

10. cvičení z PSt — 14. a 17.12.2020

1. Otestujte náhodnost kostky. Můžete házet ručně, máte-li kostku po ruce, nebo využít vhodný generátor (na webu nebo softwarový). Případně přímo v R (např. `sample(6,100, replace=T)`). Použijte Pearsonův test dobré shody.

2. Pojďme zkontrolovat, zda počty emailů jsou opravdu dobře modelovány Poissonovským rozdělením. Můžete použít svoje data, nebo použít moje, za letošní listopad. (Nejsou to všechny emaily, jenom „ty důležité“, ale to by na statistických vlastnostech nemělo nic měnit.)

```
moje_emaily11 = c(0, 6, 14, 8, 8, 9, 3, 3, 12, 12, 15, 7, 15, 2, 5, 13, 5, 17, 15, 11, 9, 2, 16, 8, 9, 11, 6, 2, 2, 9)
```

Rozmyslete, jak použít test dobré shody, zjistěte velikost λ a test proveďte. V R můžete použít buď vzorec z přednášky a dosadit do distribuční funkce (`pchisq(T, df)`), nebo použít příkaz `chisq.test`. Ideálně zkuste obojí.

Pro případné zvýšení přesnosti lze použít příkaz `options(digits = 22)`.

Zamyslete se nad výsledkem.

3. Pro hodnoty $x = 1, 5, 9, 10, 13, 16, 20, 30$ nám vyšlo $y = 6.1982, 12.9892, 23.8005, 23.8891, 30.0391, 35.7535, 49.0685, 63.1825$. Proveďte lineární regresi. Jak dobrá je? Nakreslete i obrázek, kde vynesete body a proloženou přímkou.

4. Na stránce <https://www.google.com/covid19/mobility/> najdete data o změně mobility lidí v tomto roce. Stáhněte si data (v csv) pro Českou republiku. Označme jako x vektor s hodnotami „`retail_and_recreation_percent_change_from_baseline`“ pro celé Česko a y analogický vektor pro Prahu. (Jedná se o změnu v mobilitě lidí v obchodech a rekreačních zařízeních.) Dejte pozor, abyste k sobě přiřadili údaje ze stejného dne.

(a) Proveďte lineární regresi. Pro kontrolu nakreslete i graf, který lineární závislost dokládá (nebo vyvrací). Zakreslete tam i nalezenou přímkou.

(b) Nepovinný bonus: vymyslete graf, který by ilustroval i závislost na čase (ale stále i lineární závislost).

Bude se vám hodit funkce `read.csv` a základní povědomí, co jsou data frames. Na <https://www.datacamp.com/community/tutorials/15-easy-solutions-data-frame-problems-r> je mnohem více, než potřebujete. stručnější je <https://stats.idre.ucla.edu/r/modules/subsetting-data/>.