1 Testowanie zgodności rozkładu

Poznaliśmy już sposoby testowania hipotezy o tym, czy próba pochodzi z populacji o rozkładzie normalnym. Dzisiaj zajmiemy się ogólniejszymi metodami weryfikacji zgodności z dowolnym rozkładem. Dla przykładu w pliku rozklady.sav mamy dwie próby pobrane z dwóch różnych rozkładów. Czy można stwierdzić, że pierwsza próba pobrana jest z rozkładu normalnego, a druga z lognormalnego? Dla odmiany zadanie to rozwiążemy najpierw w programie Statistica. Z meny Statystyka wybieramy "Dopasowanie rozkładów"



następnie zaś wybieramy interesujący nas rozkład oraz wybieramy stosowną zmienną

🖄 Dopasowanie rozkładu:	rozklady	? ×
Podstawowe Rozkłady ciągłe: Prostokątny Prostokątny Gamma Lognormalny Chi-kwadrat M Inne	Rozkłady dyskretne: Image: Construction Image: Construction	C Opcje
		CRSES 2 10 W



Jako wynik otrzymujemy w zależności od tego co wybraliśmy tabelę z oczekiwanym rozkładem lub odpowiednio wykres
(histogram) z nałożoną oczekiwaną

gęstością. Najistotniejsze w obu przypadkach to istotność, która wynosi p =0,00042, zatem odrzucamy hipotezę o równomierności naszego rozkładu. W analogiczny sposób weryfikujemy następujące hipotezy dla drugiej zmiennej

${\cal H}_0$: Y ma rozkład lognormalny

 H_1 : Y nie ma rozkładu lognormalnego

W tym wypadku nie ma podstaw do odrzucenia H_0 ponieważ p=0,4331. W programie SPSS test χ^2 zgodności znajduje się w sekcji test nieparame- $\operatorname{tryczne}$

Analiz	a) Wyk	resy	Narzędz	zia O	kno	Por	noc		
ŀ	aporty	warstv	vowe		×	1			
0	Opis stat	ystycz	zny		•				
- 1	abele s	pecjal	ne		•				
F	orówny	wanie	e średnicł	h	•		var	var	var
] ()gólny r	nodel	liniowy		•				
N	/lodele i	miesza	ane		•	L			
- K	Corelacje	2			►	<u> </u>			
- F	legresja				•	-			
- 4	Analizy l	oglini	owe		•	-			
ĸ	lasyfika	cja			•				
F	ledukcja	a dany	ch		•				
s	kalowa	nie							
_ 1	esty nie	paran	netryczne	e	►		Chi-kwa	drat	
- 4	Analizy p	orzeży	cia		•		Dwumiar	nowy	
- V	Vielokro	tne o	dpowied	zi	•		Serii		
							K-S dla je	dnej próby	
							Dwie pró	by niezależne.	
							K prób ni	ezależnych	
							Dwie pró	by zależne	
					_		K prób za	leżnych	



W przypadku rozkładu równomiernego nie ma właściwie żadnego problemu

W raporcie otrzymujemy interesujące nas wyniki



Jak widzimy istnieją pewne rozbieżności pomiędzy poznawanymi programami. W jednym otrzymaliśmy istotność p=0.00042, zaś w drugim $\alpha = 0.001$. Różnice są powodowane różną dokładnością obliczeń oraz błędami zaokrągleń, nie zmieniają one jednak odpowiedzi podczas weryfikacji hipotez.

2 Weryfikacja hipotez dla jednej średniej

Bardzo często podczas różnych badań w naturalny sposób pojawia się pytanie, czy w analizowanych przez nas danych dla pewnej cechy średnia jest równa pewnej zadanej wartości. Dla przykładu chcemy sprawdzić, czy w rozważanej przez nas wcześniej ankiecie średnia wieku dla badanej populacji wynosi powiedzmy 35 lat. Oczywiście większość osób stwierdzi, że wystarczy policzyć średnią i będziemy znali odpowiedź. Należy jednak pamiętać, że do wypełnienia ankiety została wybrana pewna próba. Jest oczywiste, że parametry z próby estymują parametry z populacji, to jednak wciąż są to pewne przybliżenia. Fachowej odpowiedzi na takie pytanie dokonujemy posiłkując się stosownym testem do weryfikacji hipotez

 $H_0: \mu = \mu_0 (35)$ $H_1: \mu = \mu_0 (35),$

gdzie μ_0 jest zadaną wartością. Oczywiście takiej weryfikacji można dokonać w sposób tradycyjny, jak również z zastosowaniem pakietów statystycznych. My skupimy się raczej na tym drugim sposobie. W programie SPSS odpowiedni test znajdujemy w sekcji porównanie średnich

a	Anali	za	Wykresy	Narzędzia	Okno	Pomoc							
Ŧ		Rap	orty warstv	vowe	•	1							
-		Opi	s statystycz	ny	•								
5	-	Tab	olo epocjali		<u> </u>	owość	marka	rzatalnaga	lokolizaci				
1 (Por	ównywanie	średnich		Śrec	Inie						
1,0	-	Ogć	ólny model	liniowy	<u>۲</u>	Test	t dla jednej p	róby	,0(
,(Mo	dele miesza	ine	•	Test	t t dla prób nie	zależnych	1,00				
,(Kore	elacje		•	Test	t t dla prób zal	eżnych	,0(
, 		Reg	resja		•	Jednoczynnikowa ANOVA							
1,0		Analizy logliniowe		owe	•	2.00	,	.,	1,00				
1.0		Klas	syfikacja		•	.00	1.00	1.00	.0(
1,(Red	ukcja dany	ch	•	2,00	1,00	,00	,0(
1,(Skal	lowanie		•	3,00	1,00	,00	,0(
1,(Testy nieparametryczne Analizy przeżycia Wielokrotne odpowiedzi			•	1,00	,00	,00	,0(
1,(•	3,00	1,00	1,00	1,00				
,(•	2,00	,00	,00	,0(
,(_				2,00	,00	,00	,0(

