Centrální limitní věta

Věta: Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny, které mají stejné rozdělení pravděpodobnosti, stejné střední hodnoty a stejné konečné rozptyly. Tj. $\forall i = 1, \ldots, n : EX_i = \mu \quad , DX_i = \sigma^2 \in \mathbb{R}$.

Potom platí:

$$X = \frac{\left(\sum_{i=1}^{n} X_i\right) - n \cdot \mu}{\sigma \cdot \sqrt{n}} \Rightarrow X \text{ má asymptoticky normální rozdělení } N(0, 1^2)$$

Poznámka: $X = \sum_{i=1}^{n} X_i \Rightarrow$ pro „velká" n : $X \xrightarrow{\cdot} N(n\mu, (\sigma \sqrt{n})^2)$

$X = \frac{\sum_{i=1}^{n} X_i}{n} \Rightarrow$ pro „velká" n : $X \xrightarrow{\cdot} N(\mu, (\frac{\sigma}{\sqrt{n}})^2)$

Poznámka: Centrální limitní větu můžeme použít k aproximaci rozdělení pravděpodobnosti n.v. X (skutečnou neznáme, nebo je obtížné ji vypočítat).

Ale! Ideální by bylo, kdyby $n \to \infty$, v případě, že $n < \infty$ jde jen o odhad. CLV obvykle dává dobré odhady v případě, že:

$$\boxed{n > 30}$$

Pokud tato podmínka není splněna, raději CLV nepoužijeme. Pokud ano, stále musíme mít na paměti, že jde jen o odhad! - Jeho chybu velmi ovlivňují rozdělení n.v. X_i.

Př. Kinetická energie částice je náhodná veličina, která má střední hodnotu $\mu = 100\,J$ a směrodatnou odchylku $\sigma = 40\,J$.

a) Odhadněte pravděpodobnost, že celková kinetická energie 81 částic má hodnotu v intervalu $(8000\,J, 8200\,J)$.

X_i ... kinetická energie i-té částice $\quad EX_i = \mu = 100\,J, \quad \sigma = DX_i = 40\,J$

$X = \sum_{i=1}^{81} X_i$ ... kinetická energie 81 částic $(n = 81)$

$\boxed{n > 30 \Rightarrow \text{Použijeme CLV}}$: $\quad X = \sum X_i \xrightarrow{\cdot} N(\underbrace{n \cdot \mu}_{=\mu^*}, \underbrace{(\sigma\sqrt{n})^2}_{=\sigma^{*2}})$

$$\left.\begin{aligned} \Rightarrow EX &\doteq n \cdot \mu = 81 \cdot 100 = 8100\,J \\ \sqrt{DX} &\doteq \sqrt{n} \cdot \sigma = \sqrt{81} \cdot 40 = 360\,J \end{aligned}\right\} \Rightarrow X \xrightarrow{\cdot} N(8100, 360^2)$$

$\Rightarrow P(8000 < X < 8200) = P(X < 8200) - P(X < 8000) \doteq \underline{\underline{0{,}2188}}$

pnorm(8200, 8100, 360) - pnorm(8000, 8100, 360)

b) Odhadněte pravděpodobnost, že kinetická energie 810000 částic je v intervalu $(80\,000\,000\,J, 82\,000\,000\,J)$.

$\boxed{n = 810\,000 > 30 \Rightarrow \text{Použijeme CLV:}}$ $\quad X_i$ ... kin. en. i-té částice $\Rightarrow$

$$\underset{\substack{\uparrow \\ \text{celková kin. en.}}}{X} = \sum_{i=1}^{810000} X_i \xrightarrow{\cdot} N(\underbrace{n\mu}_{=810000 \cdot 100}, \underbrace{(\sigma\sqrt{n})^2}_{=40 \cdot \sqrt{810000} = 40 \cdot 900}) = N(81\,000\,000, 36\,000^2)$$

$P(80\,MJ < X < 82\,MJ) = P(X < 82 \cdot 10^6) - P(X < 80 \cdot 10^6) \doteq \underline{\underline{1}}$

Střední hodnota součtu je úměrná n, ale rozptyl je úměrný $\sqrt{n}$.

