Zadání Cvičení #3

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 jsou ve formě tabulky kde první sloupec obsahuje identifikační kódy subjektů – labely (ve formě textového řetězce), druhý sloupec labely skupin a třetí sloupec příslušné naměřené hodnoty.

Zdroj dat: Skupina zdravých mluvčí (HC – healthy controls) a skupina pacientů s Parkinsonovou nemocí (PD) byla podrobena diadochokinetickému řečovému testu (DDK) 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. 2014).

Zadání úlohy	body
• Vizualizujte si **do jednoho obrázku** poskytnutá data takovým způsobem, aby byla distribuce dat v obou skupinách co nejzřetelnější, a aby byly mezi skupinami dobře vidět případné rozdíly. **Využijte znalosti a kód z předešlého cvičení.** Nezapomeňte obrázek dobře popsat! 😉 • **Slovy odpovězte:** *Jak byste pouhým okem zhodnotili efekt skupin, tj. rozdíly mezi skupinami, viditelné na vaší vizualizaci?* **Nápověda:** Pro vynesení více grafů do jednoho obrázku (figure) – příkaz `hold on`	1
• Vypočítejte mezi skupinami HC a PD **Cohenovo *d*** (Cohen's *d*). Výslednou hodnotu si vypište a slovy odpovězte: *Jak silný je efekt mezi analyzovanými skupinami?* $d = \frac{\mu_1 - \mu_2}{\sigma_P}\ ;\ \sigma_P = \sqrt{\frac{(n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2}{n_1+n_2+2}}$	
• Vykreslete si (**populační**) kumulativní distribuční funkce (CDF) pro každou skupinu do jednoho obrázku. • Do obrázku s oběma CDF vykreslete **subjekt X** (vzorek s labelem „X", skupina „?") libovolným způsobem, nejlépe jako vertikální čáru. • **Slovy odpovězte:** *Přiřadili byste subjekt X spíše do skupiny HC, do skupiny PD nebo byste ho nezařadili do ani do jedné?* **Nápověda:** *Populační CDF* graf je hladký, *Empirický CDF* graf má tvar schodů.	1

V této části použijete techniku "bootstrappingu" pro praktickou ukázku výpočtu a významu p-hodnoty (p-value).

Co je "bootstrapping"? Jednoduše řečeno je to umělé zvětšování datasetu pouze s využitím naměřených dat, které již máme k dispozici. „Bootstrapping" jako výraz pochází z anglického příměru kdy se „vytáhnete (např. z jámy) za vlastní kšandy", tzv. si pomůžete sami, bez cizí pomoci. V kontextu analýzy dat to má pak význam toho, že si „přiděláte" data bez dalšího reálného měření (lidí, strojů, počasí atd.).

Úloha:

1. **Vytvořte si (nulovou) hypotézu k otestování** – typicky chcete zjistit, zda vzorky od skupin PD a HC pocházejí z populací se stejnou střední hodnotou (za předpokladu stejného rozptylu obou rozdělení).
2. Vypočtěte si reálnou hodnotu vyšetřované testové statistiky: v našem případě nás zajímá rozdíl středních hodnot, takže **vypočtěte rozdíl středních hodnot obou vzorků.**
3. **Nastavte si hladinu statistické významnosti α.**
4. Začněte s *bootstrappingem*:

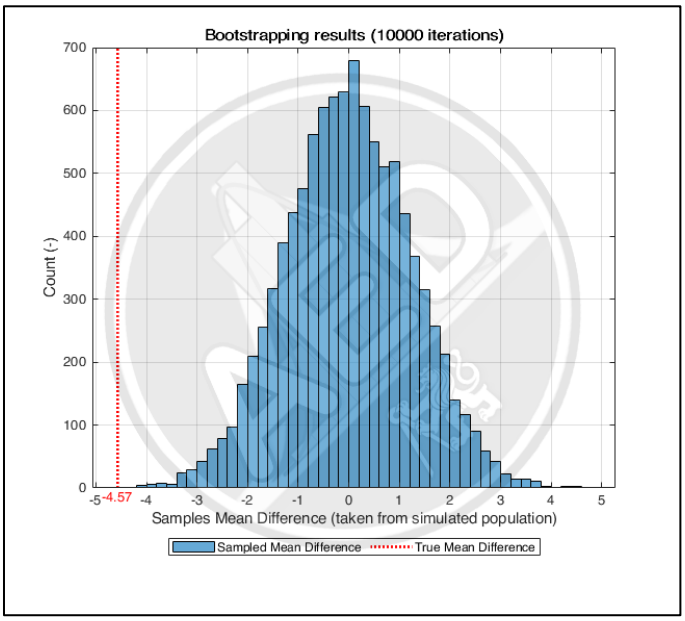
   Předpokládejte, že vaše nulová hypotéza platí. Pak můžete simulovat teoretickou společnou populaci (ze které pocházejí data obou skupin), **spojením všech dat do jednoho datasetu**.