Następnie wskazujemy interesującą nas zmienną, zadajemy wartość testowaną

Test t dla jednej próby	,			-X
 ID_ankietera numerAnkiety płeć wykształcenie miejscowość marka uczelni (mar rzetelnosc lakaliazaja 		Zmienne testowane:		OK Wklej Resetuj Anuluj Pomoc
parkingi		Wartość testowa <mark>r</mark> a:	35	Opcje

Na podstawie otrzymanego raportu

Test dla jednej próby

			Wartość t	estowana = 3	15		
			Istotność	Różnica	95% przedział ufności dla różnicy średnich		
	t	df	(dwustronna)	średnich	Dolna granica	Górna granica	
wiek	-4,282	199	,000	-2,25500	-3,2936	-1,2164	

łatwo stwierdzamy, że odrzucamy hipotezę H_0 na korzyść H_1 .

Sprawdźmy teraz jaki jest poziom istotności jeśli jako wartość testowaną obierzemy 33 lata. Jak widać z poniższego raportu

	Test dla jednej próby											
	Wartość testowana = 33											
	95% przedział ufności dla											
			Istotność	Różnica	różnicy średnich							
	t	df	(dwustronna)	średnich	Dolna granica	Górna granica						
wiek	-,484	199	,629	-,25500	-1,2936	,7836						

w takim przypadku nie podstaw do odrzucenia hipotezy zerowej świadczącej o tym, że średnia w populacji wynosi 33 lata.

Teraz prześledźmy wykonanie tego samego zadania w programie Statistica. Z menu "Statystyka" wybieramy "Statystyki podstawowe i tabele" a następnie

z okna dialogowego



jak na powyższym rysunku. Po wybraniu odpowiednie zmiennej i testowanej

wartości

🖾 Test t dla pojedynczych średnich: ankieta	? 💌
☑ Zmienne: wiek	Podsumowanie
Podstawowe Więcej Opcje	Anuluj
Podsumowanie: testy <u>t</u>	🔁 Opcje 🔻
 Wartości odniesienia Testuj średnie względem: 33 	<u>₩ G</u> rupami
Testuj średnie względem określonych wartości	CRSES S
	Momenty ważone
₩ykres ramka-wąsy	© ₩-1 ○ N-1
	Usuwanie BD
	Przypadkami
	Parami
Imiante novertal 1.00 nia Inia Ini	ia laia

otrzymujemy skoroszyt z wynikami testu.

		Test średnich względem stałej wartości odniesienia (ankieta)									
1		Średnia Odch.st. Ważnych Bł. std. Odniesienie t df p									
1	Zmienna					Stała					
	wiek	32,74500	7,448313	200	0,526675	33,00000	-0,484169	199	0,628798		

Tak samo jak wcześniej zauważamy różnice pomiędzy oboma programami wynikające z zaokrągleń.

Dość istotnie z weryfikacją hipotez dla jednej średniej jest wyznaczanie przedziału ufności dla średniej. W statystyce matematycznej przedział ufności dla średniej wyraża się wzorem

$$\mu \in \left[\overline{X} - \frac{S_0}{\sqrt{n}} t_{1-\alpha/2, n-1}; \overline{X} - \frac{S_0}{\sqrt{n}} t_{1-\alpha/2, n-1}\right]$$

gdzie \overline{X} oznacza średnią z próby;

 S_0 - nieobciążone odchylenie standardowe;

_

 \boldsymbol{n} - liczebność próby;

 $t_{1-\alpha/2,n-1}$ oznacza kwantyl rzędu $1-\alpha/2$ z rozkładu t-studenta
on-1 stopniach swobody.

W programie Statistica musimy wejść do okna dialogowego "test
t dla pojedynczych średnich" a następnie przejść na zakładkę Opcje

- 11											
		Test średr	nich wzglęc	dem stałej v	vartości od	dniesienia (ankieta)				
4		Średnia	Odch.st.	Ważnych	Bł. std.	Ufność	Ufność	Odniesienie	t	df	F
F	Zmienna					-95,000%	+95,000%	Stała			
e	wiek	32,74500	7,448313	200	0,52667	31,70642	33,78358	0,00	62,17304	199	0,
				_							
			🖉 🏹 Tes	t t dla pojed	ynczych śre	dnich: ankie	eta	2	x		
					-		-				
			- 🛃 🛃 Z	(mienne:] wie	k			Podsum	owanie		
					-	1	.	A de la			
			Pods	awowe Wię	cej Opcje			Anulu			
			-Wa	artości odniesia	enia						
				Testui średnie	wzaledem:	0					
				Testui średnie	wzaledem			🛛 🛄 🛛 <u>G</u> rupa	ami		
				określonych w	vartości	i	lk <u>r</u> eśl				
				1				SELECT CRSES S	& ∞		
				okaz drugie p	2714IU 701900						
			0 🔽	blicz granice u	ufności; Prz	zedział <mark>95,0</mark>	0 🛢 %	Momenty w	ażone		
				نصوب المتحد المت		(دو دناله ۱		DF =			
			pozio	m p dla podśv	vietlania: 闻)5		🔘 W-1 🔿	N-1		
								– Usuwanie BD)		
								Przupadk	ami		
								Preservice			
								🖉 Parami			