$$\frac{\sigma^*}{\mu^*} = \frac{\sqrt{n} \cdot \sigma}{n \cdot \mu} = \underbrace{\frac{1}{\sqrt{n}}}_{\to 0} \cdot \underbrace{\frac{\sigma}{\mu}}_{=\text{konst.}} \longrightarrow 0$$

$\Rightarrow$ S rostoucím počtem částic bude růst pravděpodobnost, že celková energie je „blízko“ střední hodnoty.

c) Odhadněte pravděpodobnost, že průměrná hodnota kin. energie 81 částic zjištěná měřením je v intervalu (95 J ; 105 J)

$\boxed{n = 81 > 30 \Rightarrow \text{Použijeme CLV:}}$

X_i ... energie jedné částice $\mu = EX_i = 100J$; $\sigma = \sqrt{DX_i} = 40J$.

$$\Rightarrow \bar{X} = \frac{\sum_{i=1}^{81} X_i}{81} \xrightarrow{\cdot} N\left(\mu ; \left(\frac{\sigma}{\sqrt{n}}\right)^2\right) = N\left(100, \left(\frac{40}{\sqrt{81}}\right)^2\right) = N\left(100, \left(\frac{40}{9}\right)^2\right)$$

$$P(95 < \bar{X} < 105) = P(\bar{X} < 105) - P(\bar{X} < 95) = 0{,}7394$$

pnorm(105, 100, 40/9) - pnorm(95, 100, 40/9)

d) Odhadněte pravděpodobnost, že průměrná hodnota kinetické energie 810 000 částic zjištěná v souboru celkem je v intervalu (95 J ; 105 J).

$\boxed{n = 810\,000 > 30 \Rightarrow \text{Použijeme CLV:}}$

X_i ... energie 1 částice $\Rightarrow$ $\mu = EX_i = 100J$; $\sigma = \sqrt{DX_i} = 40J$

$$\bar{X} = \frac{\sum_{i=1}^{810000} X_i}{}$$ ... prům. en. 1 částice z 810000 celkem $\Rightarrow \bar{X} \xrightarrow{\cdot} N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$

$$\Rightarrow \bar{X} \xrightarrow{\cdot} N\left(100, \frac{40}{\sqrt{810000}}\right) = N\left(100, \left(\frac{40}{900}\right)^2\right) = N\left(100, \left(\frac{4}{90}\right)^2\right)$$

$$P(95 < \bar{X} < 105) = P(\bar{X} < 105) - P(\bar{X} < 95) \doteq 1$$

Všimněme si: $P(99{,}05 < \bar{X} < 100{,}05) \doteq 0{,}7394$

$$\frac{\sigma_{\bar{X}} \text{ v c)}}{\sigma_{\bar{X}} \text{ v d)}} = \frac{\frac{40}{9}}{\frac{4}{90}} = \frac{40 \cdot 90}{9 \cdot 4} = 100$$ .... v c) je směrodatná odch. 100 krát větší než v d) $\Rightarrow$ rozptyl v c) je 10 000 krát větší než v d)

Př. Pan doktor Vomáčka dlouhodobým měřením zjistil, že pacient u něj v ordinaci stráví průměrně 7 minut a směrodatná odchylka této doby je 2 minuty.

a) Odhadněte pravděpodobnost, že během 4 hodin ordinační doby stihne ošetřit všech 35 pacientů, kteří mu dnes přišli do ordinace (nebude mít žádné přestávky).

X_i ... doba, kterou potřebuje na 1 pacienta

$EX_i = \mu \doteq 7$ min $\qquad \sqrt{DX_i} \doteq \sigma = 2$ min

$X = \sum_{i=1}^{35} X_i$ .... doba nutná pro ošetření 35 pacientů

$\boxed{n = 35 > 30 \Rightarrow \text{Použijeme CLV}} \Rightarrow X \dot{\to} N(\underset{35}{n} \cdot \underset{7}{\mu}, (\underset{\sqrt{35}}{\sqrt{n}} \cdot \underset{2}{\sigma})^2) = N(245, (2 \cdot \sqrt{35})^2)$

4 hodiny $= 4 \cdot 60 = 240$ minut $\Rightarrow$ Hledáme $P(X < 240)$

$P(X < 240) \doteq \underline{\underline{0,3363}} \doteq$ pnorm(240, 245, 2* sqrt(35))

b) Odhadněte pravděpodobnost, že průměrná doba ošetření jednoho pacienta bude menší než 6 minut, přijde-li $\underbrace{\text{36 pacientů}}_{n>30}$.

$\bar{X} = \dfrac{\sum_{i=1}^{36} X_i}{36}$ .... prům. doba ošetření $\quad \bar{X} \dot{\to} N(\underset{7}{\mu}, (\frac{\sigma}{\sqrt{n}})^2) = N(7, (\frac{1}{3})^2)$

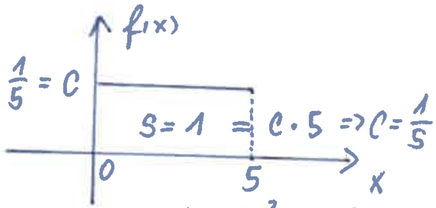
$\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{36}} = \frac{2}{6} = \frac{1}{3}$

$P(\bar{X} < 6) \doteq \underline{\underline{0,0013}} \doteq$ pnorm(6, 7, 1/3)

Př. Karel jezdí do práce autobusem, který má odjezdy co 5 minut. Karel přichází na zastávku zcela náhodně.

a) Odhadněte kolik celkem času stráví Karel čekáním na autobus do práce během 100 pracovních dnů.

X_i ... doba čekání na autobus i-tý den



$X_i \Rightarrow R_o(a,b) = R_o(0,5) \Rightarrow EX_i = 2{,}5 \qquad DX_i = \frac{(a-b)^2}{12} = \frac{5^2}{12} \Rightarrow$

$\boxed{n = 100 > 30 \Rightarrow CLV} \qquad \Rightarrow \sigma = \sqrt{\frac{5^2}{12}} = \frac{5}{\sqrt{12}}$

$X = \sum_{i=1}^{100} X_i$ ... celková doba čekání $\Rightarrow \quad X \xrightarrow{} N(\underset{100}{n} \cdot \underset{2{,}5}{\mu}, (\underset{10}{\sqrt{n}} \cdot \underset{\frac{5}{\sqrt{12}}}{\sigma})^2)$

$X \xrightarrow{} N(250, (\frac{50}{\sqrt{12}})^2)$

$\Rightarrow EX \doteq n \cdot \mu = 100 \cdot 2{,}5 = \underline{\underline{250 \text{ minut}}}$

b) Odhadněte pravděpodobnost, že během těchto 100 dní stráví Karel čekáním na autobus do práce více než 280 minut.

$P(X > 280) \doteq 1 - P(X < 280) \doteq 0{,}0188$

1 - pnorm(280, 250, 50/sqrt(12))

Př. Předpokládejme, že životnost žárovky má exponenciální rozdělení pravděpodobnosti se střední hodnotou 2000 hodin.

a) Máme 50 žárovek. Žárovku vyměníme okamžitě poté co se pokazí. Odhadněte pravděpodobnost, že nám 50 žárovek vystačí na 120000 hodin svícení (vždy svítí jen jedna).

X_i ... životnost 1 žárovky $\Rightarrow$ $EX_i = \frac{1}{\lambda} = 2000\,h$, $DX_i = \frac{1}{\lambda^2} = 2000^2$.

$\Rightarrow \mu = 2000\,h \qquad \sigma = \sqrt{DX_i} = 2000\,h$

$X = \sum_{i=1}^{50} X_i$ .... životnost 50 žárovek (po sobě) $\qquad \boxed{n = 50 > 30 \Rightarrow CLV}$

$$X \dot{\rightarrow} N(n \cdot \mu, (\sqrt{n}\,\sigma)^2) = N(\underbrace{50 \cdot 2000}_{100\,000}, (\sqrt{50} \cdot 2000)^2)$$

$$\Rightarrow P(X > 120\,000\,h) = 1 - P(X < 120\,000) \doteq \underline{\underline{0{,}0786}}$$

1 - pnorm(120000, 100000, 2000*sqrt(50))

Skutečná hodnota:

$X = \sum_{i=1}^{k} X_i$, kde $X_i \rightarrow Exp(\frac{1}{\lambda})$ $\Rightarrow$ X má tzv. Erlangovo (Gamma) rozdělení pravděpodobnosti; $\frac{1}{\lambda} = EX_i = 2000$

$X \rightarrow Erlang(k, \frac{1}{\lambda})$ ... doba do k-tého „úspěchu" - prasknutí žárovky; $k = 50$, $\frac{1}{\lambda} = 2000 \Rightarrow \lambda = \frac{1}{2000}$

$$F(x) = 1 - e^{-\lambda x} \cdot \sum_{j=0}^{k-1} \frac{(\lambda x)^j}{j!} \quad \Rightarrow \quad P(X > 120000) = 1 - P(X < 120\,000) =$$

$= 1 -$ pgamma(120000, $\underbrace{\text{shape} = 50}_{k}$, $\underbrace{\text{rate} = 1/2000}_{\lambda}$) $=$

$\doteq \underline{\underline{0{,}0844}}$

Př. 15% zákazníků si u stánku koupí hamburger (zjištěno dlouhodobým měřením).

a) Určete pravděpodobnost, že si ve středu koupilo hamburger 60 (respektive 60 nebo více) zákazníků, víte-li, že ve středu nakupovalo právě 375 zákazníků.

Přesný výpočet

X ... počet koupených hamburgerů = počet „úspěchů" při 375 pokusech, kde pravděp. úspěchu $p \doteq 0{,}15$ ($375 = n$)

$X \to Bi(n;p) = Bi(375;0{,}15)$ $\Rightarrow EX = n\cdot p = 56{,}25$, $\sqrt{DX} = \sqrt{n\cdot p(1-p)} = \sqrt{47{,}8125}$

$P(X=60) = \binom{375}{60} 0{,}15^{60}\cdot(1-0{,}15)^{315} \doteq \underline{0{,}0486} \doteq$ dbinom(60, 375, 0.15)

$P(X \geq 60) = 1 - P(X \leq 59) = 1 - \sum_{k=0}^{59} \binom{375}{k} 0{,}15^{k}(1-0{,}15)^{375-k} = \underline{\underline{0{,}3145}}$

1 − pbinom(59, 375, 0.15)

b) Odhadněte stejné pravděpodobnosti pomocí CLV.

$n = 375 > \frac{9}{0{,}15\cdot(1-0{,}15)} \doteq 71$

Doporučovaná podmínka pro použití CLV splněna

Nechť $X_i = \begin{cases} 0 \Leftrightarrow i\text{-tý zákazník nekoupí hamburger} \\ 1 \Leftrightarrow i\text{-tý zákazník koupí hamburger} \end{cases} \Rightarrow X = \underbrace{\sum_{i=1}^{375} X_i}_{\text{počet koupených hamb.}}$ ($375 = n$)

kde: $X_i \to Bi(1; 0{,}15)$ ... počet úsp. při jednom pokusu ... binom. náh. veličina

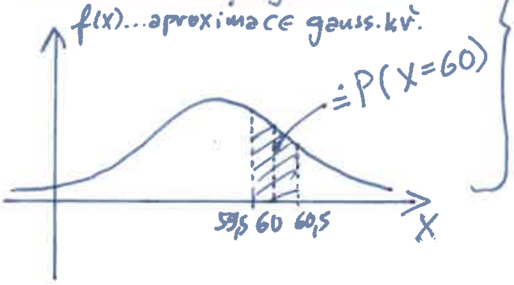
$\Rightarrow$ (viz vybrané diskr. náh. vel.) $EX_i = \mu = n\cdot p = 0{,}15 = p$

$\sqrt{DX_i} = \sigma = \sqrt{\underset{1}{n}\cdot p\cdot(1-p)} = \sqrt{0{,}15\cdot 0{,}85} = \sqrt{p\cdot(1-p)}$

$\Rightarrow X \xrightarrow{\cdot} N(n\cdot\mu, \sqrt{n}\,\sigma) = N(n\cdot p, (\sqrt{n}\sqrt{p(1-p)})^2)$

$\Rightarrow X \to Bi(n,p)$ aproximujeme $N(\underbrace{n\cdot p}_{EX}, \underbrace{\sqrt{n\cdot p\cdot(1-p)}^2}_{DX})$

Oprava na spojitost!



