Úlohy ke cvičení MAI010

1. Normální rozdělení a rozdělení od normálního odvozené

1. **(lognormální rozdělení)** Nechť X je náhodná veličina s normálním rozdělením $N(\mu, \sigma^2)$. Spočtěte hustotu a střední hodnotu náhodné veličiny $Y = \mathrm{e}^X$.
2. S využitím toho, že střední hodnota F-rozdělení s k a m stupni volnosti je $\frac{m}{m-2}$ (pro $m > 2$), určete rozptyl náhodné veličiny se Studentovým t-rozdělením o n stupních volnosti ($n > 2$).
3. Označme $U = \frac{X}{\sqrt{Y}}$, kde X, Y jsou nezávislé náhodné veličiny, X má normované normální rozdělení a Y má χ^2-rozdělení o n stupních volnosti. Určete $\operatorname{var} U$.
4. Nechť náhodná veličina X má normované normální rozdělení, náhodná veličina Y má χ^2-rozdělení s 20 stupni volnosti a obě veličiny jsou navzájem nezávislé. Pomocí tabulky kvantilů určete konstantu k tak, aby platilo
$$P(|X| \geq k\sqrt{Y}) = 0{,}05.$$

2. Náhodný výběr, výběrový průměr a rozptyl, bodové odhady parametrů

1. Nechť $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny s normovaným normálním rozdělením. Označme $\bar{X}_n$ jejich aritmetický průměr. Najděte číslo y takové, že
$$P(|\bar{X}_n| \geq y) = 0{,}05.$$
2. Uvažujte náhodný výběr z rozdělení s konečným rozptylem σ^2. Určete konstantu c tak, aby statistika $c\sum_{i=2}^{n}(X_i - X_{i-1})^2$ byla nestranným odhadem rozptylu σ^2.
3. Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení $N(\mu, \sigma^2)$. Spočtěte $\operatorname{var} S^2$, kde $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ značí výběrový rozptyl.
Návod: Jaký rozptyl má χ^2-rozdělení s $n-1$ stupni volnosti?
4. Nechť $X_1, \ldots, X_n$ je náhodný výběr z exponenciálního rozdělení s hustotou $f(x) = \lambda \mathrm{e}^{-\lambda x}$. Uvažujme odhad intenzity λ ve tvaru
$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i}.$$
Rozhodněte, zda je tento odhad nestranný a konzistentní.
5. Náhodné veličiny $X_1, \ldots, X_n$ jsou nezávislé stejně rozdělené s hustotou
$$f(x) = \begin{cases} 3\theta^3 \frac{1}{x^4} & \text{pro } x > \theta, \\ 0 & \text{pro } x < \theta, \end{cases}$$
kde $\theta > 0$ je parametr. Pokuste se odvodit vlastnosti (nestrannost a konzistenci) následujících dvou odhadů parametru θ a zjistit, který z nich je lepší (má menší střední kvadratickou chybu).
a) $\hat{\theta}_n = \frac{2}{3}\frac{1}{n}\sum_{i=1}^{n} X_i$,
b) $\tilde{\theta}_n = \frac{3n-1}{3n}\min_{1 \leq i \leq n} X_i$.

3. Metoda maximální věrohodnosti a metoda momentů

1. Nechť $X_1, \ldots, X_n$ je náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$. Určete odhad metodou maximální věrohodnosti a metodou momentů pro σ^2. Nejprve předpokládejte případ známého μ a poté neznámého μ. Rozhodněte o nestrannosti a konzistenci nalezených odhadů.

2. Najděte maximálně věrohodný odhad parametru p v binomickém rozdělení pro případ, kdy parametr n je známý.

3. Mějme geometrické rozdělení $P(X = k) = p(1-p)^k$, $k = 0, 1, 2, \ldots$. Najděte maximálně věrohodný a momentový odhad parametru p. Vyšetřete vlastnosti těchto odhadů.

4. Nechť $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny splňující pro každé $i = 1, \ldots, n$:

$$P(X_i = 0) = 1 - p - q,\ P(X_i = 1) = p,\ P(X_i = 2) = q,$$

kde $p, q \in (0, 1)$ jsou parametry takové, že $p+q < 1$. Určete odhad p a q metodou maximální věrohodnosti i metodou momentů. Liší se odvozené odhady?

