STA2 – zk (60 min) ....

jméno a příjmení

Varianta S2

Teoretické otázky:

1) Když zamítnete hypotézu $\mu = \mu_0$ při jednostranném testu, zamítnete ji při stejné hladině významnosti také při oboustranném testu? Proč?
2) Jaký je vztah mezi binomickým, hypergeometrickým a Poissonovým rozdělením?
3) K čemu slouží variační koeficient? Jaká je jeho přednost?
4) Co měří dílčí korelační koeficient a co vícenásobný korelační koeficient?
5) Co si představujete pod pojmem „dekompozice časové řady“?

Příklady:

1) Určitá linka městské autobusové dopravy jezdí v době dopravní špičky v centru města průměrnou rychlostí 8 km/h. Vedení dopravního podniku uvažovalo o tom, zda by změna trasy dané linky vedla ke zvýšení průměrné rychlosti. Nová trasa proto byla projeta v deseti náhodně vybraných dnech v době dopravní špičky a byly zjištěny tyto průměrné rychlosti (v km/h):

   8,5 9,5 7,8 8,2 9,0 7,5 8,2 7,8 9,0 8,5

   Rozhodněte na 5 %-ní hladině významnosti, zda změna trasy vede ke zvýšení průměrné rychlosti, přičemž předpokládejte normální rozdělení rychlosti autobusu v centru v době dopravní špičky. 2 BODY

2) Při silniční kontrole byly u náhodně vybraných 200 vozidel zjišťovány závady na osvětlení a pneumatikách. Posuďte na hladině významnosti 1 %, zda existuje závislost mezi závadami na pneumatikách a osvětlení, případně změřte sílu a směr závislosti vhodnou charakteristikou.

   2 BODY

Závady na pneumatikách	**Závady na osvětlení**	
	ANO	NE
ANO	32	12
NE	16	140

3) V následující tabulce máte údaje o vývozu osobních automobilů do určité země v letech 1990 – 1997. Posuďte, jakou trendovou funkci je vhodné použít k popisu trendu tohoto ukazatele (omezte svůj výběr na přímku, parabolu a exponenciálu). Své rozhodnutí zdůvodněte. Odhadněte vývoz automobilů (v tis. Kč) na rok 2000 na základě funkce, kterou jste vybrali jako nejvýstižnější. Uvažujte α = 0,05. 3 BODY

Rok	1990	1991	1992	1993	1994	1995	1996	1997
Vývoz v tis. Kč	60	58	55	53	51	48	46	44

Celkový počet bodů: 12
Minimální počet bodů pro úspěšné složení zkoušky: 7,25
Minimální počet bodů z teorie: 1,75
Minimální počet bodů z příkladů: 2,5
Každá teoretická otázka je za 1 bod.

Teoretické otázky – řešení:

1) Tato hypotéza by zamítnuta být nemusela, a to z důvodu „nepřekrývání“ se kritického oboru pro jednostranný a oboustranný test. Např. když $H_1: \mu \neq \mu_0$, potom může být kritický obor při zvolené hladině významnosti 0,05 následující: $W \equiv \left\{U; U \leq u_{\frac{\alpha}{2}} \cup U \geq u_{1-\frac{\alpha}{2}}\right\}$ a konkrétně $W \equiv \{U; U \leq u_{0,025} \cup U \geq u_{0,975}\}$, tj. $W \equiv \{U; U \leq -1{,}96 \cup U \geq 1{,}96\}$.
U jednostranného testu, kdy je např. $H_1: \mu > \mu_0$, je kritický obor $W \equiv \{U; U \geq u_{1-\alpha}\}$, konkrétně tedy: $W \equiv \{U; U \geq u_{0,95}\}$, tj. $W \equiv \{U; U \geq 1{,}645\}$.
Pokud hodnota testového kritéria bude např. U = 1,75, pak u oboustranného testu nevedla k zamítnutí H_0, ale u jednostranného ano.

