

Wstęp do testów nieparametrycznych

Wstęp do testów nieparametrycznych

Parametryczne vs nieparametryczne

Większość poznanych przez Państwa procedur statystycznych obejmuje estymację parametrów rozkładu populacji, z której pochodziła próba lub próby. Procedury te opierają na pewnych założeniach dotyczących rozkładu populacji. Na przykład, jednym z założeń testu t jest to, że populacja, z której pochodzą próby (lub próba) ma rozkład normalny. Takie testy (opierające się na estymacji parametrów i/lub założeniach dotyczących rozkładu) nazywać będziemy testami **parametrycznymi**. Testy **nieparametryczne** nie wymagają normalności rozkładu, ale mają własne założenia.

Kiedy użyć testu nieparametrycznego?

Przede wszystkim powinniśmy się zastanowić, czy nasze dane spełniają założenia testu parametrycznego, którego chcemy użyć. Jeżeli chcemy użyć testu t , to musimy zadać sobie pytanie - na ile uprawnieni jesteśmy, żeby zakładać, że rozkład danej zmiennej w populacji jest normalny.

Kiedy możemy rozsądnie założyć, że rozkład w populacji nie jest normalny?

- kiedy nasza zmienna jest na skali porządkowej
- kiedy mamy jakieś obserwacje odstające (których nie możemy się pozbyć)
- kiedy widać to na jednym z wykresów diagnostycznych (histogram, wykres kwantyl-kwantyl)

Kiedy jeszcze możemy użyć testów nieparametrycznych?

- kiedy bardziej niż średnia interesuje nas mediana
- kiedy mamy bardzo małe grupy

Kiedy nie jest rozsądnie założyć, że rozkład populacji nie jest normalny?

- kiedy mamy dużą próbę i jakiś test na normalność odrzuci nam hipotezę, że dane pochodzą z rozkładu normalnego, ale w rzeczywistości rozbieżność nie jest duża (testy parametryczne są dość odporne na lekką skośność itp.)

Testy na rangach

Test rangowanych znaków Wilcoxon dla prób zależnych

Założmy, że chcielibyśmy przetestować hipotezę zgodnie z którą długofalowe bieganie redukuje ciśnienie krwi. Aby przetestować tę hipotezę zmierzylismy ciśnienie krwi naszym uczestnikom a następnie kazalismy im systematycznie biegać przez 6 miesięcy, po których ponownie zmierzylismy ciśnienie krwi.

Naszą zmienną zależną jest w tym wypadku *zmiana ciśnienia krwi po 6 miesiącach*. Jeżeli bieganie rzeczywiście obniża ciśnienie krwi, to oczekujemy że większość badanych uzyska niższy wynik przy drugim pomiarze. Oczekujemy więc dodatniej różnicy między pomiarami. Oczekujemy również że wśród tych, u których ciśnienie krwi zwiększyło się, zmiana ta nie będzie duża.

Z drugiej strony, jeżeli bieganie jest nieefektywne, to oczekujemy że około połowa różnic będzie dodatnia, a połowa ujemna, a różnice dodatnie będą mniej-więcej takie same jak te ujemne, jeśli chodzi o ich odległość od zera.

Założmy, że opisany powyżej eksperyment dał następujące wyniki:

```
przed <- c(130, 170, 125, 170, 130, 130, 125, 160)
po <- c(120, 163, 120, 135, 143, 136, 124, 120)
roznica <- przed - po

print(przed)
```

```
[1] 130 170 125 170 130 130 125 160
```