5. Uvažujme náhodný výběr z rovnoměrného rozdělení na intervalu $(\theta - 1, \theta + 1)$. Odhadněte parametr θ metodou maximální věrohodnosti a metodou momentů.

4. Intervaly spolehlivosti

1. V roce 1961 byla zjištěna výška u 15 náhodně vybraných chlapců z populace všech desetiletých chlapců žijících v Československu. Určete 95% interval spolehlivosti pro střední výšku, pokud víte, že v roce 1951 byla zjištěna směrodatná odchylka $\sigma = 6{,}4$ cm a je známo, že variabilita výšek postavy se v různých generacích příliš nemění. V roce 1951 byla střední výška 136,1 cm, došlo za 10 let ke změně populačního průměru?

130	140	136	141	139	133	149	151
139	136	138	142	127	139	147	

2. Základní soubor tvoří deset miliónů lidí, z nichž každý hodlá hlasovat buď pro variantu A nebo B.
   a) Zkonstruujte intervalový odhad pro podíl $p = N_A/N$ lidí, kteří hodlají hlasovat pro variantu A, jestliže z 1 000 lidí, kteří byli vybráni na základě prostého náhodného výběru, 300 hodlá hlasovat pro variantu A. Koeficient spolehlivosti volte $1 - \alpha = 0{,}99$.
   b) Jaký rozsah má mít prostý náhodný výběr, aby intervalový odhad s koeficientem spolehlivosti $1-\alpha = 0{,}95$ pro podíl p měl šířku nejvýše 0,03?

3. Při kontrolních zkouškách 16 žárovek byl stanoven odhad střední hodnoty jejich životnosti $\bar{x} = 3\,000$ h a směrodatné odchylky $\hat{\sigma} = 20$ h. Za předpokladu, že životnost každé náhodné žárovky je náhodná veličina s normálním rozdělením určete:
   a) interval spolehlivosti pro střední hodnotu při $\alpha = 0{,}1$,
   b) s jakou pravděpodobností lze tvrdit, že absolutní hodnota chyby při určování $\bar{x}$ nepřekročí 10 hodin (vyjádřete pomocí distribuční funkce známého rozdělení).

4. U 100 náhodně vybraných výrobků z produkce určitého závodu byla zjištěna spotřeba materiálu na jeden výrobek. Z výběrových dat byla spočtena průměrná spotřeba $\bar{x} = 150$ a výběrový rozptyl $s_x^2 = 16$. Stanovte intervalový odhad pro průměrnou spotřebu se spolehlivostí 0,99!

5. Pro zácvik laboranta na určitém optickém přístroji je důležitým měřítkem variabilita při měření určitého objektu. Za předpokladu normality rozdělení sestrojte interval o spolehlivosti $1 - \alpha = 0{,}95$ pro rozptyl, jestliže pokus vedl k výsledkům

6,42 6,44 6,38 6,21 6,38 6,60 6,51.

5. Testování hypotéz

1. Dle výrobce má mít auto na 100 km průměrnou spotřebu 9 l. U 20 náhodně vybraných aut byla zjištěna následující spotřeba:

8.8 8.9 9.0 8.7 9.3 9.0 8.7 8.8 9.4 8.6
8.9 9.2 9.4 8.9 9.1 8.8 9.4 9.3 9.1 8.9

Potvrzují naměřené hodnoty tvrzení výrobce? Volte hladinu testu $\alpha = 0{,}05$.

2. Mějme náhodný výběr z normálního rozdělení $N(\mu, 1)$ a předepsané pravděpodobnosti chyb 1. a 2. druhu α a $1-\beta$. Testujme nulovou hypotézu $H_0 : \mu = \mu_0$ proti alternativní hypotéze $H_1 : \mu = \mu_1$, kde $\mu_0 < \mu_1$. Jaký je třeba zvolit rozsah výběru?

3. Pro porovnání dvou metod učení nazpaměť bylo mezi 18 pokusnými dvojicemi vybráno 9 párů se stejnou nebo velmi podobnou hodnotou IQ. Náhodně zvolená osoba z každého páru použila při učení metodu A, druhá osoba metodu B.

