Zadání Cvičení #2

Popis dat: Pracovní data jsou uložena v souboru data.csv, který je k dispozici ke stažení na Moodle stránce tohoto předmětu, ve složce příslušného cvičení. Pro načtení dat do Matlabu využijte funkce `readtable`.

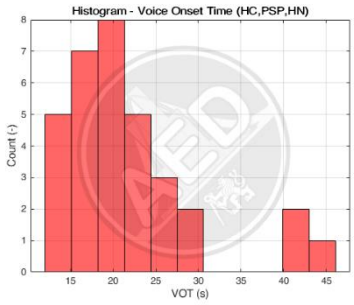
```
T = readtable('data.csv','ReadVariableNames',true);
```

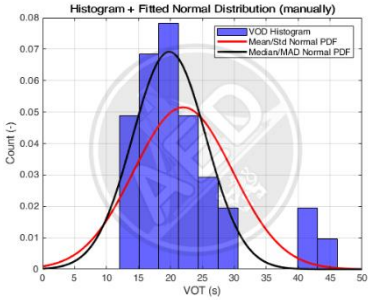
Data musíte mít samozřejmě stažená a musíte se nacházet ve složce kde jsou, nebo mít složku kde se nacházejí přidanou do pracovní cesty Matlabu (*Matlab Path*). Data jsou ve formě tabulky kde první sloupec obsahuje identifikační kódy subjektů – labely (ve formě textového řetězce) a druhý sloupec příslušné naměřené hodnoty.

Zdroj dat: Skupina zdravých mluvčí (HC – healthy controls) a několik pacientů s velmi rozvinutou dysartrií (labely PSP – Progressive supranuclear palsy a HN – Huntington disease) byli podrobeni diadochokinetickému řečovému testu (DDK – test, který ověřuje schopnost řečového traktu provádět rychlé změny v „konfiguraci“, typicky prováděný jako rychlý sled hlásek *pa-ta-ka* opakovaně za sebou) a kvalita jejich artikulace byla vyhodnocena příznakem s názvem *voice onset time* (VOT), který byl získán z řečových nahrávek pomocí automatizovaného počítačového programu (detaily k měření a k podstatě naměřeného parametru viz Novotný et al. 2015).

Zadání úlohy	body

Z poskytnutých dat si **vykreslete běžný histogram**. Vyzkoušejte různá nastavení počtu *binů*, aby byla distribuce co nejvíce zřetelná. Nezapomeňte graf s histogramem správně popsat – osy, titulek atd.
 	
Vypočtěte na datech následující:
• Obyčejný průměr
• Standardní směrodatnou odchylku
• Medián
• Mediánovou absolutní odchylku (MAD)
• *Trim-mean*
 	
Do obrázku s histogramem přidejte křivky (*ploty*) pro:
• **Průběh hustoty pravděpodobnosti** (PDF) normálního rozdělení s parametry odhadů průměru a směrodatné odchylky
• PDF pro normální rozdělení, kde argumenty:
  ○ Průměr nahradíte mediánem
  ○ Směrodatnou odchylku nahradíte hodnotou MAD*1.48
Nápověda: Pro správně naškálování histogramu vzhledem k PDF funkcím bude potřeba změnit parametr histogramu `Normalization`. PDF funkce nepište manuálně podle definice, vyhledejte si příslušnou funkci v Matlabu.
	1



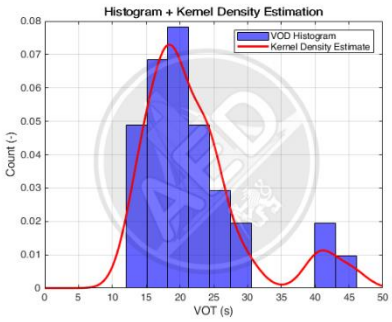


V LiveScriptu naprogramujte řešení, nechte si vykreslit požadované obrázky a správně je popište. Hodnoty deskriptivních veličin, které jste měli vypočítat, vypište pomocí funkce `fprintf` spolu s popisky.

Manuálně **implementujte algoritmus pro výpočet *kernel density estimation*** (KDE). Pro odhad šířky pásma použijte následující pravidlo (Silverman 1986):

$$band\ width = min\left(\sigma, \frac{IQR}{1{,}35}\right) \cdot 0.9 \cdot n^{-1/5}$$

kde σ je směrodatná odchylka Vašich dat, *IQR* je mezikvartilové rozpětí vašich dat a *n* je počet pozorování ve vzorku.



V LiveScriptu naprogramujte algoritmus pro výpočet KDE. Pomocí něj vypočtěte KDE distribuci nad vašimi daty, a zobrazte jí do jednoho obrázku společně s histogramem.

1.5

Zkopírujte si obrázek s histogramem a odhadem normální PDF z první úlohy. Do tohoto obrázku vykreslete vertikální čáry reprezentující hodnoty vzdálené 1, 2 a 3 směrodatné odchylky od střední hodnoty *na každou stranu*. Pomocí pravidla *68-95-99.7* určete, které hodnoty v datech byste považovali za extrémní (outliery).

Vykreslete si **kumulativní empirickou distribuční funkci** (CDF) a s její pomocí odhadněte vizuálně hranici za kterou se by se hodnoty daly považovat za outliery. Do grafu s empirickou CDF vykreslete taktéž CDF ideálního normálního rozdělení (s parametry odhadu průměru a směrodatné odchylky vašich dat). Pokud chcete, můžete si do grafu s CDF vykreslit i vertikální čáry pro průměr, medián či 68-95-99.7 hranice.

V LiveScriptu vykreslete oba výše popsané grafy (s PDF a CDF) a zaznamenejte do nich, kde byste sami umístili hranici pro lokalizaci outlierů.

Vámi identifikované extrémní hodnoty z datasetu **vyjměte.** Vypočtěte všechny deskriptivní statistiky jako v první úloze a také si znovu vykreslete histogram a do grafu opět vykreslete průběhy PDF funkcí stejně jako v první části (teoreticky stačí jen *copy+paste* kódu z první části a změna vstupních dat). Srovnejte hodnoty deskriptivních statistik a vzhledy grafů před a po úpravě datasetu.

V LiveScriptu vytvořte výše popsané obrázky se všemi náležitostmi. Také odpovězte stručně 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?*

1.5

Zadání úlohy	body
## Nepovinný bonus: Vygenerujte si libovolný počet `M` náhodných vektorů o délce `L` z libovolného nenormálního rozdělení[1]. Tyto vektory složte do matice o velikosti `L x M` nebo `M x L` (záleží, zda jsou vaše vektory sloupcové nebo řádkové). Vektory samozřejmě můžete vygenerovat i jedním příkazem jako matici `L x M`. Hodnoty přes jednotlivé vektory zprůměrujte a vykreslete histogram výsledného vektoru. Vyzkoušejte si různé počty vektorů `M` (např. 2,3,5, 100) a různé délky `L`. Měli byste tím získat aproximaci normálního rozdělení. V praxi si tak zkusíte efekt *centrální limitní věty* 😊, která je podle mnohých odpovědí na otázku, proč je normální rozdělení v přírodě tak hojné. Zodpovězte následující otázku: *Jak souvisí počet zprůměrovaných vektorů s tím, jak moc se výsledné rozdělení bude blížit normálnímu rozdělení?*	0.5

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. 😊