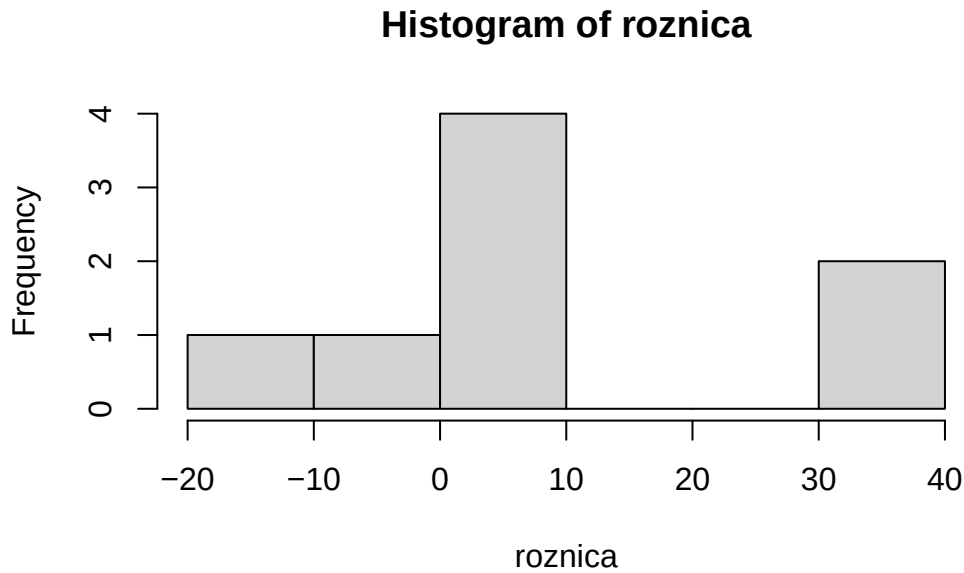
```
print(po)
```

```
[1] 120 163 120 135 143 136 124 120
```

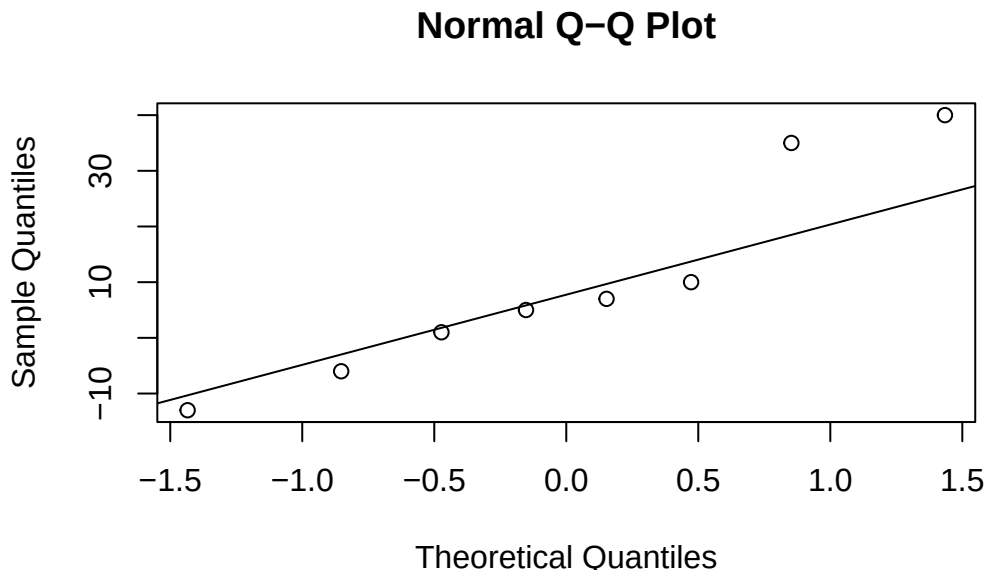
```
print(roznica)
```

```
[1] 10 7 5 35 -13 -6 1 40
```

```
hist(roznica)
```



```
qqnorm(roznica)  
qqline(roznica)
```



Przyjmijmy, że nie możemy (albo nie chcemy!) zakładać, że nasze obserwacje pochodzą z rozkładu normalnego. Co możemy w takiej sytuacji zrobić? Czy jedyne co nam pozostaje, to poddać się? Nie! Tak jak powiedzieliśmy sobie wcześniej, jeśli bieganie nie ma wpływu na ciśnienie, to oczekujemy (odrobinę upraszczając), że połowa naszych różnic będzie ujemna, a

połowa będzie dodatnia. Mówiąc bardziej precyzyjnie, nasza hipoteza zerowa głosi, że różnice będzie symetrycznie rozłożone wokół zera, hipoteza alternatywna (którą chcemy potwierdzić!) mówi zaś, że nie będą symetrycznie rozłożone wokół zera.

