Úloha III.S ... limitní

10 bodů; průměr 7,81; řešilo 26 studentů

a) *Zkuste vlastními slovy popsat postup konstrukce intervalových odhadů střední hodnoty v případě obecného rozdělení měřených dat (postačí vlastními slovy popsat následující: centrální limitní věta a předpoklady jejího použití, kovariance a korelace (a jejich odhady), vícerozměrná centrální limitní věta a předpoklady jejího použití, zákon šíření nejistot a kdy ho lze použít). Není potřeba uvádět přesná matematická odvození, stačí požadované pojmy a vlastnosti stručně popsat.*

b) *V přiloženém datovém souboru* mereni3-1.csv *najdete výsledky měření určité fyzikální veličiny v. Předpokládejme, že si nemůžeme být jisti, zda mají měřená data normální rozdělení. Vyjádřete nejistotu měření této fyzikální veličiny (nejistotu typu B neuvažujte), zkonstruujte intervalové odhady na základě CLV a stručně interpretujte jeho význam. Jak by se změnily výsledky a interpretace, pokud bychom měli k dispozici jen čtvrtinu měření (řekněme první čtvrtinu dat z datového souboru)?*

c) *Předpokládejme, že naším cílem je naměřit fyzikální veličiny x a y, které budeme chtít využít pro dosazení do vzorce*

$$v = \frac{1}{2}xy^2 .$$

Předpokládejme, že díky znalosti způsobu měření jsme si jisti, že jsou všechna měření na sobě nezávislá a ze zpracování naměřených dat měření máme následující výsledky, které jsou založeny na velkém počtu měření (více než 30 měření každé fyzikální veličiny)

$$x = (5{,}2 \pm 0{,}1) ,$$
$$y = (12{,}84 \pm 0{,}06) .$$

Určete odhad fyzikální veličiny v a nejistotu měření fyzikální veličiny v.
Nápověda: *Mohly by se vám hodit následující vztahy:*

$$\frac{\partial}{\partial x}\left(\frac{1}{2}xy^2\right) = \frac{1}{2}y^2 ,$$
$$\frac{\partial}{\partial y}\left(\frac{1}{2}xy^2\right) = xy .$$

d) *Pomocí simulace ve výpočetním prostředí* R *demonstrujte platnost centrální limitní věty. Tj. generujte n-tice nezávislých realizací náhodné veličiny, která nemá normální rozdělení (pro tento případ použijte exponenciální, rovnoměrné a Poissonovo rozdělení s libovolně zvolenými parametry) a na histogramu ukažte, že pokud na data provedeme následující transformaci*

$$\sqrt{n}\frac{\overline{x_n - \mu}}{S_n} ,$$

takto transformovaná data už budou rozdělena přibližně podle normálního rozdělení $N(0,1)$. (Součástí hodnocení bude i hodnocení vzhledu grafů – zejména vhodně zvolené popisky os a legenda.)

Bonus: *Předpokládejme, že naším cílem je naměřit fyzikální veličiny x a y, které budeme chtít dosadit do vzorce*

$$v = x^2 \sin y .$$

Uvažujme nejobecnější model měření (tj. měřená data nemají normální rozdělení a měření různých fyzikálních veličin na sobě mohou být závislá). V datovém souboru mereni3-2.csv *máme výsledky měření fyzikálních veličin x a y, určete nejistotu určení veličiny v a zkonstruujte pro ni intervalový odhad. Michal se pokusil vymyslet limitně těžké zadání seriálové úlohy.*

a) Detailní odpověď na tuto otázku dostanete pouze přečtením 3. dílu seriálu, v tomto vzorovém řešení uvedeme jen ty nejdůležitější věci.
Centrální limitní věta je důležitá pro zpracování měřených dat (zejména pro intervalový odhad střední hodnoty), o kterých si nemůžeme být jisti, že mají normální rozdělení. Pokud si označíme měřená data jako $x_1, \dots x_n$, potom centrální limitní věta říká, že následující transformace našich dat (musíme si uvědomit, že jde o náhodnou veličinu, neboť závisí na náhodných datech a nemůžeme tedy dopředu znát její hodnotu)

$$\sqrt{n}\frac{\overline{x_n} - \mu}{S_n}$$

konverguje v distribuci k rozdělení $N(0,1)$, nezávisle na tom, jaké rozdělení[1] měla naše původní data. Matematicky zapsáno platí

$$\sqrt{n}\frac{\overline{x_n} - \mu}{S_n} \xrightarrow{D} N(0,1).$$

Pokud chceme v praxi používat aproximace založené na CLV, musíme si uvědomit, že potřebujeme mít dostatečně velký počet měření, aby byla takováto aproximace přesná. Obecné pravidlo zní:

- Pokud máme alespoň 30 měření, potom je aproximace pomocí CLV velice přesná.
- Pokud máme alespoň 10 měření, potom je aproximace pomocí CLV pouze přibližná, ale stále poměrně přesná.
- Pokud máme méně než 10 měření, potom může být aproximace pomocí CLV značně nepřesná.

