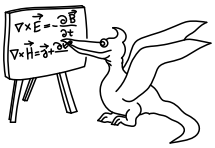


Seriál: Zpracování dat fyzikálních měření

Úvod

Letošní seriál bude věnován statistice a zpracování dat fyzikálních měření. Na první pohled by se mohlo zdát, že toto téma není příliš fyzikální. „K čemu potřebuji statistiku?“, mohli by si někteří fyzikové a fyzičky pomyslet. Statistika je velmi důležitá pro experimentální fyziku a nakonec i pro tu teoretickou. Vždyť každý fyzikální zákon musí být potvrzen (přesnější význam slova „potvrzen“ si rozebereme v tomto seriálu) experimentálními měřeními, bez toho bychom nemohli určit, zda skutečně platí. Na konci tohoto seriálu budeme schopni pomocí matematické statistiky a naměřených dat potvrzovat nebo vyvracet různá tvrzení aspirující na fyzikální zákony a osvojíme si i mnoho dalších ve fyzikální praxi užitečných znalostí a dovedností.

V rámci zadávaných příkladů se zaměříme hlavně na praktickou práci s daty, neboť ta je mnohem důležitější než znalost teoretických pouček. V dnešní době si už nelze představit provádění složitých matematických výpočtů tužkou na papíře, proto i my budeme používat matematický software, kterým bude výpočetní prostředí a programovací jazyk *R*. Doufáme, že se vám podaří si ho zdárně nainstalovat ze stránky `https://cran.rstudio.com/` podle návodu distributora. Dále doporučujeme používat editor RStudio (volně ke stažení ze stránky `https://www.rstudio.com/products/rstudio/download2/`). V případě problémů s instalací nás kontaktujte na e-mailové adrese nozicka@fykos.cz. Součástí úloh v každé sérii bude také práce s reálnými daty. Syntaxi jazyka *R* se budeme učit z připravených skriptů, které budou doplněny vysvětlujícími komentáři.

Nyní už se zaměřme na samotný obsah seriálu. Začneme vysvětlením základních pojmů, které budeme používat.

Náhoda

Náhoda je takové často používané slovo, ale když se pořádně zamyslíme, uvědomíme si, že je vlastně velmi těžké definovat, co přesně znamená. Na začátek si proto musíme vyjasnit, co budeme pod pojmem náhoda ve zbytku tohoto seriálu mít na mysli. Je dobré začít z druhé strany, tedy uvědomit si, co je opakem slova náhoda nebo náhodný. Opakem slova náhodný je slovo *deterministický* (česky něco jako „určený“), které říká, že je vývoj popisovaného systému jednoznačně dán jeho aktuálním stavem. Jako *náhodné jevy* tedy označíme ty jevy, u kterých nelze předem určit, jakým způsobem dopadnou. Jako takový typický náhodný jev si můžeme označit výsledek hodu kostkou. Nikdo dopředu neví, jaké číslo padne na kostce. Kdyby někdo takový existoval, už by jistě byl díky všem možným hazardním hrám velice bohatý. Slovo náhoda tedy pro nás bude mít význam procesu, který zařídí, že při stejných počátečních podmínkách nějakého pokusu, můžeme na konci dostat různé výsledky.

Jak přesně ta náhoda funguje pro nás teď nebude podstatné. Například u zmíněného hodu kostkou můžeme říci, že počáteční podmínky nejsou vždy úplně stejné. Výsledek hodu kostkou závisí na mnoha parametrech, například rychlosti hodu, rotaci kostky, tvaru podložky atd. Kdybychom všechny tyto parametry znali, byli bychom schopni podle fyzikálních zákonů spočítat

výsledek hodu. Problém je, že takovýchto parametrů je příliš velké množství na to, abychom je mohli všechny přesně určit. I kdybychom ale znali hodnoty všech těchto parametrů, příslušné výpočty by byly natolik složité, že bychom se rozhodně ani za použití moderní techniky nedopočítali výsledku dříve, než by kostka dopadla. Radši proto řekneme, že výsledek hodu kostkou je náhodný. I tak jsme totiž za pomoci matematické statistiky schopni přesně popsat a formulovat mnoho tvrzení, která budou v těchto případech platit. Například že každé číslo bude padat stejně často nebo že při velkém počtu hodů kostkou bude průměr hozených čísel přibližně 3,5.

