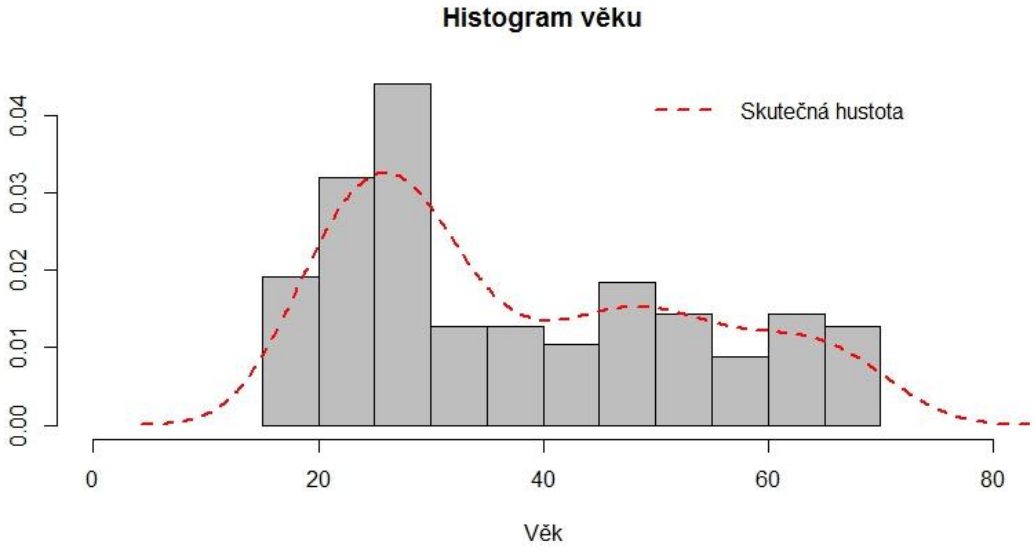
I Grafická reprezentace dat

Tabulky je vhodné při prezentaci vhodné doplnit správným typem grafu. Představíme si ty nejzákladnější, jejich konstrukci a jak se v nich orientovat.

Histogram a sloupcový graf (bar chart)

Histogram je graf, který se používá pro vizualizaci dat intervalového typu, kdy intervaly jsou stejně široké a navazují na sebe. Každému intervalu na ose x odpovídá obdélník, jehož výškou je četnostní hustota dat z daného intervalu v náhodném výběru. Například pro data o věku respondentů zadané tabulkou níže bude histogram relativních četností vypadat následovně (pozor, je třeba zachovat shodnou šířku všech intervalů, tj. rozšířit první na $\langle 15; 20)$.

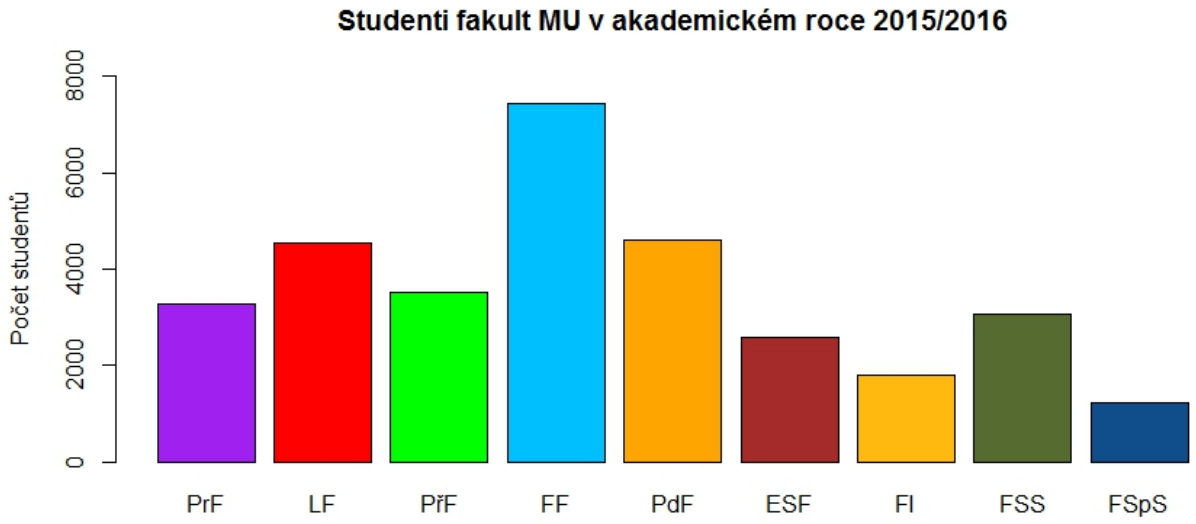
Věk respondenta					
Věk	Střed intervalu	Absolutní četnost	Kumulativní četnost	Relativní četnost	četnostní hustota
$\langle 18, 20)$	19	10	10	0,040	0,02
$\langle 20, 25)$	22,5	43	53	0,172	0,0344
$\langle 25, 30)$	27,5	49	102	0,196	0,0392
$\langle 30, 35)$	32,5	30	132	0,120	0,024
$\langle 35, 40)$	37,5	14	146	0,056	0,0112
$\langle 40, 45)$	42,5	15	161	0,060	0,012
$\langle 45, 50)$	45,5	24	185	0,096	0,0192
$\langle 50, 55)$	52,5	18	203	0,072	0,0144
$\langle 55, 60)$	57,5	12	215	0,048	0,0096
$\langle 60, 65)$	62,5	17	232	0,068	0,0136
$\langle 65, 70\rangle$	67,5	18	250	0,072	0,0144