Pokud měříme více fyzikálních veličin najednou, musíme se zabývat také tím, zda nejsou naše měření závislá. Závislost našich měření (tedy vlastně náhodných veličin) měříme pomocí kovariance a korelace, které jsou pro dvě náhodné veličiny X a Y definovány následovně

$$\operatorname{cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}X)(Y - \mathrm{E}Y)\right],$$
$$\operatorname{corr}(X,Y) = (X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\cdot\operatorname{var}(Y)}}.$$

Korelace je vhodně znormovaná kovariance (může nabývat jen hodnot od -1 do 1) a vyjadřuje, jak moc jsou náhodné veličiny lineárně závislé (hodnoty kolem 0 značí malou závislost, hodnoty blízké 1 nebo -1 značí velkou závislost). Pokud jsou veličiny X a Y nezávislé, potom je kovarianční i korelační koeficient roven 0 (obrácená implikace ale neplatí). V praxi je

[1] Ve skutečnosti je zde ještě podmínka na konečný rozptyl.

důležité vědět, jak z naměřených dat odhadovat kovarianční a korelační koeficient. K tomu slouží výběrový kovarianční koeficient (resp. výběrový korelační koeficient) definovaný jako

$$\widehat{\operatorname{cov}}(X,Y) = \sum_{i=1}^{n} \left(x_i - \overline{x_n}\right)\left(y_i - \overline{y_n}\right),$$

$$\widehat{\operatorname{corr}}(X,Y) = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x_n}\right)\left(y_i - \overline{y_n}\right)}{\sqrt{S_{X,n}^2 \cdot S_{Y,n}^2}}.$$

Na tomto místě musíme poznamenat, že při výpočtu výběrového kovariančního (resp. korelačního) koeficientu musíme vždy používat dvojice odpovídajích si měření. Nelze postupovat tak, že si měření libovolně popárujeme.

Existuje i vícerozměrná verze CLV, která se zaobírá případem, kdy chceme změřit k fyzikálních veličin $v^{(1)}, \ldots, v^{(k)}$, dosadit je do vzorce

$$v = f\left(v^{(1)}, \ldots, v^{(k)}\right)$$

a následně chceme určit nejistotu měření fyzikální veličiny v a konstruovat pro ni intervalové odhady. Vícerozměrná centrální limitní říká, že v tomto případě platí

$$\frac{f\left(\overline{v_n^{(1)}}, \ldots, \overline{v_n^{(k)}}\right) - f\left(v^{(1)}, \ldots, v^{(k)}\right)}{\sqrt{S^2}} \xrightarrow{D} N(0,1),$$

kde $\overline{v_n^{(i)}}$ je výběrový průměr měření i-té fyzikální veličiny, $v^{(i)}$ je skutečná hodnota i-té fyzikální veličiny a S^2 je vhodný normalizační koeficient, který se spočte podle vzorce

$$S^2 = \begin{pmatrix} \frac{\partial f}{\partial v^{(1)}}(\overline{v}) & \cdots & \frac{\partial f}{\partial v^{(k)}}(\overline{v}) \end{pmatrix} \begin{pmatrix} s_{n_1}^2 & \cdots & \frac{\widehat{\operatorname{cov}}\left(v^{(1)}, v^{(k)}\right)}{\sqrt{n_1 n_k}} \\ \frac{\widehat{\operatorname{cov}}\left(v^{(2)}, v^{(1)}\right)}{\sqrt{n_2 n_1}} & \cdots & \frac{\widehat{\operatorname{cov}}\left(v^{(2)}, v^{(k)}\right)}{\sqrt{n_2 n_k}} \\ \vdots & \ddots & \vdots \\ \frac{\widehat{\operatorname{cov}}\left(v^{(k)}, v^{(1)}\right)}{\sqrt{n_k n_1}} & \cdots & s_{n_k}^2 \end{pmatrix} \begin{pmatrix} \frac{\partial f}{\partial v^{(1)}}(\overline{v}) \\ \vdots \\ \frac{\partial f}{\partial v^{(k)}}(\overline{v}) \end{pmatrix}, \quad (1)$$

kde

$$\overline{v} = \left(\overline{v_{n_1}^{(1)}}, \ldots, \overline{v_{n_k}^{(k)}}\right).$$

