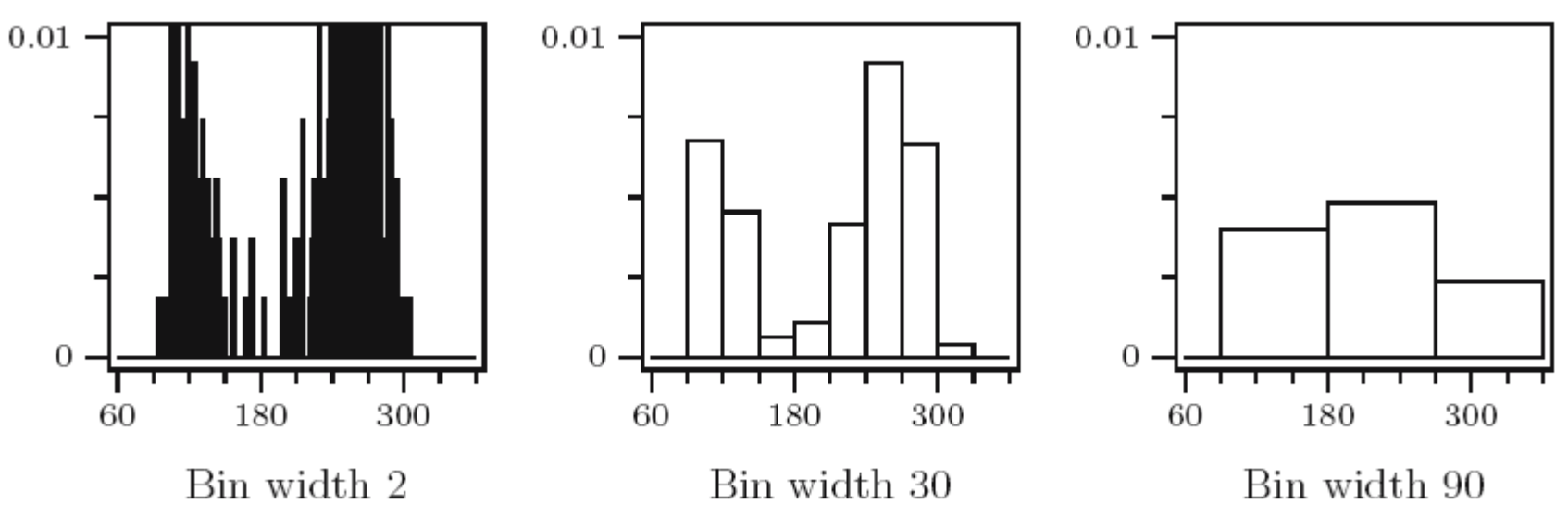
216	108	200	137	272	173	282	216	117	261
110	235	252	105	282	130	105	288	96	255
108	105	207	184	272	216	118	245	231	266
258	268	202	242	230	121	112	290	110	287
261	113	274	105	272	199	230	126	278	120
288	283	110	290	104	293	223	100	274	259
134	270	105	288	109	264	250	282	124	282
242	118	270	240	119	304	121	274	233	216
248	260	246	158	244	296	237	271	130	240
132	260	112	289	110	258	280	225	112	294
149	262	126	270	243	112	282	107	291	221
284	138	294	265	102	278	139	276	109	265
157	244	255	118	276	226	115	270	136	279
112	250	168	260	110	263	113	296	122	224
254	134	272	289	260	119	278	121	306	108
302	240	144	276	214	240	270	245	108	238
132	249	120	230	210	275	142	300	116	277
115	125	275	200	250	260	270	145	240	250
113	275	255	226	122	266	245	110	265	131
288	110	288	246	238	254	210	262	135	280
126	261	248	112	276	107	262	231	116	270
143	282	112	230	205	254	144	288	120	249
112	256	105	269	240	247	245	256	235	273
245	145	251	133	267	113	111	257	237	140
249	141	296	174	275	230	125	262	128	261
132	267	214	270	249	229	235	267	120	257
286	272	111	255	119	135	285	247	129	265
109	268								