2) Všechno to jsou rozdělení nespojitých náhodných veličin. Hypergeometrické rozdělení můžeme za určitých podmínek aproximovat buď Binomickým, nebo Poissonovým rozdělením. Binomické rozdělení lze za určitých podmínek aproximovat Poissonovým rozdělením.

3) Slouží k měření relativní variability číselných proměnných. Jeho výhodou je, že jde o bezrozměrné číslo, můžeme tak porovnávat variabilitu souborů, kde hodnoty jsou v různých měrných jednotkách.

4) Dílčí korelační koeficient měří sílu lineární závislosti závisle proměnné y na proměnné např. x_1 za předpokladu, že všechny ostatní vysvětlující proměnné (uvedené za tečkou) jsou konstantní. Vícenásobný korelační koeficient měří těsnost lineární závislosti závisle proměnné y na všech vysvětlujících proměnných dohromady.

5) Dekompozicí časové řady rozumíme rozklad časové řady na složky, které charakterizují různé druhy pohybů v čase. Rozlišujeme složku trendovou, sezónní, cyklickou a náhodnou.

Příklady – řešení:

1) $H_0: \mu = 8$
$H_1: \mu > 8$

$$t = \frac{\bar{x} - \mu}{\frac{s'_x}{\sqrt{n}}}$$

$W \equiv \{t; t \geq t_{1-\alpha}(n-1)\}$
$W \equiv \{t; t \geq t_{0,95}(9)\}$
$W \equiv \{t; t \geq 1{,}833\}$

Z dat vypočítáme výběrové charakteristiky:
$\bar{x} = 8{,}4$
$s'_x = 0{,}629$

$$t = \frac{\bar{x} - \mu}{\frac{s'_x}{\sqrt{n}}} = \frac{8{,}4 - 8}{\frac{0{,}629}{\sqrt{10}}} = 2{,}011$$

$t \in W$, proto zamítáme H_0 a přijímáme H_1.

Na hladině významnosti 5 % jsme prokázali, že změna trasy autobusu MHD vede ke zvýšení průměrné rychlosti.

Poznámka: Lze řešit i ve statistickém programu. Rozhodnutí provádíme podle tzv. significance level, někde uváděné také jako P-Value. Zde je P- Value = 0,037591. Protože P-Value < α, zamítáme H_0 a přijímáme H_1.

2) Jde o kategoriální (nominální) znaky, proto je vhodnou metodou zkoumání závislostí mezi nimi test nezávislosti kategoriálních znaků. Zde obě proměnné nabývají pouze 2 obměn, jsou to tedy alternativní proměnné. Lze použít zjednodušenou verzi testu nezávislosti kategoriálních znaků.

H_0: Závady na pneumatikách a na osvětlení na sobě nezávisí
H_1: non H_0

$$G = n \cdot \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}$$

$$W \equiv \{G; G \geq \chi^2_{1-\alpha}[(r-1)(s-1)]\}$$
$$W \equiv \{G; G \geq \chi^2_{0,99}(1)\}$$
$$W \equiv \{G; G \geq 6{,}639\}$$

$$G = 200 \cdot \frac{(32 \cdot 140 - 16 \cdot 12)^2}{48 \cdot 152 \cdot 44 \cdot 156} = 73{,}43$$

$G \in W$, tj. zamítáme H_0, přijímáme H_1.

Na hladině významnosti 1 % jsme prokázali, že mezi závadami na pneumatikách a na osvětlení existuje závislost.

Pozn.: Pokud řešíme ve statistickém programu, posuzujeme výsledek testu podle tzv. P-Value. Zde je P-Value = 0,0000, což je méně než α, proto zamítáme H_0 a přijímáme H_1.

Těsnost závislosti změříme pomocí koeficientu asociace:
$$r_{AB} = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}}$$

$$r_{AB} = \frac{32 \cdot 140 - 16 \cdot 12}{\sqrt{48 \cdot 152 \cdot 44 \cdot 156}} = 0{,}606$$

Závislost mezi závadami je středně vysoká a přímá, tj. závady na pneumatikách se vyskytují častěji, když se zároveň vyskytnou závady na osvětlení.