Histogram také slouží jako odhad hustoty – skutečnou hustotu, ze které data pocházejí obvykle neznáme, nicméně tvar histogramu nám může dát informaci, jakým rozdělením pravděpodobnosti skutečné rozdělení modelovat.

Pokud data nejsou intervalového typu, namísto histogramu používáme sloupcový graf (bar chart). U něj jsou mezi jednotlivými sloupci mezery, také je třeba důsledně pojmenovat všechny veličiny, které vynášíme na osu x. Jednotlivé obdélníky opět mají stejnou šířku a jejich výška odpovídá četnosti dané kategorie v náhodném výběru. Na obrázku je zachycen sloupcový graf studentů jednotlivých fakult MU v roce akademickém roce 2015/2016.

Počty studentů jednotlivých fakult MU v roce 2015/2016		
Fakulta	počet studenů	relativní četnost
Právnická fakulta	3267	0,102
Lékařská fakulta	4529	0,142
Přírodovědecká fakulta	3526	0,110
Filozofická fakulta	7422	0,232
Pedagogická fakulta	4607	0,144
Ekonomicko-správní fakulta	2589	0,081
Fakulta informatiky	1797	0,056
Fakulta sociálních studií	3063	0,096
Fakulta sportovních studií	1214	0,038
součet	32 014	1



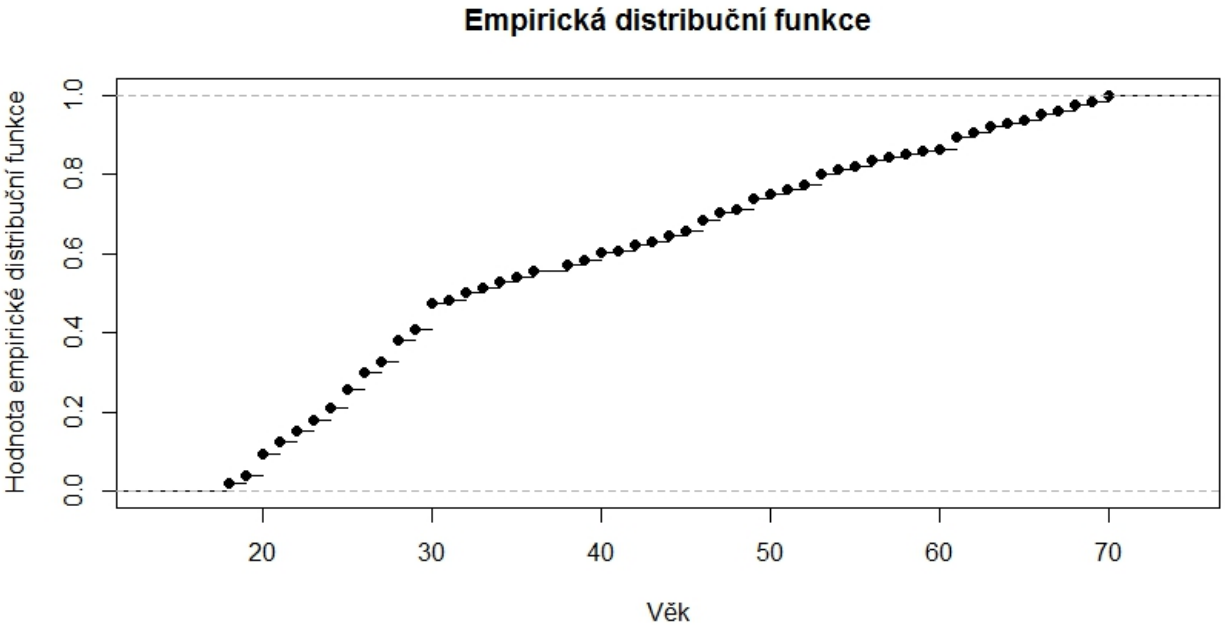
II Empirická distribuční funkce

Empirická distribuční funkce slouží jako odhad skutečné distribuční funkce náhodné veličiny. Označíme-li card(M) velikost množiny M (velikostí míníme počet jejích prvků), pak empirickou distribuční funkci můžeme spočítat následovně.