Náhodná veličina

Ústřední pojem prvního dílu seriálu bude *náhodná veličina*. Ze střední školy znáte pojem proměnná, která označuje nějaké číslo. Proměnná může mít známou i neznámou hodnotu, ale vždy je to jen jedna hodnota. Pod pojmem náhodná veličina budeme rozumět něco jako proměnnou, která ovšem nemá jen jednu hodnotu, ale může nabývat různých hodnot s různou pravděpodobností.

Náhodné veličiny se často značí velkými tiskacími písmeny. Jako příklad takové náhodné veličiny můžeme uvést výsledek hodu kostkou. Pokud si jako X označíme výsledek hodu férovou šestistěnnou kostkou, nemůžeme dopředu (tj. před samotným hodem) mluvit o tom, jakou hodnotu má náhodná veličina X. Jediné, o čem můžeme s jistotou něco říci, jsou pravděpodobnosti, že náhodná veličina X nabude určité hodnoty.

Toto nás přivádí k otázce, čím je taková náhodná veličina určená. Podobně jako u klasické proměnné se zajímáme o její hodnotu, u náhodné veličiny se budeme zajímat o tzv. *rozdělení náhodné veličiny*. Rozdělení můžeme představit jako předpis, který nám udává, jakých hodnot a s jakou pravděpodobností náhodná veličina nabývá.

Podle typu rozdělení dělíme náhodné veličiny na dvě hlavní skupiny, kterými jsou *diskrétní* a absolutně *spojité* náhodné veličiny.[1]

Diskrétní náhodné veličiny

Diskrétní náhodná veličina je taková veličina, která může nabývat jen konečně (nebo spočetně) mnoha hodnot. Typickým příkladem diskrétní náhodné veličiny je už několikrát zmíněný výsledek hodu kostkou, který může nabývat pouze hodnot z množiny $\{1, 2, 3, 4, 5, 6\}$. Diskrétní náhodné veličiny se popisují velmi snadno, stačí nám určit pravděpodobnosti, že náhodná veličina nabude konkrétních hodnot z množiny možných hodnot. Pro hod kostkou bychom mohli psát

$$\mathrm{P}(X = 1) = \mathrm{P}(X = 2) = \cdots = \mathrm{P}(X = 6) = \frac{1}{6}\,.$$

Toto je rozdělení náhodné veličiny X, která představuje výsledek hodu kostkou.

Spojité náhodné veličiny

Jistě každý hned poznal, že musí existovat i jiný druh náhodných veličin než pouze diskrétní náhodné veličiny. Pokud bychom si za náhodnou veličinu Y označili například teplotu, která

[1] Ve skutečnosti existují ještě i jiné typy náhodných veličin (něco mezi diskrétními a absolutně spojitými), ale těmi se v tomto seriálu zabývat nebudeme. Také místo absolutně *spojitá* náhodná veličina budeme používat zkráceně jen *spojitá* náhodná veličina.

bude zítra v 8:00 na meteorologické stanici v Rudolfinu v Praze (ne tu, kterou naměříme digitálním teploměrem, který zaokrouhluje na několik desetinných míst, ale tu, která skutečně bude), jistě bychom si nevystačili s konečně mnoha možnými hodnotami. Teplota může nabývat libovolné hodnoty od absolutní nuly výše, nejenom celých čísel. Takovýchto hodnot je samozřejmě nekonečně (dokonce nespočetně) mnoho, proto nemůžeme rozdělení takovéto náhodné veličiny určit pouze výčtem pravděpodobností, protože těch pravděpodobností by muselo být nekonečně (nespočetně) mnoho.

Spojité náhodné veličiny se používají na modelování těchto situací, pro které jsou diskrétní náhodné veličiny nedostačující. Rozdělení spojitých náhodných veličin bude určeno tzv. hustotou pravděpodobnosti. Hustota pravděpodobnosti je funkce, která udává, jak pravděpodobné jsou jednotlivé hodnoty. Přesněji je to funkce s vlastností, že pravděpodobnost nabývání nějaké hodnoty z intervalu $[a, b]$ náhodnou veličinou je rovna obsahu plochy[2] pod křivkou hustoty mezi body a a b (viz obrázek 1). Jistě si už každý sám rozmyslí, že nejvíce pravděpodobné jsou právě ty body, kde hustota pravděpodobnosti nabývá největších hodnot.