Histogram

- Jak zkonstruovat histogram?
- Mějme x_1, x_2, x_3,…, x_n naměřených dat. Histogram budeme normovat na jedničku tzn., že plocha histogramu = 1.
- Rozdělíme si statistický soubor na intervaly – sloupce: B_1, B_2, B_3,…, B_m.
- Délka intervalu $B_i = |B_i|$ se nazývá jako šířka sloupce. Plocha každého sloupce B_i reprezentuje počet dat v B_i. Protože plocha všech sloupců reprezentuje počet dat n a je rovna 1, tak plocha sloupce B_i = (počet x_j v B_i)/n.
- Výška sloupce H_i = (počet x_j v B_i)/($n \cdot B_i$).

Histogram

- Zcela obecně šířka sloupců v histogramu nemusí být stejná.
- Jak široký sloupec vybrat?



Histogram

- Pokud všechny sloupce budou stejně široké, pak délka intervalu je: $B_i = (r + (i-1)b, r + ib] \quad \text{for } i = 1, 2, \ldots, m,$

  kde *r* je referenční bod menší než minimum v datovém souboru a *b* je šířka sloupce.
- Výběr vhodného *b* (potažmo počtu sloupců *m*) mi určuje jak histogram bude vypadat. Buď to bude spleť lokálních izolovaných maxim nebo jen graf, kde ztratíme příliš mnoho informací.
- Pro náš příklad se jeví jako nejlepší šířka sloupce 30.
- Šířku sloupce můžeme vybrat metodou pokus-omyl, dokud graf nevypadá rozumně.
- Nicméně matematici vyvinuly polo-empirický postup, jak spočítat optimální šířku sloupce.

Histogram

- Efektivní počet sloupců je dán vztahem: $m = 1 + 3.3 \log_{10}(n)$
- Případně šířka sloupce: $b = 3.49\, sn^{-1/3}$, kde *s* je tzv. výběrová směrodatná odchylka (určuje mi, jak moc se jednotlivé vzorky statistického souboru od sebe liší).
- Šířka sloupce je odvozena na základě požadavku, aby byl minimalizován rozdíl mezi výškou sloupce H_n a hustotou pravděpodobnosti *f*, která generuje náš datový soubor.
- Tato minimalizace je realizována skrze tzv. střední integrovanou kvadratickou odchylku (MISE):

$$\mathrm{E}\left[\int_{-\infty}^{\infty} (H_n(x) - f(x))^2 \, \mathrm{d}x\right].$$

- Takové *b*, které mi minimalizuje MISE pro $n \to \infty$ je dané vztahem:

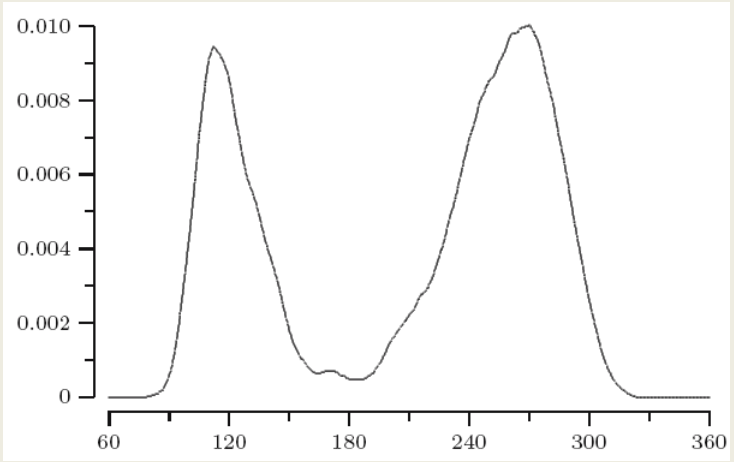
$$b = C(f)n^{-1/3} \quad \text{where } C(f) = 6^{1/3}\left(\int_{-\infty}^{\infty} f'(x)^2 \, \mathrm{d}x\right)^{-1/3}$$

Histogram

- Pokud *f* bude normální rozdělení N(μ,σ²), pak konstantu *C*(*f*) lze kvantifikovat: $C(f) = (24\sqrt{\pi})^{1/3}\sigma$
- Směrodatnou odchylku *σ* pak lze nahradit výběrovou směrodatnou odchylkou *s*.
- Výhoda histogramu leží v jeho jednoduchosti.
- Nevýhodou je diskrétní charakter grafu.
- Další problém spočívá se skutečnosti, že malá změna šířky sloupce nebo malý posuv sloupců při fixaci jejich šířky vede ke grafům, které mají už jiný význam.
- Tyto problémy lze řešit pomocí metody tzv. jádrového odhadu hustoty.

