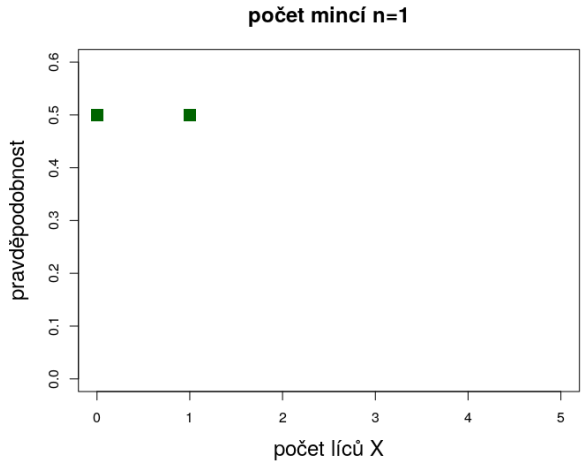
Marie Turčičová

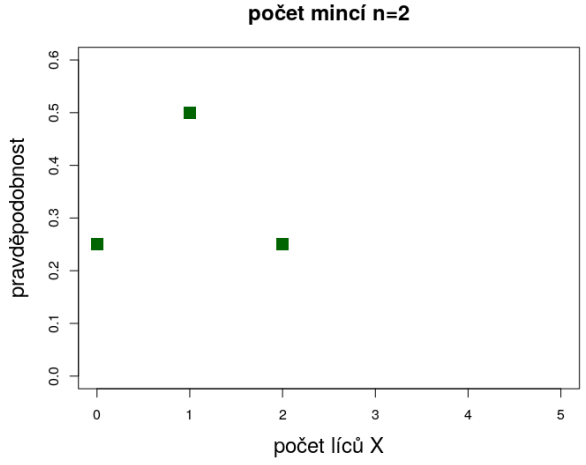
Centrální limitní věta

Centrální limitní věta (dále jen CLV, angl. *central limit theorem*) je úzce spjata s normálním rozdělením. Tomuto rozdělení se někdy také říká Gaussovo, ale Carl Friedrich Gauss nebyl první, kdo ho vymyslel. Autorem normálního rozdělení byl Abraham de Moivre[1], který ho roku 1733 publikoval ve svém článku právě společně s tvrzením, kterému dnes říkáme centrální limitní věta. Podívejme se nyní na situaci, která odpovídá jeho přístupu.

Představme si, že budeme házet sadou mincí. Počet těchto mincí bude různý, označme si ho jako n. Mince vždy hodíme a spočítáme, na kolika z nich padl líc. Počet líců si označme X. Dále si uvědomme, jaké jsou pro dané n pravděpodobnosti různých počtů líců. Je-li $n = 1$, tj. házím pouze jednou mincí, tak X může nabývat hodnoty 0 nebo 1, přičemž $P(X = 0) = \frac{1}{2}$ a $P(X = 1) = \frac{1}{2}$. Zakresleme si to do grafu:



Budeme-li házet dvěma mincemi ($n = 2$), tak počet líců X může nabývat hodnot 0, 1, nebo 2. Přitom $P(X = 0) = P(X = 2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$ a $P(X = 1) = 2 \cdot \frac{1}{4} = \frac{1}{2}$ (rozmyslete si, že to tak opravdu je). Vynesme si to opět do grafu:

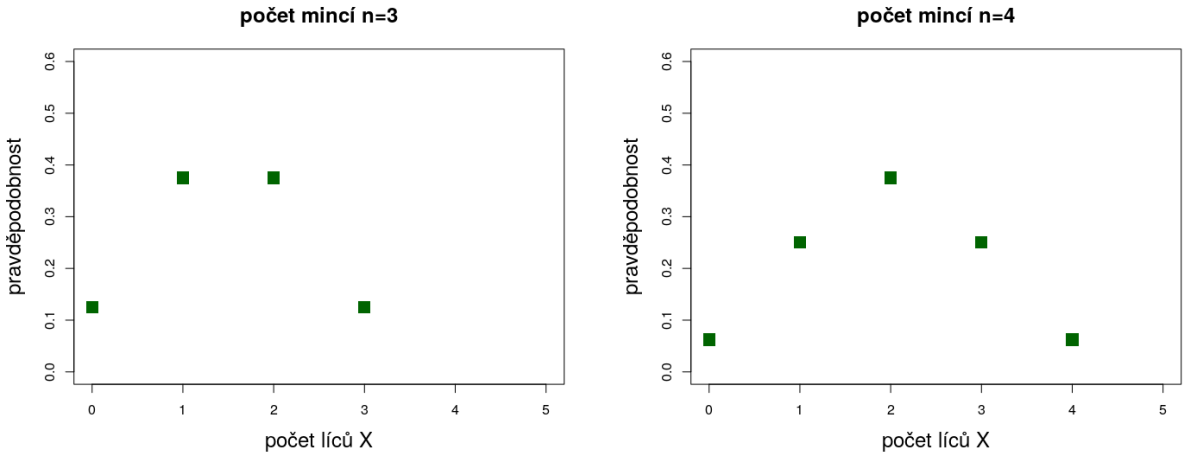


Pro $n = 3$ je $P(X = 0) = P(X = 3) = \frac{1}{8}$ a $P(X = 1) = P(X = 2) = \frac{3}{8}$.

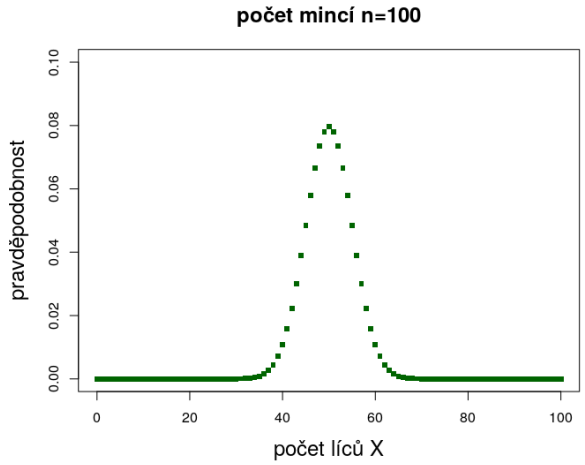
Pro $n = 4$ pak máme $P(X = 0) = P(X = 4) = \frac{1}{16}$, $P(X = 1) = P(X = 3) = \frac{1}{4}$ a $P(X = 2) = \frac{3}{8}$.

[1] Moivre je autorem řady dalších matematických výsledků, například věty o umocňování komplexních čísel, kterou si možná pamatujete ze střední školy.

Graficky to vypadá takto:



Kdybychom takhle pokračovali dál a dál... stále s větším počtem mincí, začal by náš graf vypadat takto:



... tedy začal by připomínat hustotu normálního rozdělení, tzv. Gaussovu křivku. Toto pozorování je vlastně obsahem centrální limitní věty. Ta říká, že pravděpodobnostní rozdělení součtu veličin (v tomto případě součet líců z n mincí) se pro velká n (tj. když mincí bude hodně) bude blížit normálnímu rozdělení.

Pokusme se toto pozorování formulovat trochu přesněji. Asi jste si uvědomili, že výsledek hodu mincí je náhodná veličina s alternativním rozdělením. Označme si tuto veličinu jako Y a „zakódujme“ si $0 =$ rub, $1 =$ líc. Pak $P(Y = 1) = \frac{1}{2}$. Symbolicky to můžeme zapsat jako $Y \sim alt(\frac{1}{2})$. Mincí máme celkem n, tedy máme vlastně n veličin: $Y_1, Y_2, \ldots, Y_n$. Předpokládáme, že všechny mince jsou stejné a spravedlivé, tedy všechny veličiny $Y_1, \ldots, Y_n$ mají stejné rozdělení $alt(\frac{1}{2})$. Navíc to, co padne na jedné minci, nijak neovliňuje výsledky na dalších mincích, tedy $Y_1, \ldots, Y_n$ jsou navíc nezávislé. Náš součet líců je pak veličina $X = Y_1 + Y_2 + \ldots + Y_n$, tedy součet nezávislých stejně rozdělených náhodných veličin. Na základě našich obrázků bychom teď mohli vyslovit jakousi neformální verzi CLV:

CLV neformálně:
Uvažujme součet n nezávislých stejně rozdělených náhodných veličin. S rostoucím n se rozdělení tohoto součtu (za určitých předpokladů) blíží k normálnímu rozdělení.