Náhodné veličiny ve fyzice

Po malém teoretickém úvodu se nyní zaměříme na praktické využití náhodných veličin ve fyzice. Náhodné veličiny potkáme prakticky při každém měření fyzikálními přístroji. Jelikož jsou všechny měřicí přístroje nedokonalé, nenaměříme vždy skutečnou hodnotu měřené fyzikální veličiny, tj. naše měření budou nepřesná.

Jako příklad si můžeme uvést měření doby kyvu kyvadla stopkami. Doba kyvu kyvadla nebude vždy naprosto stejná kvůli vlivu nepatrných pohybů vzduchu, které budou na kyvadlo působit, nedokonalostem závěsu, další nepřesnosti vzniknou na straně člověka mačkajícího stopky vlivem jeho reakčního času atd. Číslo, které uvidíme na stopkách, rozhodně nebude přesně doba kyvu kyvadla za ideálních podmínek. V našem modelu budeme hodnoty, které získáváme měřeními, považovat za náhodné veličiny – podobně jako při hodu kostkou.

Realizace náhodné veličiny – Měření

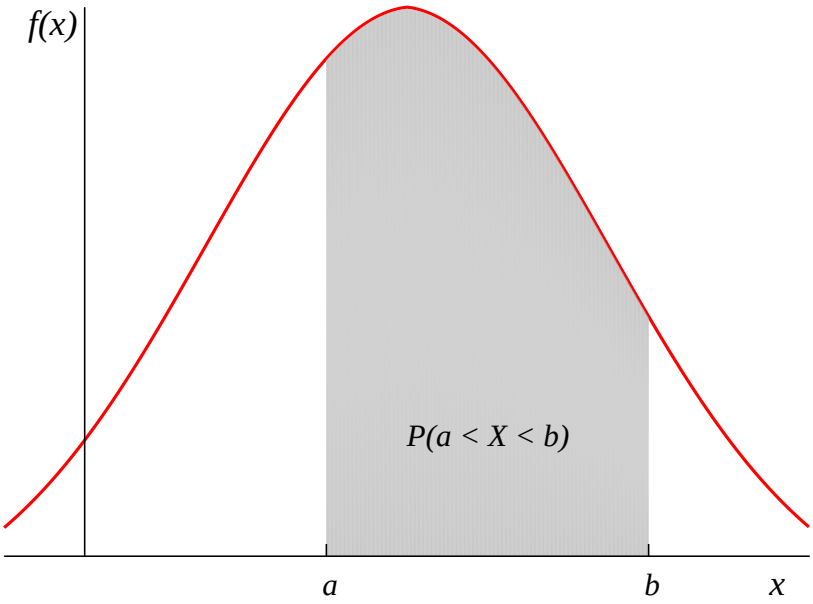
Je důležité si uvědomit rozdíl mezi náhodnou veličinou a realizací náhodné veličiny. Náhodnou veličinou budeme rozumět funkci, která nám teoreticky popisuje výsledek nějakého procesu (hod kostkou, fyzikální měření atd.). Realizací náhodné veličiny budeme rozumět výsledek tohoto procesu v konkrétním případě (např. výsledek jednoho konkrétního hodu kostkou).

Zejména je důležité pochopit, že realizace náhodné veličiny už je jen konstanta (je to jedna konkrétní hodnota), není na ní nic náhodného. Rozdělení náhodné veličiny potom popisuje (určuje pravděpodobnosti), jak budou vypadat jednotlivé realizace této náhodné veličiny.

Fyzikální měření potom můžeme chápat tak, že získáváme realizace nějaké náhodné veličiny (o této náhodné veličině typicky nebudeme mít moc informací). Statistické zpracování naměřených dat[3] se potom snaží získat co nejvíce informací o skutečné hodnotě měřené fyzikální veličiny.

[2] Pro čtenáře seznámené s integrálním počtem můžeme uvést, že se jedná o určitý integrál z hustoty pravděpodobnosti od a do b. Znalost integrálního počtu nicméně není vyžadována, v tomto seriálu si plně vystačíme s grafickou představou.

[3] Od této chvíle budeme naměřená data chápat jako realizace nějaké náhodné veličiny a nebudeme už to v jednotlivých případech připomínat.