Na początek wykluczmy z naszej próby wszystkie te obserwacje, w których różnica wynosiła dokładnie zero. W naszej próbie nie ma takich obserwacji, ale przeprowadzając test Wilcozona należy to zrobić.

```
print(roznica)
```

```
[1] 10  7  5 35 -13 -6  1 40
```

```
roznica <- roznica[roznica != 0]
```

Następnie uporządkujemy nasze obserwacje od tych, gdzie wartość bezwzględna różnic jest najmniejsza do tej, gdzie jest największej.

```
sorted_roznica <- roznica[order(abs(roznica))]  
print(sorted_roznica)
```

```
[1]  1  5 -6  7 10 -13 35 40
```

Następnie przypiszemy każdej obserwacji 1 jeśli różnica jest dodatnia i -1 jeśli różnica jest ujemna (jest to funkcja *signum* nazywana także w skrócie *sgn*).

```
sgn_roznica <- ifelse(sorted_roznica > 0, 1, -1)  
print(sgn_roznica)
```

```
[1]  1  1 -1  1  1 -1  1  1
```

Następnie każdej naszej obserwacji przypiszemy rangę według wartości bezwzględnej różnicy. Jeżeli dwie lub więcej wartości są takie same, przypiszemy im średnią z zajmowanych rang. Na przykład dwie jednakowe wartości na miejscach 2 i 3 otrzymają rangę 2.5. Funkcja `rank()` zrobi to za nas.

```
ranks <- rank(abs(sorted_roznica))  
print(ranks)
```

```
[1] 1 2 3 4 5 6 7 8
```

Naszą statystyką testową (którą nazwiemy V) będzie suma iloczynów *signum* różnicy i jej rangi dla tych obserwacji, dla których *signum* jest dodatnie.

```
V <- sum(sgn_roznica[sgn_roznica > 0] * ranks[sgn_roznica > 0])
print(V)
```

```
[1] 27
```

Jaki rozkład ma tak zdefiniowana zmienna losowa? Sprawa nie jest prosta - dla małych liczebności prób rozkład ten możemy wyliczyć (ale jest to dość zasobochłonne!) dla dużych ($n > 50$) korzysta się czasami z przybliżenia z rozkładu normalnego. My jednak w ten problem nie będziemy się wgłębiać i posłużymy się odpowiednią funkcją wbudowaną w R w celu obliczenia wartości p :

```
psignrank(V-1, length(roznica), lower.tail = F)
```

```
[1] 0.125
```

Powyższy sposób obliczania wartości p za pomocą `psignrank()` dotyczy naszego przykładu bez remisów, czyli bez jednakowych wartości bezwzględnych różnic. Przy remisach skorzystajmy z funkcji `wilcox.test()`, która je uwzględnia.

Nic oczywiście nie stoi na przeszkodzie, aby posłużyć się tutaj funkcją `wilcox.test` wbudowaną w R:

```
wilcox.test(przed, po, paired = TRUE)
```

Wilcoxon signed rank exact test

data: przed and po

V = 27, p-value = 0.25

alternative hypothesis: true location shift is not equal to 0

```
wilcox.test(przed, po, paired = TRUE, alternative = 'greater')
```

Wilcoxon signed rank exact test

data: przed and po

V = 27, p-value = 0.125

alternative hypothesis: true location shift is greater than 0

UWAGA W literaturze spotkają się Państwo z nieco innym sposobem obliczania statystyki testowej w tym przypadku (i inną jej nazwą, W). Proszę uwierzyć, że te sposoby są ekwiwalentne (oczywiście jeśli korzystamy z właściwego rozkładu!)