Jako wynik otrzymujemy przedział

$\mu \in [31.70642; 33.78358]$

W programie SPSS postępujemy dość podobnie. Wybieramy test dla jednej

próby



wskazujemy zmienną i wchodzimy w opcje, gdzie podajemy poziom ufności i otrzymujemy raport

-				
1651	i di	a iei	dne	I Droby
		- 1-		1 1 1 1 1 1 1

	Wartość testowana = 0										
			Istotność	Różnica	95% przedział ufności dla różnicy średnich						
	t	df	(dwustronna)	średnich	Dolna granica	Górna granica					
wiek	62,173	199	,000	32,74500	31,7064	33,7836					

3 Weryfikacja hipotez o równości dwóch średnich

Zajmiemy się teraz problem stwierdzenia faktu, że w dwóch próbach średnia jest taka sama. Musimy tutaj rozważyć dwa przypadki. W przypadku pierwszym zakładamy niezależność dwóch zmiennych(prób), w drugim przypadku nie zakładamy niezależności. W większości przypadków intuicja sama nam podpowiada, który przypadek zastosować w danej sytuacji.

W przypadku zmiennych niezależnych w programie SPSS musimy dysponować

ſ	Anal	liza	Wykresy	Narzędzia	Okno	Pon	юс			
		Rap	orty warst	vowe	•	1				
-		Opis statystyczny			•	-				
-		Tab	ele specjal	ne	•		Nor	NOT	Vor	
		Porównywanie średnich			•		Średnie			Г
-		Ogolny model liniowy		•		Test t dla	jednej próby.			
		Mo	dele miesz	ane	•		Test t dla	prób niezależ	nych	
_		Korelacje			•		Test t dla	prób zależny	ch	-
_		Regresja			•	Jednoczynnikowa ANOVA				
_		Ana	lizy loglini	owe	•	F				-
-		Klasyfikacja		•					_	

zmienną grupującą dla badanej cechy. Wybieramy następującą metodę

Następnie wskazujemy zmienną testowaną oraz zmienną grupującą, wskazujemy podział dla zmiennej grupującej i następnie otrzymujemy raport, w którym mamy interesującą nas odpowiedź.



W raporcie oprócz interesującej nas istotności otrzymujemy niejako w gratisie przedział ufności dla różnicy średnich.

Test dla	prób	niezal	leżny	/ch

	Test Le jednoro wari	evene'a Idności ancji		Test t równości średnich					
				Błąd 95% przedział ufności dla Istotność Różnica standardowy różnicy średnich					ł ufności dla średnich
	F	Istotność	t	df	(dwustronna)	średnich	różnicy	Dolna granica	Górna granica
Założono równość wariancji	,074	,786	,552	98	,582	,15320	,27771	-,39790	,70430
Nie założono równości wariancji			,552	97,982	,582	,15320	,27771	-,39790	,70430

W programie Statistica dla prób niezależnych nie musimy mieć zmiennej grupującej, ale dla danych ze zmienną grupującą jest również stosowna metoda.



Następnie wskazujemy zmienną grupującą oraz testowaną (jak na rysunku poniżej) i otrzymujemy wynik.

Testy t; Grupująca: osoba (zmienneNiezalezne)								
Grupa 1: 1								
Grupa 2 2								
Średnia	Średnia	t	df	р	N ważnyc	N ważnych		
Średnia 1	Średnia 2	t	df	р	N ważnyc 1	N ważnych 2		

🌇 Test t dla prób niezależnych, w grupach: zmienne	Niezalezne	? 💌
☑ Zależna: czasSzukaniaKlucza Grupująca: osoba		Podsumowanie Anuluj
Kod grupy 1: 1 Kod grupy 2:	2	🔊 Opcje 🔻
Podstawowe Więcej Opcje		un <u>G</u> rupami
Podsumowanie: testy <u>t</u>		SELECT CRSES S
📴 Wykres ramka-wąsy		Momenty ważone
		DF =
		Usuwanie BD
		Przypadkami
		Parami

W programie Statistica próby niezależne możemy mieć podane w dwóch zmiennych. W takim przypadku należy zastosować inną metodę.

🖾 Statystyki podstawowe i tabele: zmienneZalezne	? <mark>×</mark>
Podstawowe	
Ant. Statystyki opisowe	
Macierze korelacji	Anuluj
🚰 Test tidla próbiniezależnych (wzgl. zmn.)	
I st t dla pojedynczej próby	

W odpowiednim oknie dialogowym wskazujemy, które zmienne mają być porównywane. Następnie otrzymujemy interesujące nas wyniki.



Badanie równości średniej dla prób zależnych wykonuje się w bardzo podobny sposób. Musimy jedynie dokładnie czytać wszystkie polecenia.

