MSP - Statistika a pravděpodobnost
2021 / 2022

Projekt

Vypracoval: Jan Lorenc (xloren15) Datum: 8. 12. 2021

K projektu

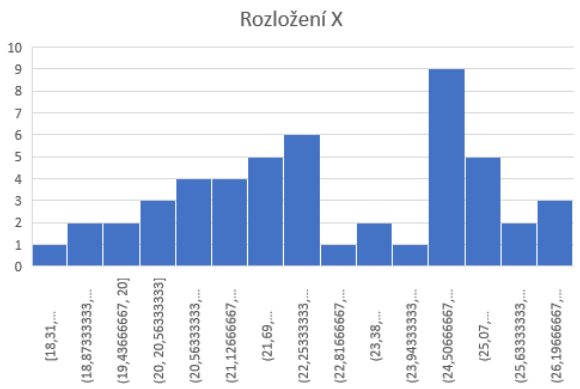
Výpočty všech tří úkolů se nacházejí v souboru xloren15_vypocty.xlsx. V tomto excelovském souboru se nachází list pro každý úkol. Všechny úkoly byly počítány ručně, a proto jsou postupy zřetelně vidět v daném excelu. V tomto protokolu se tedy zabývám již jen výsledky výpočtů. Testy ve všech úlohách pak byly prováděny na hladině významnosti $\alpha = 0{,}05$.

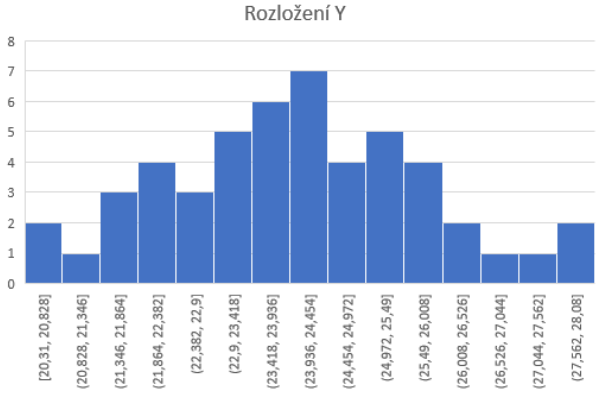
Úkol 1

V tomto úkolu jsou 2 výběry dat (X a Y), které musí být porovnány. Není zadán předpoklad normality, proto musí být test na normalitu proveden. Pokud se by se oba výběry ukázaly být z normálního rozložení, šel by použít dvouvýběrový studentův test, v opačném případě je nutné sáhnout po neparametrickém Mann Whitney U testu.

Test normality:

Z prvního pohledu na histogramy se zdá, že data X z normálního rozložení nevycházejí a Y ano. Tuto domněnku potvrdil Pearsonův test dobré shody, ve kterém jsem si data rozdělil do 15 intervalů (histogramových košů) a porovnával jsem četnosti v těchto intervalech s teoretickými četnostmi.





Test normality - Pearson									
	X	Y		X			Y		
n	50	50	koš	četnost n	teoretická četnost ñ	$(n-ñ)^2/ñ$	četnost n	teoretická četnost ñ	$(n-ñ)^2/ñ$
max	26,76	28,08	1	1	0,233059401	2,523811	2	0,233059401	13,39607
min	18,31	20,31	2	2	0,462112974	5,118005	1	0,462112974	0,626086
šířka	8,45	7,77	3	2	1,10134358	0,733271	3	1,10134358	3,27318
košů	15		4	3	2,241317006	0,256813	4	2,241317006	1,379977
rozsah koše	0,56333333	0,518	5	4	3,894929735	0,002834	3	3,894929735	0,205626
α	0,05		6	4	5,779893191	0,54811	5	5,779893191	0,105233
$\chi^2_{0,95}(49)$	66,33864886		7	5	7,324358641	0,737627	6	7,324358641	0,239465
$\bar{w}$	<0; 66,338>		8	6	7,925970944	0,468001	7	7,925970944	0,108179
t	92,622486	35,008188	9	1	7,324358641	5,460889	4	7,324358641	1,50885
normalita	ne	ano	10	2	5,779893191	2,471948	5	5,779893191	0,105233
	(t neleží v $\bar{w}$)	(t leží v $\bar{w}$)	11	1	3,894929735	2,151674	4	3,894929735	0,002834
			12	9	2,241317006	20,38078	2	2,241317006	0,025982
			13	5	1,10134358	13,80089	1	1,10134358	0,009325
			14	2	0,462112974	5,118005	1	0,462112974	0,626086
			15	3	0,233059401	32,84982	2	0,233059401	13,39607
			Σ	50	50	92,62249	50	50	35,00819

Poněvadž neplatí, že by oba výběry měly normální rozložení, je nutné použít Mann Whitney U test.

Mann Whitney U test

Test slouží k porovnání dvou nezávislých výběrů. Jeho výsledkem je testové kritérium U, které pro dostatečně velký výběr aproximuje normální rozložení (našich 50 vzorků splňuje) a tedy na něj lze použít například Z test pro určení strannosti. Jelikož medián výběru X je menší než medián výběru Y, tak jsem zvolil levostranný test, kterým testuji alternativní hypotézu, že X je menší než Y.

H_0: X - Y = 0 (X = Y)

H_A: X - Y < 0 (X < Y)

Výsledky testu jsou poté následující (postup je přiložením v excelu):

n_X	50
n_Y	50
$n_X * n_Y$	2500
$n_X + 1$	51
$n_Y + 1$	51
T_X	2197
T_Y	2853
U_X	1578
U_Y	922
U	**922**

Z test	
μ_U	1250
σ_U	145,057
z	-2,26117
$\phi(z)$	0,01191
α	0,05
Pro levostranný test je p-hodnota = $\phi(z)$	
p	**0,01191**

P-hodnota je menší než alfa (0,05), a tak zamítám nulovou hypotézu. Na hladině významnosti 0,05 tedy potvrzuji alternativní hypotézu, že X < Y. Poněvadž se jedná o hodnoty odezvy, tak čím menší, tím lepší. Potvrdil jsem tedy, že Y má větší odezvu než X, z čehož plyne, že X je lepší poskytovatel.

Úkol 2

Tento úkol spočíval ve výpočtu dvoufaktorové nevyvážené anovy s cílem zjistit, zda doba vyřešení úlohy závisí na denní době, hlučnosti okolí nebo jejich kombinaci. Výpočet jsem provedl celkem dvakrát. Nejprve jsem počítal manuálně dle vzorců z tohoto instruktážního videa. Správnost postupu jsem si ověřil ve zcela nezávislém online nástroji, který výpočet provádí stejným způsobem. Pro zajímavost jsem výpočet provedl i pomocí funkce z excel rozšíření real-statistics, která nevyváženou dvoufaktorovou anovu počítá pomocí regrese postupem uvedeným v dokumentaci.

Ruční výpočet:

ANOVA	k	SS	MS	F	F krit
Doba	2	31,29808	15,64904	1,185587	3,340386
Hlučnost	3	415,6922	138,5641	10,49775	2,946685
Interakce	6	19,80142	3,300236	0,250029	2,445259
Chyba	28	369,5833	13,1994		
Celkem	39	836,375			

Z výsledku vyplývá, že doba řešení úlohy nezávisí na denní době, neboť testové kritérium F nepřekročilo kritickou hranici, tedy spadá do doplňku testového kritéria. Naopak testové kritérium faktoru hlučnosti svou kritickou hranici přesáhlo výrazně, a proto je zde závislost patrná. Testové kritérium interakce pak kritickou hranici taktéž nepřesáhlo, a tak lze závislost opět vyloučit.

Výpočet pomocí regrese:

ANOVA				Alpha	0,05	
	SS	*df*	*MS*	*F*	*p-value*	*p eta-sq*
Rows	11,56693	2	5,783465	0,438161	0,64957	0,030347
Columns	370,6119	3	123,5373	9,359308	0,000189	0,500695
Inter	39,34038	6	6,556731	0,496744	0,805278	0,096205
Within	369,5833	28	13,1994			
Total	836,375	39	21,44551			

„Rows“ v tomto případě značí faktor denní doby, „Columns“ poté hlučnost. Lze vidět, že p-hodnoty denní doby a interakce překračují hladinu významnosti α, což značí, že na nich doba řešení úlohy nezávisí. Nicméně, p-hodnota hluku je značně nižší, tedy zde již závislost s dobou výpočtu existuje.

Závěr:

Odlišné postupy výpočtů anovy počítají sumy čtverců různými způsoby, proto jejich odlišnost není příliš podstatná. Důležité je, že se oba výpočty dostaly ke stejné celkové SS a že se shodují na závěru. Z toho důvodu mohu na hladině významnosti 0,05 zamítnout hypotézy, že by doba vyřešení úlohy závisela na denní době nebo její interakci s hlučností, nicméně nezamítám hypotézu, že doba řešení úlohy závisí na hlučnosti.

Úloha 3

V rámci tohoto úkolu jsem testoval, zda spolu souvisí studijní výsledky a lidské povahové rysy. Existují 4 základní druhy (cholerik, sangvinik, flegmatik, melancholik) a dle mého názoru tyto mohou ovlivňovat prospěch. Cholerik by mohl mít sklon k lepší známkám vzhledem k jeho sebevědomí a vůdcovským schopnostem, zatímco sangvinik má potenciál brát školu volněji a více si užívat. Flegmatik by taktéž mohl mít horší prospěch, neboť mu může být ukradený, ovšem melancholik zase může nabýt lepších známek díky větší snaživosti pro svou úzkostlivost. Jako ohodnocení studia byla brána průměrná známka všech předmětů za celé dokončené bakalářské studium (tedy bez F) a to bez ohledu na univerzitu či fakultu.