3) Doporučeno řešit ve statistickém programu.

Přímka

Simple Regression - vyvoz vs. t

Dependent variable: vyvoz
Independent variable: t
Linear model: Y = a + b*X

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Intercept	62,3214	0,232829	267,671	0,0000
Slope	-2,32143	0,0461069	-50,3488	0,0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	226,339	1	226,339	2535,00	0,0000
Residual	0,535714	6	0,0892857		
Total (Corr.)	226,875	7			

Correlation Coefficient = -0,998819
R-squared = 99,7639 percent
R-squared (adjusted for d.f.) = 99,7245 percent

Rovnice: $\hat{T}_t = 62{,}3214 - 2{,}3214t$

Celkový F-test (posouzení vhodnosti přímky k vyrovnání daných hodnot):

$H_0: \alpha_0 = c, \alpha_1 = 0$ (přímka není vhodná k vyrovnání daných hodnot)
H_1: non H_0

P-Value = 0,0000 < α, tj. zamítáme H_0, přijímáme H_1.
Na hladině významnosti 5 % jsme prokázali, že přímka je vhodným modelem k vyrovnání daných hodnot.

Individuální t-testy o nulové hodnotě parametrů trendové funkce (posouzení vhodnosti jednotlivých parametrů):

$H_0: \alpha_0 = 0$ (parametr α_0 není pro funkci přínosný)
H_1: non H_0

P-Value = 0,0000 < α, tj. zamítáme H_0, přijímáme H_1.

Na hladině významnosti 5 % jsme prokázali, že parametr α_0 je statisticky významný.

$H_0: \alpha_1 = 0$ (parametr α_1 není pro funkci přínosný)
H_1: non H_0

P-Value = 0,0000 < α, tj. zamítáme H_0, přijímáme H_1.

Na hladině významnosti 5 % jsme prokázali, že parametr α_1 je statisticky významný.

Závěr: Vzhledem k tomu, že celkový F-test je významný a oba t-testy také, je možné přímku použít k vyrovnání uvedených hodnot.

Parabola

Polynomial Regression - vyvoz versus t

Dependent variable: vyvoz
Independent variable: t
Order of polynomial = 2

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	62,5893	0,433234	144,47	0,0000
t	-2,48214	0,220881	-11,2375	0,0001
t^2	0,0178571	0,0239579	0,745356	0,4896

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	226,393	2	113,196	1173,89	0,0000
Residual	0,482143	5	0,0964286		
Total (Corr.)	226,875	7			

R-squared = 99,7875 percent
R-squared (adjusted for d.f.) = 99,7025 percent

Rovnice: $\hat{T}_t = 62{,}5893 - 2{,}48214t + 0{,}0178571t^2$

Celkový F-test (posouzení vhodnosti paraboly k vyrovnání daných hodnot):

$H_0: \alpha_0 = c, \alpha_1 = \alpha_2 = 0$ (parabola není vhodná k vyrovnání daných hodnot)
H_1: non H_0

P-Value = 0,0000 < α, tj. zamítáme H_0, přijímáme H_1.
Na hladině významnosti 5 % jsme prokázali, že parabola je vhodným modelem k vyrovnání daných hodnot.

Individuální t-testy o nulové hodnotě parametrů trendové funkce (posouzení vhodnosti jednotlivých parametrů):

$H_0: \alpha_0 = 0$ (parametr α_0 není pro funkci přínosný)
H_1: non H_0

P-Value = 0,0000 < α, tj. zamítáme H_0, přijímáme H_1.

Na hladině významnosti 5 % jsme prokázali, že parametr α_0 je statisticky významný.

$H_0: \alpha_1 = 0$ (parametr α_1 není pro funkci přínosný)
H_1: non H_0

P-Value = 0,0001 < α, tj. zamítáme H_0, přijímáme H_1.

Na hladině významnosti 5 % jsme prokázali, že parametr α_1 je statisticky významný.

