Kurz Biostatistiky pro zaměstnance FNO

1. Představme si, že provádíme test na okultní krvácení ve stolici (FOB) u 2 030 osob ke zjištění chorobných změn v dolní části zažívacího traktu. Pak můžeme popsat možné stavy pomocí níže uvedené tabulky. Určete senzitivitu a specificitu testu, prediktivní hodnoty a přesnost testu.

	má rakovinu tlustého střeva	**nemá rakovinu tlustého střeva**	**celkem**
test pozitivní	20	180	**200**
test negativní	10	1 820	**1 830**
celkem	**30**	**2 000**	**2 030**

2. Předpokládejme, že prevalence choroby je 0,005. Máme k dispozici dva diagnostické testy (T_1, T_2). První test (T_1) má senzitivitu 0,95 a specificitu 0,90, druhý test (T_2) má senzitivitu 0,92 a specificitu 0,99.
   a) Určete prediktivní hodnoty a přesnost testu T_1.

   b) U osob pozitivně testovaných testem T_1 byl proveden test T_2. Určete pravděpodobnost, že osoba, která má pozitivní i test T_2, skutečně nemoc má.

c) Jak se změní výsledky testování, změníme-li pořadí testů? Určete prediktivní hodnoty a přesnost testu T_2 a poté určete pravděpodobnost, že osoba, která má pozitivní test T_2 a poté i test T_1, skutečně nemoc má. V čem je rozdíl mezi jednotlivými přístupy?

3. Vylučovatelství skupinově specifických substancí ABH je podmíněno dominantní alelou *Se*, nevylučovatelství je podmíněno recesivní alelou *se*. Jestliže rodiče jsou heterozygotní vylučovatelé (*Se*, *se*), jejich potomek může být nevylučovatel (*se*, *se*), homozygotní vylučovatel (*Se*, *Se*) nebo heterozygotní vylučovatel (*Se*, *se*). Počet alelol *Se* lze modelovat náhodnou veličinou s distribuční funkcí

$$F(x) = \begin{cases} 0{,}00, & x \in (-\infty; 0\rangle \\ 0{,}25, & x \in (0; 1\rangle \\ 0{,}75, & x \in (1; 2\rangle \\ 1{,}00, & x \in (2; \infty). \end{cases}$$

   a) Určete pravděpodobnostní funkci počtu alel *Se*.

   b) Určete střední hodnotu a modus počtu alel *Se*.

4. Modelujeme výšku chlapců ve věku 3,5 – 4 roky. Vysvětlete:
   a) 2% kvantil modelované náhodné veličiny je 93 cm.

   b) Střední hodnota modelované NV je 102 cm, směrodatná odchylka je 4,5 cm. V jakém rozpětí lze očekávat výšku chlapců ve věku 3,5 – 4 roky? Pro interpretaci využijte Čebyševovu nerovnost.

   c) Střední hodnota modelované NV je 102 cm, směrodatná odchylka je 4,5 cm. Posuďte variabilitu modelované NV. (Není příliš vysoká? Pro posouzení použijte variační koeficient.)

   d) Víme, že pro distribuční funkci modelované NV platí: $F(111) = 0{,}98$. Co jsme se dozvěděli?

5. Rentgenové vyšetření pacienta trvá 10 minut. V čekárně v současné chvíli není žádný pacient, 1 pacient je ve vyšetřovně. Vypočtěte pravděpodobnost, že pacient, který právě přišel do čekárny, bude na vyšetření čekat déle než 7 minut.

6. Doba přežití (měsíce) pacienta má Weibullovo rozdělení s lineárně rostoucí rizikovou funkcí (tj. parametrem tvaru 2) a parametrem měřítka 10.
   a) V jakém rozmezí očekáváte dobu přežití pacientů? (Posuďte na základě grafu hustoty pravděpodobnosti.)

   b) S jakou pravděpodobností bude doba přežití pacienta delší než 1 rok?

   c) Jakou dobu přežije alespoň polovina pacientů?

   d) Jaká je hodnota rizikové funkce v 10 měsících? Jaká je pravděpodobnost, že pacient, který přežil 10 měsíců, zemře v následujících 14 dnech?