4 Analiza wariancji

Na wstępie zapoznamy się z metodą pozwalającą porównywać średnie w kilku grupach. Do tego typu analiz służy jednoczynnikowa analiza wariancji, tzw. jednoczynnikowa ANOVA. W pliku czasDojazdu.sav mamy informacje o czasie dojazdu na uczelnie na kolejne zjazdy. Chcemy zweryfikować hipotezę o równości średnich czasów dojazdu w poszczególnych okresach czasu. Postawmy zatem hipotezy

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

wobec hipotezy alternatywnej, która jest zaprzeczeniem H_0 , tzn.

$$H_1: \mu_1 \neq \mu_2 \lor \mu_1 \neq \mu_3 \lor \mu_1 \neq \mu_4 \lor \mu_2 \neq \mu_3 \lor \mu_2 \neq \mu_4 \lor \mu_3 \neq \mu_4$$

Zgodnie ze stosowną teorią powinniśmy sprawdzić, czy poszczególne próby pochodzą z populacji o rozkładzie normalnym. Przypomnijmy sobie jak sprawdzić,

czy nasz próba ma rozkład normalny. Wybieramy Analiza\Opis statystyczny\Eksploracja danych. Następnie postępujemy jak na poniższych rysunkach



Jako wynik otrzymujemy raport, w którym najistotniejszym punktem jest

tabela

	Kołmogorow-Smirnow ^a			Shapiro-Wilk			
	Statystyka	df	Istotność	Statystyka	df	Istotność	
czas dojazdu na 1 zjazd	,041	200	,200*	,986	200	,054	
czas dojazdu na drugi zajzd	,036	200	,200*	,993	200	,438	
czas dojazdu na trzeci zjazd	,033	200	,200*	,997	200	,988	
czas dojazdu na 4 zjazd	,038	200	,200*	,994	200	,637	

Testy normalności rozkładu

Dolna granica rzeczywistej istotności.

a. Z poprawką istotności Lillieforsa

Latwo jest stwierdzić, że założenie o normalności rozkładu jest spełnione. W naszym przekonaniu umacnia nas dodatkowo analiza wykresów. W klasycznej analizie wariancji dodatkowym założeniem jest równość wariancji. Warunek ten sprawdzamy za pomocą chociażby testu Levene'a. Sprawdzenia tego dokonamy równolegle z testowaniem naszej głównej hipotezy. Zanim przystąpimy do testowania hipotezy musimy przygotować sobie nasze dane. W programie SPSS w jednej zmiennej powinny znajdować się wartości obserwacji natomiast w innej zmiennej czynnik grupujący. Dlatego dalszej analizy dokonamy na pliku czasDojazdu1.sav, w którym nasze wartości są już odpowiednio przygotowane. Wybieramy z menu jednoczynnikową ANOVA jak na poniższym rysunku

	Ana	liza Wykresy Narzę	dzia Okno	Pomoc
Ī		Raporty warstwowe	•	1
		Opis statystyczny	•	
_		Tabele specialne		
7		Porównywanie średni	ich	Średnie
		Ogoiny model liniow	y P	Test t dla jednej próby
]	Modele mieszane	•	Test t dla prób niezależnych
_		Korelacje	•	Test t dla prób zależnych
_		Regresja	•	Jednoczynnikowa ANOVA –
_		Analizy logliniowe	•	

Jako zmienną zależną obieramy czas, natomiast jako czynnik zmienną czyn-

nik jak na rysunku i przechodzimy na kartę opcje

Jednoczynnikowa ANO	VA	×
	Zmienne zależne:	OK Wklej Resetuj Anuluj Pomoc
	Kontrasty	Opcje

W oknie Opcje wybieramy interesujące nas wskaźniki

Jednoczynnikowa ANOVA: Opcje	×
Statystyki Opisowe	Dalej
📃 Efekty stałe i losowe	Anuluj
🔽 Test jednorodności wariancji	Pomoc
Brown-Forsyme	
Wykres średnich Braki danych	
Wyłączanie obserwacji analiza po analizie	
🔘 Wyłączanie wszystkich obserwacji z brakami	

Jako wynik otrzymujemy raport, w którym mamy następujące dane

Jednoczynnikowa analiza wariancji (ONEWAY)

[ZbiórDanych2] D:\Users\Adam\Desktop\statistica\sad\czasDojazdu1.sav

Test jednorodności wariancji

czas			
Test Levene'a	df1	df2	Istotność
1,649	3	796	,177

Jednoczynnikowa ANOVA

czas					
	Suma		Średni	_	
	kwadratów	df	kwadrat	F	Istotność
Między grupami	46,926	3	15,642	,633	,594
Wewnątrz grup	19657,498	796	24,695		
Ogółem	19704,424	799			

z których wynika, że założenie jednorodności wariancji jest spełnione oraz nie ma podstaw do odrzucenia hipotezy H_0 o równości wariancji. Ponadto zgodnie, z tym co wybraliśmy na karcie opcje otrzymaliśmy wykres reprezentujący poszczególne średnie. Możemy na nim łatwo sprawdzić słuszność naszego osądu.



Teraz spróbujemy stwierdzić, w której grupie jest najwyższa średnia i pomiędzy, którymi parami występuje statystycznie istotna różnica pomiędzy średnimi. W tym celu wykorzystamy test post-hoc

Zmienne zależne:	×
)K 'klej setuj nuluj moc
Czynnik:	ə

Mamy do dyspozycji wiele testów skorzystamy jednak z testu zaproponowanego przez twórcę analizy wariancji, tj testu NIR. Wybór ten jest jak najbardziej uzasadniony ponieważ już wiemy, że wariancje są równe.