Nejspíš jste si všimli, že veličina X má binomické rozdělení s parametry n=*počet mincí* a $p = 1/2$, symbolicky zapsáno $X \sim Bi(n, \frac{1}{2})$. Tento náš příklad s házením mincemi tedy ilustruje i aproximaci binomického rozdělení normálním, o níž jste také slyšeli na přednášce.

Nyní se podíváme, co jsou to ty „určité předpoklady“ z našeho tvrzení v rámečku. Základní verze CLV požaduje pouze to, aby veličiny byly nezávislé, měly stejné pravděpodobnostní rozdělení, konečnou střední hodnotu a nenulový rozptyl[2]. Zejména ty poslední dva jmenované předpoklady jsou jen technické požadavky, které jsou potřeba k formálnímu matematickému důkazu tohoto tvrzení, ale v praxi nepředstavují výrazné omezení.

CLV formálně:

- Mám $Y_1, \ldots, Y_n$ nezávislé, stejně rozdělené a takové, že

$$\mathrm{E}\, Y_i = a \qquad \mathrm{var}\, Y_i = b^2, \qquad \text{kde } 0 < b^2 < \infty$$

  pro všechna $i = 1, \ldots, n$.

- Pak platí, že

$$\sum_{i=1}^{n} Y_i \overset{as.}{\sim} N(na, nb^2)$$

  (čti: rozdělení součtu veličin se pro velká n blíží normálnímu rozdělení se střední hodnotou na a rozptylem nb^2.)

- Ekvivalentně lze tvrzení formulovat i pro průměr $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$:

$$\bar{Y} \overset{as.}{\sim} N(a, \frac{b^2}{n})$$

$$\text{nebo} \quad \frac{\bar{Y} - a}{b}\sqrt{n} \overset{as.}{\sim} N(0, 1).$$

Této verzi CLV se říká Lindebergova, ale existuje spousta dalších verzí, které se zpravidla liší v předpokladech. Ještě doplňme, že zkratka *as.* znamená „asymptoticky“, to jest „pro velká n“. Značka $\overset{as.}{\sim}$ pak znamená „pro velká n má rozdělení“.

CLV je vzácným příkladem matematického tvrzení, které za zcela minimálních předpokladů dává silný výsledek. Takových případů moc není. Je to jako byste do výdejního automatu vhodili pětikorunu a dole vám vypadl dvoupatrový dort :o)

Poznámka 1 *Normální rozdělení se dříve s úspěchem používalo k modelování chyb fyzikálních měření a samotná CLV se tehdy nazývala „zákon chyb“. Vycházelo to z představy, že každé měření má chybu, která je součtem velkého množství drobných náhodných chybiček. Dle CLV má pak výsledná chyba přibližně normální rozdělení.*

Poznámka 2 ***(pro zájemce)*** *Centrální limitní větu lze velmi pěkně demonstrovat pomocí tzv. Galtonova prkna (angl. Galton board)*[3]*. Jedná se o svislé prkno, do něhož jsou v několika řadách zatlučeny hřebíky. Ze zásobníku nahoře se pak vysypou kuličky, které se během pádu dolů odrážejí od hřebíků. Každá kulička se na daném hřebíku odrazí vpravo nebo vlevo, obě strany mají přitom shodnou pravděpodobnost $p = \frac{1}{2}$. Ve spodní části prkna jsou přihrádky (představující jednotlivé „šuplíčky“ histogramu), ve kterých se kuličky hromadí. Názorně si to můžete prohlédnout na jednom z videí, kterých lze na internetu nalézt celou řadu:* `https://www.youtube.com/watch?v=4HpvBZnHOVI`*. Označíme-li si počet řad hřebíků nad přihrádkami jako n, pak každá kulička provede právě n odrazů. Každý odraz je přitom náhodná veličina s alternativním rozdělením $alt(\frac{1}{2})$ (zakódujme se například, že $0 =$ vlevo, $1 =$ vpravo). Označme si výsledek odrazu na i-tém hřebíku jako Y_i. Výsledná poloha kuličky (tj. to, do které přihrádky nakonec spadne) je dáno tím,*

[2] Splňuje náš experiment s házením mincemi tyto předpoklady?