7. Nechť náhodná veličina modelující IQ (inteligenční kvocient) evropské populace má normální rozdělení se střední hodnotou 100 bodů a směrodatnou odchylkou 15 bodů.
   a) V jakém rozmezí očekáváte IQ evropské populace? (Posuďte na základě grafu hustoty pravděpodobnosti.)

   b) Kolik procent Evropanů má IQ v rozmezí 85-115 bodů?

   c) Kolik procent Evropanů má IQ vyšší než 115 bodů?

   d) Jakou hodnotu IQ překračuje maximálně 5% evropské populace?

8. V datovém souboru *tlak.xlsx* jsou uvedeny hodnoty diastolického tlaku pacientů s diabetem a pacientů bez diagnostikovaného diabetu. Data analyzujte, graficky prezentujte a doplňte následující tabulky a text.

Explorační analýza

Diastolický tlak krve (mm Hg)				po odstranění odlehlých pozorování	
	Diabetes_ano	Diabetes_ne		Diabetes_ano	Diabetes_ne
počet pacientů					
Míry polohy					
minimum					
dolní kvartil					
medián					
průměr					
horní kvartil					
maximum					
Míry variability					
směrodatná odchylka					
variační koeficient (%)					
Míry šikmosti a špičatosti					
šikmost					
špičatost					

Identifikace odlehlých pozorování		
Vnitřní hradby		
dolní mez		
horní mez		

Zde vložte vhodnou vizualizaci datového souboru:

Pacienti s diabetem

Byly analyzovány záznamy o diastolickém tlaku ……. pacientů, u nichž byl diagnostikován diabetes. Hodnoty diastolického tlaku se pohybovaly v rozmezí …… až …… mm Hg. Hodnoty diastolického tlaku ležící mimo interval …… až …… mm Hg byly identifikovány jako odlehlá pozorování a příslušní pacienti byli z dalšího zpracování vyřazeni. Níže uvedené výsledky pocházejí z analýzy datového souboru o rozsahu …… pacientů.

Průměrná hodnota diastolického tlaku byla …… mm Hg, směrodatná odchylka …… mm Hg. Polovině pacientů byl zjištěn diastolický tlak nižší než …… mm Hg. (Podrobněji: U čtvrtiny pacientů s diabetem byl zjištěn diastolický tlak nižší než …… mm Hg, u čtvrtiny pacientů diastolický tlak vyšší než …… mm Hg.) Vzhledem k hodnotě variačního koeficientu (…….%) lze / nelze analyzovaný soubor považovat za homogenní.

Obdobně lze popsat výsledky analýzy diastolického tlaku pacientů, u nichž nebyl diagnostikován diabetes.

Ověření normality

Na základě grafického zobrazení (viz histogram) a výběrové šikmosti a špičatosti (výběrová šikmost i špičatost leží / neleží v intervalu $\langle -2; 2 \rangle$) lze / nelze předpokládat, že diastolický tlak pacientů s diabetem má normální rozdělení. Na hladině významnosti 0,05 zamítáme / nezamítáme předpoklad normality diastolického tlaku pacientů s diabetem ($p - hodnota =$ ……, …… test). Dle pravidla 3σ / Čebyševovy nerovnosti lze tedy očekávat, že 95% / více než 75% pacientů s diabetem bude mít hodnotu diastolického tlaku v rozmezí …… až …… mm Hg.

Méně podrobněji pro pacienty bez diagnostikovaného diabetu: Na hladině významnosti 0,05 zamítáme / nezamítáme předpoklad normality diastolického tlaku pacientů s nediagnostikovaným diabetem ($p - hodnota =$ ……, …… test).

Zobecnění výsledků výběrového šetření na celou populaci pacientů s diabetem (statistická indukce)

Vzhledem k tomu, že předpoklad normality lze / nelze považovat za splněný, můžeme / nemůžeme určit intervalový odhad střední hodnoty diastolického tlaku obou skupin pacientů. Střední hodnotu diastolického tlaku pacientů s diabetem lze se spolehlivostí 95% očekávat v rozmezí …… až …… mm Hg. Střední hodnotu diastolického tlaku pacientů s nediagnostikovaným diabetem lze se spolehlivostí 95% očekávat v rozmezí …… až …… mm Hg.