Obr. 1: Grafický význam hustoty pravděpodobnosti spojité náhodné veličiny.

Zjednodušený popis náhodných veličin

Jak už jsme řekli na začátku, náhodnou veličinu budeme popisovat jejím rozdělením. To je ovšem v některých případech velmi těžké a nešikovné. Představme si například náhodnou veličinu udávající součet čísel, které hodíme na 10 kostkách. Kolika možných hodnot může takováto náhodná veličina nabývat? Jaké je přesné rozdělení takovéto náhodné veličiny (tj. pravděpodobnosti jednotlivých možných výsledků)? Jistě to není nic, čím bychom se chtěli dlouze zabývat, neboť je to poměrně složité a pro účely experimentální fyziky neužitečné. V praxi bychom určitě narazili i na mnohem komplikovanější problémy. Musíme si tedy zavést nějaký zjednodušený popis náhodných veličin.

Pro snazší popis se zavádí pojem *střední hodnoty* a *rozptylu* náhodné veličiny. Je ovšem nutné poznamenat, že střední hodnota a rozptyl nemusí plně určovat rozdělení náhodné veličiny[4], jen ho zjednodušeně popisují.

[4] Tj. existují dvě náhodné veličiny s různými rozděleními, ale obě mají stejnou střední hodnotu i rozptyl.

Střední hodnota

Střední hodnota udává, jaké hodnoty náhodná veličina v průměru nabývá (jakou hodnotu bychom měli v průměru očekávat). Matematicky je pro diskrétní náhodné veličiny definována jako

$$\mathrm{E}X = \sum_{i=1}^{n} k_i \mathrm{P}(X = k_i)\,,$$

kde $\{k_1, \ldots, k_n\}$ jsou všechny možné hodnoty náhodné veličiny X (značení střední hodnoty E pochází z anglického *Expectation*). Střední hodnotu diskrétní náhodné veličiny lze interpretovat jako vážený průměr všech možných hodnot, kterých náhodná veličina může nabývat, kde váhy jsou právě pravděpodobnosti nabývání.

Pro úplnost uvedeme i definici střední hodnoty pro spojité náhodné veličiny. Poznamenejme ovšem, že pokud neznáte integrální počet, nemusí vás trápit, že této definici nebudete rozumět, dále v seriálu to nebude potřeba. Pro spojité náhodné veličiny je střední hodnota zadefinována jako

$$\mathrm{E}X = \int_{-\infty}^{\infty} x f(x)\mathrm{d}x\,,$$

kde f je hustota náhodné veličiny. Opět si tento výraz lze vyložit jako vážený integrální průměr všech hodnot, kterých může náhodná veličina nabývat.

Rozptyl

Rozptyl udává, jak moc je náhodná veličina rozptýlena okolo své střední hodnoty. Matematicky je pro diskrétní náhodnou veličinu rozptyl definován jako

$$\mathrm{var}X = \sum_{i=1}^{n} (k_i - \mathrm{E}X)^2 \mathrm{P}(X = k_i)\,,$$

kde $\{k_1, \ldots, k_n\}$ jsou možné hodnoty náhodné veličiny (značení pochází z anglického *Variance*). Tento výraz nám říká, že při výpočtu rozptylu děláme vážený průměr druhých mocnin vzdálenosti možných hodnot od střední hodnoty. Druhá mocnina je zavedena proto, aby zvyšovala váhu odlehlých pozorování.

Pro úplnost opět uvedeme i definici rozptylu pro spojité náhodné veličiny, která ale nebude dále v textu seriálu vyžadována.

$$\mathrm{var}X = \int_{-\infty}^{\infty} (x - \mathrm{E}X)^2 f(x)\mathrm{d}x\,,$$

kde f je hustota pravděpodobnosti náhodné veličiny X.

Nejčastěji se vyskytující rozdělení náhodných veličin

Na tomto místě uvedeme čtyři v praxi nejčastěji se vyskytující rozdělení náhodných veličin, které bude každý fyzik pravidelně potkávat. U každého rozdělení uvedeme jeho hustotu v případě,

že se bude jednat o spojité rozdělení, nebo uvedeme výčet pravděpodobností v případě, že se bude jednat o diskrétní rozdělení. U každého rozdělení také uvedeme střední hodnotu a rozptyl, správnost těchto údajů si lze ověřit dosazením do příslušných vzorců (výpočty bývají zpravidla velmi náročné, proto je zde nebudeme uvádět). Nakonec také zmíníme, kde v praxi můžeme které rozdělení potkat.