[3] Jeho autorem je Francis Galton (1822-1911), všestranný anglický vědec, který za pomoci statistiky posunul vpřed řadu vědních oborů, zejména vědy o člověku (antrolopogie, daktyloskopie, sociologie, psychologie a další).

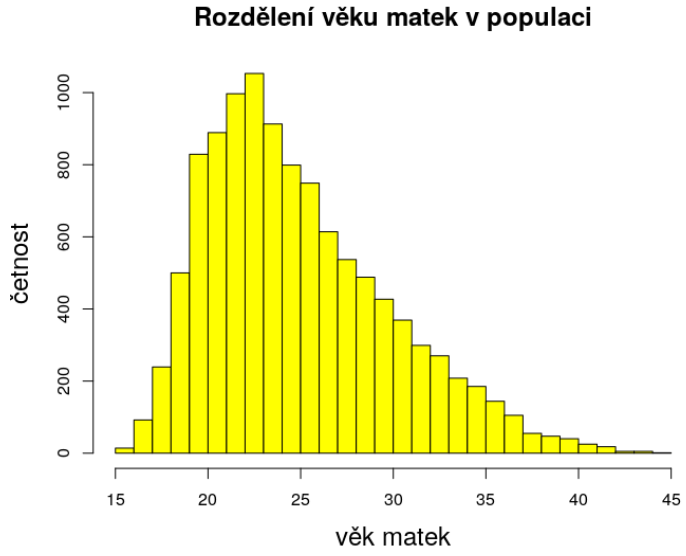
kolikrát nastal odraz vpravo a jde tedy o součet veličin $Y_1 + \ldots + Y_n$, což je opět veličina s binomických rozdělením. Počet kuliček v dané přihrádce pak odpovídá tomu, jak pravděpodobná je hodnota příslušného součtu. Dle CLV lze (pro velká n) tuto pravděpodobnost modelovat Gaussovou křivkou. Je to zcela analogický příklad jako při házení mincemi, akorát místo n mincí máme nyní n hřebíků. V příkladu s mincemi jsme také uvažovali přesné pravděpodobnosti, zde je máme pouze odhadnutí pomocí podílu kuliček.

Pro zajímavost dodejme, že sám Galton si při popisu tohoto prkna ve své práci poznamenal: „Statistika je jediným nástrojem, jímž lze prosekat otvor do hrozivého houští potíží bránících v cestě těm, kde se věnují vědě o člověku.“

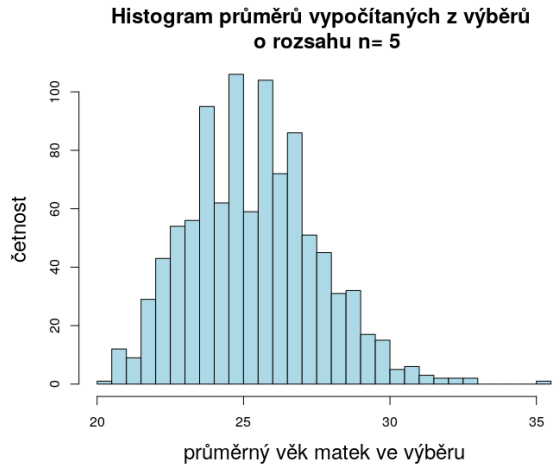
To, že chování průměrů (nebo součtů) z rozsáhlých výběrů lze dobře popsat pomocí normálního rozdělení jste viděli i na příkladu z přednášky (slide 166). Vezměme si velký soubor reálných dat naměřený v 90. letech, který obsahuje přes 10 tisíc údajů o novorozených dětech a jejich matkách. Tento soubor je tak velký, že ho nyní ze studijních důvodů prohlásíme za celou vyšetřovanou populaci. Z této populace budeme postupně pořizovat výběry s různým rozsahem.