$\Rightarrow P(X=60) \doteq P(X \leq 60{,}5) - P(X < 59{,}5) = \underline{0{,}0498}$

pnorm(60.5, 56.25, sqrt(375*0.15*0.85))...

$P(X \geq 60) \doteq 1 - P(X \leq 59{,}5) = \underline{0{,}3145}$

1 − pnorm(59.5, 56.25, sqrt(375*0.15*0.85))

$\boxed{\text{Př: V populaci má 6 \% lidí krevní skupinu AB.}}$

a) Odhadněte pravděpodobnost, že mezi 200 testovanými lidmi se najde více než 10 se skupinou AB.

$$X_i \ldots \begin{cases} 0 & i\text{-tý dárce nemá AB} \\ 1 & i\text{-tý dárce má AB} \end{cases} \qquad X_i \to Bi(1, 0{,}06) \Rightarrow \begin{array}{l} EX_i = \mu = 0{,}06 \\ DX_i = 1 \cdot 0{,}06 \cdot 0{,}94 \\ \sigma = \sqrt{0{,}06 \cdot 0{,}94} \end{array}$$

$$\Rightarrow X = \sum_{i=1}^{200} X_i \quad \boxed{490 > \frac{9}{0{,}06 \cdot 0{,}94} \doteq 160} \Rightarrow CLV \Rightarrow X \xrightarrow{\cdot} N(n \cdot \mu; (\sqrt{n}\,\sigma)^2) = N(\underbrace{200 \cdot 0{,}06}; (\sqrt{200 \cdot 0{,}06 \cdot 0{,}94})^2)$$

$$P(X > 10) = 1 - P(X \leq 10) \doteq \underline{\underline{0{,}6724}} \doteq 1 - \text{pnorm}(\underbrace{10.5}_{\text{opravа na spojitost}}, 12, \sqrt{200 \cdot 0{,}06 \cdot 0{,}94})$$

„Přesný" výpočet: $X \to Bi(\overset{200}{\overset{\shortparallel}{n}}, \overset{0{,}06}{\overset{\shortparallel}{p}}) \Rightarrow P(X > 10) = 1 - \text{pbinom}(10, 200, 0.06) = \underline{\underline{0{,}6593}}$

b) Odhadněte pravděpodobnost, že mezi 490 testovanými bude méně než 7 % lidí se skupinou AB

$$X_i < \begin{cases} 0 & i\text{-tý dárce nemá AB} \\ 1 & i\text{-tý dárce má AB} \end{cases} \Bigg\} \Rightarrow X_i \to B(1; \overset{p}{0{,}06}) \Rightarrow \begin{array}{l} \mu = EX_i = 0{,}06 = p \\ \sigma = \sqrt{DX_i} = \sqrt{\underset{p}{0{,}06} \cdot \underset{(1-p)}{0{,}94}} \end{array}$$

$$\bar{X} = \frac{\sum_{i=1}^{490} X_i}{490} \quad \ldots \text{relativní četnost dárců s AB}$$

$$\boxed{n = 490 > \frac{9}{p(1-p)} = \frac{9}{0{,}06 \cdot 0{,}94} \doteq 160 \quad \Rightarrow \text{Použijeme CLV}} \Rightarrow$$

$$\bar{X} \xrightarrow{\cdot} N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right) = N\left(0{,}06; \left(\frac{\sqrt{0{,}06 \cdot 0{,}94}}{\sqrt{490}}\right)^2\right)$$

$$P(\bar{X} < 0{,}07) \doteq \underline{\underline{0{,}8244}} \doteq \text{pnorm}\left(0{,}07, 0.06, \sqrt{\frac{0{,}06 \cdot 0{,}94}{490}}\right)$$