Jedno	czynnikowa A	NOVA: Wielokrot	ne porównania post hoc 🗾 🐱
Zəł	ożenie o równo: NIR Bonferroni Sidak Scheffe F R-E-G-W	ści wariancji S-N-K GT2 Hochber Tukey b Tukey'a Duncan	Waller-Duncan Stosunek powagi błędów Typ 1/Typ 2: 100 Dunnett Kategoria kontrolna: Ostatnia
	Q R-E-G-W	🔲 Gabriel	Dwustronny
Bra	k założenia o ró T2 Tamhane'a	wności wariancji T3 Dunnetta	🔲 Games-Howell 📄 C Dunnetta
Poz	iom istotności:	,05	Dalej Anuluj Pomoc

W wynikowym raporcie z łatwością odnajdujemy tabelę

IEST NIK						
		Różnica	Bład	95% przedział ufnośc		
(I) czynnik	(J) czynnik	średnich (I-J)	standardowy	Istotność	Dolna granica	Górna granica
1,00	2,00	-,33935	,49694	,495	-1,3148	,6361
	3,00	,34300	,49694	,490	-,6325	1,3185
	4,00	,05055	,49694	,919	-,9249	1,0260
2,00	1,00	,33935	,49694	,495	-,6361	1,3148
	3,00	,68235	,49694	,170	-,2931	1,6578
	4,00	,38990	,49694	,433	-,5856	1,3654
3,00	1,00	-,34300	,49694	,490	-1,3185	,6325
	2,00	-,68235	,49694	,170	-1,6578	,2931
	4,00	-,29245	,49694	,556	-1,2679	,6830
4,00	1,00	-,05055	,49694	,919	-1,0260	,9249
	2,00	-,38990	,49694	,433	-1,3654	,5856
	3,00	,29245	,49694	,556	-,6830	1,2679

Porównania wielokrotne

Zmienna zależna: czas

Analiza powyższej tabeli pozwala nam stwierdzić, że pomiędzy żadną parą nie występuje statystycznie istotna różnica dla średnich. Ponadto najwyższa średnia jest w drugiej grupie a najniższa w 3.

Teraz nadszedł czas na analogiczne rozważania w programie Statistica. Skorzystamy z już przygotowanego pliku czasDojazdu. W programie Statistica testy ANOVA można odszukać w kilku miejscach. Nam w zupełności wystarczy na razie moduł dostępny menu Statystyka\Statystyki podstawowe i tabele.



Po wskazaniu odpowiednich zmiennych

🕼 Statystyki w grupach (przekroje): c	zasDojazdu1	? ×
Pojedyncze tabele Listy tabel	ОК	
🖳 Zmienne	Anuluj	
Zależna: brak Grupująca: brak	🔈 Opcje 🔻	
Wybierz zmienne zależne i zmien	? ×	
<mark>1 - czas</mark> 2 - czynnik	1 - czas 2 - czynnik	OK Anuluj [Zestawy]

przechodzimy dalej i mamy okienko, w którym mamy kilka testów związanych

z analizą wariancji

🔀 Statystyki w grupach - wyniki: czasDojazdu1	? ×
ZALEŻNA:1 zmienna: czas	
GRUPUJĄCA 1-czynnik (4): 1 2 3 4	Brs. ★
Podstawowe Statystyki opisowe Testy ANOVA Post-hoc Podsum.: tabela statystyk WW Wykresy interakcji Dokładne tabele dwudzielcze Skategoryzow. wykresy ramka-wąsy Analiza wariancji	Podsumowanie Anuluj Opcje V Grupami

Na zakładce "Testy ANOVA" mamy m. in. do dyspozycji testy jednorodności wariancji (1), jak również skategoryzowane wykresy normalności (2) pozwalające sprawdzić nam założenia klasycznej analizy wariancji.

🖾 Statystyki w grupach - wyniki: czasDojazdu1	? ×
ZALEŻNA:1 zmienna: czas	
GRUPUJĄCA 1-czynnik (4): 1 2 3 4	the table
Podstawowe Statystyki opisowe Testy ANOVA Post-hoc	Emergina Podsumowanie
🛄 <u>A</u> naliza wariancji 🚼 <u>S</u> kateg. wykres normalności	Anuluj
🔲 Test F Welcha 🧧 Skateg. wykres <u>n</u> orm. połówkowej	🔁 Opcje 🔻
Test jednorodności wariancji – E Skateg. wykres odchyleń od normalności	🛄 Grupami
Test Levene'a Średnie <u>w</u> zgędem. odch. std.	2
I est Browna- <u>+</u> orsytha	
poziom p dla podświetlania: 05	

Domyślnie przy wyborze wykresów skategoryzowanych mamy jedynie same wykresy, jednak po dwukrotnym kliknięciu w obszar wykresu możemy wybrać opcję wyświetlającą wyniki testu Shapiro-Wilka.



Oczywiście jeśli ktoś woli mieć wyznaczony histogramy z nałożonymi wynikami testu badającego normalność to można skorzystać z zakładki Statystyki opisowe. Dość ciekawym sposobem wizualizacji danych jest wykres interakcji dostępny na zakładce podstawowe. Jako wynik otrzymujemy wykres, na którym oprócz śred-

nich zaznaczone są95% przedziały ufności dla średniej.



Jeśli będziemy dysponowali zapasem czasu to powrócimy do tematu i omówimy nieco bardziej skomplikowane zagadnienie jakim jest wieloczynnikowa analiza wariancji.