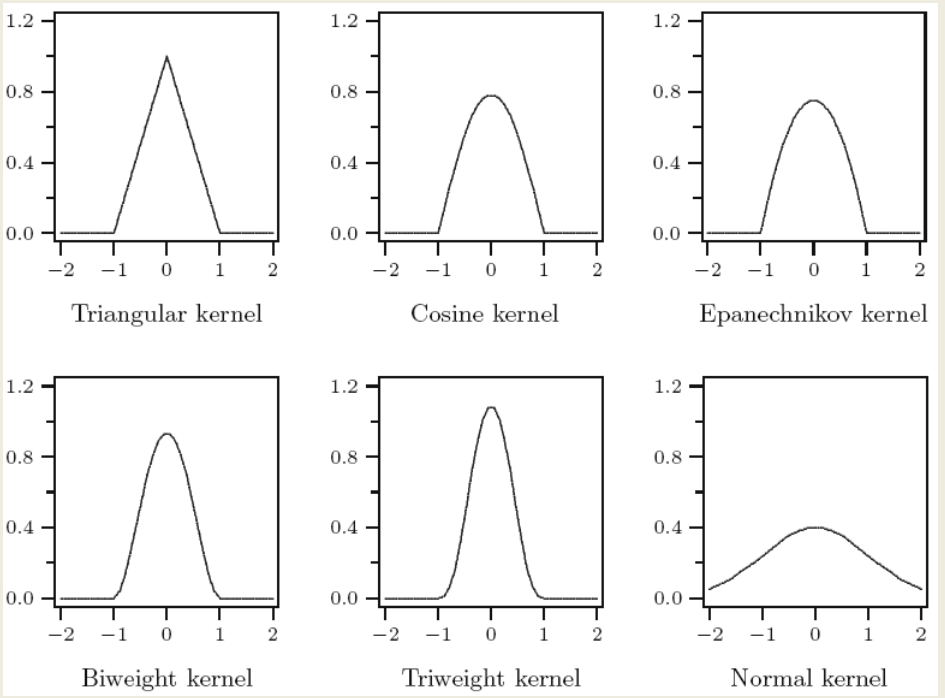
Jádrový odhad hustoty

- Metoda byla navržena Rosenblattem a Parzenem v 50. letech. Díky velmi vysoké výpočetní náročnosti se stává zajímavou až v poslední době výkonných počítačů.
- Graf je mnohem hladší a snadněji detekujeme maxima s největší četností dat.
- Princip spočívá v „sypání písku“ okolo prvků datového souboru. Hromada písku roste tam, kde se akumulují prvky.
- Graf konstruuji na základě vybrání jádra *K* a šířky pásma *h*. Jádro *K* odráží tvar „hromady písku“ a parametr *h* mi ladí šířku „hromady písku“.