   Proveďte výběry vzorků z vaší vytvořené společné populace (iterujte):

   - Pro získání relevantního odhadu je potřeba z populace vybrat alespoň 10000 vzorků. Simulujete tak 10000 jednotlivých měření.
   - Protože nás zajímá rozdíl středních hodnot, musíte vytáhnout v každé iteraci vždy dvojici vzorků.
   - Každý vytažený vzorek musí být stejně velký jako původní dva naměřené vzorky. Vzorky ze společné populace vytahujte náhodně a s opakováním!
5. Pro každou vytaženou dvojici vzorků **vypočtěte rozdíl jejich středních hodnot a uložte si ho**. Získáte tak 10000 hodnot, které dohromady tvoří pravděpodobnostní rozdělení rozdílů (🙂) středních hodnot dvou náhodně vytažených vzorků z vaší populace.
6. Toto rozdělení si vykreslete pomocí histogramu.
7. **Podívejte se,** kolik pozorovaných rozdílů středních hodnot z vytvořeného setu nabývá hodnoty **stejné nebo větší**, než je **absolutní hodnota** skutečného rozdílu středních hodnot ze začátku („*true mean difference*"). **Tuto hodnotu vydělte celkovým počtem hodnot** (10000), a získáte tak pravděpodobnost s jakou byste mohli získat stejnou nebo větší hodnotu rozdílu středních hodnot dvou vzorků, pokud by pocházeli ze stejné populace. Toto je vaše *p*-value. Skutečný rozdíl středních hodnot si vyznačte do obrázku s histogramem.

2

8. **Porovnejte** *p*-value s nastavenou hladinou statistické významnosti a vyhodnoťte vaše testování. Odpovídá výsledná *p*-value očekávání z první úlohy?



1 *Příklad výsledného obrázku s bootstrappingovým histogramem a vyznačenou skutečnou hodnotou rozdílu středních hodnot*

Na konec proveďte ještě standardní oboustranný dvouvzorkový t-test (*two-tailed, two-sample t-test)* pomocí Matlabovské funkce `ttest2`. Výslednou hodnotu testovací t-statistiky, společně se stupni volnosti a *p*-hodnotou, **reportujte korektním způsobem**. Příklady naleznete v přednáškách nebo rychlým vyhledáním na internetu *(„how to report t-test APA style"*).

Výstupem úlohy bude:

- **Obrázek s bootstrappingovým histogramem a skutečnými hodnotami rozdílů středních hodnot skupin.**
- **Slovní znění vašich hypotéz, vaše nastavení hladiny statistické významnosti a vypočtená hodnota *p*-value z bootstrappingu.**
- **Korektní report výsledku t-testu.**
- **Velmi stručná interpretace výsledků.**
- **Porovnání výsledků bootstrappingu s výsledky funkce ttest2.**

Zadání úlohy	body

Nepovinný bonus:

Do jaké ze dvou skupin, se kterými jste pracovali v tomto cvičení, byste zařadili subjekt X, pokud víte, že výskyt Parkinsonovy nemoci v populaci je 0.3 %?

Pro řešení využijte **Bayesovu větu.** Pokud nadefinujeme událost A jako příslušnost vzorku k modelu nějaké skupiny a událost B jako naměřenou hodnotu parametru pro tento vzorek, pak množiny událostí obsahují:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}; \quad \begin{matrix} A = \{X \in HC, X \in PD\} \\ B = VOT_X \end{matrix}$$

HC a PD reprezentují populační modely vypočtené na základě vzorků (z odhadů středních hodnot a odhadů směrodatných odchylek vypočtených z dat). **V našem cvičení tyto modely tvoří funkce hustoty pravděpodobnosti PDF stojící na zmíněných odhadech.**

B pak reprezentuje výsledek měření parametru VOT, v našem případě bude množina možností obsahovat jen jeden případ, a to naměřené VOT pro subjekt X.