Za nulovou hypotézu pokládám to, že studijní výsledky a lidská povaha jsou navzájem nezávislé. Poté zamítnu-li tuto hypotézu, potvrdím tím své tvrzení o tom, že spolu tyto faktory souvisí.

Ke sběru dat jsem sestavil následující jednoduchý formulář:

1) Jaká z následujících čtyř povah u tebe převládá?
   a) cholerik b) sangvinik c) flegmatik d) melancholik

2) Jaký byl tvůj studijní průměr za celé bakalářské studium (průměr ze všech předmětů)?
   a) A b) B c) C d) D e) E

Tázaní lidé byli známí a přátelé, zejména spolužáci z gymnázia, kteří studují zároveň se mnou, neboť tato skupina lidí mi mohla poskytnout aktuální data a zároveň jsem k nim měl snadný přístup. Dotazník byl položen pouze těm lidem, kteří studium dokončili, neboť někdo s výborným prospěchem studium nemusel dokončit z jiných důvodů, takže zde by mohla nastala nekonzistence. Podmínkou dokončeného studia získávám izolovanější data a lépe se chráním.

Formuláře jsem formou Google dotazníků rozeslal osobně svým bývalým spolužákům (nejen ze své třídy, ale i paralelních ročníků, jichž bylo celkem 5, tedy 150 lidí) prostřednictvím Facebook messengeru. V průběhu prvního listopadového týdne jsem takto oslovil 82 přátel. Zároveň jsem je poprosil, jestli by dotazníky nemohli přeposlat některým svým novým spolužákům. Po týdnu sbírání dat jsem získal celkem 112 vzorků:

Nasbíraná data

Povaha \ Známka	A	B	C	D	E	Σ
Cholerik	4	7	8	7	3	29
Sangvinik	1	5	5	11	6	28
Flegmatik	2	3	8	10	2	25
Melancholik	4	11	9	5	1	30
Σ	11	26	30	33	12	112

Teoretické četnosti

Povaha \ Známka	A	B	C	D	E	Σ
Cholerik	2,848214	6,732143	7,767857	8,544643	3,107143	29
Sangvinik	2,75	6,5	7,5	8,25	3	28
Flegmatik	2,455357	5,803571	6,696429	7,366071	2,678571	25
Melancholik	2,946429	6,964286	8,035714	8,839286	3,214286	30
Σ	11	26	30	33	12	112

Ukázalo se, že pro splnění požadavku, aby všechny teoretické četnosti byly větší než 5, mi ještě dost dat chybělo. Z toho důvodu jsem sbírání dat prodloužil o další týden, kdy jsem se ptal lidí, kteří studovali ročník nade mnou. Poněvadž jsem se bál, že to stále nemusí stačit, rozšířil jsem dotazování i na přátele z fakulty. Po dalším týdnu jsem již dosáhl dostatečného počtu dat.

Nasbíraná data

Povaha \ Známka	A	B	C	D	E	Σ
Cholerik	6	14	15	9	5	49
Sangvinik	4	8	12	14	8	46
Flegmatik	3	6	17	18	6	50
Melancholik	9	18	14	8	3	52
Σ	22	46	58	49	22	197

Teoretické četnosti

Povaha \ Známka	A	B	C	D	E	Σ
Cholerik	5,472081	11,44162	14,4264	12,18782	5,472081	49
Sangvinik	5,137056	10,74112	13,54315	11,44162	5,137056	46
Flegmatik	5,583756	11,67513	14,72081	12,43655	5,583756	50
Melancholik	5,807107	12,14213	15,30964	12,93401	5,807107	52
Σ	22	46	58	49	22	197

Následně jsem nad daty provedl test dobré shody, jenž měl následující výsledky.

Testové kritérium χ2

Povaha \ Známka	A	B	C	D	E	Σ
Cholerik	0,050931	0,572059	0,022807	0,833798	0,040727	1,520322
Sangvinik	0,25168	0,699529	0,175831	0,572059	1,595554	3,294653
Flegmatik	1,195575	2,758605	0,352881	2,488793	0,031029	6,826883
Melancholik	1,755533	2,826078	0,112032	1,882205	1,356932	7,93278
Σ	3,253719	6,856272	0,663551	5,776855	3,024242	19,57464

Doplněk kritického oboru je následující: $<0; \chi^2_{0,95}(12)> = <0; 21,026>$

Testové kritérium je poté výsledek 19,57464 a spadá tedy do doplňku kritického oboru, tedy H_0 nezamítám. Na hladině významnosti 0,05 tedy nezamítám nezávislost faktorů, a proto nepotvrzuji své původní tvrzení, že by lidská povaha měla vliv na studijní výsledky (faktory by byly závislé).