Niestety, okazuje się, że 6 miesięcy biegania nie sprawi, że nagle staniemy się okazami zdrowia! A szkoda...

Test U Manna-Whitneya/Test sumy rank Wilcoxona dla prób niezależnych

Rozważmy następujące dane dotyczące liczby stresujących wydarzeń w ostatnim czasie wśród pacjentów oddziału kardiologicznego i ortopedycznego. Wiemy, że stresującym zdarzeniom w życiu (małżeństwo, nowa praca, śmierć partnera, urodzenie dziecka) często towarzyszą choroby i jest rozsądnie oczekiwać, że pacjenci oddziału kardiologicznego przeważnie doświadczyli większej liczby stresujących wydarzeń, niż pacjenci oddziału ortopedycznego. Mamy pewne podejrzenia, że wyniki na skali stresu mogą nie mieć symetrycznego rozkładu w populacji (szczególnie w przypadku pacjentów kardiologicznych, jeżeli nasza hipoteza potwierdziłaby się) dlatego użyjemy testu nieparametrycznego.

Tym razem nie będziemy przeprowadzać obliczeń ręcznie - są dość podobne do obliczeń wykonanych w poprzednim zadaniu. Zainteresowani doczytają w jaki sposób przeprowadzić ten test “na papierze”. My posłużymy się funkcją `wilcox.test`.

```
orto <- c(10, 12, 17, 13, 19, 20)
kardio <- c(30, 26, 25, 33, 18, 27)
data_ex <- data.frame(stres = c(orto, kardio),
                      grupa = c(rep('orto', length(orto)), rep('kardio', length(kardio))))

wilcox.test(stres ~ grupa, alternative = "greater", data = data_ex)
```

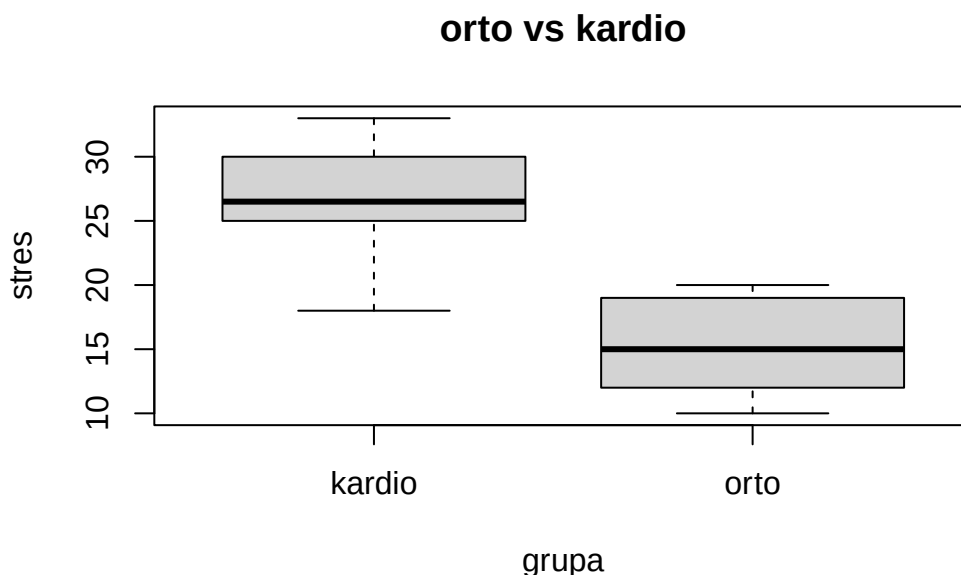
Wilcoxon rank sum exact test

data: stres by grupa

W = 34, p-value = 0.004329

alternative hypothesis: true location shift is greater than 0

```
boxplot(stres ~ grupa, data = data_ex, main = "orto vs kardio")
```



Zgodność sędziów

Kappa Cohena

Założmy, że poprosiliśmy sędziego ze znacznym doświadczeniem klinicznym o przeprowadzenie wywiadów z 30 nastolatkami i zakwalifikowanie każdego z nich jako przejawiających: (1) brak problemów behawioralnych, (2) internalizację problemów behawioralnych, (3) eksternalizację problemów behawioralnych.