5 Testowanie niezależności

W wielu przypadkach interesuje nas sprawdzenie, czy istnieją zależności pomiędzy pewnymi próbami. Możemy np. wyobrazić sobie sytuację, że chcemy sprawdzić czy istnieje zależność pomiędzy ocenami z "Podstaw statystyki matematycznej" oraz "Podstaw statystyki opisowej". W pliku ocenyPodstawy.sav mamy informacje o ocenach pewnej grupy studentów. Chcemy dokonać weryfikacji następujących hipotez

 $H_0:$ oceny z "Podstaw statystyki matematycznej" i oceny z "Podstaw statystyki opisowej" są niezależne

 H_1 : istnieje zależność pomiędzy ocenami

W klasycznej statystyce do weryfikacji hipotez o niezależności stosuje się test niezależności \mathcal{X}^2 . W programie SPSS odnajdujemy go w nieco zaskakującym miejscu



Jedną z naszych zmiennych wskazujemy jako zmienną w wierszu, drugą jako kolumnę (jest to bez znaczenia), w oknie statystyki wskazujemy test chikwadrat.



Jako wynik otrzymujemy raport, w którym mamy odpowiedź na interesujące nas pytanie, tabele krzyżowe zależności pomiędzy poszczególnymi ocenami oraz dość interesujący wykres reprezentujący zależności opisane w tabeli krzyżowej.

Testy Chi-kwadrat

			Istotność asymptotyczn
			a
	Wartość	df	(dwustronna)
Chi-kwadrat Pearsona	7,138	9	,623
Iloraz wiarygodności	7,691	9	,566
Test związku liniowego	,196	1	,658
N Ważnych obserwacji	100		

a. 37,5% komórek (8) ma liczebność oczekiwaną mniejszą niż 5. Minimalna liczebność oczekiwana wynosi 2,24.

Tabela krzyżowa Podstawy statystyki matematycznej * Podstawy statystyki opisowej Liczebność

		Poc	Podstawy statystyki opisowej				
		2,00	3,00	4,00	5,00	Ogółem	
Podstawy	2,00	2	6	7	2	17	
statystyki	3,00	2	13	12	5	32	
matematycznej	4,00	7	8	12	8	35	
	5,00	3	7	3	3	16	
Ogółem		14	34	34	18	100	



Wykres słupkowy

W programie Statistica test ten odnajdujemy w części Statystyki podsta-

wowe i tabele.

🔀 Statystyki podstawowe i tabele: ocenyPodstawy	? ×
Podstawowe	
🚕 Statystyki opisowe	
Macierze korelacji	Anuluj
EE Test t dla prób niezależnych (wzgl. grup)	Docie 🗸
🛐 Test t dla prób niezależnych (wzgl. zmn.)	· ····
Test t dla prób zależnych	
🗐 🗵 Test t dla pojedynczej próby	
🚋 Przekroje, prosta ANOVA	
Przekroje uproszczone	
Sizza Takala Kasua (a)	
Tabele wielodzielcze	
🕅 Inne testy istotności	🗁 O <u>t</u> wórz dane
🦾 Kalkulator prawdopodobieństwa	

Po określeniu interesujących nas zmiennych w zakładce opcje wybieramy stosowny test



oraz w zakładce Więcej wybieramy dokładne tabele dwudzielcze



Dość interesujące wyniki otrzymujemy wybierając na zakładce Więcej Histogramy skategoryzowane oraz Wykresy interakcji liczności. W pierwszym przypadku otrzymujemy histogramy dla poszczególnych wartości jednej zmiennej.



W naszym konkretnym przypadku są to histogramy zmiennej p
so dla poszczególnych wartości psm. W drugim przypadku mamy natomiast



zobrazowane w sposób liniowy analogiczne zależności.

6 Regresja

Do tej pory poznaliśmy już metody pozwalające stwierdzić, czy dwie próby są niezależne. W tym miejscu spróbujemy opisać rodzaj zależności pomiędzy badanami cechami. Najprostszym sposobem zależności jest zależność liniowa pomiędzy dwiema cechami. Rozważmy przykładową zależność pomiędzy czasem nauki a wynikiem z egzaminu. Wykres rozrzutu danych z pliku "nauka.sav" pozwala nam oszacować rodzaj zależności. W tym celu wybieramy prosty wykres



```
rozrzutu
```



Następnie wskazujemy interesujące nas zmienne, teoretycznie kolejność zmiennych jest bez znaczenia, ale w sposób naturalny odczuwamy, że wynik zależy od czasu nauki, a nie odwrotnie.

Prosty wykres rozrzutu			×
	\rightarrow	0śY:	OK Wklej
	\rightarrow	OśX: 🛷 czas nauki w godzianch [c	Resetuj
	\rightarrow	Ustaw znaczniki według:	Pomoc
		Użyj do opisu obserwacji:	
	Zmienn	ne panelu	
		Wiersze:	
	\rightarrow		
		📃 Zagnieżdżaj bez pustych wier:	szy
		Kolumny:	
	\rightarrow		
		Zagnieżdżaj bez pustych kolu	mn
Szablon Zastosuj szablon wykre:	su z:	ſ	Tytuły
Plik			Opcje

Jako wynik otrzymujemy raport, w którym mamy w sposób graficzny zaprezen-

towane występujące zależności



Latwo zauważyć, że nasze obserwacje rozkładają się wokół pewnej prostej, naszym celem będzie wyznaczenie równania tej prostej. W tym celu skorzystamy z modułu regresja liniowa



Następnie określamy zmienną zależną i niezależną

💷 Regresja liniowa		×
🖋 czas nauki w godzianc	Zmienna zależna: Iczba punktów [wynik] Blok 1 1 Poprzedni Następny	OK Wklej Resetuj
	Zmenne niezależne: Czas nauki w godzianch [czas] Metoda: Wprowadzania	Pomoc
	Wprowadzania Krokowa Usuwania Eliminacji wstecznej Selekcji postępującej cyklety obserwacji.	Filtr
	WNK Waga:	Opcje

W wynikowym raporcie najważniejsza jest następująca tabelka.