Pro použití vícerozměrné CLV opět platí podobná pravidla jako pro použití jednorozměrné CLV, tedy:

- Pokud máme alespoň 30 měření každé fyzikální veličiny $v^{(1)}, \ldots, v^{(k)}$, potom je aproximace pomocí vícerozměrné CLV velmi přesná.
- Pokud máme alespoň 10 měření každé fyzikální veličiny, potom je aproximace pomocí vícerozměrné CLV pouze přibližná, ale stále poměrně přesná.
- Pokud máme méně než 10 měření nějaké fyzikální veličiny, potom může být aproximace pomocí vícerozměrné CLV značně nepřesná.

V případě, že si ze znalosti průběhu experimentu můžeme být jistí, že jsou všechna měření různých fyzikálních veličin na sobě nezávislá (což je v praxi velmi časté), se vzorec na výpočet členu S^2 velmi zjednoduší na tvar

$$S = \sqrt{\left(\frac{\partial f\left(\overline{v_{n_1}^{(1)}}\right)}{\partial v^{(1)}}\right)^2 {s_{n_1}^{(1)}}^2 + \cdots + \left(\frac{\partial f\left(\overline{v_{n_k}^{(k)}}\right)}{\partial v^{(k)}}\right)^2 {s_{n_k}^{(k)}}^2}.$$

Tomuto vzorci se často říká zákon šíření nejistot (případně zákon propagace nejistot).
V obou případech můžeme konstruovat intervaly spolehlivosti pro výslednou fyzikální veličinu v. Ze znění vícerozměrné CLV lze odvodit, že platí

$$\mathrm{P}\left(f\left(\overline{v_n^{(1)}}, \ldots, \overline{v_n^{(k)}}\right) - u_{1-\frac{\alpha}{2}} S < f\left(v^{(1)}, \ldots, v^{(k)}\right) < f\left(\overline{v_n^{(1)}}, \ldots, \overline{v_n^{(k)}}\right) + u_{1-\frac{\alpha}{2}} S\right) \simeq 1 - \alpha.$$

Tedy intervalový odhad pro skutečnou hodnotu výsledné fyzikální veličiny v bude mít tvar

$$\left(f\left(\overline{v_n^{(1)}}, \ldots, \overline{v_n^{(k)}}\right) \pm u_{1-\frac{\alpha}{2}} S\right)$$

Fyzikové tento intervalový odhad zapisují zkráceně jen jako

$$\left(f\left(\overline{v_n}\right) \pm S\right)$$

a členu S říkají standardní odchylka.

b) V přiloženém datavém souboru se nachází 40 naměřených dat, se kterými budeme pracovat. Nejprve musíme spočítat výběrový průměr našich dat podle vzorce

$$\overline{x_{40}} = \frac{1}{40} \sum_{i=1}^{40} x_i = 9{,}826.$$

Dále spočítáme výběrovou směrodatnou odchylku průměru podle vzorce

$$s_{40} = \sqrt{\frac{1}{40(40-1)} \sum_{i=1}^{40} \left(x_i - \overline{x_n}\right)^2} = 0{,}036.$$

V našem případě, kdy neuvažujeme žádnou nejistotu typu B, vyjadřuje výběrová směrodatná odchylka nejistotu měření.
Podle teorie odvozené v seriálu bude mít asymptotický intervalový odhad založený na CLV o spolehlivosti $(1-\alpha)$ tvar

$$\left(\overline{x_n} \pm s_n u_{1-\frac{\alpha}{2}}\right).$$

Pokud použijeme standardní zkrácený zápis, vyjde nám následující výsledek

$$(9{,}83 \pm 0{,}04).$$

Jelikož jsme použili velké množství dat (tj. více než 30), bude aproximace pomocí CLV už velice přesná. Tedy takovýto interval bude mít pravděpodobnost pokrytí skutečné hodnoty měřené fyzikální veličiny velice blízkou 68%.