Nový pojem: Empirická distribuční funkce

Hodnota empirická distribuční funkce $F(x)$ náhodného výběru $X_1, X_2, \ldots, X_n$ v bodě x je

$$F(x) = \frac{\text{card}\{i, X_i \leq x\}}{n}.$$

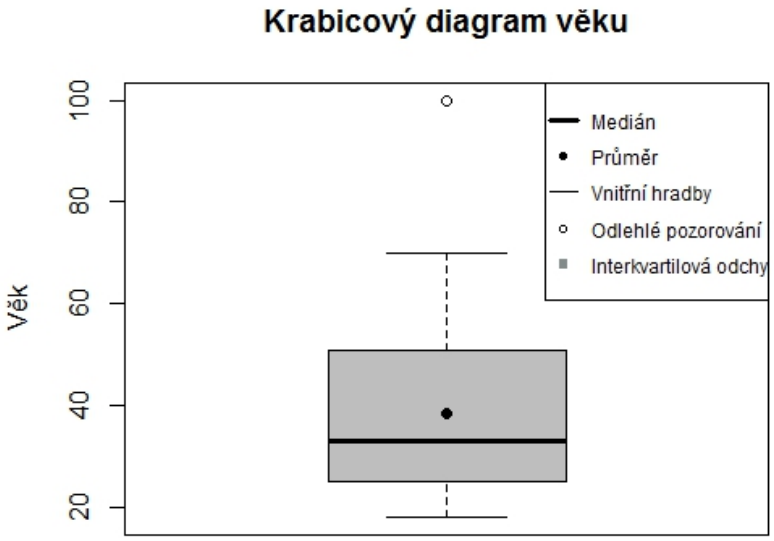


Interpretační poznámka. Zjednodušeně, spočítáme kolik náhodných veličin v našem náhodném výběru nabylo hodnoty menší nebo rovno x a podělíme velikostí náhodného výběru n. Kdybychom konstruovali tabulku kumulativních relativních četností pro všechny realizace (ne intervalovou), pak právě tyto hodnoty jsou i hodnotami empirické distribuční funkce.

Empirická distribuční funkce je schodovitá, se zvětšující se velikostí náhodného výběru n se blíží teoretické distribuční funkci. Pro data z příkladu o věku respondenta je empirická distribuční funkce zachcena na tomto obrázku.

Krabicový diagram (box plot)

Krabicový diagram možná vypadá nepřehledně, nicméně nám dává spoustu informací o datovém souboru. Existuje více typů krabicových diagramů (proto je potřeba si vždy nejdřív rozmyslet, co vlastně krabicový diagram znázorňuje, o tom nám dává informaci legenda grafu), my se seznámíme pouze s jedním z nich.
Začneme ukázkou krabicového diagramu (jedná se o data „Věk respondenta“ s přidanou jednou hodnotou 100).



Nejprve si všimneme šedého obdélníčku. Jeho horní strana je horní kvartil, spodní strana dolní kvartil. Mezi těmito hodnotami je obsaženo 50% získaných dat. Uvnitř je vyznačen medián a výběrový průměr. Širší interval ohraničují tzv. vnitřní hradby. Horní vnitřní hradba má hodnotu „horní kvartil + 1,5 interkvartilová odchylka“, (nebo-li $x_{0,75} + 1,5q$) , dolní vnitřní hradba analogicky $x_{0,25} - 1,5q$. Hodnoty, které leží mimo vnitřní hradby označujeme jako odlehlé. Dále se počítají vnější hradby, horní vnější hradba je $x_{0,75}+3q$, dolní vnější hradby $x_{0,25}-3q$. Pokud máme hodnoty mimo vnější hradby, říkáme jim extrémní. Je třeba zamyslet se, zda-li jsme případná odlehlá a extrémní data získali přirozeným způsobem, nebo se jedná o chybu. Pokud chybí odlehlé a extrémní hodnoty, pak se do krabicového diagramu namísto vnitřních hradeb nanáší maximum, resp. minimum z dat.

Krabicový diagram nám umožňuje odhadnout pomocí interkvartilové odchylky variabilitu dat a podle symetrie šikmost – v našem případě data nejsou symetricky rozdělená kolem průměru, více se jich nachází pod průměrem (medián $<$ průměr).