Jádrový odhad hustoty

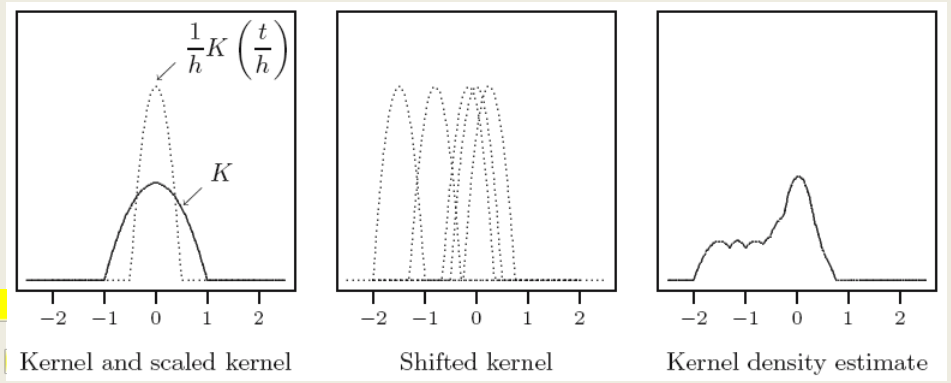
- Jádrová funkce *K* musí splňovat podmínky:
  - *K* je hustota pravděpodobnosti $K(u) \geq 0 \text{ and } \int_{-\infty}^{\infty} K(u)\,\mathrm{d}u = 1$
  - *K* je symetrická kolem nuly $K(u) = K(-u)$
  - *K*(u) = 0 pro |u|>1
- Příklady používaných jádrových funkcí:



Jádrový odhad hustoty

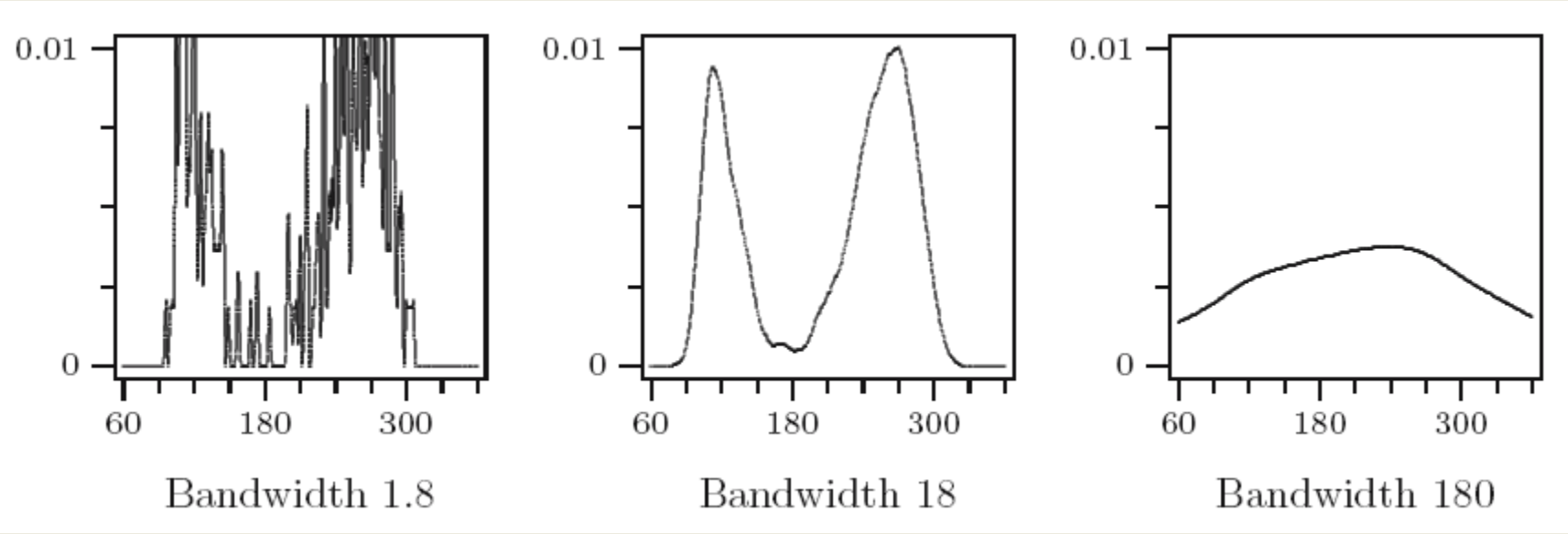
- Př.: mějme datový soubor x_1, x_2, x_3,…, x_n, jádrová funkce bude Epanechnikov a $h = 0{,}5$.
- První transformujeme *K* na šířku pásma *h*, tzn., že *K* bude kladné na intervalu [-h, h] místo [-1, 1]. Takové *K* aplikujeme kolem každého elementu x_i a dostaneme funkci: $t \mapsto \frac{1}{h} K\left(\frac{t - x_i}{h}\right)$
- Jednotlivé transformované *K* funkce každého prvku výběrového souboru se překrývají, pokud se v těch místech akumuluje více prvků.
- Výsledný odhad jádrové hustoty $f_{n,h}$ dostaneme součtem všech *K* funkcí dělených *n*:

$$f_{n,h}(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - x_i}{h}\right)$$