Pokud bychom uvažovali, že máme k dispozici jen první čtvrtinu měření (tj. prvních 10 měření), dostali bychom jiné hodnoty výběrového průměru a výběrové směrodatné odchylky průměru. Konkrétně bychom dostali

$$\overline{x_{10}} = 9{,}92,$$
$$s_{10} = 0{,}06.$$

Intervalový odhad pro měřenou fyzikální veličinu by v tomto případě byl následující (používáme zkrácený zápis)

$$(9{,}92 \pm 0{,}06).$$

Je zřejmé, že za použití jiných vstupních dat dostaneme jiné číselné výsledky (i na tomto je vidět, že výběrový průměr a výběrová směrodatná odchylka průměru jsou náhodné veličiny). V tomto případě jsme provedli méně měření, takže jsme vcelku logicky dostali větší nejistotu měření. V tomto případě je ale nutné poznamenat, že náš intervalový odhad byl založen na poměrně málo měřeních (jen 10 měření), takže v tomto případě aproximace pomocí CLV nemusí být tolik přesná. Tento interval má tedy pravděpodobnost pokrytí skutečné hodnoty měřené fyzikální veličiny pouze přibližně 68%.

c) Odhad fyzikální veličiny v se zkonstruuje jednoduše tak, že do funkce f dosadíme odhady fyzikálních veličin x a y. V našem případě dostaneme

$$\overline{v_n} = f\left(\overline{x_n}, \overline{y_n}\right) = \frac{1}{2} \cdot 5{,}2 \cdot 12{,}84^2 = 428{,}65.$$

V tomto speciálním případě, kdy můžeme měření různých fyzikálních veličin považovat za nezávislé, můžeme na výpočet nejistoty měření fyzikální veličiny v použít zákon šíření nejistot. Všechny potřebné údaje máme zadané, takže stačí dosadit. V tomto vzorovém řešením budeme dosazování dělat postupně, aby bylo všem jasné, jak jsme na náš výsledek přišli.

$$S = \sqrt{\left(\frac{\partial f\left(\overline{x_n}\right)}{\partial x}\right)^2 {s_n^{(x)}}^2 + \left(\frac{\partial f\left(\overline{y_n}\right)}{\partial y}\right)^2 {s_n^{(y)}}^2} = \sqrt{\left(\frac{1}{2}\overline{y_n}^2\right)^2 {s_n^{(x)}}^2 + \left(\overline{x_n y_n}\right)^2 {s_n^{(y)}}^2} =$$
$$= \sqrt{\left(\frac{1}{2} \cdot 12{,}84^2\right)^2 \cdot 0{,}1^2 + \left(5{,}2 \cdot 12{,}84\right)^2 \cdot 0{,}06^2} = 9{,}165.$$

Když bychom měli zapsat intervalový odhad pro fyzikální veličinu v ve zkráceném tvaru, vypadal by následovně

$$(429 \pm 9).$$

d) S využitím výpočetního prostředí R provedeme přesně to, co se píše v zadání. Nejprve si zvolíme přirozené číslo n. Toto číslo bude v našem modelu představovat počet měření, které při konkrétním experimentu provádíme (budeme volit hodnoty v řádu nízkých desítek nebo jednotek, což je typický počet měření při fyzikálních experimentech). Dále zvolíme počet opakování naší simulace (ideálně zvolit co nejvyšší, v našem případě zvolíme 10 000).
Nyní už můžeme začít se samotnou simulací. V každé jednotlivém cyklu vygenerujeme n nezávislých realizací náhodné veličiny s určitým rozdělením, které není normální (v našem případě používáme rozdělení $Exp(2)$, $R(0, 10)$ a $Poiss(2)$). Z těchto n realizací spočteme následující transformaci

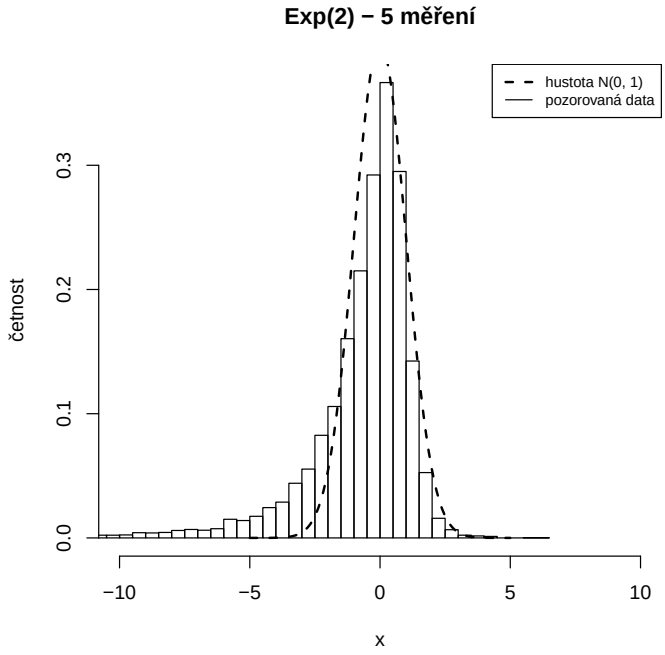
$$t = \sqrt{n}\frac{\overline{x_n} - \mu}{S_n}, \tag{2}$$

