Statystyka opisowa

Wykład 4:
Elementy wnioskowania statystycznego
Analiza zależności pomiędzy dwiema zmiennymi

Idea wnioskowania statystycznego

Celem analizy statystycznej nie jest zwykle tylko opisanie (prezentacja) posiadanych danych, czyli tzw. próby statystycznej.

Najczęściej informacje z próby powinny mają pozwolić na uzyskanie wniosków o całej populacji, która nie mogła (z różnych względów) być poddana badaniu w całości.

Zdefiniowane w poprzednim zdaniu cele tzw. wnioskowania statystycznego są możliwe do realizacji poprzez użycie odpowiednich narzędzi analitycznych – testów statystycznych i/lub przedziałów ufności.

Po co nam wnioskowanie statystyczne?

Rozważmy następującą informację – w sondażu partia *ABC* uzyskała 25% poparcia. Sondaż był przeprowadzony na 1000-osobowej próbie – zakładamy, że reprezentatywnej.

Nadchodzą wybory… Zakładamy, że preferencje wyborców się znacząco nie zmienią. Czy oznacza to, że w wyborach partia *ABC* otrzyma też dokładnie 25%? Zapewne nie, ale pewnie będzie to wartość zbliżona? Ale jak bardzo – czy na przykład wynik poniżej 21% jest też realny?

Odpowiedzi na takie pytania to już tak zwana statystyka matematyczna, służąca do wnioskowania statystycznego.

W Internecie znaleźć można proste narzędzia statystyczne *on-line*, które pozwolą nam udzielić odpowiedzi na to pytanie.

Na przykład na stronie *https://www.sample-size.net/confidence-interval-proportion/* można wyliczyć tzw. przedział ufności (p.u.) dla poparcia partii ABC w całej populacji wyborców.

Proszę sprawdzić, że otrzymamy wynik (dla poziomu ufności 95%):

95% p.u. dla oczekiwanego wyniku wyborów – (22,3%; 27,8%).

A zatem uzyskanie wyniku poniżej 21% jest mało prawdopodobne.

Testy statystyczne

Testy statystyczne służą do oceny, czy zależności (ogólniej – pewne prawidłowości) zaobserwowane w próbie są efektem ogólniejszej zasady obowiązującej w całej populacji czy tylko przypadkowym rezultatem.

Wynikiem testu statystycznego jest prawdopodobieństwo testowe (*p*), którego niskie wartości świadczą o istotności statystycznej rozważanej zależności.

Przyjmuje się przy tym najczęściej następujące reguły:

- **gdy $p < 0{,}05$ mówimy o statystycznie istotnej zależności (oznaczamy ten fakt za pomocą *);**
- **$p < 0{,}01$ to wysoce istotna statystycznie zależności (**);**
- **$p < 0{,}001$ to bardzo istotna statystycznie zależność (***).**

Konstrukcja testu statystycznego

Ideę wnioskowania statystycznego można w uproszczeniu opisać jako następujący proces:

1) **przyjmujemy, iż badane zjawisko podlega pewnemu modelowi;**

2) **realizujemy eksperyment (na przykład badanie ankietowe) i stwierdzamy na ile jego wyniki są zgodne z założonym modelem;**

3) **jeżeli zgodność wyników otrzymanych w eksperymencie z założonym modelem jest mała, przyjmujemy, iż założenie przyjęte w p. 1) było błędne – badane zjawisko nie funkcjonuje zgodnie z założonym modelem.**

Prosty test statystyczny

Celem analizy jest zweryfikowanie, czy moneta jest symetryczna, czyli prawdopodobieństwo wyrzucenia orła i reszki jest jednakowe ($p = ½$).

1) **Model badanego zjawiska: prawdopodobieństwo wyrzucenia orła i reszki jest jednakowe i wynosi ½;**

2) **Wynik przykładowego eksperymentu (wykonano 8 rzutów monetą):**

O O O O O O O O

3) **Zgodność uzyskanego wyniku z założonym modelem jest bardzo mała – prawdopodobieństwo uzyskania takiego wyniku wynosi: $(1/2)^8 = 0{,}0039$. Ponieważ jest ono bardzo niewielkie, więc można domniemywać, iż moneta nie jest symetryczna.**

Wątpliwości...

Procedura testowania hipotez statystycznych niesie ze sobą wiele ograniczeń, o których należy wiedzieć i pamiętać.

1) Wniosek płynący z wyniku testu statystycznego nie musi być prawdziwy – nawiązując do przedstawionego przykładu, nawet dla symetrycznej monety jest możliwe uzyskanie na przykład następującego wyniku:

O O O O O O O O O O O O O O O O O O O O O O O O O O

2) Procedury wielu testów wymagają przyjęcia pewnych dodatkowych założeń o rozkładzie wartości badanej cechy (lub kilku cech) w całej populacji – bardzo często słuszności tych założeń nie da się jednoznacznie wykazać.

Ważne pojęcia i informacje

Oto ważne pojęcia, które są związane z procedurą testowania hipotez:

1) **model opisujący funkcjonowanie badanego zjawiska nazywany jest hipotezą zerową;**

2) **alternatywę modelu badanego zjawiska zawiera hipoteza alternatywna (najczęściej będąca prostym zaprzeczeniem hipotezy zerowej);**

3) **błąd polegający na odrzuceniu hipotezy zerowej mimo tego, że jest ona prawdziwa nazywa się błędem I rodzaju (prawdopodobieństwo tego błędu jest oznaczane symbolem α i nazywane jest też poziomem istotności);**

4) **Testy statystyczne są skonstruowane w taki sposób, że poziom istotności α jest ustalony na poziomie zbliżonym do 0 (zwyczajowo: $\alpha = 0{,}05$).**

Ważne pojęcia i informacje

5) przyjęcie hipotezy zerowej mimo, że jest ona nieprawdziwa nazywane jest błędem II rodzaju (prawdopodobieństwo popełnienia tego błędu zależy od poziomu istotności α i rodzaju hipotezy alternatywnej i jest oznaczane symbolem β).

6) Prawdopodobieństwo odrzucenia hipotezy zerowej (co oznacza na przykład stwierdzenie, że moneta jest niesymetryczna jeżeli tak faktycznie jest) jeżeli jest ona rzeczywiście nieprawdziwa wynosi $1-\beta$ i jest nazywane mocą testu

Test niezależności chi-kwadrat

Test niezależności chi-kwadrat jest najpopularniejszym testem statystycznym służącym do badania zależności między dwiema cechami zmierzonymi na skali nominalnej. W teście tym stawiana jest hipoteza zerowa, że wystąpienie wariantu jednej cechy nie zależy od wartości przyjętej dla drugiej cechy (brak związku pomiędzy obiema cechami).

Niskie wartości prawdopodobieństwa testowego *p* pozwalają hipotezę tę odrzucić i wnioskować o istnieniu zależności w całej populacji pomiędzy dwiema rozważanymi cechami.

Przykład testowanie hipotez z programem *STATISTICA*

Analiza dotyczy zbioru danych *Opinie o integracji z UE (2004)*.

Celem analizy jest zbadanie wpływu wybranych czynników społeczno-ekonomicznych na sposób głosowania w referendum akcesyjnym.

Pod uwagę wzięto płeć, wykształcenie oraz miejsce zamieszkania respondentów.

Analizę przeprowadzono za pomocą analizy *TABELE WIELODZIELCZE* uzupełniając wyniki analizy opisowej próbą odpowiedzi na pytanie, czy wnioski z próby są wiarygodne, czy dadzą się uogólnić na całą populację. W tym celu zastosowano test niezależności chi-kwadrat.