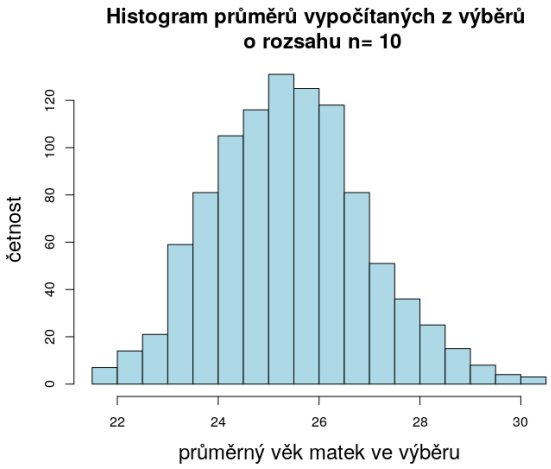
Ze všech naměřených údajů nás nyní bude zajímat pouze věk matek. Z histogramu (viz žlutý histogram níže) je vidět, že rozdělení věku matek není symetrické, takže určitě nejde o normální rozdělení.

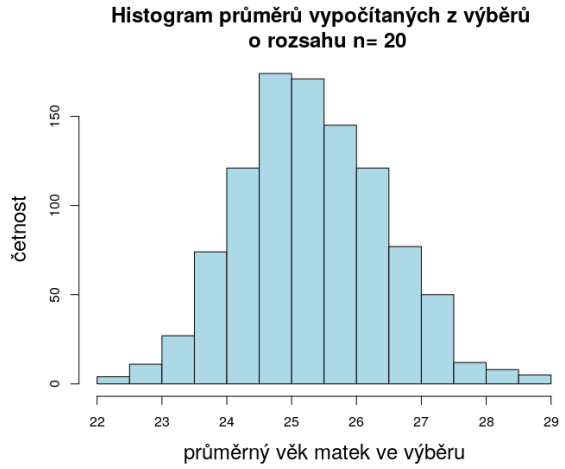
Nyní si zvolíme nějaké n a necháme z naší populace pořídit 1000 náhodných výběrů s rozsahem n. U každého takového výběru spočítáme průměr. Spočítané průměry si zapamatujeme a následně vykreslíme do histogramu. Postupně zvolíme $n = 5, 10, 20, 100$ a pokaždé vykreslíme tento histogram průměrů (viz modré histogramy níže). Z obrázků je patrné, že pro velké n histogram průměrů již velmi dobře připomíná normální rozdělení, přestože samotný věk matek normálnímu rozdělení neodpovídá. To je opět projev centrální limitní věty.

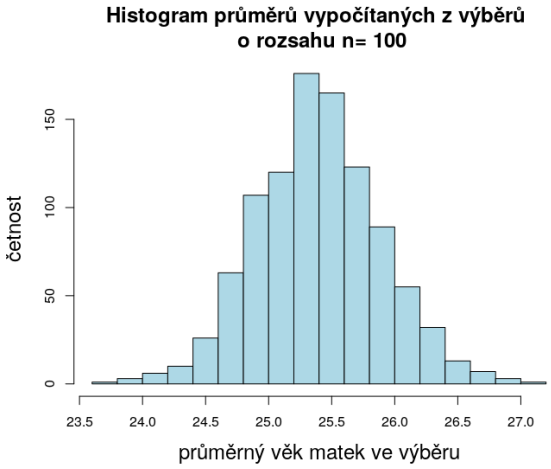


Poznámka 3 *Centrální limitní věta má dalekosáhlé důsledky. Na jejím základě je například odvozena řada statistických testů, se kterými se potkáme později v semestru.*


Histogram průměrů vypočítaných z výběrů
o rozsahu n= 5
četnost
0
20
40
60
80
100
20
25
30
35
průměrný věk matek ve výběru


Histogram průměrů vypočítaných z výběrů
o rozsahu n= 10
četnost
0
20
40
60
80
100
120
22
24
26
28
30
průměrný věk matek ve výběru


Histogram průměrů vypočítaných z výběrů
o rozsahu n= 20
četnost
0
50
100
150
22
23
24
25
26
27
28
29
průměrný věk matek ve výběru


Histogram průměrů vypočítaných z výběrů
o rozsahu n= 100
četnost
0
50
100
150
23.5
24.0
24.5
25.0
25.5
26.0
26.5
27.0
průměrný věk matek ve výběru