Każdy, kto recenzowałby naszą pracę zastanawiałby się nad wiarygodnością naszych pomiarów - skąd wiemy, że sędzia jest w tym co robi lepszy, niż gdyby po prostu zgadywał na chybił-trafił? Aby wykluczyć tę możliwość, poprosiliśmy drugiego sędziego o porównywalnym wykształceniu oraz doświadczeniu zawodowym o przeprowadzenie niezależnej klasyfikacji na podstawie transkrypcji tych samych wywiadów. Po zebraniu danych, to znaczy przeprowadzeniu wywiadów oraz ich ocenie przez dwóch sędziów, utworzyliśmy tabelę krzyżową pokazującą zgodę (oraz jej brak) między naszymi oceniającymi.

```
# Potrzebny nam jest pakiet `irr` (interrater reliability)
#install.packages('irr')

judge_1 = c(rep('brak problemu',15), rep('internalizacja', 2), rep('eksternalizacja', 3),
            rep('brak problemu', 1), rep('internalizacja', 3), rep('eksternalizacja', 2),
            rep('brak problemu', 0), rep('internalizacja', 1), rep('eksternalizacja', 3))
judge_2 = c(rep('brak problemu',20), rep('internalizacja', 6), rep('eksternalizacja',4))
```

```
kategoryzacja = data.frame(przypadek = 1:30, sedzia_1 = judge_1, sedzia_2 = judge_2)
kategoryzacja
```

	przypadek	sedzia_1	sedzia_2
1	1	brak problemu	brak problemu
2	2	brak problemu	brak problemu
3	3	brak problemu	brak problemu
4	4	brak problemu	brak problemu
5	5	brak problemu	brak problemu
6	6	brak problemu	brak problemu
7	7	brak problemu	brak problemu
8	8	brak problemu	brak problemu
9	9	brak problemu	brak problemu
10	10	brak problemu	brak problemu
11	11	brak problemu	brak problemu
12	12	brak problemu	brak problemu
13	13	brak problemu	brak problemu
14	14	brak problemu	brak problemu
15	15	brak problemu	brak problemu
16	16	internalizacja	brak problemu
17	17	internalizacja	brak problemu
18	18	eksternalizacja	brak problemu
19	19	eksternalizacja	brak problemu
20	20	eksternalizacja	brak problemu
21	21	brak problemu	internalizacja
22	22	internalizacja	internalizacja
23	23	internalizacja	internalizacja
24	24	internalizacja	internalizacja
25	25	eksternalizacja	internalizacja
26	26	eksternalizacja	internalizacja
27	27	internalizacja	eksternalizacja
28	28	eksternalizacja	eksternalizacja
29	29	eksternalizacja	eksternalizacja
30	30	eksternalizacja	eksternalizacja

```
table(kategoryzacja$sedzia_1, kategoryzacja$sedzia_2)
```

	brak problemu	eksternalizacja	internalizacja
brak problemu	15	0	1
eksternalizacja	3	3	2

Moglibyśmy po prostu policzyć, w jakiej liczbie przypadków nasi sędziowie się zgadzali i zobaczyć, czy jest to więcej niż 50% przypadków. Byłoby to jednak złe podejście. Gdyby np. każdy z sędziów losowo w 80% przypadków stawiał określoną diagnozę, to wówczas mielibyśmy dość dużą szansę na to, że między nimi wystąpi dość duża zgoda, ale będzie to wciąż wynik przypadku a nie faktycznej zbieżności opinii.

```
library(irr)
```

Loading required package: lpSolve

```
agree(kategoryzacja[, c("sedzia_1", "sedzia_2")])
```

Percentage agreement (Tolerance=0)

```
Subjects = 30  
Raters = 2  
%-agree = 70
```

Sposobem na poradzenie sobie z tym jest obliczenie wartosci Kappa (κ) Cohena. Jest to zgodność między sędziami z poprawką na to, że jej część może być efektem przypadku. Możemy nawet potraktować ją jako statystykę testową i sprawdzić jakie jest jej prawdopodobieństwo przy hipotezie zerowej ($\kappa = 0$, czyli cała zaobserwowana zgodność jest wynikiem przypadku) ale proszę pamiętać, że nawet małe wartości będą statystycznie istotne.