		Współczynniki niestandaryzowane		Współczynniki standaryzowa ne		
Model		в	Błąd standardowy	Beta	t	Istotność
1	(Stała)	-,153	1,392		-,110	,913
	czas nauki w godzianch	3,010	,055	,992	54,651	,000

Współczynniki

a. Zmienna zależna: liczba punktów

z której możemy odczytać interesujący nas wzór prostej. Okazuje się, że zależność wyniku w zależności od czasu nauki opisuje prosta

$$y = 3.010x - 0.153$$

Za pomocą tej prostej możemy szacować wyniki znając czas poświęcony na naukę i dla przykładu osoba ucząca się 50 godzin powinna otrzymać około 150 punktów (na 100 możliwych ;)). Zauważmy jeszcze, że w programie mamy kilka możliwych sposobów wyznaczania równania (1). Mamy również możliwość wyboru z kilku dostępnych statystyk oraz wykresów (2).

💷 Regresja liniowa		×
czas nauki w godzianc	Zmienna zależna: <pre></pre>	OK Wklej Resetuj Anuluj Pomoc
	Statystyki Wykresy Zapisz	Opcje

Teraz wyznaczymy równanie prostej regresji za pomocą programu Statistica. Wybieramy z menu statystyka regresję wieloraką

	<u>S</u> ta	tystyka	Data <u>M</u> ining	<u>W</u> ykresy	<u>N</u> arzędzia	Dar					
ľ	æ	<u>⊾ontynu</u> uj Ctrl+R									
h	50	Statystyki podstawowo i tabolo									
٦	1	<u>R</u> egresja wieloraka									
ľ	鶗	<u>ANOVA</u>				1					
l	2 71	Statystyki <u>n</u> ieparametryczne									
	20	<u>D</u> opasov	vanie rozkłado	ów							

Następnie wskazujemy zmienne

	📈 Regresja wieloraka: na	iuka	?	×
	Podstawowe Więcej	_	40 📰 I	
	🖳 Zmienne		Anuluj	
	Zależna: brak Niezależne: brak		Dpcje	
			🔁 O <u>t</u> wórz	dane
🔥 Wybie	rz listy zmiennych zależnyc	n i niezależnych:	Mar 1	? ×
1 - czast 2 - wynik	Nauki	1 - czasNauki 2 - wynik		ОК
			T	Anuluj
				[Zestawy]
				Włącz opcję "Pokazuj tylko
				zmienne o odpowiedniej skali" aby na listach w
Wezvetk	rie Rozwiń Przybliż	Wszystkie Bozwiń	Przybliż	zależności od potrzeby, pojawiały
7	nieżne:	Lista amjennych niezale	żnych:	się tylko zmienne jakościowe albo ilościowe. Naciścii
2		1		F1 aby uzyskać więcej informacji.
Poka:	zuj tylko zmienne o odpowiedni	ej skali		

i jako wynik otrzymujemy skoroszyt, w którym mamy interesujący nas wynik.

)		Podsumowanie regresji zmiennej zależnej: wynik (nauka) R= ,99206004 R^2= ,98418313 Skoryg. R2= ,98385361 F(1,48)=2986,7 p<0,0000 Błąd std. estymacji: 1,0535								
2		b*	Bł. std.	b	Bł. std.	t(48)	р			
r	N=50		z b*		zb					
	W. wolny			-0,152589	1,391552	-0,10965	0,913141			
	czasNauki	0,992060	0,018153	3,009528	0,055068	54,65103	0,000000			

Oczywiście w statystyce można rozważać inne, bardziej złożone typy regresji. W klasyczny sposób (na kartce) wyznaczanie innych krzywych regresji jest dość trudne i pracochłonne. Na szczęście za pomocą programów statystycznych jest stosunkowo proste. Musimy jedynie zasugerować jakiego rodzaju krzywej regresji się spodziewamy. Przy wyborze mogą nam pomóc wykresy rozrzutu. W pliku regresja.sav mamy kilka przykładowych zestawów danych. Pierwsza zmienna jest zmienną niezależną, jak łatwo odczytać z poniższego wykresu rozrzutu druga ze zmiennych zależy od pierwszej w sposób liniowy



Trzecia zmienna jest związana z pierwszą za pomocą trójmianu kwadratowego (chociaż może tego nie widać na rysunku), czyli jak ktoś woli funkcji kwadratowej.



Trzecia ze zmiennych zależnych (u nas jest to zmienna o nazwie zależna3) jest związana wielomianem stopnia trzeciego. Wykres rozrzutu przyjmuje następującą postać.