kde μ představuje skutečnou střední hodnotu rozdělení, ze kterého jsme generovali použité realizace náhodné veličiny (tedy $\frac{1}{2}$ u exponenciálního rozdělení, 5 u rovnoměrného rozdělení a 2 u Poissonova rozdělení). Toto opakujeme celkem 10 000krát, čímž získáme transformace

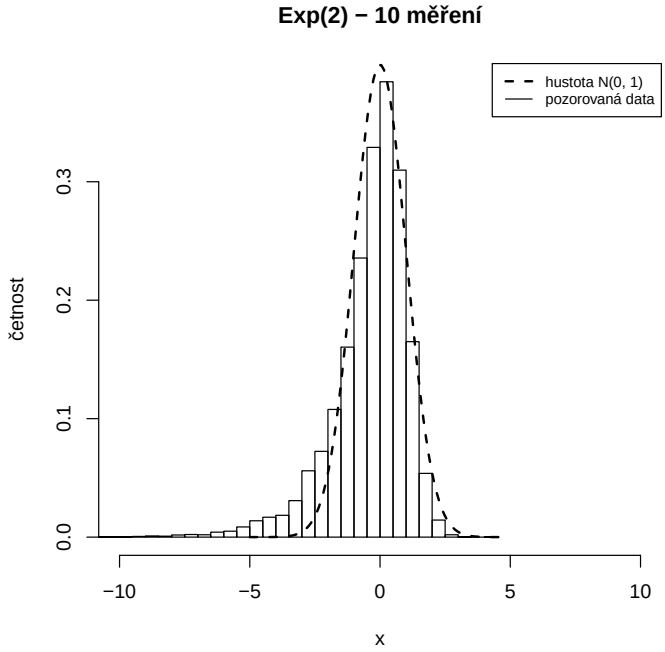
$$t_1, \ldots, t_{10\,000}.$$

Pokud centrální limitní věta platí, potom by mělo být rozdělení takto transformovaných dat velice podobné rozdělení $N(0,1)$. Toto ověříme na histogramu a budeme zkoumat závislost podoby s rozdělením $N(0,1)$ na volbě n.

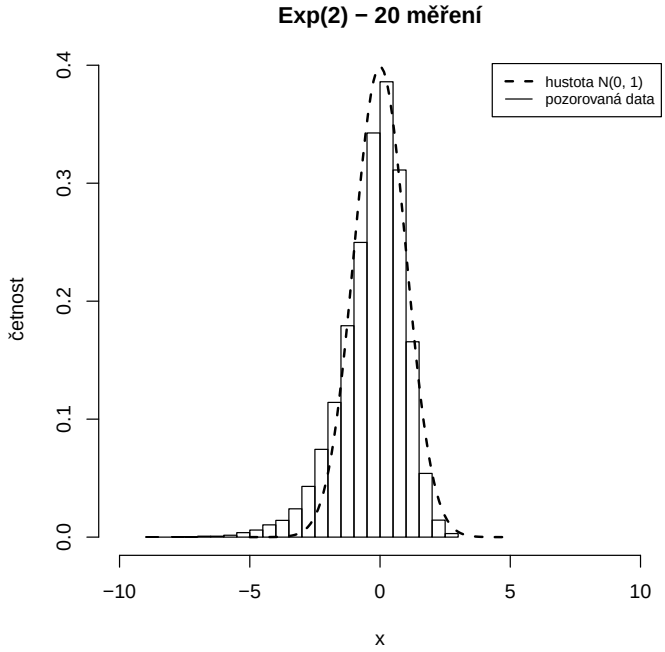
Nyní už k samotným výsledkům. Příslušné histogramy můžeme vidět na obrázcích 1, 2, 3, 4, 5, 6, 7, 8 a 9. Ze všech těchto obrázků je vidět, že limitní rozdělení transformace měřených dat (2) je právě rozdělení $N(0,1)$, přesně jak říká CLV. Je dobré si povšimnout, že rychlost konvergence je pro různá rozdělení měřených dat různá. Pokud ale provedeme alespoň 30 měření, je už rozdělení transformace měřených dat velice podobné rozdělení $N(0,1)$, tedy aproximace pomocí CLV už bude v takovýchto případech velice přesná.



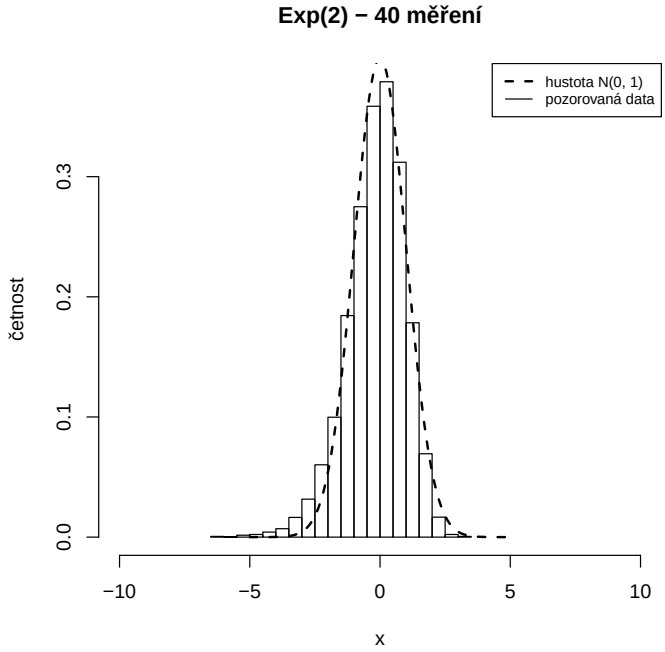
Obr. 1: Histogram transformace měřených dat pocházejících z exponenciálního rozdělení pro 5 měření.



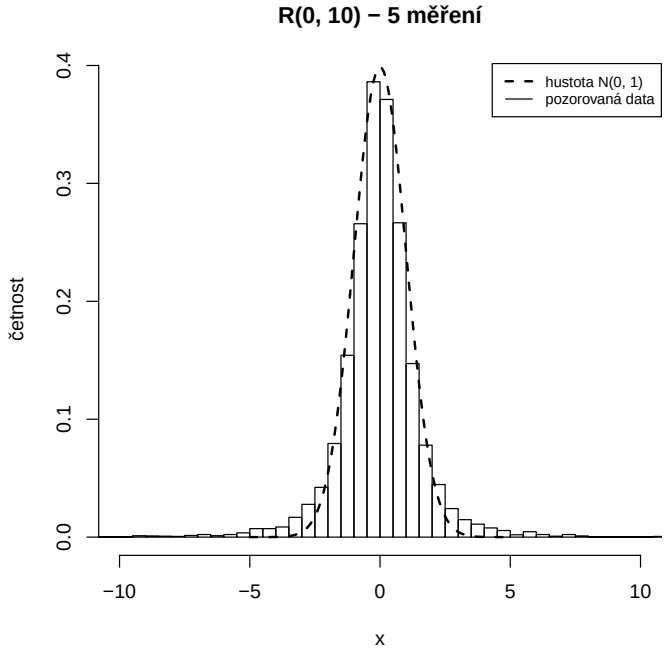
Obr. 2: Histogram transformace měřených dat pocházejících z exponenciálního rozdělení pro 10 měření.



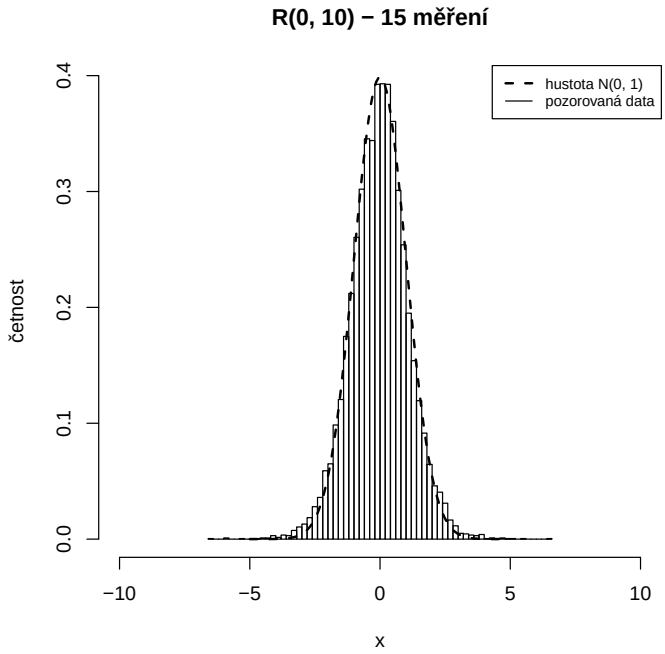
Obr. 3: Histogram transformace měřených dat pocházejících z exponenciálního rozdělení pro 20 měření.



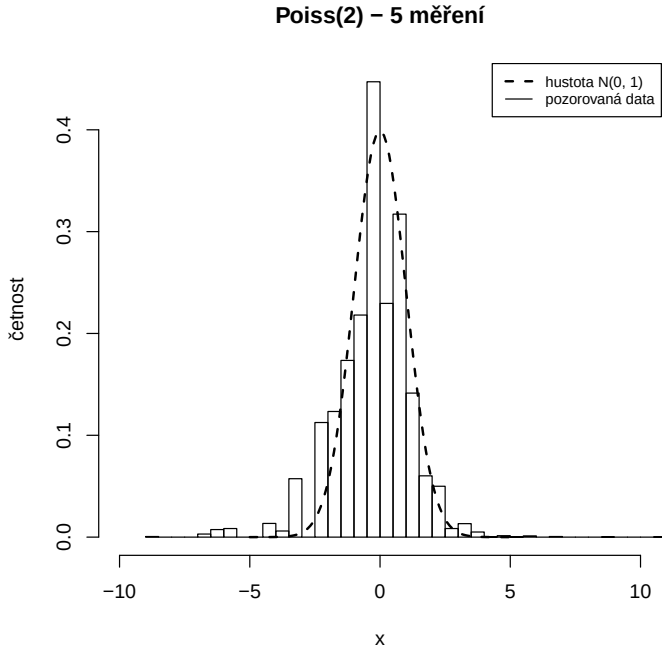
Obr. 4: Histogram transformace měřených dat pocházejících z exponenciálního rozdělení pro 40 měření.



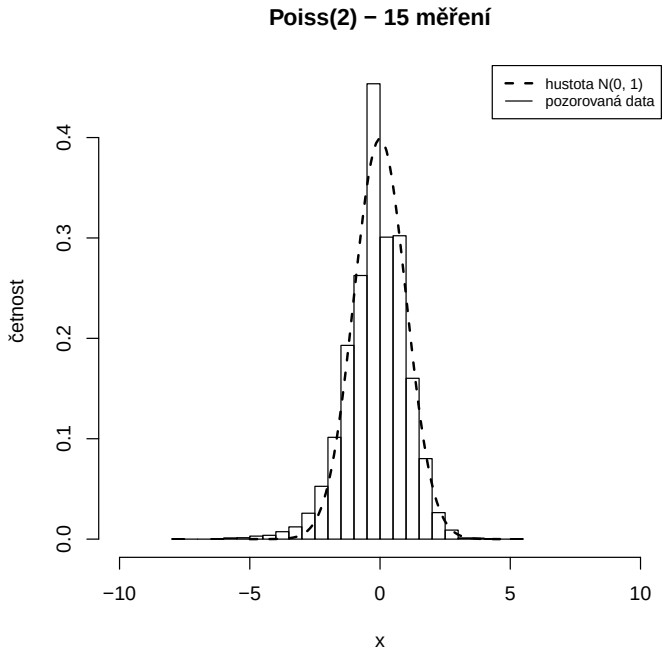
Obr. 5: Histogram transformace měřených dat pocházejících z rovnoměrného rozdělení pro 5 měření.