Pak můžeme přepsat vzorec Bayesovy věty:

$$P(X \in HC|VOT_X) = \frac{P(VOT_X|X \in HC) \cdot P(X \in HC)}{P(VOT_X)}$$

$$P(X \in PD|VOT_X) = \frac{P(VOT_X|X \in PD) \cdot P(X \in PD)}{P(VOT_X)}$$

Vzhledem k tomu, že pracujeme se spojitými daty a pravděpodobnostními funkcemi, nebudou nás zajímat konkrétní hodnoty, ale pouze jejich poměr – ostatně chceme jen vzorek přiřadit do jedné, či druhé skupiny:

$$\frac{P(X \in HC|VOT_X)}{P(X \in PD|VOT_X)} = \frac{P(VOT_X|X \in HC) \cdot P(X \in HC)}{\cancel{P(VOT_X)}} \cdot \frac{\cancel{P(VOT_X)}}{P(VOT_X|X \in PD) \cdot P(X \in PD)} =$$

$$= \frac{P(X \in HC)}{P(X \in PD)} \cdot \frac{P(VOT_X|X \in HC)}{P(VOT_X|X \in PD)} = \frac{P(X \in HC)}{P(X \in PD)} \cdot \frac{\mathcal{L}(X \in HC|VOT_X)}{\mathcal{L}(X \in PD|VOT_X)}$$

$$= \frac{P(X \in HC)}{P(X \in PD)} \cdot \frac{f_{HC}(VOT_X)}{f_{PD}(VOT_X)}$$

V této úpravě jsme převedli pravděpodobnosti v druhém zlomku na *likelihoody* a využili toho, že hodnota *likelihood* funkce $\mathcal{L}$ je v bodě VOT_X totožná s hodnotou funkce hustoty pravděpodobnosti pro daný model f_{Model} v bodě VOT_X.

- První zlomek obsahuje apriorní pravděpodobnosti $P(X \in HC)$ a $P(X \in PD)$, že vzorek X patří do jedné ze dvou skupin bez znalosti hodnot měření. **Zde využijete hodnoty výskytu skupin v populaci.** Pamatujte, že pokud máme jen dvě skupiny, platí:
$$P(X \in HC) = 1 - P(X \in PD)$$
- Druhý zlomek se také nazývá likelihood ratio nebo Bayesovský faktor (pokud pracujeme s pravděpodobnostmi) a reprezentuje informace, které do výsledku – rozhodnutí – dodávají nové důkazy - měření.

0.5

Pokud bude hodnota výsledného poměru větší než 1, výsledek podporuje hypotézu v čitateli (tj. přiřazení vzorku do skupiny reprezentované prvním modelem), pokud bude menší než 1, je pravděpodobnější hypotéza ve jmenovateli (tj. přiřazení vzorku do skupiny reprezentované druhým modelem). Pro sílu rozhodnutí se můžete inspirovat zde. **Vyzkoušejte si podle tohoto návodu, jak ovlivní apriorní informace o rozdělení skupin vaše rozhodnutí o přiřazení neznámého vzorku do jedné ze skupin. Napište výslednou hodnotu poměru pravděpodobností a sílu vašeho rozhodnutí.** Tip: Video od kanálu 3Blue1Brown pro intuitivní pochopení Bayesovy věty.	

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.

Testování hypotéz - teorie

- ***Nulová* hypotéza** je formálním popisem či vyjádřením nějakého faktu, který si statistickou analýzou ověřujeme. Ke každé nulové hypotéze pak existuje tzv. ***alternativní* hypotéza**, která je (a musí být) jejím přesným logickým opakem (negací).
- Hypotézy musí být jednoznačné a musí být definovány slovy nebo ekvivalentním zápisem pomocí logických výrazů.
- Většina statistických testů používá tzv. ***p*-hodnotu** (*probability*-value) pro vyjádření pravděpodobnosti s jakou můžeme pozorovat naměřená data za předpokladu, že je nulová hypotéza pravdivá. Jinými slovy: *Pokud je nulová hypotéza pravdivá, jaká je pravděpodobnost, že výběrem z populace dostaneme taková data, jako ty naše?*
- Abychom dokázali interpretovat výslednou *p*-hodnotu libovolného testu, musíme nejdříve testování správně „nastavit", provést ho a následně vyhodnotit.

1. Vytvoříme hypotézy: nulovou hypotézu H_0 a její přesný logický opak, alternativní hypotézu H_1

- *Příklad*: Budeme mít dva vzorky naměřených dat od zdravých lidí a od nemocných lidí. Chceme zjistit, jestli lze na základě nějaké veličiny (*feature*) vypočtené z našich dat rozlišit tyto dvě skupiny. **Vytvoříme hypotézy**:

  Nulová hypotéza H_0: Oba naměřené vzorky (jeden od skupiny zdravých, druhý od skupiny nemocných) pocházejí z populací jejichž pravděpodobnostní rozdělení mají stejnou střední hodnotu.

  Alternativní hypotéza H_1: Pravděpodobnostní rozdělení populací, ze kterých naše dva vzorky pocházejí nemají stejnou střední hodnotu.