Ostatnia zaś zmienna zależna jest związana za pomocą funkcji wykładniczej. Postaramy się wyznaczyć poszczególne funkcje opisujące nasze zależności. W programie SPSS skorzystamy z modułu estymacja krzywej...

[Ana	liza Wykresy Narzędzia	Okno	Pomoc			
		Raporty warstwowe	•				
=		Opis statystyczny	•				
2		Tabele specjalne	•	var	var	var	
50		Porównywanie średnich	•				
50		Ogólny model liniowy	•				
52		Modele mieszane	•				
94	_	Korelacie	•				
36		Regresja	•	Liniowa			ŀ
24		Analizy logliniowe	•	Estymacj	a krzywej		
Ĩ.		Klasyfikacja	•	1. 1.			H

gdzie musimy wybrać zmienne oraz rodzaj krzywej regresji

💽 Estymacja krzywej	—
Iniczależna Zmienne zależne: Inicwa [zalezna1] Inicwa [zalezna2] Imiczależna Imiczależna Imiczależna <td< td=""><td>OK Wklej Resetuj Anuluj Pomoc</td></td<>	OK Wklej Resetuj Anuluj Pomoc
Etykiety obserwacji: 📝 Uwzględnij	stałą w równaniu rezentacja modeli
Modele	
🔽 Liniowy 📃 Kwadratowy 🔄 Złożony 📃 Wykładniczy	
🗖 Logarytmiczny 🔲 Sześcienny 👘 Krzywa S 👘 Logistyczny	
🔲 Odwrotny 🔲 Potęgowy 📄 Wzrostu Górna granica	
Wyswieti tabelę ANUVA	Zapisz

Dla pierwszej zmiennej zależnej zgonie z sugestią wykresu rozrzutu wyznaczamy prostą regresji. Otrzymujemy raport w którym mamy wyznaczoną interesującą nas prostą oraz dodatkowo poprzez wybranie stosownego pola mamy na wykres rozrzutu naniesioną naszą prostą.

Podsumowanie modelu i oszacowań parametrów

Zmienna zależna: liniowa

		Podsu	Oszacowania parametrów						
Równanie	R-kwadrat	F	df1	df2	Istotność	Stała	b1		
Liniowy	,899	425,710	1	48	,000	9,359	1,730		
Zmienna niezależna jest niezależna									

Zmienną niezależną jest niezależna.





W tym miejscu pojawia się naturalne pytanie z jaką dokładnością nasza wyznaczona prosta pokrywa się ze stanem faktycznym. Ponieważ dane do przykładu zostały dobrane w taki sposób, że zmienna zależna wyraża się za pomocą wzoru

$$y = 2x + 5 + \delta$$

gdzie δ przyjmuje wartości ±1 z prawdopodobieństwem $\frac{1}{2}.$ Natomiast SPSS wyznaczył następujący wzór

$$y = 1.73x + 9.359$$

Może nie jest to idealne przybliżenie, jest ono spowodowane stosunkowo małą liczbą obserwacji. Dla 500 obserwacji nasza estymowana krzywa przyjmuje postać.

Zmienna zależna: zalezna											
	Podsumowanie modelu Oszacowania parametrów										
Równanie	R-kwadrat	F	df1	df2	Istotność	Stała	b1				
Liniowy	,896	4313,028	1	498	,000	4,472	2,044				
7	Zerieses sizelaise isst sizelases										

Zmienną niezależną jest niezalezna.

i widzimy, że dopasowanie jest już znacznie lepsze.

Dla zmiennej zależnej2 otrzymujemy następujący wynik

Podsumowanie modelu i oszacowań parametrów

Zmienna zależna: kwadratowa

		Podsu	umowanie mo	Oszacowania parametrów						
Równanie	R-kwadrat	F	df1	df2	Istotność	Stała	b1	b2		
Kwadratowy	,998	10267,690	2	47	,000	22,257	-17,149	1,008		
Zmienną niezależną jest niezależna.										

oraz stosowną krzywą



kwadratowa

W rzeczywistości wzór ma następującą postać

$$y = x^2 - 17x + 22 + \delta.$$

Dla trzeciej zmienne prawdziwy związek zadany jest za pomocą funkcji

$$y = x^3 - 2x^2 - 5x - 2 + \delta$$

program natomiast proponuje nam następujący wzór oraz jego reprezentację

graficzną.

Podsumowanie modelu i oszacowań parametrów

Zmienna zależna: trzeciego stopnia wielomian

		Podsu	imowanie mo	odelu	Oszacowania parametrów							
Równanie	R-kwadrat	F	df1	df2	Istotność	Stała	b1	b2	b3			
Sześcienny	,970	750,114	2	47	,000	149,042	,000	-21,754	1,583			
	Zeriesen einerleiten inst einerleiten											

Zmienną niezależną jest niezależna.



trzeciego stopnia wielomian

Latwo spostrzec, że sugerowany wzór funkcji nie jest zbytnio zbliżony do stanu faktycznego, jest to spowodowane tym, że nasze wartości nie w pobliżu punktów charakterystycznych jakimi są wierzchołki. W ostatnim przypadku funkcja

zadane jest wzorem

$$y = \frac{e^x}{100000}$$

natomiast program szacuje nam krzywą w następujący sposób

$$y = e^{0.095x}$$

której reprezentacja graficzna przyjmuje postać



wykładnicza

W czasie ćwiczeń prześledzimy również inne przykładowe dane, w tym również inne rodzaje krzywych regresji oraz ich kombinacje.