W tym artykule znajdują Państwo pewne reguły interpretacji Kappy dla nauk biomedycznych (a podobne artykuły dotyczące innych dziedzin łatwo znaleźć w internecie):

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/>

```
kappa2(kategoryzacja[, c("sedzia_1", "sedzia_2")])
```

Cohen's Kappa for 2 Raters (Weights: unweighted)

```
Subjects = 30  
Raters = 2  
Kappa = 0.473  
  
z = 3.68  
p-value = 0.000232
```

Kappa Fleissa

A co by się stało, gdybyśmy mieli nie 2 a 3 sędziów? Musielibyśmy wówczas użyć innej miary - zamiast Kappy Cohena użyjemy wtedy Kappy Fleissa. Jej interpretacja jest analogiczna.

```
judge_3 = c(rep('brak problemu',17), rep('internalizacja', 8), rep('eksternalizacja',5))
kategoryzacja2 = data.frame(sedzia_1 = judge_1, sedzia_2 = judge_2, sedzia_3 = judge_3)
kappam.fleiss(kategoryzacja2)
```

Fleiss' Kappa for m Raters

Subjects = 30

Raters = 3

Kappa = 0.589

z = 7.6

p-value = 2.86e-14

Nieparametryczna “analiza wariancji”

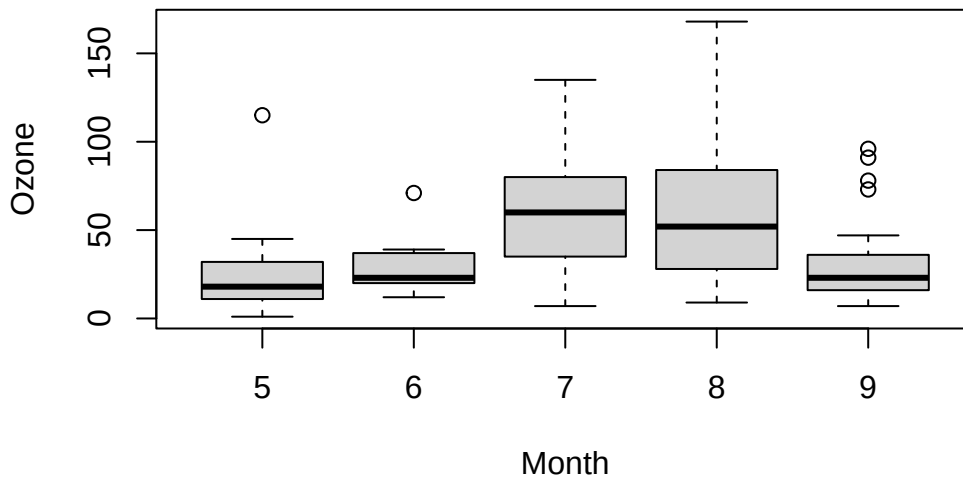
Test Kruskala-Wallisa

Jest to nieparametryczny odpowiednik analizy wariancji, z tą różnicą, że zamiast na wartościach pracujemy na rangach, tak jak w przypadku testu sumy rang Wilcoxa (test Kruskala-Wallisa jest jego uogólnieniem).