- Mějte na vědomí, že hypotézy nepojednávají o rozdílech mezi naměřenými *vzorky*, ale o rozdílech mezi teoretickými *populacemi*, ze kterých vzorky pocházejí. Kdybychom testovali rozdíly mezi vzorky, případná informace, že mezi zdravými a nemocnými skupinami lze najít rozdíl, by nám byla k ničemu, protože by byla platná a použitelná jen pro náš naměřený vzorek. My chceme tuto informaci generalizovat a využít například pro klasifikaci nebo regresi na všech možných jiných subjektech, které bychom mohli vybrat z obecné populace. Proto na základě naměřených vzorků vytváříme statistické modely populací a hledáme rozdíly mezi nimi.
- Dále můžeme předpokládat, že vzorky mají normální rozdělení, že populace, ze kterých vzorky pochází, mají normální rozdělení, nebo že oba vzorky mají stejný nebo přibližně stejný rozptyl. Tyto předpoklady pak určí, jaký specifický statistický test použijeme.

2. Nastavíme hladinu statistické významnosti α (*level of statistical significance*)

- Tato hladina určujte hodnotu, se kterou budeme náš výsledek, tj. naši výslednou p-hodnotu, porovnávat. Pravděpodobnosti hypotéz, které vyjdou menší, než je tato hodnota, budeme již považovat za *statisticky nevýznamné* (nepravděpodobné).
- Typické úrovně jsou α=0.05, 0.01 či 0.001, což odpovídá pravděpodobnosti 5 %, 1 % a 0.1 %, resp. 95%, 99% a 99.9%. Čím nižší je hladina α, tím větší musí být pravděpodobnost, že námi pozorovaný výsledek není jen dílem statistické náhody.
- Hodnoty, na které se úroveň nastavuje se liší mezi obory, nicméně hodnota 5% je zdaleka nejběžnější.

3. Provedeme statistický test a získáme *p*-hodnotu

Statistické testy většinou vypočtou nějakou hodnotu tzv. *testovací statistiky* (například u t-testu je to hodnota *t*, tedy „*t-statistic*“) a k ní příslušnou hodnotu *p-value*.

Příklad*: *Dvou-vzorkový t-test (také *Studentův t-test*, ang. *„two-sample t-test“*), který ve standardním provedení testuje hypotézy z bodu 1, vypočte hodnotu testovací *t*-statistiky pomocí vzorce který bere v potaz střední hodnoty a rozptyly naměřených vzorků.*

Příslušná výsledná *p*-hodnota pak v tomto případě přímo říká: pokud by populace, ze kterých naše naměřené vzorky pochází, *opravdu* měly stejnou střední hodnotou, jaká je pravděpodobnost, že budeme pozorovat (naměříme) takové hodnoty, jaké jsme naměřili v našich vzorcích?

* Hodnota *p*-value se určí odečtením hodnoty pravděpodobnosti z funkce hustoty pravděpodobnosti pro Studentovo t-rozdělení, které popisuje (matematicky) rozdělení dat při odhadu střední hodnoty normálního rozdělení.

4. Porovnáme p-hodnotu a zamítneme/nezamítneme hypotézu

- Podíváme se, zdali je *p*-hodnota větší než námi určená hranice statistické významnosti. Pokud ano, je pravděpodobnost nulové hypotézy příliš, a nemůžeme jí zamítnout. Alternativní hypotézu pak nemůžeme přijmout.
- Pokud je hodnota *p*-value menší než α, je pravděpodobnost pozorování našich dat za předpokladu nulové hypotézy příliš malá, a nulovou hypotézu zamítáme. V tom případě můžeme přijmout hypotézu alternativní.

Pokud by byla hodnota například $p = 0.01$, pak pravděpodobnost, že vytáhneme námi naměřená data z populací, která mají stejnou střední hodnotu, je pouze 1 %. Jinými slovy je taková pravděpodobnost velmi malá, proto je velmi nepravděpodobné, že vzorky pocházejí z rozdělení se stejnou střední hodnotou. Nulovou hypotézu můžeme zamítnout a místo toho přijmout hypotézu alternativní.

Na druhou stranu, pokud by byla výsledná hodnota $p = 0.5$, pak je 50% šance, že z populací se stejnou střední hodnotou vytáhneme takové vzorky jako ty naše. Taková šance je příliš velká, a proto nelze nulovou hypotézu zamítnout (*to neznamená že automaticky platí*, jen jí nelze zamítnout – a naopak alternativní hypotézu nelze přijmout).

5. Interpretujeme naše výsledky

Výsledky musíme vždy interpretovat běžným jazykem. Jaké jsou důsledky našich zjištění? Můžeme například použít nějaký řečový test pro to, abychom rozlišili zdravé a nemocné jedince, bez jakékoli další informace? Jak silný je náš test? Jaké jsme pro náš výsledek museli mít předpoklady? Ve vědeckých článcích se tato část vždy objevuje v sekci diskuze, tedy „*discussion*“.