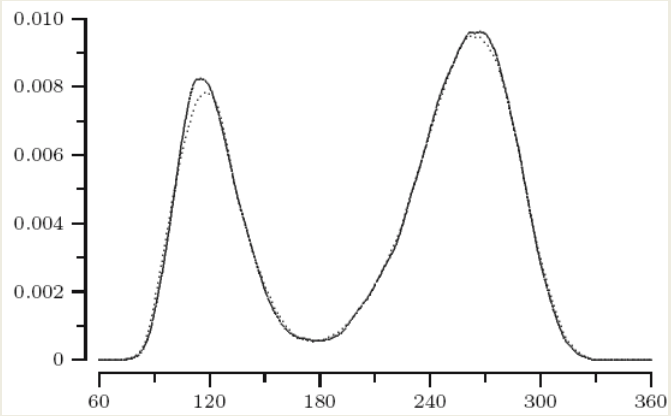
Jádrový odhad hustoty

- Při výpočtu $f_{n,h}$ větší váhu přisuzujeme těm prvkům, které jsou nejblíže proměnné t. Na rozdíl od histogramu, kde jen prostě počítáme množství prvků spadajících do sloupce se středem t.
- Výběr šířky pásma h hraje stejnou roli jako výběr šířky sloupce při konstrukci histogramu.



Jádrový odhad hustoty

- Je třeba vybrat vhodné h tak, aby graf byl srozumitelný. Buď použijeme metody pokus-omyl, nebo použijeme vodítka na základě spočítané optimální šířky pásma h.
- Podobně jako u histogramu lze použít vztah: $h = 1{,}06 \cdot s \cdot n^{-1/5}$.
- Ukazuje se, že správný výběr jádrové funkce K není až tak kritický.
- Na obrázku je Epanechnikovo a trojváhové K.



Bodový graf

- Mějme situaci, kdy máme statistický soubor obsahující dvě náhodné proměnné. Dostáváme tedy prvky statistického souboru jako dvojice proměnných.
- V takovém případě nás často zajímá, jestli proměnná *y* nějak souvisí s proměnnou *x*. Pokud ano, jestli můžeme tuto vzájemnou závislost nějak popsat.
- Jednotlivé prvky statistického souboru (x_i, y_i) vyneseme do **bodového grafu**.

Bodový graf

- Data zkoumající vztah mezi tvrdostí dřeva a jeho hustotou.