Hodnoty diastolického tlaku pacientů s diabetem a pacientů, u nichž diabetes nebyl diagnostikován, lze srovnat na výše uvedeném vícenásobném grafu (*doufám, že jste jej použili pro prezentaci analyzovaných dat* ☺). Mezi průměrnými hodnotami / mediány diastolického tlaku pacientů s diabetem a pacientů bez diagnostikovaného diabetu byl pozorován rozdíl tlaků …… mm Hg. S 95% spolehlivostí lze střední hodnotu / medián diastolických tlaku pacientů s diabetem očekávat o …… až …… mm Hg vyšší než u pacientů s nediagnostikovaným diabetem. S 95% spolehlivostí lze tvrdit, že střední diastolický tlak /medián diastolického tlaku pacientů s diabetem je / není statisticky významně vyšší než střední diastolický tlak /medián diastolického tlaku pacientů s nediagnostikovaným diabetem ($p - hodnota =$ ……, …… test). Pozorovaný rozdíl tlaků je / není možno považovat za prakticky významný.

9. V letech 1965 až 1968 bylo v kohortové studii kardiovaskulárních onemocnění v rámci Honolulu Heart Program“ zahájeno sledování 8 006 mužů, z nichž 7 872 nemělo při zahájení studie v anamnéze mrtvici (apoplexii). Z tohoto počtu bylo 3 435 kuřáků a 4 437 nekuřáků. Při jejich sledování po dobu 12 let dostalo mrtvici 171 mužů ve skupině kuřáků a 117 mužů ve skupině nekuřáků. (Zdroj: Malý, M., Zvárová, M., *Statistické metody v epidemiologii*, Praha, 2003, ISBN: 8 024 607 654)
   a) Zapište zjištěné výsledky do asociační tabulky.

   b) Na základě vizuálního posouzení vhodného grafu a $Cramerova\ V$ odhadněte vliv kouření na výskyt kardiovaskulárních onemocnění.

   c) Určete absolutní riziko vzniku kardiovaskulárních onemocnění u kuřáků a nekuřáků.

   d) Určete relativní riziko (včetně 95% intervalového odhadu) vzniku kardiovaskulárních onemocnění u kuřáků a nekuřáků. Vysvětlete praktický význam zjištěných výsledků.

e) Určete absolutní šance vzniku kardiovaskulárních onemocnění u kuřáků a nekuřáků (včetně 95% intervalových odhadů).

f) Určete relativní šanci (včetně 95% intervalového odhadu) vzniku kardiovaskulárních onemocnění u kuřáků a nekuřáků. Vysvětlete praktický význam zjištěných výsledků.

g) Rozhodněte na hladině významnosti 0,05 o závislosti výskytu kardiovaskulárních chorob na kouření. Použijte χ^2 test nezávislosti.

10. 122 pacientů, kteří podstoupili operaci srdce, bylo náhodně rozděleno do tří skupin.
    Skupina 1: Pacienti dostali 50 % oxidu dusného a 50 % kyslíkové směsi nepřetržitě po dobu 24 hodin.
    Skupina 2: Pacienti dostali 50 % oxidu dusného a 50 % kyslíkové směsi pouze během operace.
    Skupina 3: Pacienti nedostali žádný oxid dusný, ale dostali 35-50 % kyslíku po dobu 24 hodin.
    Data v souboru *kyselina_listova.xls* odpovídají koncentracím soli kyseliny listové (pmol/l) v červených krvinkách ve všech třech skupinách po uplynutí 24 hodin ventilace. Ověřte, zda pozorované rozdíly mezi koncentracemi soli kyseliny listové jsou statisticky významné, tj. zda existuje vliv složení směsi na sledovaný parametr. (Nezapomeňte na explorační analýzu, identifikaci odlehlých pozorování, ověření předpokladů testu a případnou post hoc analýzu).