A :	90	86	72	65	44	52	46	38	43
B :	85	87	70	62	44	53	42	35	46

Testujte hypotézu, že obě metody jsou stejně dobré. Volte hladina testu $\alpha = 0{,}05$.

4. Vážením jsme získali údaje o přesném množství automaticky balených potravinářských výrobků určitého druhu náhodně vybraných před a po seřízení balicího automatu.

Před seřízením:	243,2	244,8	253,1	247,5	251,0	251,7	254,0
	252,5	252,8	250,1	247,3	250,9	253,2	252,7
Po seřízení:	250,4	250,2	251,1	249,3	249,9	250,2	251,1

Zjistěte, zda je možné na hladině $\alpha = 0{,}05$ prokázat, že:
a) před seřízením automatu střední hodnota překračuje 250 g,
b) před seřízením automatu směrodatná odchylka překračuje 1 g,
c) kolísavost množství se seřízením automatu snížila,
d) střední hodnota se seřízením změnila.
Doplňte předpoklady.

5. Byla sledována účinnost léku na snížení tlaku krve. Snížení tlaku nastalo u 140 z 225 pacientu. Rozhodněte, zda je léčba efektivní! Volte hladinu testu $\alpha = 0{,}01$.
Návod: Použijte centrální limitní větu. Označíme-li p pravděpodobnost snížení tlaku u jednotlivého pacienta, pak $p = 1/2$ znamená, že léčba není efektivní, $p > 1/2$ znamená, že léčba je efektivní.

6. Korelační koeficient

1. Nechť X je počet líců při třech hodech korunovou mincí, nechť Y je počet líců při čtyřech hodech pětikorunovou mincí. Označme celkový počet líců v těchto pokusech jako W. Určete korelační koeficient X a W.

2. Náhodná veličina X má rovnoměrné rozdělení
a) na intervalu $(0, 2)$,
b) na intervalu $(-1, 1)$.
Označme $Y = X^2$. V obou případech spočtěte kovarianci a korelační koeficient veličin X a Y.

3. V tabulce je uvedena míra nezaměstnanosti a míra inflace v roce 2003 v zemích EU (v procentech).

Země	Nezam.	Inflace	Země	Nezam.	Inflace
Belgie	8,1	1,5	Malta	8,2	2,5
Česká republika	7,8	−0,1	Německo	9,6	1,0
Dánsko	5,6	2,0	Nizozemsko	3,8	2,2
Estonsko	10,1	1,4	Polsko	19,2	0,7
Finsko	9,0	1,3	Portugalsko	6,3	3,3
Francie	9,4	2,2	Rakousko	4,1	1,3
Irsko	8,6	4,0	Řecko	9,3	3,4
Itálie	8,6	2,8	Slovensko	17,1	8,5
Kypr	4,4	4,0	Slovinsko	6,5	5,7
Litva	12,7	−1,1	Spojené království	5,0	1,4
Lotyšsko	10,5	2,9	Španělsko	11,3	3,1
Lucembursko	3,7	2,5	Švédsko	5,6	2,3
Maďarsko	5,8	4,7			

Spočtěte výběrový korelační koeficient.

4. U 20 rodin byl sledován jejich měsíční příjem a měsíční výdaje za stravu. Označíme-li X_i příjem i-té rodiny a Y_i její výdaje (v tisících), tak bylo zjištěno, že $\sum X_i = 546$, $\sum Y_i = 95{,}58$, $\sum X_i Y_i = 3331{,}98$, $\sum X_i^2 = 18830{,}88$, $\sum Y_i^2 = 605{,}5812$. Odhadněte korelační koeficient a na hladině 5% testujte hypotézu, že je roven nule.

7. Regresní přímka, lineární model

1. Prodej daného výrobku byl následující (v tis.)

1998	30
1999	34
2000	39
2001	47
2002	53
2003	57
2004	65

Na základě regresní přímky odhadněte prodej v letech 2005 a 2006.