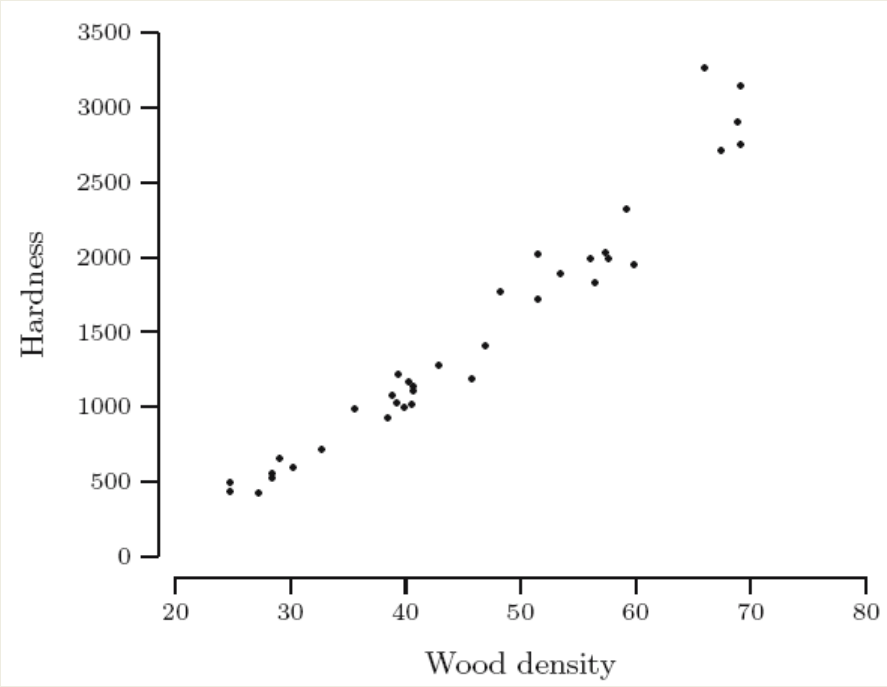


Table 15.5. Density and hardness of Australian timber.

Density	Hardness	Density	Hardness	Density	Hardness
24.7	484	39.4	1210	53.4	1880
24.8	427	39.9	989	56.0	1980
27.3	413	40.3	1160	56.5	1820
28.4	517	40.6	1010	57.3	2020
28.4	549	40.7	1100	57.6	1980
29.0	648	40.7	1130	59.2	2310
30.3	587	42.9	1270	59.8	1940
32.7	704	45.8	1180	66.0	3260
35.6	979	46.9	1400	67.4	2700
38.5	914	48.2	1760	68.8	2890
38.8	1070	51.5	1710	69.1	2740
39.3	1020	51.5	2010	69.1	3140

Numerické charakteristiky statistických souborů

Střed

- Základní vlastnosti statistického souboru lze popsat několika číselnými charakteristikami.
- Často nás zajímá najít tzv. **střed (průměr)** základního statistického souboru (populace), pokud by byl vzestupně uspořádán.
- Zkoumané vlastnosti obvykle studujeme na vybraných *n* prvcích základního statistického souboru – **výběrovém souboru**.
- Nejjednodušší cesta je spočítat tzv. **výběrový průměr**:

$$\bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Je to obdoba střední hodnoty (aritmetického průměru) náhodné proměnné.
- Jiný způsob nalezení středu statistického souboru je spočítání **výběrového mediánu** Med_n.
- Je definován jako prostřední prvek vzestupně seřazeného stat. souboru.
  - Pokud je počet prvků lichý, tak je to zřejmé.
  - Pokud je počet prvků sudý, tak je to průměr dvou prostředních prvků.

Střed

- Př.: mějme sadu měření teploty během nějaké doby ve stupních F.
- Statistický soubor: 43, 43, 41, 41, 41, 42, 43, 58, 58, 41, 41
- Výběrový průměr je: 44,7 F
- Výběrový medián: 42 F
- Vidíme, že jsme dostali značný rozdíl ve stanovení středu statistického souboru.
- Výběrový průměr je velmi citlivý na **mimořádné hodnoty** na rozdíl od mediánu.
- Je zřejmé, že v naměřených datech jsou čísla 58 mimořádné hodnoty, které se značně odlišují od většiny prvků.
- Pokud mimořádné hodnoty odstraníme ze statistického souboru, tak výběrový průměr bude 41,8 a výběrový medián bude 41.
- Tedy vidíme, že výběrový medián se příliš nezměnil odstraněním mimořádných hodnot ze statistického souboru.
- Z toho plyne, že výběrový medián je více robustní proti výskytu mimořádných hodnot.
- V realitě je třeba si dávat pozor na mimořádné hodnoty a být obezřetný, protože mohou naznačovat třeba pochybení při měření. Pak jim přidělíme menší váhu nebo je ze statistického souboru odstraníme, nebo opravíme experiment.

Střed

- Z příkladu na str. 40 je zřejmé, že mimořádné hodnoty budou odpovídat nějaké systematické chybě měření, protože jde o měření teploty meteorologické stanice v noci a teplota by měla postupně klesat.
- Ukázalo se, že po půlnoci automatický zapisovač teploty se přepnul do °C a tedy hodnoty 58 F a 41 F jsou ve skutečnosti 5,8 °C a 4,1°C.
- Tedy místo mimořádných hodnot dáme správné hodnoty 42 F a 39 F.
- Pak výběrový průměr je 41,5 F a výběrový medián 42 F.

Výběrový rozptyl

- Další numerický parametr, který nás zajímá je variabilita mezi prvky naměřeného statistického souboru.
- K charakterizaci se používá tzv. **výběrový rozptyl** definovaný:

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

- Je vidět, že je to střední hodnota druhých mocnin odchylek od průměru.
- Protože s_n^2 je v jiných jednotkách než výběrový průměr, tak zavádíme tzv. **výběrovou směrodatnou odchylku**:

$$s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2},$$

- Výběrová směrodatná odchylka je vyjádřena ve stejných jednotkách jako statistický soubor.
- Je zřejmé, že výběrový rozptyl bude stejně jako výběrový průměr silně závislý na přítomnosti mimořádných hodnot.

Medián absolutních odchylek

- Pro příklad na str. 40 máme výběrovou směrodatnou odchylku rovnou 6,62 resp. 0,97 pokud odstraníme mimořádné hodnoty.
- Mnohem robustnější charakteristikou je tedy **medián absolutních odchylek** (MAD).
- Je definován následně: mějme absolutní odchylku každého prvku x_i s ohledem na výběrový medián $|x_i - \mathrm{Med}_n|$.
- Potom MAD je roven mediánu všech absolutních odchylek: $\mathrm{MAD}(x_1, x_2, \ldots, x_n) = \mathrm{Med}(|x_1 - \mathrm{Med}_n|, \ldots, |x_n - \mathrm{Med}_n|)$
- MAD je velmi těžce ovlivnitelný mimořádnými hodnotami. Pro příklad na str. 40 je MAD rovno 1.

Empirické kvantily

- Medián mi dělí statistický soubor na dvě stejně velké části prvků.
- Obecně statistický soubor můžeme rozdělit tak, že jen určitá procentní část *p* statistického souboru bude menší než nějaké číslo a druhá procentní část 1-*p* bude větší.
- Prvek statistického souboru odpovídající poměru *p*/(1-*p*) nazýváme jako **empirický kvantil**. Zapisujeme ho jako $q_n(p)$ a představující prvek statistického souboru!!!
- **Uspořádaný statistický soubor** se skládá ze stejných prvků jako originální statistický soubor, ale je vzestupně uspořádaný. Musí tedy platit:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

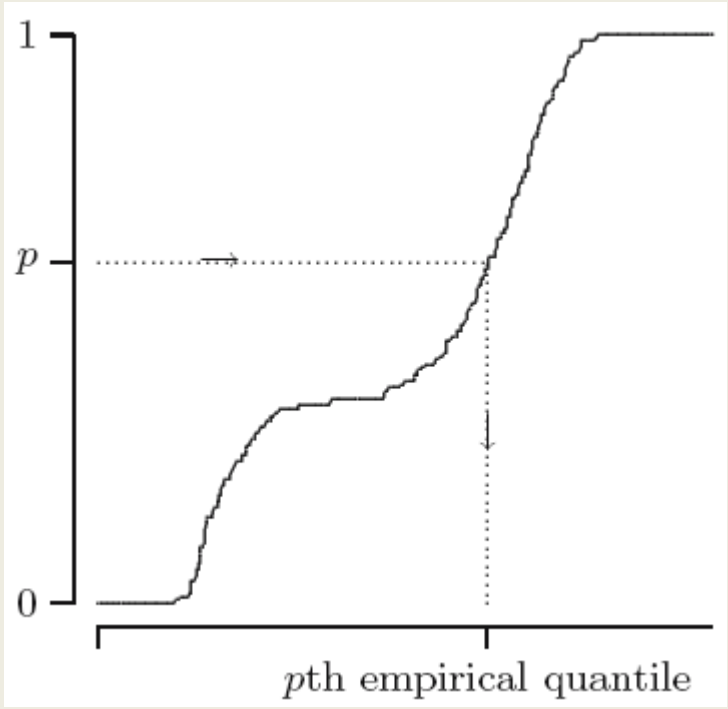
- Hledání empirického kvantilu je vlastně lineární interpolace mezi uspořádanými prvky statistického souboru.
- Nechť $0 < p < 1$. Pro výpočet $q_n(p)$ empirického kvantilu požadujeme, aby *i*-tý prvek uspořádaného statistického souboru byl $i/(n+1)$ kvantil.
- Nechť celá část obecného reálného čísla *a* je označena jako $\lfloor a \rfloor$. Potom výpočet $q_n(p)$ je dán:

$$q_n(p) = x_{(k)} + \alpha(x_{(k+1)} - x_{(k)})$$

kde $k = \lfloor p(n+1) \rfloor$ a $\alpha = p(n+1) - k$.

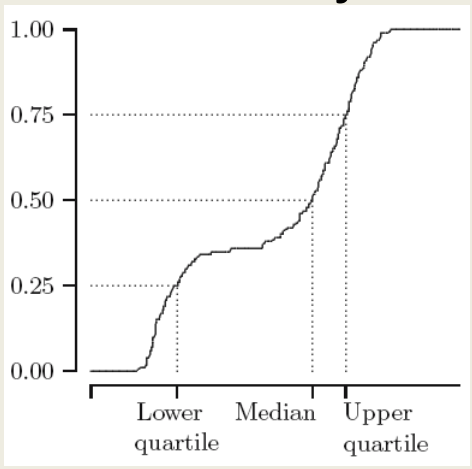
Empirické kvantily

- Na obrázku je empirická distribuční funkce statistického souboru ze str. 25.
- Je zde ilustrován empirický $q_n(p)$ kvantil.



Mezikvartilové rozpětí

- Místo hledání středu (průměru) statistického souboru bylo navrženo 5 číselných charakteristik shrnujících vlastnosti statistického souboru:
  - Minimum
  - Maximum
  - výběrový medián
  - 0,25 empirický kvantil
  - 0,75 empirický kvantil



- Empirický kvantil $q_n(0{,}25)$ se nazývá **první kvartil** a $q_n(0{,}75)$ se nazývá jako **třetí kvartil**.
- Společně s výběrovým mediánem mi 1. a 3. empirický kvartil rozdělují statistický soubor na 4 více méně stejné části obsahující každá čtvrtinu všech prvků – viz obrázek.

Mezikvartilové rozpětí

- Vzdálenost mezi 1. kvartilem a mediánem vzhledem ke vzdálenosti mezi mediánem a 3. kvartilem mi charakterizuje míru šikmosti statistického souboru.
- Vzdálenost mezi 1. a 3. kvartilem se nazývá jako **mezikvartilové rozpětí** (IQR):

$$\mathrm{IQR} = q_n(0.75) - q_n(0.25).$$

- Specifikuje mi rozsah prvků statistického souboru, které vymezují střední polovinou statistického souboru.
- Je to velmi silná míra variability statistického souboru.
- IQR je odolné proti mimořádným hodnotám.

Krabicový graf

- K vizualizaci 5 základních číselných charakteristik statistického souboru se používá tzv. **krabicový graf**.
- Je to symbolické zobrazení statistického souboru.
- Na svislé ose jsou číselné charakteristiky statistického souboru. Horizontální šířka grafu je libovolná.
- Obdélník je vymezen 1. a 3. empirickým kvartilem, tedy jeho výška je rovna IQR.
- Uprostřed obdélníku je vyznačen výběrový medián.
- Nad a pod obdélníkem vyznačujeme vzdálenost rovnou 1,5×IQR.
- Horizontální čarou označíme nejvyšší (nejnižší) hodnoty prvků statistického souboru, které ještě leží v intervalu 1,5×IQR.
- Všechny ostatní prvky statistického souboru ležící mimo obdélník a mimo vzdálenost 1,5×IQR nazýváme jako mimořádné (odlehlé) hodnoty.
- Poloha výběrového mediánu uvnitř obdélníku naznačuje šikmost souboru.
- Krabicové grafy jsou důležité pro rychlé a názorné porovnání více statistických souborů.
- Pro zobrazení dat jednoho statistického souboru je vhodnější histogram.

Krabicový graf

- Krabicový graf statistického souboru ze str. 25.