11. U výběru 24 jedinců ve věku 21 až 70 let, náhodně vybraných ze stejné etnické skupiny, byl po dobu 2 týdnů denně vždy v 8 hodin ráno měřen systolický tlak (*vek_tlak.xlsx*). Nalezněte lineární regresní model závislosti krevního tlaku y (mm) na stáří jedince x (rok).
    a) Dle korelačního pole posuďte, zda je vhodné použít model $Y = \beta_0 + \beta_1 x$. (proložení naměřených hodnot přímkou)

    b) Jaké je rozdělení věku a tlaku sledovaných pacientů? (průměr, medián, variační koeficient, normalita)

    c) Proveďte korelační analýzu.

d) Nalezněte lineární regresní model, v případě potřeby model optimalizujte.

e) Na základě indexu determinace posuďte kvalitu modelu.

f) Verifikujte model na základě analýzy reziduí.

g) Vykreslete graf závislosti tlaku na věku (včetně regresní přímky).

h) Na základě nalezeného modelu určete, o kolik mm se zvýší krevní tlak jedince každým rokem.

i) Na základě nalezeného modelu určete, jaký bude průměrný krevní tlak jedinců ve věku 65 let. (Určete bodový i intervalový odhad.)

j) Na základě nalezeného modelu určete, jaký bude krevní tlak jedince ve věku 65 let. (Určete bodový i intervalový odhad.)

12. Mějme celkem 202 pacientů, u kterých zkoumáme, zda daná nemoc propukne nebo nikoliv. Na základě dlouhodobých pozorování víme, že vznik nemoci by mělo ovlivňovat 8 faktorů (4 v osobní a 4 v rodinné anamnéze pacienta). Sestavili jsme logistický regresní model a program RkWard nám vrátil níže uvedené výsledky.

```
Call:
glm(formula = nemoc ~ RA.ASTMA + RA.RYMA + RA.EKZEM + RA.OSTATN. +
    OA.ASTMA + OA.RYMA + OA.EKZEM + OA.OSTATN., family = binomial("logit"),
    data = kuze.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5902  -0.7881  -0.6782   1.1151   1.9331

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8116     1.5550  -0.522  0.60169
RA.ASTMA      0.1994     0.3755   0.531  0.59550
RA.RYMA       0.1877     0.3617   0.519  0.60376
RA.EKZEM      0.1550     0.3595   0.431  0.66643
RA.OSTATN.   -0.3481     0.4810  -0.724  0.46929
OA.ASTMA      0.4146     0.7062   0.587  0.55717
OA.RYMA       1.3133     0.4181   3.141  0.00168 **
OA.EKZEM     -0.5411     1.5368  -0.352  0.72478
OA.OSTATN.    0.7842     0.3937   1.992  0.04641 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 250.72  on 201  degrees of freedom
Residual deviance: 233.05  on 193  degrees of freedom
AIC: 251.05
```

a) Určete, které proměnné jsou statisticky významné.

b) Napište předpis funkce modelu ve tvaru:

$$ln\left(\frac{R}{1-R}\right) = \cdots$$

c) Pro ověření kvality modelu byla využita následující kontingenční tabulka. Vypočítejte senzitivitu a specificitu modelu. Co tyto informace znamenají?

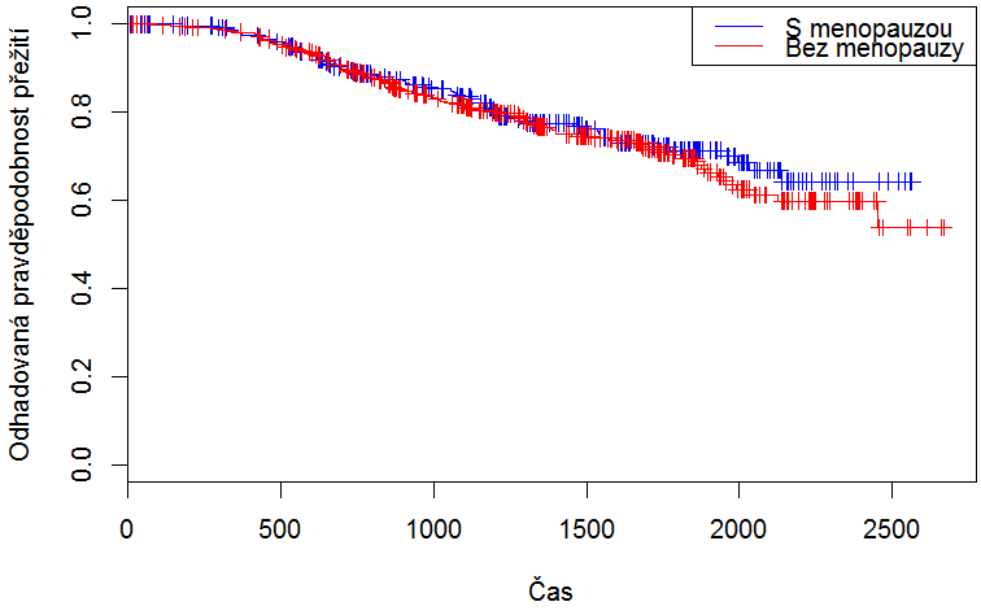
Predikce	Skutečnost	
	Y=1	Y=0
Y=1	19	7
Y=0	44	132