2. V roce 2003 byl v jednotlivých krajích České republiky zjištěn následující počet lidí starších 15 let, kteří mají doma přístup k internetu (údaje jsou v tisících):

Kraj	Obyvatel	Internet	Kraj	Obyvatel	Internet
Hl. m. Praha	1 018	354	Královehradecký	464	84
Středočeský	962	231	Pardubický	426	92
Jihočeský	528	80	Vysočina	434	89
Plzeňský	468	73	Jihomoravský	954	227
Karlovarský	256	57	Olomoucký	538	75
Ústecký	689	87	Zlínský	502	98
Liberecký	359	61	Moravskoslezský	1 061	189

Zkoumejte závislost počtu obyvatel starších 15 let s internetem na počtu všech obyvatel starších 15 let. Použijte model jednoduché lineární regrese a odhadněte jeho parametry. Spočtěte reziduální součet čtverců a koeficient determinace.

3. Sledujme závislost spotřeby benzínu na výkonu motoru u různých značek aut. Nechť Y_i značí počet ujetých kilometrů na 1 litr pohonné hmoty a x_i je výkon v kW.

x_i	77	137,9	115,5	76,3	99,4	133,7	80,5	178,5	108,5	140	49	56,7
Y_i	14,4	9,2	8,0	11,6	8,4	11,2	12,0	10,0	12,4	10,8	17,2	13,2

Nejprve uvažujte lineární model ve tvaru $Y_i = \beta_0 + \beta_1 x_i + \sigma\varepsilon_i$ a najděte odhady parametrů.

Ze zkušenosti se dá očekávat, že se silnějším autem roste spotřeba (převracená hodnota Y_i). Zkuste nyní zvolit model $Y_i = \beta_0 + \frac{\beta_1}{x_i} + \sigma\varepsilon_i$. Opět odhadněte parametry. Lze na 5% hladině prokázat závislost Y_i na x_i? V kterém případě je reziduální součet čtverců menší (koeficient determinace větší)?

4. Při běžeckých testech byl kromě dosaženého času (v sekundách) zaznamenán i věk a hmotnost daného člověka.

Čas	Věk	Hmotnost
481	37	84
297	24	65
359	31	73
313	26	76
438	42	69
419	19	82
386	33	74
463	49	81

Vyšetřete závislost dosaženého času na dvojici veličin hmotnost a věk.

8. Multinomické rozdělení a testy dobré shody

1. Ve městě je k hospod, n lidí si náhodně vybírá nezávisle na sobě jednu hospodu, kterou navštíví. Pravděpodobnost, že si vyberou j-tou hospodu je pro všechny lidi stejná a rovna p_j, přitom $\sum_{j=1}^{k} p_j = 1$. Určete sdružené rozdělení počtu lidí v jednotlivých hospodách a marginální rozdělení počtu lidí v jedné zvolené hospodě (například v první).

2. Sto vybraných lidí určilo svou oblíbenou číslici.

0	1	2	3	4	5	6	7	8	9
3	14	9	19	10	7	8	17	5	8

Posuďte, zda je některá číslice preferována.

3. Při 600 hodech hrací kostkou byly zjištěny následující četnosti jednotlivých stran: 103, 99, 91, 108, 119, 80. Lze na 5% hladině považovat tuto hrací kostku za symetrickou?

4. Uvažujme následující dvě realizace náhodných výběrů:

−8,77	−6,52	−7,06	−4,36	−5,74	−7,71	−3,91	2,68	2,42	2,22
−8,91	9,34	−6,95	−2,65	9,32	5,88	−0,85	2,29	−4,38	4,95
0,77	0,64	0,37	−2,67	4,43	−3,93	0,32	−2,66	0,63	−0,84
−1,77	0,73	−1,15	0,48	1,04	−2,87	−1,24	−0,81	0,66	−2,56

Nejedná se o reálná ale o nasimulovaná data z dvou různých rozdělení s nulovou střední hodnotou.

Lze pomocí Kolmogorovova-Smirnovova testu na 5% hladině zamítnout nulovou hypotézu tvrdící, že výběry pocházejí ze stejných rozdělení?

Všimněte si, že dvouvýběrový t-test v tomto případě vede k chybnému závěru, není schopen rozeznat odlišnost výběrů (ve skutečnosti první skupina dat nepochází z normálního rozdělení, a proto nejsou nesplněny předpoklady dvouvýběrového t-testu).