1) Definujte pojem stredni hodnota náhodné veličiny a aritmeticky prumer. Vysvětlete rozdíl mezi nimi.

Střední hodnota

$$\mathrm{EV}(x) = \sum_{i=1}^{n} p(x_i)x_i$$

where

x_i = outcome i

$p(x_i)$ = probablilty of outcome i.

Aritmetický průměr $$A = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

Aritmetický průměr je jednoduše průměrná hodnota ze všech hodnot, střední hodnota je průměrná hodnota náhodné veličiny zatížená pravděpodobností.

2) Uvažujte náhodný vektor X. Definujte kovarianční a korelační matici. Jake maji tyto matice vlastnosti? K cemu se dají použít?

Kovarianční matice

$s_j^2 = (1/n)\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the variance of the j-th variable

$s_{jk} = (1/n)\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ is the covariance between the j-th and k-th variables

$\bar{x}_j = (1/n)\sum_{i=1}^{n} x_{ij}$ is the mean of the j-th variable

$$\mathbf{S} = \begin{pmatrix} s_1^2 & s_{12} & s_{13} & \cdots & s_{1p} \\ s_{21} & s_2^2 & s_{23} & \cdots & s_{2p} \\ s_{31} & s_{32} & s_3^2 & \cdots & s_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & s_{p3} & \cdots & s_p^2 \end{pmatrix}$$

- symetrická a pozitivně semidefinitní
- využití v PCA k nalezení optimální báze prostoru nižší dimenze

Korelační matice

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}$$

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ r_{31} & r_{32} & 1 & \cdots & r_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{pmatrix}$$

- symetrická a pozitivně semidefinitní
- korelace je “normovaná kovariance”

3) Co je to distribuční funkce? Co je to kvantil a kvantilová funkce? Co je to funkce pravděpodobnostní hustoty? Definujte pojmy formalne a uveďte jaky je mezi nimi vztah.

- distr funkce je funkce F(x) = P[x < a] …. udává pravdepodobnost, ze je hodnota NV je menší než nějaká zadana hodnota ‘a’
  - zleva spojita, neklesajici, limity v 0,1,..
  - pokud je NV spojita s hustotou f, tak platí F(x) = integral f(t) dt od -inf do x

- kvantilová funkce je inverzní distribucni funkce - x = F^-1(y)
- pokud je distr fce rostouci, tak lze psát jako Q(p) = F^1(p)
- udava, pro jake x bude vysledek nahodneho pokusu s pravdepodobnosti y mensi nebo roven x (opak distribucni - ta udava, s jakou pravdepodobnosti bude x < nejaka hodnota)
- **kvantil** = cisla, ktere deli statisticke soubory na stejne velke casti (median, kvartil, kvintil, decil, blabla)
    - kvantilova fce dostane na vstupu alfa <0,1> a produkuje alfa-kvantil - to je cislo takovy, ze alfa% dat je mensich a 1-alfa% dat je vetsich
    - P(x < q(p)) = p

fce pst hutoty - pokud f(x) je hustota pravdepodobnosti NV z intevalu <a,b>, tak hustota

4) Vysvětlete vlastními slovy vyznam p-hodnoty statistickeho testu. Jaky je výklad hladiny vyznamnosti α? Predpokladejte, ze p-hodnota nejakeho testu je 0.045? Jaka je pravdepodobnost, ze nulová hypotéza tohoto testu neplatí? Jak to souvisí s α?

p-hodnota je nejmenší hodnota hladiny významnosti, kdy ještě můžeme zamítnout nulovou hypotézu. Hladiny vyznamnosti α nám určuje s jakoou pravděpodobnostní zamítáme nulovou hypotézu pokud je pravdivá.

5) Vysvětlete pojem interval spolehlivosti (confidence interval). Co nejpresneji popiste, jak byste spočítal/a 99% interval spolehlivosti odhadu střední hodnoty normálního rozdělení z maleho vzorku velikosti m.