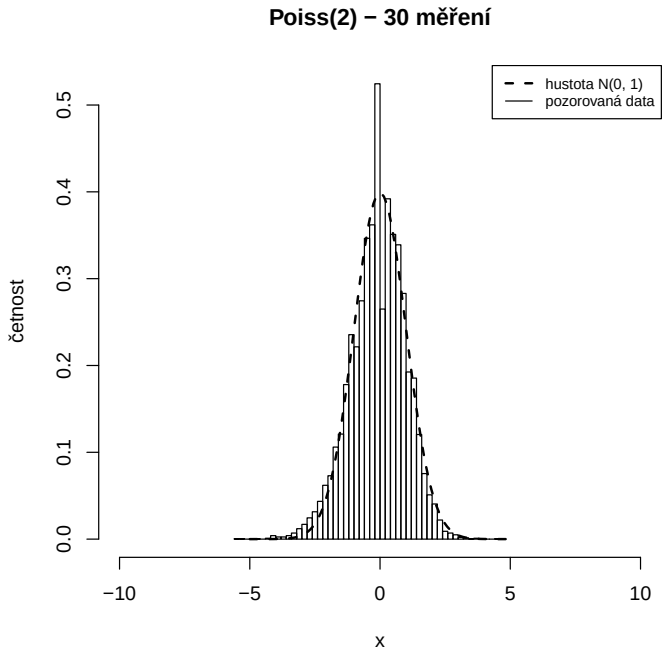
Obr. 6: Histogram transformace měřených dat pocházejících z rovnoměrného rozdělení pro 15 měření.



Obr. 7: Histogram transformace měřených dat pocházejících z Poissonova rozdělení pro 5 měření.



Obr. 8: Histogram transformace měřených dat pocházejících z Poissonova rozdělení pro 15 měření.



Obr. 9: Histogram transformace měřených dat pocházejících z Poissonova rozdělení pro 30 měření.

Bonus: Jediné, co pro vyřešení tohoto úkolu potřebujeme, je dosadit do vícerozměrné centrální limitní věty. Nejprve vypočítáme nejistotu určení fyzikální veličiny podle vzorce (1). K tomu budeme potřebovat znát parciální derivace funkce f, které mají tvar

$$\frac{\partial f(x,y)}{\partial x} = 2x \sin y,$$
$$\frac{\partial f(x,y)}{\partial y} = x^2 \cos y.$$

Nyní už můžeme psát, čemu se rovná nejistota určení fyzikální veličiny v

$$S^2 = \left(2\overline{x_n} \sin \overline{y_n} \quad \overline{x_n}^2 \cos \overline{y_n}\right) \begin{pmatrix} {s_n^{(x)}}^2 & \frac{\widehat{\mathrm{cov}(x,y)}}{n} \\ \frac{\widehat{\mathrm{cov}(x,y)}}{n} & {s_n^{(y)}}^2 \end{pmatrix} \begin{pmatrix} 2\overline{x_n} \sin \overline{y_n} \\ \overline{x_n}^2 \cos \overline{y_n} \end{pmatrix},$$

kde používáme standardní značení. Za použití matematického softwaru (výpočetní prostředí R) dostáváme, že nejistota určení veličiny v je

$$S = 6{,}03.$$

Odhad veličiny v získáme pouze dosazením výběrových průměrů do naší funkce, tedy

$$\overline{v_n} = \overline{x_n}^2 \sin \overline{y_n}.$$

Za použití matematického softwaru dostáváme výsledek

$$\overline{v_n} = 117{,}29.$$

Vícerozměrná centrální limitní věta potom říká, že intervalový odhad pro fyzikální veličinu v bude tvaru

$$\left(\overline{v_n} \pm S u_{1-\frac{\alpha}{2}}\right).$$

Pokud použijeme zkrácený zápis intervalového odhadu a dosadíme naše konkrétní číselné výsledky, dostaneme

$$v = (117 \pm 6).$$

Jelikož máme dostatečně velký počet měření obou fyzikálních veličin (tj. více než 30) bude tento intervalový odhad už velice přesný.

Jen pro zajímavost na tomto místě uvedeme, že pokud bychom předpokládali nezávislá data a použili bychom pouze zákon šíření nejistot, dostali bychom výsledek

$$v = (117 \pm 7).$$

Vidíme, že rozdíl v získaných výsledcích není velký, nicméně je nutné poznamenat, že takovýto postup není správný, neboť zanedbává korelaci v našich měřených datech. V jiných příkladech z praxe už může být chyba, které bychom se takovýmto zanedbáním dopustili, velmi velká.

Michal Nožička
nozicka@fykos.cz

Fyzikální korespondenční seminář je organizován studenty MFF UK. Je zastřešen Oddělením pro vnější vztahy a propagaci MFF UK a podporován Ústavem teoretické fyziky MFF UK, jeho zaměstnanci a Jednotou českých matematiků a fyziků.