$H_0: \alpha_2 = 0$ (parametr α_2 není pro funkci přínosný)
H_1: non H_0

P-Value = 0,4896 > α, tj. nezamítáme H_0, nepřijímáme H_1.

Na hladině významnosti 5 % jsme neprokázali, že parametr α_2 je statisticky významný. To naznačuje, že počet parametrů je příliš velký a stačily by pravděpodobně dva.

Závěr: Vzhledem k tomu, že celkový F-test je významný a většina t-testů také, je možné parabolu použít k vyrovnání uvedených hodnot.

Exponenciála

Simple Regression - vyvoz vs. t

Dependent variable: vyvoz
Independent variable: t
Exponential model: Y = exp(a + b*X)

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Intercept	4,14587	0,00497813	832,817	0,0000
Slope	-0,0449638	0,000985816	-45,6107	0,0000

NOTE: intercept = ln(a)

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	0,084913	1	0,084913	2080,34	0,0000
Residual	0,000244902	6	0,000040817		
Total (Corr.)	0,0851579	7			

Correlation Coefficient = -0,998561
R-squared = 99,7124 percent
R-squared (adjusted for d.f.) = 99,6645 percent

Rovnice: $\hat{T}_t = e^{4{,}14587-0{,}0449638t}$

Celkový F-test (posouzení vhodnosti exponenciály k vyrovnání daných hodnot):

$H_0: \alpha_0 = c, \alpha_1 = 0$ (exponenciála není vhodná k vyrovnání daných hodnot)
H_1: non H_0

P-Value = 0,0000 < α, tj. zamítáme H_0, přijímáme H_1.
Na hladině významnosti 5 % jsme prokázali, že exponenciála je vhodným modelem k vyrovnání daných hodnot.

Individuální t-testy o nulové hodnotě parametrů trendové funkce (posouzení vhodnosti jednotlivých parametrů):

$H_0: \alpha_0 = 0$ (parametr α_0 není pro funkci přínosný)
H_1: non H_0

P-Value = 0,0000 < α, tj. zamítáme H_0, přijímáme H_1.

Na hladině významnosti 5 % jsme prokázali, že parametr α_0 je statisticky významný.

$H_0: \alpha_1 = 0$ (parametr α_1 není pro funkci přínosný)
H_1: non H_0

P-Value = 0,0000 < α, tj. zamítáme H_0, přijímáme H_1.

Na hladině významnosti 5 % jsme prokázali, že parametr α_1 je statisticky významný.

Závěr: Vzhledem k tomu, že celkový F-test je významný a oba t-testy také, je možné exponenciálu použít k vyrovnání uvedených hodnot.

Z výše uvedeného vyplývá, že všechny uvedené funkce jsou použitelné k vyrovnání daných hodnot. Otázkou je, která z těch funkcí je nejvýstižnější. K řešení použijeme vybraná matematicko-statistická kritéria:

Kritérium	**Přímka**	**Parabola**	**Exponenciála**
F	2535	1173,89	2080,335
s_R^2	0,08929	0,09643	0,00004
I_{adj}^2	0,997	0,997	0,997

Podle hodnoty testového kritéria F se jako nejlepší jeví trendová přímka.
Podle hodnoty reziduálního rozptylu bychom jako nejvýstižnější funkci vybrali trendovou exponenciálu. Podle hodnoty upraveného indexu determinace bychom se mohli přiklonit k jakékoli z těchto tří funkcí. Záleží na tom, které kritérium si vybereme. Tu funkci pak zvolíme. Zde bychom pravděpodobně vybrali přímku, protože má nejvyšší hodnotu F a je zároveň ze všech tří uvedených funkcí nejjednodušší.

Předpověď vývozu automobilů na rok 2000 na základě trendové přímky:
$\hat{T}_t = 62{,}3214 - 2{,}3214t$
$\hat{T}_{11} = 62{,}3214 - 2{,}3214 \cdot 11$
$\hat{T}_{11} = 36{,}786$

Za předpokladu, že se trend vývoje nezmění, lze v roce 2000 očekávat vývoz automobilů v hodnotě 36,786 tis. Kč.