Normální (Gaussovo) rozdělení

Jedná se o spojité rozdělení určené dvěma parametry $\mu \in \mathbb{R}, \sigma^2 > 0$ (značíme ho $\mathrm{N}(\mu, \sigma^2)$). Hustota pravděpodobnosti tohoto rozdělení má tvar

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} .$$

Střední hodnota a rozptyl tohoto rozdělení jsou[5]

$$\begin{aligned} \mathrm{E}X &= \mu \,, \\ \mathrm{var}X &= \sigma^2 \,. \end{aligned}$$

Normální rozdělení mívají hlavně náhodné veličiny vzniklé měřením nějaké fyzikální veličiny, která může teoreticky nabývat víceméně libovolné hodnoty[6] (například hodnota elektrického proudu v obvodu měřená ampérmetrem, měření teploty teploměrem atd.)

Exponenciální rozdělení

Jedná se o spojité rozdělení určené jedním parametrem $\lambda > 0$, značíme ho $\mathrm{Exp}(\lambda)$. Hustota pravděpodobnosti tohoto rozdělení má tvar

$$f(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x} & \text{pro } x > 0 \\ 0 & \text{jinak} \,. \end{cases}$$

Střední hodnota a rozptyl tohoto rozdělení jsou

$$\begin{aligned} \mathrm{E}X &= \frac{1}{\lambda} \,, \\ \mathrm{var}X &= \frac{1}{\lambda^2} \,. \end{aligned}$$

Exponenciální rozdělení mívají hlavně náhodné veličiny vyjadřující čas uplynulý mezi dvěma po sobě jdoucími opakujícími se událostmi (například čas mezi jednotlivým radioaktivními rozpady atomů v radioaktivní látce).

[5] Všimněme si, že v případě normálního rozdělení platí, že je jednoznačně určeno střední hodnotou a rozptylem. U ostatních rozdělení to ale platit nemusí!

[6] Ale třeba čas mezi dvěma událostmi nemůže být záporný apod.

Rovnoměrné rozdělení

Jedná se o další spojité rozdělení určené dvěma parametry $a < b$, značíme ho $\mathrm{R}(a, b)$. Hustota pravděpodobnosti tohoto rozdělení má tvar

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{pro } x \in (a, b) \\ 0 & \text{jinak}\,. \end{cases}$$

Střední hodnota a rozptyl tohoto rozdělení jsou

$$\mathrm{E}X = \frac{a\ + b}{2}\,,$$
$$\mathrm{var}X = \frac{(b-a)^2}{12}\,.$$

Rovnoměrné rozdělení se používá na modelování událostí, které mají všechny výsledky stejně pravděpodobné.

Poissonovo rozdělení

Jedná se o diskrétní rozdělení určené jedním parametrem $\lambda > 0$ (značíme ho $\mathrm{Poiss}(\lambda)$). Pravděpodobnosti mají následující tvar

$$\mathrm{P}(X = k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}.$$

Střední hodnota a rozptyl tohoto rozdělení jsou

$$\mathrm{E}X = \lambda\,,$$
$$\mathrm{var}X = \lambda\,.$$

Poissonovo rozdělení mají zejména náhodné veličiny vyjadřující počet opakujících se událostí v nějakém časovém intervalu (například počet rozpadů atomů v radioaktivní látce za určitý časový interval, počet branek ve fotbalovém zápase atd.).

Odhadování typu rozdělení

Když jsme si nyní popsali, co je to náhodná veličina, realizace náhodné veličiny (pokud nevíte, jaký je rozdíl mezi náhodnou veličinou a její realizací, vraťte se k odstavci *Realizace náhodné veličiny*) a uvedli jsme si seznam nejčastěji se vyskytujících rozdělení náhodných veličin, mohlo by se zdát, že už známe všechno ze základů statistiky. Bohužel, není tomu tak – ve skutečnosti známe jen polovinu toho, co potřebujeme znát. V praxi nám totiž většinou někdo dá naměřená data (tedy vlastně realizace nějaké náhodné veličiny), ale už nám neřekne (protože to ani sám neví) jaké rozdělení tato náhodná veličina měla. Jak se tedy dá poznat, z jakého rozdělení pocházejí naměřená data?

K odhadu typu rozdělení pomáhá histogram. Histogram je vlastně forma grafického znázornění dat, kde si osu x rozdělíme na několik intervalů a nad každým takovým intervalem zkonstruujeme sloupec takové výšky, která odpovídá tomu, kolik naměřených dat padne do tohoto intervalu. Příklad histogramu můžeme vidět na obrázku.

Intuitivně lze tvrdit, že pro velký počet naměřených dat bude tvar histogramu velmi podobný, jako je tvar hustoty pravděpodobnosti (případně pravděpodobností nabývání jednotlivých

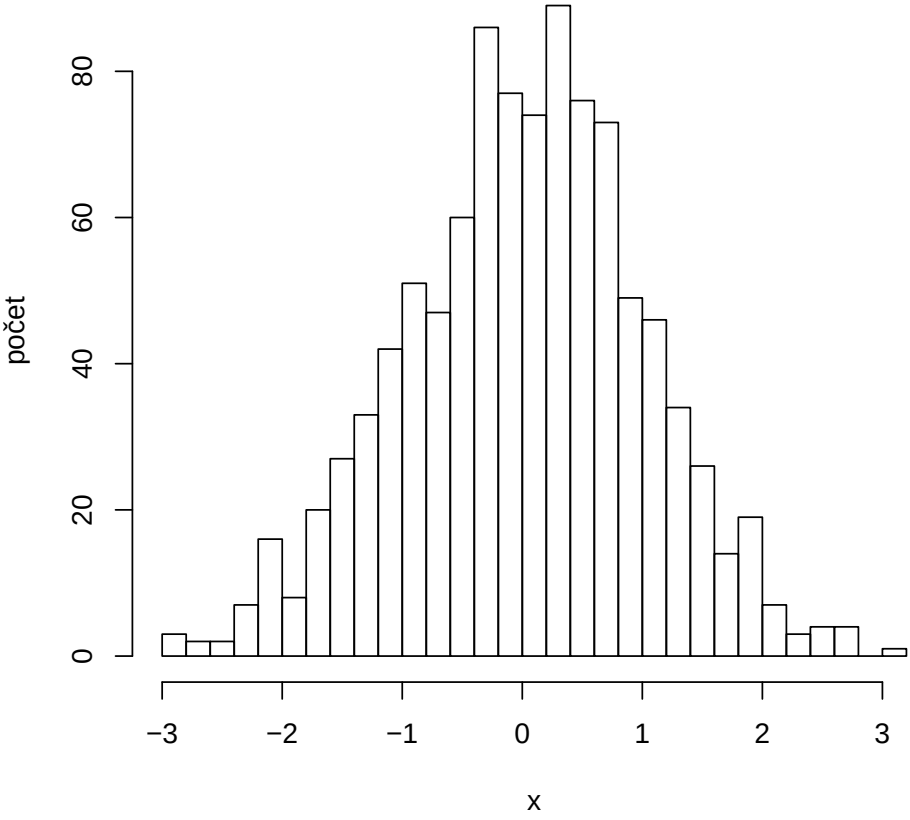
hodnot) rozdělení, ze kterého měřená data pochází. Nikdy však nemůžeme očekávat, že histogram bude mít přesně takový teoretický tvar – bude se jen podobat!

Problém, na který při konstrukci histogramu narazíme, je volba počtu a šířky sloupců. Toto je velmi složité téma, o kterém bylo napsáno spousty knih a článků, my si ovšem vystačíme s následujícími dvěma obecnými poučkami:

- Intervaly volíme všechny stejně široké. Začátek prvního intervalu volíme jako nejmenší naměřenou hodnotu, konec posledního intervalu jako největší naměřenou hodnotu.
- Počet sloupců volíme jako
$$P_n \approx \sqrt{n}\,,$$
kde n je počet měření.

Prostý pohled na histogram a jeho srovnání s nejběžnějšími typy hustot pravděpodobností (a pravděpodobností nabývání) často odhalí, ze kterého rozdělení naměřená data pochází. Toto si také vyzkoušíte v zadaných úlohách.

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.



Obr. 2: Ukázka histogramu – v tomto případě naměřená data pocházejí pravděpodobně z normálního rozdělení.