Interval spolehlivosti nám dává nějaký rozsah hodnot, kde se nejspíše nachází neznámý populační parametr.

Úroveň C intervalu spolehlivosti(typicky 0.90, 0.95, 0.99) nám dává pravděpodobnost s jakou je v tomto intervalu skutečná hodnota nějakého parametru, který se snažíme zjistit. (Typicky třeba aritmetický průměr populace.)

Example:
Suppose a student measuring the boiling temperature of a certain liquid observes the readings (in degrees Celsius) 102.5, 101.7, 103.1, 100.9, 100.5, and 102.2 on 6 different samples of the liquid. He calculates the sample mean to be 101.82. If he knows that the standard deviation for this procedure is 1.2 degrees, what is the confidence interval for the population mean at a 95% confidence level?
In other words, the student wishes to estimate the true mean boiling temperature of the liquid using the results of his measurements. If the measurements follow a normal distribution, then the sample mean will have the distribution $N(\mu, \frac{\sigma}{\sqrt{n}})$. Since the sample size is 6, the standard deviation of the sample mean is equal to 1.2/sqrt(6) = 0.49.

The critical value for a 95% confidence interval is 1.96, where (1-0.95)/2 = 0.025. A 95% confidence interval for the unknown mean μ is ((101.82 - (1.96*0.49)), (101.82 + (1.96*0.49))) = (101.82 - 0.96, 101.82 + 0.96) = (100.86, 102.78).
As the level of confidence decreases, the size of the corresponding interval will decrease. Suppose the student was interested in a 90% confidence interval for the boiling temperature. In this case, C= 0.90, and (1-C)/2 = 0.05. The critical value z^* for this level is equal to 1.645, so the 90% confidence interval is ((101.82 - (1.645*0.49)), (101.82 + (1.645*0.49))) = (101.82 - 0.81, 101.82 + 0.81) = (101.01, 102.63)

6) Vysvětlete význam pojmu chyba prvního druhu a chyba druheho druhu pouzivanych k popsan konkretnich chyb v procesu testování statistickych hypotez. Vysvětlete jak spolu tyto chyby souvis. Popiste, jak se da výskyt těchto chyb ovlivnit.

Type 1: Zamítnutí nulové hypotézy, i když je pravdivá.

- Šance erroru prvního typu je úroveň významnosti a značí se α. (Nazývá se také alpha level). Typicky je nastaven na 0.05

Type 2: Nezamítnutí nulové hypotézy, i když není pravdivá.

- Šance erroru druhého stupně se značí β.
- Vztahuje se k síle testu, kde síla testu je 1-β

Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision About Null Hypothesis (H_0)	**Fail to reject**	Correct inference (True Negative) (Probability = 1 - α)	Type II error (False Negative) (Probability = β)
	Reject	Type I error (False Positive) (Probability = α)	Correct inference (True Positive) (Probability = 1 - β)

- Většinou jdou oba errory ruku v ruce. Když jeden zmenšíme tak šance druhého se zvětší.

7) Formulujte centralni limitni vetu. Kde se dá využít?

8) Motivujte zavedení Studentova t-rozdělení. Ke kteremu rozdělení se t-rozdělení asymptoticky blíží s rostoucím počtem stupňů volnosti? Vysvetlete za jakych okolnost a jak se od tohoto rozdělení odlisuje. K čemu se používá?

Motivací pro jeho zavedení je zavedení je to, že v realitě nemáme vždy dostatečně velký počet vzorků a a nebo nevíme přesně standardní odchylku populace. To je se projevuje na t-rozdělení tak, že je kratší a má tlustší konce než normální rozdělení.

Čím více máme vzorků (stupňů volnosti), tím víc se podobá normálnímu rozdělení. Pro dostatečně velký počet vzorků jsou praktický identické.

Používá se pro zjištění zda přijmout nebo zamítnout nulovou hypotézu.

9) Definujte věrohodnostní funkci (likelihood). K cemu se pouziva metoda maximální věrohodnosti? Pojmenujte alespoň dvě metody, které se k maximalizaci verohodnosti používají. Vysvetlete, proc se často používá logaritmus verohodnosti. jak se dá věrohodnosti využít při testování statistických hypotéz?

Věrohodnostní funkce popisuje vztah parametrů rozdělení a dat. Pro náhodnou veličinu, která má hustotu f a parametr q je věrohodnostní fce L(q|x) = f_q(x).

Metoda maximální věrohodnosti se používá k nalezení modelu, který nejvěrohodněji vysvětluje pozorovaná data. Protože se na pravé straně výpočtu objeví součin (jeden člen pro každý datový bod), celá rovnice se obvykle zlogaritmuje. Tím se součin převede na příjemnější součet, a v případě normálního rozdělení se hezky upraví exponenty.

Použití: lineární regrese, polynomiální regrese

10) Formalne definujte multivariátní normální rozdělení. Kolik parametru budeme obecne potřebovat pro popis tohoto rozdělení v d dimenzích? Lze pocet techto parametru nejak omezit? Jake dusledky toto omezení může mít? \

tak, to rozdělený je popsaný vektorem středních hodnot (rozměru d pro d-rozměrný rozdělení), a kovarianční maticí velikosti dxd

pro vektor středních hodnot potřebuješ znát d parametrů (střední hodnoty jednotlivejch dimenzí)

pro kovarianční matici ale nepotřebuješ všech dxd parametrů, protože je symetrická

stačí (napůl tip, radši si to někde ověř) d(d+1)/2 parametrů

takže dohromady by to mělo bejt d + d(d+1)/2 parametrů k popisu

počet parametrů lze omezit, jinak by se neptali :D

to co sem popsal tady je nejobecnější formulace, jinak se uvažujou 2 způsoby zjednodušení

první je že tu kovarianční matici uvažuješ jenom diagonální, tzn necháš si z ní rozptyly na diagonále, ale nestaráš se o kovariance mimo diagonálu

pak potřebuješ pro popis jenom d (za střední hodnotu) + d (za diagonálu kovariance) parametrů

důsledek je že nejsi schopnej modelovat korelace mezi jednotlivejma složkma toho náhodnýho vektoru

ale plus je úspora paměti (stačí uložit diagonálu), a že neodhaduješ tolik parametrů (už jenom lineární počet místo kvadrátu), takže se ten gaussián naučíš z míň dat

druhý zjednodušení je ještě brutálnější, uvažuješ kovarianční matici tvar s*I, kde s by měla bejt sigma na druhou, ale nechce se mi to psát, a I je d-rozměrná jednotková matice

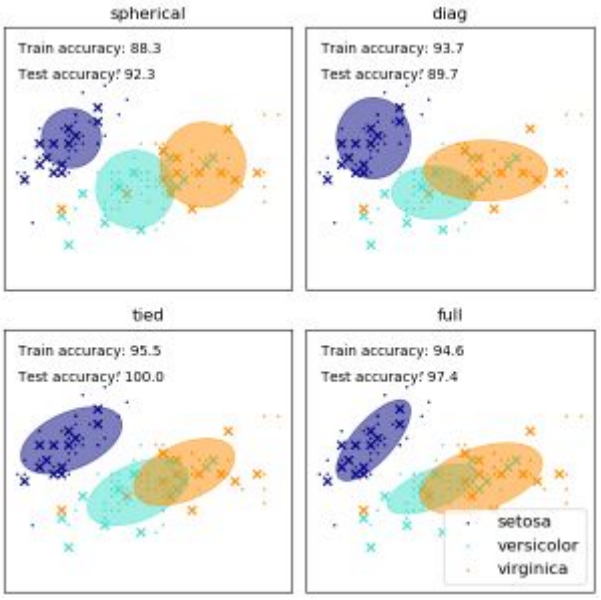
tím říkáš že nejenom že sereš na kovariance, ale dokonce i sereš na různý rozptyly těch souřadnic, a modeluješ jedinej společnej rozptyl (to číslo s) všech souřadnic dohromady

a pro popis pak stačí jenom d (za střední hodnotu) + 1 (za to číslo s) parametrů

---

ilustrace:

https://scikit-learn.org/stable/_images/sphx_glr_plot_gmm_covariances_001.png



na ten graf "tied" kašli, to je mimo to co tu řešíme

tohle je pohled zeshora na 2D gaussián, ty kruhy/elipsy jsou místa s nejvyšší hustotou pravděpodobnosti

respektive 3 gaussiány na každym obrázku, ale to je fuk

řekněme ten tmavě modrej

gaussovský data ležej v ellipsoidu (to se dá ukázat s nějakou analytickou geometrií, nechme), jehož speciálním případem je koule

ten graf "spherical" je ten případ s maticí typu s*I, kde uvažuješ stejný rozptyly všech souřadnic -- proto je to kruh, rozptyl ve obou souřadnicích (nahoru-dolu, doleva-doprava) je stejnej

graf "diag" je to předchozí zjednodušení -- diagonální matice, ale diagonální prvky jsou různy; přidals tam možnost aby každej směr měl vlastní rozptyl (tady 2 směry), ale pořád žádný kovariance -- proto je to elipsa, ačkoli je to vidět jen na tom žlutym Gaussovi

graf "full" je ten nejobecnější případ plný matice, umožníš ještě kovariance -- interakce mezi těma souřadnicema, takže se ta elipsa může natáčet

tim že se natočí mezi dvě osy tak spolu ty dva směry korelujou

zdůrazňuju že na každym grafu jsou 3 gaussy, jedno vícerozměrný rozdělení === jeden kruh/elipsa

11) Definujte alespoň dvě často používaná rozdělení diskrétní náhodné veličiny. Pojmenujte jejich parametry. Na příkladech naznačte, kdy se tato rozdělení dají využít.

Bernoulli rozdělení - hodnota 1 má pravděpodobnost p, hodnota 0 má pravděpodobnost 1-p
Binomické rozdělení - n - počet opakování

12) Vysvětlete rozdíl mezi parametrickým a neparametrickým statistickym testem. Pojmenujte základní vyhody a nevyhody obou přístupů. Jmenujte alespoň jeden parametricky a jeden neparametrický test.

Parametrický test předpokládá nějaký vzhled a parametry rozdělení populace ze kterého jsme vzali vzorek, typicky normální rozdělení. Neparametrický nic nepředpokládá o rozdělení populace.

Výhodou parametrického testu je, že má větší sílu, ale pokud se naše předpoklady rozdělení liší od pravdy moc, mohou jeho výsledky vést na špatné závěry. Neparametrický test nemá takovou sílu a jejich výsledky je těžší interpretovat, jelikož využívají rankování dat a né data samotná. Sílou testu se bere jestli nám test řekne, že dvě proměnné mají nějaký vztah, když ho opravdu mají. Test se slabou silou nám nic říct nemusí.

Příklad: parametrický -> two sample t-test, neparametrický -> Wilcoxon rank-sum test

13) K nasledujici statisticke uloze priradte vhodny test. Bylo testováno 11 automobilu urcite znacky.
Ověřte, zda se jejich prave a leve predni pneumatiky ojizdej srovnatelne. (Predpokladejte, ze ojet pneumatik [mm] ma normalni rozdeleni. Z nasledujicich testu vyberte ten nejlepší, svou volbu vysvetlete (dvouvýběrový t-test, Friedmanuv test, jednovýběrový t-test, jednovýběrový Wilcoxonuv test, test o parametru alternativního rozdělení, test o rozptylu normálního rozdělení, parovy t-test). Co by se stalo pokud byste pouzili druhy nejlepší test z daného seznamu?

tady máš jakoby 2 populace - vzorek levejch pneumatik, a vzorek pravejch pneumatik, takže potřebuješ dvouvýběrovej test

a nebo párovej

párovej se tady hodí víc

obecně párovej test použiješ pokud na 1 individuálovi měříš 2 věci, a zajímá tě jestli se lišej