Statistická analýza dat

Jméno: _____

Podpis: _____

Cvičení	
Zkouška (písemná + ústní)	≥ 25
Celkem	≥ 50
Známka	

Pokyny k vypracování: doba řešení je 120min, jasně zodpovězte pokud možno všechny otázky ze zadání, pracujte s pojmy používanými v předmětu, můžete používat kalkulátory.

Statistické minimum. *(10 b)* Zodpovězte následující otázky:

(a) *(6 b)* Definujte věrohodnostní funkci (likelihood). K čemu se metoda maximální věrohodnosti používá? Jak se dá věrohodnosti využít při testování statistických hypotéz?

(b) *(4 b)* Vysvětlete rozdíl mezi bodovým a intervalovým statistickým odhadem parametrů.

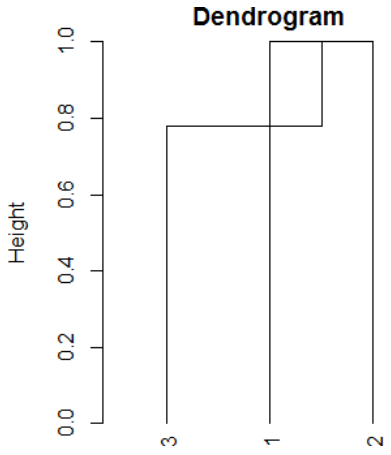
Hierarchické shlukování. *(10 b)* Níže diskutujte vlastnosti hierarchického shlukování.

(a) *(3 b)* Formálně definujte hierarchické aglomerativní shlukování (vstupy, výstup, algoritmus a jeho parametry).

(b) *(2 b)* Co to je taxonomie a dendrogram? Důsledně a detailně interpretujte jejich význam v hierarchickém shlukování (proč je taxonomie informativnější než prostý rozklad na shluky, kde lze v dendrogramu nalézt podobnost mezi objekty a jejich shluky, apod.).

(c) *(3 b)* Odhadněte složitost hierarchického aglomerativního shlukování. Důležitý je výsledek, ale i postup, kterým odhad odůvodníte.

(d) *(2 b)* Na dendrogramu níže je zobrazena tzv. inverze, tj. situace, kdy podobnost v průběhu aglomerativního shlukování neklesá ale roste. Může tato situace nastat? Za jakých podmínek?



Multivariátní regrese. *(10 b)* Jste strojní zámečník a snažíte se zjistit, jak souvisí chyba obrábění hřídele s nastavením parametrů obráběcího stroje. Sestavili jste multivariátní lineární model. Model vyjadřuje vztah mezi výrobní chybou (rozdíl mezi cílovým ideálním průměrem hřídele a skutečným průměrem hřídele, ProdError) a nastevením deseti různých spojitých parametrů stroje (P1-P10). Níže je uveden výstup, který jste obdrželi:

```
summary(lm(ProdError ~ P1+P2+P3+P4+P5+P6+P7+P8+P9+P10))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2638     0.2556   1.032    0.329
P1            0.2471     0.2564   0.964    0.360
P2           -0.6112     0.1979  -3.089    0.013 *
P3            0.2728     0.2341   1.165    0.274
P4            0.1093     0.2061   0.530    0.609
P5           -0.3165     0.4674  -0.677    0.515
P6           -0.4419     0.2660  -1.661    0.131
P7            0.1244     0.3152   0.395    0.702
P8            0.2452     0.2657   0.923    0.380
P9            0.1287     0.3093   0.416    0.687
P10           0.3544     0.2956   1.199    0.261
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7549 on 9 degrees of freedom
Multiple R-squared:  0.7237,        Adjusted R-squared:  0.4168
F-statistic: 2.358 on 10 and 9 DF,  p-value: 0.1062
```

(a) *(2 b)* Rozhodněte, zda je alespoň jeden z parametrů stroje (nezávisle proměnných) užitečný pro odhad výrobní chyby. Jinými slovy, formálně rozhodněte, zda lze zamítnout $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{10} = 0$. Správně odůvodněte.

(b) *(2 b)* Dá se z dat o modelu usoudit z kolika vzorků různých hřídelů byl model sestaven? Pokud ano, kolik jich bylo?

(c) *(2 b)* Jakým způsobem zohledníme tento počet při hodnocení užitečnosti modelu? Jak se změní chyba I druhu a chyba II druhu s rostoucím počtem vzorků pokud zachováme konstantní hladinu významnosti α?

(d) *(4 b)* Podrobně popište způsob, jak byste svůj model ověřili nad vzorky, které máte aktuálně k dispozici. Můžete tvořit další pomocné modely. Popište metodu validace, definujte chybovou funkci a určete, s čím budete vypočtenou chybu srovnávat.

Logistická regrese. *(10 b)* Máte data o loňské skupině studentů kurzu SAN. U každého studenta známe počet hodin, po který se připravoval na zkoušku (*hours*), jeho studijní průměr z posledního ročníku bakalářského studia (*avg*) a údaj o tom, zda přišel z programu OI či nikoli (*OI*). Cílovou binární veličinou Y je to, zda daný student získal známku A. Naučíte logistický model a získáte koeficienty $\beta_0 = -1$, $\beta_{hours} = 0.05$, $\beta_{avg} = -1$ a $\beta_{OI} = 1$.

(a) *(3 b)* Jaká je interpretace koeficientů v logistickém modelu (srovnejte s lineární regresí, kde koeficient vyjadřuje průměrnou změnu výstupu při jednotkové změně dané nezávisle proměnné a zafixování hodnot ostatních nezávisle proměnných)? Vysvětlete postupně pro β_0, β_{hours} a β_{OI}.

(b) *(2 b)* Vypočítejte, jak bude podle vašeho modelu klasifikován student, který přišel z OI, připravoval se 20h a jeho *avg* bylo 2?

(c) *(2 b)* Jak dlouho by se výše uvedený student musel připravovat na zkoušku, aby měl právě 50% pravděpodobnost, že dostane známku A?

(d) *(3 b)* Na uvedeném modelu/příkladu ilustrujte pojem matoucí proměnná (confounding variable). Definujte pojem, ukažte jednoduchý vztah mezi proměnnými. Model můžete libovolně upravit.

Robustní statistika. *(10 b)* Odhadněte třemi různými metodami robustně polohu (location) ze vzorku {-1.84, 1.18, 0.0499, -0.751, -0.00707, -2.05, -1.47, -0.0520, -0.991, -0.945}.

(a) *(2 b)* Metoda 1:

(b) *(2 b)* Metoda 2:

(c) *(2 b)* Metoda 3:

(d) *(2 b)* Popište kritéria, jež jsou určující pro kvalitu robustního odhadu polohy.

(e) *(2 b)* Diskutujte výhody a nevýhody vámi zvolených metod podle kritérií popsaných v předchozím bodě.