Założmy, że chcielibyśmy dowiedzieć się czy w którymś miesiącu roku poziom ozonu jest większy, niż w innych. W tym celu od maja do września przeprowadzaliśmy pomiary stężenia ozonu w powietrzu. Założmy również, że mamy dobre powody aby uznać, że rozkład, z którego pochodzą nasze obserwacje odbiega od rozkładu normalnego. Dodatkowo mamy pewne powody, żeby mniejszą wagę przyznawać obserwacjom skrajnym - to, że w jakimś miesiącu jeden czy dwa dni charakteryzowały się dużym stężeniem ozonu nie powinno mieć dużego znaczenia.

W tym przypadku hipoteza zerowa głosi, że obserwacje w każdym miesiącu pochodzą z tego samego rozkładu.

```
boxplot(Ozone ~ Month, data = airquality)
```



```
kruskal.test(Ozone ~ Month, data = airquality)
```

Kruskal-Wallis rank sum test

data: Ozone by Month

Kruskal-Wallis chi-squared = 29.267, df = 4, p-value = 6.901e-06

Sam test Kruskala-Wallisa nie powie nam które miesiące statystycznie odbiegają od innych. Mówi nam tylko, że przynajmniej jedna próba pochodzi z innego rozkładu niż inna próba. Jak dowiedzieć się które to próby i które miesiące są istotnie różne od których? Musimy posłużyć się testem *post-hoc*. W przypadku testu Kruskala-Wallisa musimy wybrać taki, który również operuje na rangach. W tym przypadku odpowiedni jest test Dunna.

```
#install.packages('dunn.test')
library(dunn.test)
dunn.test(airquality$Ozone, airquality$Month)
```

Kruskal-Wallis rank sum test

data: x and group

Kruskal-Wallis chi-squared = 29.2666, df = 4, p-value = 0

Dunn's Pairwise Comparison of x by group
(No adjustment)

Col Mean-

Row Mean	5	6	7	8
6	-0.925159 0.1774			
7	-4.419470 0.0000*	-2.244208 0.0124*		
8	-4.132813 0.0000*	-2.038635 0.0207*	0.286657 0.3872	
9	-1.321202 0.0932	0.002539 0.4990	3.217199 0.0006*	2.922827 0.0017*

= 0.05

Reject H_0 if $p < \alpha/2$, where $p = \Pr(Z > |z|)$

W tabeli w każdej komórce mamy (na górze) wartość statystyki testowej oraz (na dole) wartość p . Z tabeli możemy, analizując przecięcia kolumn i wierszy, odczytać, że statystycznie istotną różnicę możemy zaobserwować między 5 i 7 miesiącem, między 5 i 8, między 6 i 7, 6 i 8, 7 i 9 oraz 8 i 9.

Wariacje na temat χ^2

Test McNemara

Wyobraźmy sobie, że chcemy sprawdzić, czy uczestnictwo w kursie “Statystyka z R” zwiększa sympatię do statystyki. W tym celu przeprowadziliśmy badanie - spytaliśmy grupę 64 studentów kognitywistyki przed i po kursie o ich uczucia do statystyki.

Moglibyśmy skorzystać ze zwykłego testu niezależności χ^2 , ale nie uwzględnilibyśmy wtedy faktu, że pomiary można połączyć w pary. Zrobilibyśmy to wtedy tak:

```
statystyka <- matrix(c(36, 28, 45, 19),
  nrow = 2,
  dimnames = list("Przed statystyką" = c("Przed statystyką", "Po statystyce"),
                  "Po statystyce" = c("Lubię", "Nie lubię")), byrow = TRUE)
statystyka
```

	Po statystyce
Przed statystyką	Lubię Nie lubię

Przed statystyką	36	28
Po statystyce	45	19

```
chisq.test(statystyka)
```

Pearson's Chi-squared test with Yates' continuity correction

```
data:  statystyka
X-squared = 2.1518, df = 1, p-value = 0.1424
```

Możemy również sparować obserwacje i postawić trochę inną hipotezę - że częściej zmiana była w kierunku nie lubię -> lubię niż lubię -> nie lubię.