13. V rozsáhlé studii German Breast Cancer byl sledován čas přežití pacientek s rakovinou prsu v závislosti na tom, zda pacientka prodělala nebo neprodělala menopauzu.
    a) Dopište nulovou a alternativní hypotézu pro posouzení vlivu menopauzy na čas přežití pacientek.
    b) Zjistěte vizuálním posouzením grafu křivek přežití a pomocí výsledku log-rank testu, zda existuje statisticky významný rozdíl v čase přežití u žen s a bez prodělané menopauzy.
    c) Co jste výsledky studie zjistili? (Dopište závěr.)

H_0:

H_A:

log-rank test: p-hodnota = 0,484



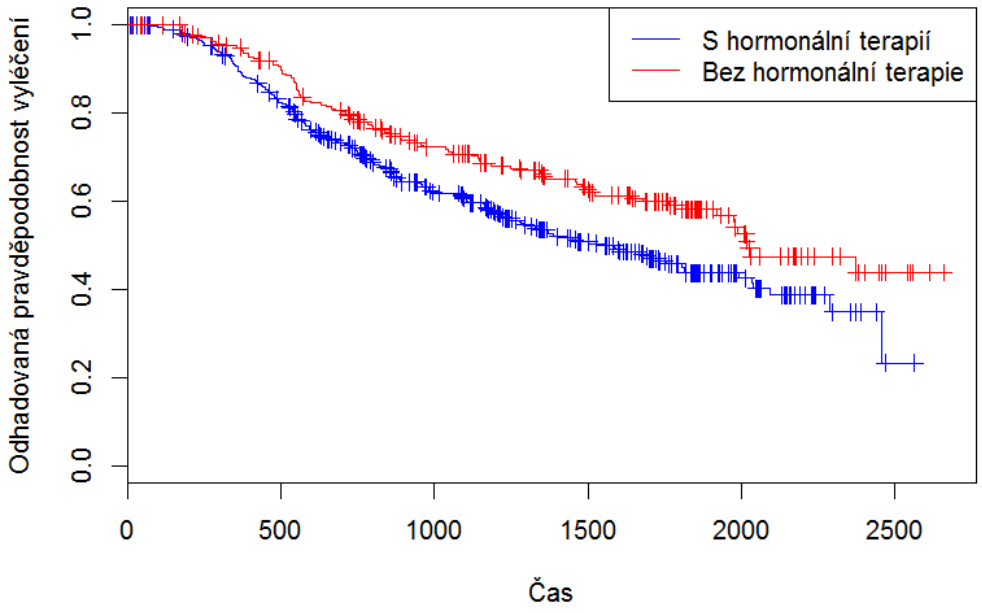
Závěr:

14. V rozsáhlé studii German Breast Cancer byl sledován čas do vyléčení pacientek s rakovinou prsu v závislosti na tom, zda byla pacientkám po dobu léčby podávána hormonální terapie.
    a) Dopište nulovou a alternativní hypotézu pro posouzení vlivu hormonální terapie na čas do vyléčení pacientek.
    b) Zjistěte vizuálním posouzením grafu křivek léčení a pomocí výsledku log-rank testu, zda existuje statisticky významný rozdíl v čase vyléčení u žen s a bez podávané hormonální terapie.
    c) Co jste výsledky studie zjistili? (Dopište závěr.)

H_0:

H_A:

log-rank test: p-hodnota = 0,003



Závěr: