ROOTbits

Tento repozitář obsahuje ROOTovský kód a vysvětlení k některým statistickým algoritmům pro potřeby výuky na MFF UK.

Peter Kvasnička, MFF UK Praha, jaro 2021

peter.kvasnicka@mff.cuni.cz

Úvod

Kód k jednotlivým témám najdete v adresáři `code`.

Bootstrap: Odhad konfidenčního intervalu pro korelační koeficient

Představte si, že máme data, zobrazená na následujícím grafu, a zajímá nás (Pearsonův) korelační koeficient mezi x a z

$$\rho(x,z) = \frac{cov(x,z)}{\sigma_x \sigma_z} = \frac{\sum_i (x_i - \bar{x})(z_i - \bar{z})}{(\sum_i (x_i - \bar{x})^2)^{1/2} (\sum_i (z_i - \bar{z})^2)^{1/2}}$$

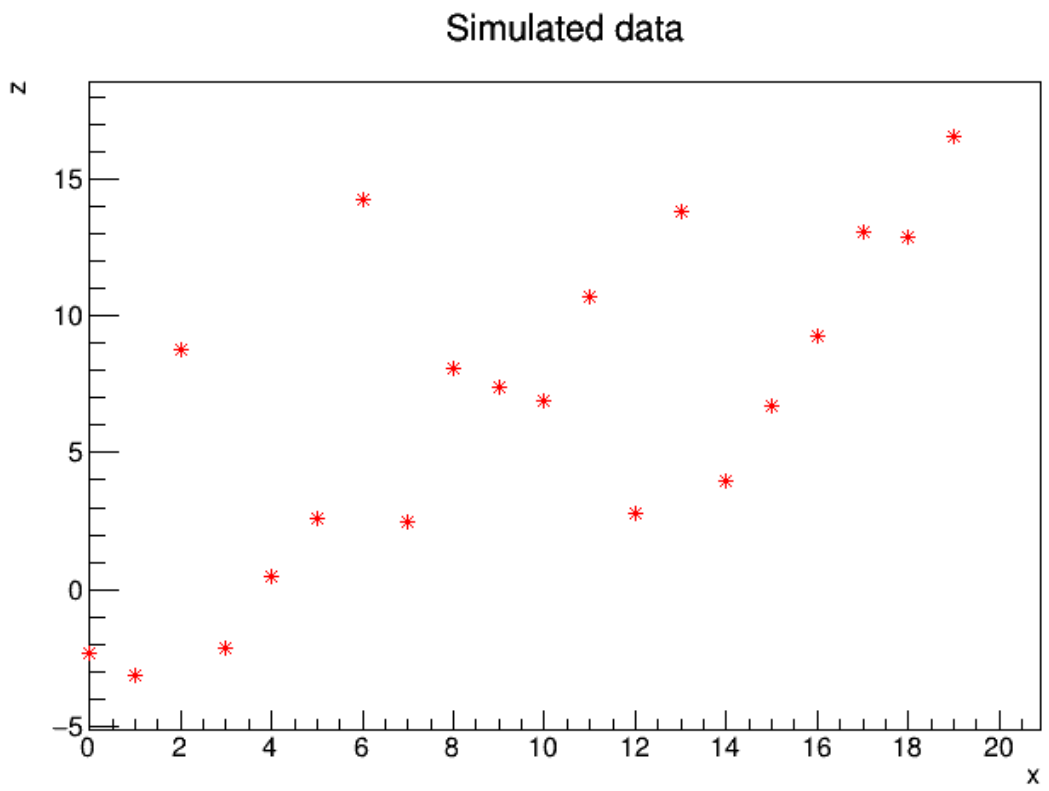
a jeho přesnost.

Níže se dovíte, jak jsme tato data nasimulovali. Prozatím uveďme, že pro ně určitě neplatí, že dvojice (x_i, z_i) byly vzorky z dvourozměrného Gaussova rozdeělení. Pak ale také nebudou platit vzorečky pro odhad chyby korelačního koeficientu, které najdete v učebnicích nebo na Wikipedii[1]:

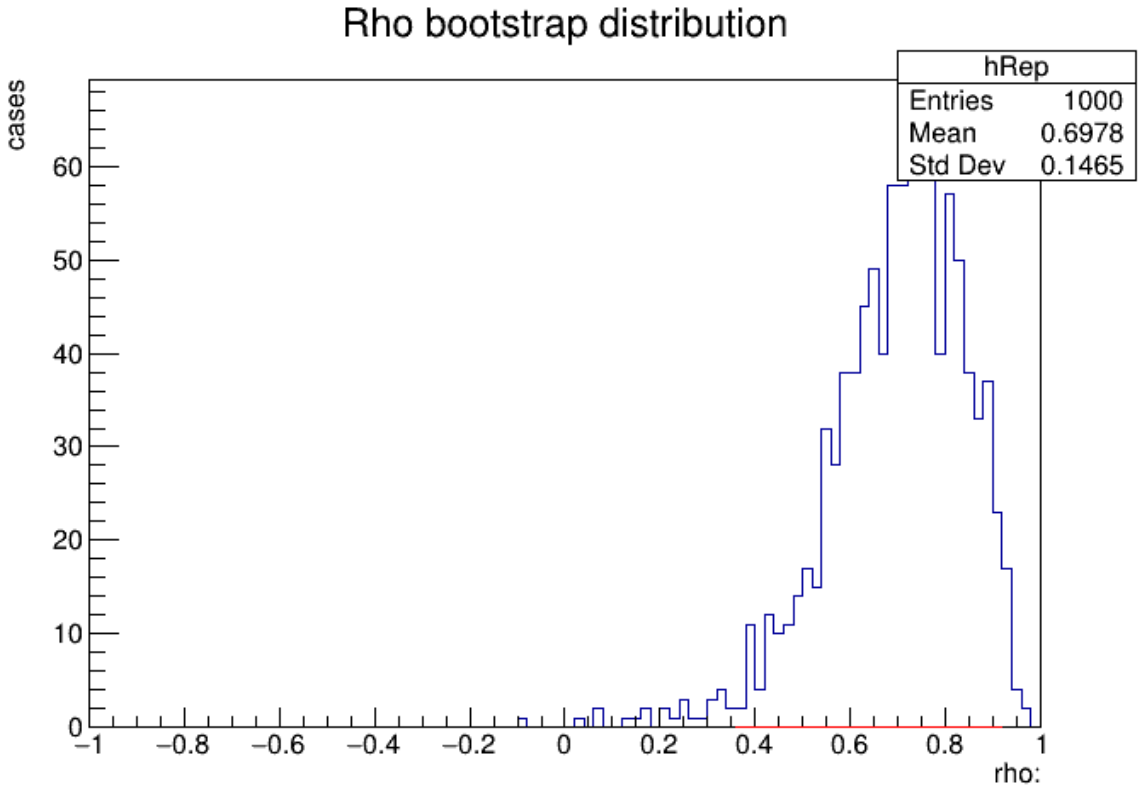
Proto si v této kapitolce ukážeme jiný postup, založený na principu, nazývaném *bootstrap*[2], který nám umožní z dat získat přiblížné rozdělení pravděpodobnosti odhadu Pearsonova korelačního koeficientu.

Při *bootstrapu* vytváříme repliky dat tak, že generujeme náhodné vzorky z odhadu distribuční funkce, získaného přímo z dat. To znamená, že repliky dat vytváříme jako náhodné výběry s vracením z původních dat.

Ukažme si to na našem případě:



Naše data tvoří $n = 20$ dvojic $(x_i, z_i), i = 1, \ldots n$, nasimulovaných tak, aby korelační koeficient byl $\rho_0 = 0.8$. Náhodným výběrem n dvojic s vracením (tedy jednu hodnotu můžeme pro danou repliku vybrat i víckrát) můžeme vytvořit velký počet replik, u každé spočíst korelační koeficient ρ a hodnoty vynést do histogramu. Z velkého množství hodnot pak lehko určíme i kvantily rozdělení pravděpodobnosti a tedy odhad konfidenčního intervalu:



Na obrázku je červeně vyznačený 95%-ní konfidenční interval

$$95\%\ CI \approx (0.36, 0.92)$$

Tento princip je velice mocný, ale má určitá omezení.

1. Především, přesnost rekonstruovaného rozdělení pravděpodobnosti pro danou statistiku závisí od toho, jak dobře umožňují výchozi data rekonstruovat ty charakteristiky distribuční funkce, které jsou důležité pro odhad statistiky. (Příklad: metoda bude fungovat špatně například pro medián, protože ten silně závisí na lokálním tvaru distribuční funkce kolem hodnoty $F = 0.5$).
2. Konfideční intervaly, získané touto metodou, budou přesnější než standardní odhady, ale nejsou to přesné konfidenční intervaly (tedy nemusí mít deklarované pokrytí) Existují metody, které umožňují tyto intervaly významně zpřesnit.

Generování dat

Pro tuto kapitolku jsme potřebovali generovat data s daným korelačním koeficientem. Přesněji, ať je x nějaký výchozí vektor, například $x_i = i, i = 0, 1, \dots n$, a chceme sestrojit nějaký vektor z tak, aby korelační koeficient mezi x a z měl předepsanou hodnotu ρ_0.

K tomu použijeme pomocné tvrzení[3]: Nechť $cov(x, y) = 0, \ \ \sigma_x = \sigma_y$. Pak pro vektor

$$z = \rho_0 x + \sqrt{1 - \rho_0^2} y$$

platí $\rho(x, z) = \rho_0$. Důkaz se provede jednoduše dosazením.

Zůstává otázka, jak pro vektor x vytvořit (nějaký) nekorelovaný vektor y. Uděláme to tak, že vytvoříme nějaký y_0 s náhodnými složkami (například gaussovskými) s nulovou střední hodnotou ($\bar{y_0} = 0$), vyprojektujeme z něj složku ve směru x, a kolmou část y nanormujeme tak, aby měla stejnou varianci jako x:

$$\begin{aligned} w &= x - \bar{x}, \\ y_\perp &= y_0 - \frac{y_0 \cdot w}{|w|^2} w \\ y &= \frac{\sigma_x}{\sigma_{y_\perp}} y_\perp \end{aligned}$$

Výhoda této metody je, že je čistě geometrická a nespoléhá na žádné předpoklady o rozdělení pravděpodobnosti komponent zůčastněných vektorů.

Literatura

1. Pearsonův korelační koeficient na Wikipedii
2. Bootstrap na Wikipedii
3. Stránka ze StackExchange s vysvětlením metody