STATYSTYKA OPISOWA (lab. 6)
ZALEŻNOŚCI POMIĘDZY DWIEMA CECHAMI NOMINALNYMI

WPROWADZENIE DO ZAJĘĆ NR 6-8

Analiza danych nie kończy się zwykle na rozważaniu każdej cechy z osobna – tak jak to robiliśmy do tej pory za pomocą *Tabel liczności* czy *Statystyk opisowych*. Analitykowi danych stawia się bardziej skomplikowane pytania. Oto przykładowe problemy, które wymagają takiej złożonej analizy:

- w badaniach ankietowych (np. typowych sondażach „politycznych") grupuje się odpowiedzi respondentów według popieranej przez nich partii politycznej, ale także bada się rozkład preferencji politycznych względem płci, grupy wiekowej, wykształcenia i innych cech respondentów – po to, by znaleźć grupy wyborców, do których należy skierować szczególnie intensywną kampanię polityczną;
- w badaniach przekrojowych, dotyczących na przykład poziomu przestępczości w krajach Unii Europejskiej, bada się wpływ na tę cechę innych czynników, takich jak poziom bezrobocia, PKB per capita czy wskaźnik migracji po to, by określić wpływ polityki gospodarczej i społecznej na poziom bezpieczeństwa publicznego;
- w medycynie, w badaniach klinicznych porównuje się wyniki leczenia lekiem testowanym (w grupie badanej) w stosunku do leku referencyjnego (w tzw. grupie kontrolnej) – tu cel badania jest oczywisty i nie wymaga komentarza.

WYBÓR METODY ANALIZY ZALEŻNOŚCI POMIĘDZY DWIEMA CECHAMI:

Wybór metody badania relacji pomiędzy dwiema cechami zależy od ich **charakteru** – **liczbowego** bądź **nominalnego** (tekstowego). Istotne jest opanowanie następujących zasad:

- dla **dwóch cech nominalnych** tworzy się dwuwymiarową tabelę liczności, w której wyznacza się strukturę procentową odpowiedzi na jedno pytanie względem wariantów drugiej cechy – w programie *STATISTICA* jest to analiza ***Tabele wielodzielcze***;
- w przypadku, gdy **jedna cecha ma charakter liczbowy a druga nominalny**, wyznacza się statystyki opisowe (średnią, medianę i inne) dla cechy liczbowej względem wariantów cechy nominalnej – w programie *STATISTICA* służy do tego analiza ***Przekroje (ANOVA)***;
- kiedy **obie cechy mają charakter liczbowy**, wyznacza się współczynnik korelacji, który pozwala określić siłę i kierunek zależności – w programie *STATISTICA* służy do tego analiza ***Macierze korelacji***.

INFORMACJA O TESTACH STATYSTYCZNYCH:

W praktyce, nasze wnioski wyciągane na podstawie analizy zebranych danych powinny wykraczać poza te dane. Na przykład, robiąc sondaż polityczny ankietujemy 1000 osób, ale wnioski chcemy odnieść do całej populacji wyborców. Do takiego uogólniania wyników służą tzw. **testy statystyczne (oraz metody estymacji przedziałowej)**. Ich zastosowanie pozwala stwierdzić, czy zależności zaobserwowane w próbie są efektem ogólniejszej prawidłowości panującej w całej populacji czy tylko przypadkowym rezultatem.
Wynikiem testu statystycznego jest tzw. **prawdopodobieństwo testowe (*p*)**, którego niskie wartości świadczą o **istotności statystycznej** rozważanej zależności. Przyjmuje się przy tym następujące reguły:

- gdy $p \geq 0{,}05$ nie stwierdzamy istotnej statystycznie zależności, nasza analiza nie pozwala wysnuć takich wniosków i pozostajemy przy hipotezie, że analizowane cechy nie są powiązane.
- $p < 0{,}05$ mówimy o statystycznie istotnej zależności (często oznacza się ten fakt za pomocą symbolu *);
- $p < 0{,}01$ to wysoce istotna zależności (**);
- $p < 0{,}001$ to bardzo wysoko istotna statystycznie zależność (***).

Przykład 1 (*Opinie o integracji z UE 2004*)

Badanie przeprowadzono bezpośrednio po wejściu Polski do UE w lecie 2004 r. Zanim przejdziemy do poznania nowego narzędzia analizy danych, proszę stosując poznane już metody grupowania danych (*Tabele liczności*) proszę odpowiedzieć na następujące pytania:

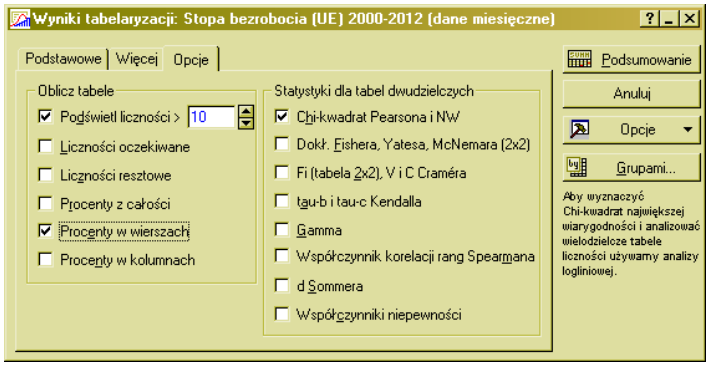
- ile osób …… i jaki procent ankietowanych osób ……% obawiał się podwyżek cen żywności po wejściu Polski do UE;
- ile osób …… i jaki procent ankietowanych osób ……% obawiał się podwyżek cen energii elektrycznej po wejściu Polski do UE;