Podstawowe wyniki analiz

Wybieramy polecenie *STATYSTYKA* / *STATYSTYKI PODSTAWOWE I TABELE* a następnie *TABELE WIELODZIELCZE*. W oknie wyboru zmiennych na jednej liście wskazujemy czynnik *WYKSZTAŁCENIE* a na drugiej *SPOSÓB GŁOSOWANIA W REFERENDUM*…

Następnie przechodzimy do okna *WYNIKÓW TABELARYZACJI* i wywołujemy wstępną tabelę pokazującą rozkład liczbowy sposobu głosowania w zależności od poziomu wykształcenia respondentów.

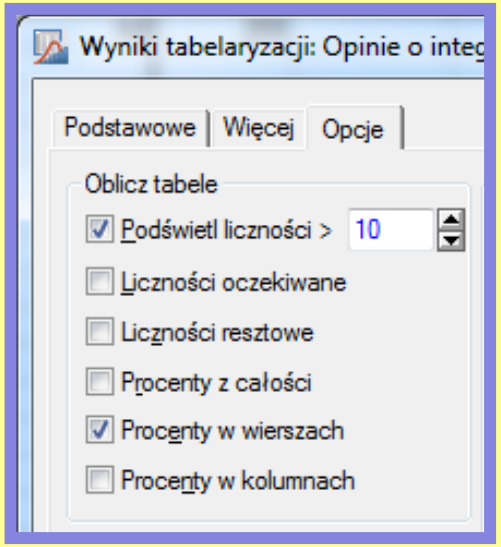
Wykształcenie	Sposób głosowania w referendum europejskim tak	Sposób głosowania w referendum europejskim nie	Wiersz Razem
podstawowe	5	11	16
średnie	87	39	126
wyższe	53	14	67
Ogół	145	64	209

Przedstawione wyniki nie pozwalają jednak w łatwy sposób porównać częstości głosowania „na tak” i „na nie” w referendum europejskim, choć widzimy, że na pewno wyróżnia się grupa osób z wykształceniem podstawowym.

Aby ułatwić wnioski, należy wyznaczyć strukturę procentową odpowiedzi.

Struktura procentowa

Przywracają okno analiz (zminimalizowane u dołu ekranu lub korzystając z użytecznego skrótu CTRL + R) włączamy zakładkę opcje i dokonujemy wyboru sposobu wyznaczania struktury procentowej.



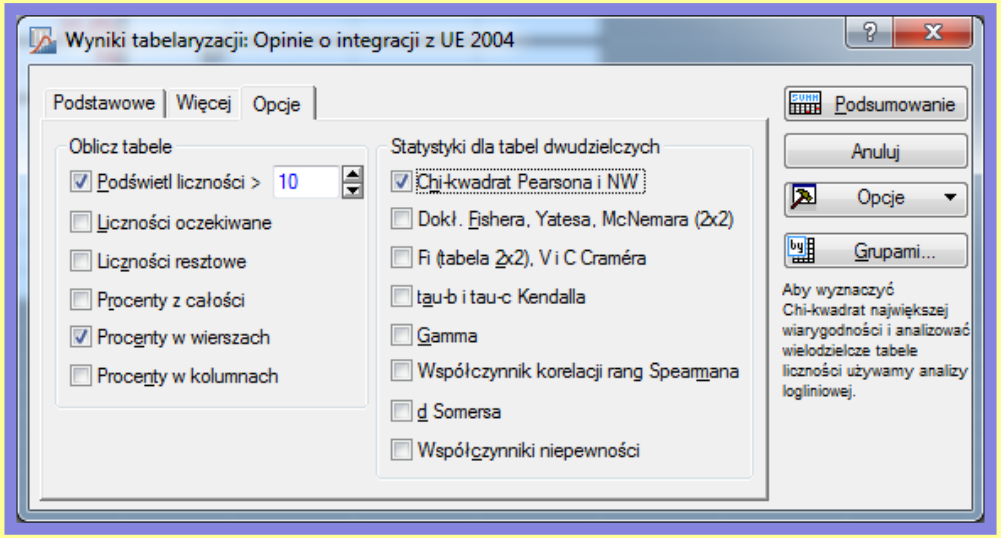
Procenty wierszowe

Wykształcenie	Sposób głosowania w referendum europejskim tak	Sposób głosowania w referendum europejskim nie	Wiersz Razem
podstawowe	5	11	16
%wiersza	31,3%	68,8%	
średnie	87	39	126
%wiersza	69,0%	31,0%	
wyższe	53	14	67
%wiersza	79,1%	20,9%	
Ogół	145	64	209

