Datum:	Příjmení, jméno:	Body celkem:

Analýza experimentálních dat, cvičení 2

Soubor dat: prijmeni_jmeno.csv

Popis dat: Skupina zdravých mluvčí a několik pacientů s velmi rozvinutou dysartrií byli podrobeni diadochokinetickému testu (DDK) a kvalita jejich artikulace byla vyhodnocena příznakem *voice onset time* (VOT) měřeného automatickou metodou (Novotný et al. **2015**).

Zadání úlohy	body
Stáhněte si a načtěte tabulku dat. Pro načtení využijte funkci readtable('data.csv', 'ReadVariableNames', true).	
Vykreslete si histogram. Vyzkoušejte různá nastavení počtu binů, aby byla distribuce co nejvíce zřetelná. Určete: průměr, standarni směrodatnou odchylku, medián, mediánovou absolutni odchylku a trimmean. Do obrázku společně s histogramem vykreslete průběh hustoty pravděpodobnosti normálního rozdělení (N) pro N(průměr, směrodatná odchylka) a N(medián, mediánová absolutní odchylka · 1.48).	0.4
Implementujte kernel density estimation. Pro odhad šířky pásma užijte pravidlo (Silverman **1986**): $$band\ width = min\left(\sigma, \frac{IQR}{1{,}35}\right) \cdot 0.9 \cdot n^{-1/5}\ , \quad (1)$$ kde σ je směrodatná odchylka Vašich dat, IQR je mezikvartilové rozpětí Vašich dat a n je počet pozorování. Vykreslete si průběh získaný pomocí Vaší implementace kernel density estimation do předchozího obrázku s histogramem.	0.7
Určete které hodnoty jsou extremní pomocí pravidla 68–95–99.7. Zkuste si vykreslit kumulativní empirickou distribuční funkci a odhadnout, kde u ní nejspíše nalezneme ony extrémní hodnoty.	0.2
Extrémní hodnoty vyjměte a znovu vykreslete histogram, znovu vypočtěte parametry (mean, standarni odchylku, medián, mediánovou absolutni odchylku a trimmean) a vykreslete si průběh hustoty normálního rozdělení N(průměr, směrodatná odchylka) a N(medián, mediánová absolutní odchylka · 1.48).	0.2
Odpovězte na následující otázky: *Jak moc se odhadnuté distribuce liší pro data s a data bez extrémních hodnot?* *Které statistické parametry jsou citlivé na výskyt extrémních hodnot?* *Které statistické parametry jsou naopak velmi robustní?*	0.5

Zadání úlohy	body
Nepovinný bonus: Vygenerujte si libovolný počet M náhodných nezávislých vektorů o délce L libovolného nenormálního rozdělení[1]. Tyto vektory složte do matice o velikosti LxM nebo MxL (záleží, zda jsou Vaše vektory sloupcové nebo řádkové). Vektory samozřejmě můžete vygenerovat i jedním příkazem jako matici LxM. Hodnoty přes jednotlivé vektory zprůměrujte a vykreslete histogram. Zkuste různé počty vektorů M (např. 2,3,5, 100) a různé délky L. Měli byste tím získat aproximaci normálního rozdělení. Právě jste si vyzkoušeli centrální limitní větu, která je podle mnohých odpovědí na otázku, proč je normální rozdělení tak hojné. Zodpovězte následující otázku: *Jak souvisí počet průměrovaných výběrů s tím, jak se výsledné rozdělení bude blížit normálnímu rozdělení?*	1

Reference

Novotný, M., Rusz, J., Čmejla, R., and Růžička, E. (2014). Automatic evaluation of articulatory disorders in Parkinson's disease. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 22, 1366-1378.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. CRC press, 47.

[1] Následující postup Vám bude samozřejmě dávat stejné výsledky i pro normální rozdělení. Nebude to však tak překvapivé.