Hipoteza zerowa stwierdza więc, że prawdopodobieństwo zaklasyfikowania do komórki $[i,j]$ i $[j,i]$ jest takie samo.

```
statystyka <- matrix(c(33, 12, 3, 16),
  nrow = 2,
  dimnames = list("Przed statystyką" = c("Lubię", "Nie lubię"),
    "Po statystyce" = c("Lubię", "Nie lubię")))
statystyka
```

	Po statystyce	
Przed statystyką	Lubię	Nie lubię
Lubię	33	3
Nie lubię	12	16

```
mcnemar.test(statystyka)
```

McNemar's Chi-squared test with continuity correction

```
data:  statystyka
McNemar's chi-squared = 4.2667, df = 1, p-value = 0.03887
```

Jak widzimy, w przypadku testu niezależności χ^2 różnica nie jest statystycznie istotna, a w przypadku testu MacNemara jest. Jest tak, ponieważ używając tego drugiego, używamy dodatkowej informacji o obserwacjach. W praktyce skutkuje to większą mocą testu. Warto więc użyć testu McNemara, jeśli nasz schemat eksperymentalny zakłada dwa pomiary!

Regresja logistyczna (od strony R)

Statystyczna istotność współczynników w regresji logistycznej

Większość z państwa będzie uczyć się regresji logistycznej na zajęciach ze Statystyki II. Tutaj chciałbym Państwu pokazać tylko przykład, jak można za pomocą R dopasować odpowiedni model do danych i jak odczytać wyniki odpowiednich testów dotyczących współczynników.

Założmy, że chcielibyśmy dowiedzieć się, które czynniki miały wpływ na to, czy w katastrofie Titanica ktoś przeżył, czy nie. Tak się składa, że dysponujemy dość dokładnymi danymi z katastrofy (pakiet `titanic`) i możemy spróbować odpowiedzieć na to pytanie. Poniżej znajduje się objaśnienie poszczególnych zmiennych.

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes

pclass: A proxy for socio-economic status (SES)

1st = Upper

2nd = Middle

3rd = Lower

```
#install.packages('titanic')
library(titanic)
head(titanic_train)
```

PassengerId	Survived	Pclass	
1	1	0	3
2	2	1	1
3	3	1	3
4	4	1	1
5	5	0	3

6	6	0	3			
				Name	Sex	Age SibSp Parch
1				Braund, Mr. Owen Harris	male	22 1 0
2				Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38 1 0
3				Heikkinen, Miss. Laina	female	26 0 0
4				Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35 1 0
5				Allen, Mr. William Henry	male	35 0 0
6				Moran, Mr. James	male	NA 0 0
	Ticket	Fare	Cabin	Embarked		
1	A/5 21171	7.2500		S		
2	PC 17599	71.2833	C85	C		
3	STON/O2. 3101282	7.9250		S		
4	113803	53.1000	C123	S		
5	373450	8.0500		S		
6	330877	8.4583		Q		

```
model = glm(Survived ~ Sex + Age + Pclass + Fare + Embarked + SibSp,
             family = 'binomial',
             data = titanic_train)
summary(model)
```

Call:

```
glm(formula = Survived ~ Sex + Age + Pclass + Fare + Embarked +
     SibSp, family = "binomial", data = titanic_train)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	17.926254	607.935670	0.029	0.97648
Sexmale	-2.616787	0.217336	-12.040	< 2e-16 ***
Age	-0.043310	0.008229	-5.263	1.42e-07 ***
Pclass	-1.210216	0.162953	-7.427	1.11e-13 ***
Fare	0.001152	0.002429	0.474	0.63536
EmbarkedC	-12.288495	607.935446	-0.020	0.98387
EmbarkedQ	-13.104016	607.935649	-0.022	0.98280
EmbarkedS	-12.692074	607.935424	-0.021	0.98334
SibSp	-0.379557	0.124713	-3.043	0.00234 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 964.52 on 713 degrees of freedom
Residual deviance: 632.58 on 705 degrees of freedom
(177 observations deleted due to missingness)
AIC: 650.58

Number of Fisher Scoring iterations: 13