Na schemacie pokazano sposób wyznaczania procentów wierszowych. W analogiczny sposób (w układzie wertykalnym) wyznacza się procenty kolumnowe. Wybór rodzaju struktury procentowej nie jest rzeczą łatwą i spotyka się tu dużo błędów – generalnie, strukturę wyznacza się względem tej cechy, która ma charakter sprawczy, jest czynnikiem niezależnym (w naszym przykładzie, to wykształcenie może wpływać na sposób głosowania a nie na odwrót).

Wnioskowanie statystyczne

Przywracając ponownie okno analiz (zminimalizowane u dołu ekranu lub korzystając z użytecznego skrótu *CTRL + R*) włączamy zakładkę opcje i zaznaczamy opcję, wywołującą wyniki testu niezależności chi-kwadrat.



Aby wywołać wyniki testu musimy włączyć zakładkę *WIĘCEJ* i wybrać polecenie *DOKŁADNE TABELE DWUDZIELCZE*.

	Statystyka: Wykształcenie(3) x Sp		
statystyka	Chi-kwadr.	df	p
Chi^2 Pearsona	13,93845	df=2	p=,00094
Chi^2 NW	13,02965	df=2	p=,00148

Wynik testu niezależności chi-kwadrat wynosi *p* = 0,0009*. Zgodnie z wcześniej podanymi regułami, oznacza to, iż fakt zróżnicowania sposobu głosowania ze względu na posiadane wykształcenie nie jest przypadkowy i zapewne znajduje swoje odzwierciedlenie w całej populacji. Możemy wnioskować, iż wyższe wykształcenie było czynnikiem stymulującym poparcie dla członkostwa Polski w UE.**

Prezentacja zależności pomiędzy dwiema cechami

<u>Dla dwóch cech jakościowych:</u> TABELE WIELODZIELCZE

Analiza powinna obejmować stworzenie dwuwymiarowej tabeli liczebności, w której dodatkowo zostałyby wyznaczone struktury procentowe wg wierszy i/lub kolumn. Porównanie tych struktur pozwala na wyciągnięcie wniosków o istnieniu lub braku zależności pomiędzy obiema cechami.

Prezentacja graficzna w postaci skategoryzowanego wykresu kołowego, histogramu lub histogramu trójwymiarowego.

Metoda została omówiona na poprzednich stronach oraz w materiałach laboratoryjnych nr 6

Prezentacja zależności pomiędzy dwiema cechami

<u>Dla dwóch cech liczbowych:</u> ANALIZA KORELACJI

Analiza polega na wyznaczeniu współczynnika korelacji liniowej (r) i interpretacji siły oraz kierunku zależności.
Prezentacja graficzna w postaci wykresu rozrzutu.

Dokładne omówienie metody znajduje się w materiałach do zajęć laboratoryjnych nr 8

Prezentacja zależności pomiędzy dwiema cechami

Dla cechy jakościowej i liczbowej: ANALIZA PRZEKROJÓW

Idea analizy sprowadza się do wyznaczenia statystyk opisowych dla cechy liczbowej (zwanej też zależną) w kategoriach wyznaczonych przez wartości cechy jakościowej (niezależnej, grupującej). Porównanie wartości średnich (a także innych miar) pozwala wyciągnąć wnioski o tym, czy pomiędzy obiema cechami występuje jakiś związek.

Ilustracja graficzna za pomocą skategoryzowanego wykresu średnich lub wykresu ramkowego.

Dokładne omówienie metody znajduje się w materiałach do zajęć laboratoryjnych nr 7