ANALIZA ZALEŻNOŚCI POMIĘDZY DWIEMA CECHAMI NOMINALNYMI:

Celem dalszej analizy będzie porównanie poglądów na te kwestie wśród mieszkańców miast i wsi. Ponieważ rozważane cechy mają charakter **nominalny** (tekstowy), więc właściwym narzędziem będzie analiza ***Tabel wielodzielczych***.

1) Za pomocą polecenia ***Statystyka*** wybieramy grupę analiz ***Statystyki podstawowe i tabele*** a następnie ***Tabele wielodzielcze***. Określamy tabelę (wybieramy zmienne) – na pierwszej liście wskazując *Miejsce zamieszkania* a na drugiej *Czy obawiasz się wzrostu cen żywności?* – i przechodzimy do okna wyników za pomocą przycisku *OK*.

2) Wywołujemy tabelę wynikową za pomocą przycisku ***Podsumowanie***. Dowiadujemy się, jak mieszkańcy miast i wsi odpowiadali na rozważane pytanie – jednak porównanie opinii obu tych grup jest trudne, bez wyznaczenia struktury procentowej. Wznawiamy więc analizę (Ctrl + R) i ponieważ miejsce zamieszkania ustawione jest „w wierszach", w zakładce ***Opcje*** zaznaczamy ***Procenty w wierszach***. Zaznaczamy także test ***Chi-kwadrat Pearsona i NW***. Aby wywołać tabele wynikowe nie posługujemy się przyciskiem ***Podsumowanie*** lecz w zakładce ***Więcej*** naciskamy ***Dokładne tabele dwudzielcze***.



3) W dwóch oddzielnych tabelach wynikowych znajdują się liczności i procenty oraz wynik testu niezależności chi-kwadrat. Wyniki zapisujemy w poniższej tabeli i opisujemy charakter zależności, jeżeli takowa występuje (wartość *p* interpretujemy według reguł podanych na poprzedniej stronie).

Uwaga: W arkuszu wynikowym podane są dwie (zwykle zresztą zbliżone do siebie) wartości *p* dla dwóch wersji testu niezależności – umawiamy się, że podawać będziemy wartość pierwszą. Wartość *p* w arkuszu jest podana w sposób niezbyt czytelny, dlatego proszę podawać ją w formie ułamka dziesiętnego zaokrąglonego do czterech miejsc po przecinku.

Uwaga: Wartości procentowe proszę podawać zaokrąglone **do jednego miejsca po przecinku!**

Miejsce zamieszkania	**Czy obawiasz się wzrostu cen żywności? (*p* =)**		Razem
	tak	nie	
miasto	…… (…… %)	…… (…… %)	……
wieś	…… (…… %)	…… (…… %)	……
Razem	……	……	……

Opis i interpretacja wartości *p* oraz struktury procentowej odpowiedzi .....
.....
.....
.....

Przykład 2 (*Opinie o integracji z UE 2004*)

W analogiczny sposób proszę zbadać zależności pomiędzy:

- *płcią* i *obawą przed wzrostem cen żywności*;
- *sytuacją finansową* i *sposobem głosowania w referendum europejskim*.

Liczności i wartości procentowe proszę umieścić w poniższych tabelach, podobnie wartości *p* – proszę zinterpretować uzyskane wyniki. Dla każdej tabeli należy przemyśleć sposób wyznaczania struktury procentowej – według wierszy czy według kolumn? Istotny jest również sposób wyboru zmiennych – proszę tak wybierać zmienne na liście 1 i 2, żeby układ tabeli był zgodny z tym zamieszczonym poniżej.

Czy obawiasz się wzrostu cen żywności?	**Płeć (*p* =)**		Razem
	mężczyzna	kobieta	
tak	…… (…… %)	…… (…… %)	……
nie	…… (…… %)	…… (…… %)	……
Razem	……	……	……

Opis i interpretacja wartości *p* oraz charakteru analizowanej zależności .....
.....
.....

Sposób głosowania w referendum europejskim	**Sytuacja finansowa (*p* =)**					Razem
	raczej dobra	dobra	średnia	raczej zła	zła	
tak	…… (…… %)	…… (…… %)	…… (…… %)	…… (…… %)	…… (…… %)	……
nie	…… (…… %)	…… (…… %)	…… (…… %)	…… (…… %)	…… (…… %)	……
Razem	……	……	……	……	……	……

Opis i interpretacja wartości *p* oraz charakteru analizowanej zależności .....
.....
.....