„Přesný" výpočet: 7 % z 490 = 34,3 $\Rightarrow$ Hledáme $P(X \leq 34)$, $X \to Bi(\overset{490}{\overset{\shortparallel}{n}}, \overset{0{,}06}{\overset{\shortparallel}{p}})$

$$\Rightarrow P(X \leq 34) = \text{pbinom}(34, 490, 0.06) = \underline{\underline{0{,}8347}}$$

Poznámka: Jestliže $X_i \to N(\mu, \sigma^2)$ a jsou navzájem nezávislé a $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, pak

$$Z = \frac{\bar{X} - \mu}{S} \sqrt{n} \longrightarrow t_{n-1}$$

S je výběrová směrodatná odchylka – vypočtená z hodnot X_i.

Tj. Z má tzv. Studentovo rozdělení pravděp. s n-1 stupni volnosti.

Př. Předpokládejme, že hmotnost částice je normální n.v. se střední hodnotou $\mu = 20$ (ve vhodných jednotkách). Změřili jsme hmotnosti 12 částic a zjistili, že směrodatná (výběrová) odchylka $S = 3$. Určete pravděpodobnost, že naměřená průměrná hmotnost $\bar{X}$ bude menší než 21.

$n = 12 \Rightarrow$ NEPOUŽIJEME CLV pro odhad rozdělení pravděp. $\bar{X}$

ale:

$$\bar{X} < 21 \qquad /-\mu = 20$$

$$\bar{X} - \mu < 21 - 20 = 1 \qquad /\cdot \frac{\sqrt{n}}{S} = \frac{\sqrt{12}}{3}$$

$$Z = \frac{\bar{X} - \mu}{S}\sqrt{n} < \frac{\sqrt{12}}{3}$$

$\Rightarrow$ hledáme: $P\left(Z < \frac{\sqrt{12}}{3}\right)$ a

víme, že: $Z \longrightarrow t_{\underset{n-1}{11}}$

$$\Rightarrow P(\bar{X} < 21) = P\left(Z < \frac{\sqrt{12}}{3}\right) = \underline{\underline{0{,}8637}} = pt(sqrt(12)/3, 11)$$

Pozn.: Tento příklad je jen ilustrační!? Když známe hodnoty X_i a byli jsme schopni určit S, nejspíše umíme určit i $\bar{X}$!?

Mnohem zajímavější úlohu dostaneme úpravou zadání: Známe naměřené hodnoty $\bar{x}$ a S, co nám to říká o střední hodnotě μ?! (Tu obecně neznáme, ale chtěli bychom alespoň odhadnout do jakého intervalu s nějakou pravděpodobností patří.)

viz. později „testování hypotéz"

Poznámka: Jestliže $X_i \to N(\mu, \sigma^2)$ a S^2 je n.v. jejíž hodnotami jsou výběrové rozptyly s^2 vypočtené z hodnot $x_1, x_2, \dots, x_n$ (tj. výběr o rozsahu n), pak:

$$Z = \frac{(n-1)\,S^2}{\sigma^2} \longrightarrow \chi^2_{n-1}$$ ... Chí kvadrát rozdělení pravděp. s $n-1$ stupni volnosti.

nabývá jen nezáporných hodnot

Př: Předpokládejme, že výška člověka má normální rozdělení pravděpodobnosti se směrodatnou odchylkou $\sigma = 8$ cm. Změřili jsme 15 lidí. S jakou pravděpodobností bude naměřená výběrová směrodatná odchylka $s < 9$ cm?

$$0 < S < 9$$

$\Updownarrow$ (protože $s > 0$, $S^2 > 0$ vždy)

$$0 < S^2 < 81 \qquad \Big| \cdot \frac{(n-1)}{\sigma^2} = \frac{14}{8^2}$$

15 lidí − 1

$$0 < \underbrace{\frac{(n-1)\,S^2}{\sigma^2}}_{=Z} < \frac{81 \cdot 14}{8^2} = 17{,}71875$$

Víme, že $\frac{(n-1)\,S^2}{\sigma^2} = Z \to \chi^2_{14}$

$$\Rightarrow \quad P(S < 9) = P(Z < 17{,}71875) \doteq \underline{\underline{0{,}78}} \doteq \text{pchisq}(17.71875, 14)$$