Instytut Sterowania i Systemów Informatycznych
Uniwersytet Zielonogórski

Analiza danych

Ćwiczenie 2: Statystyka opisowa

Zadanie 1. (plik **bol_glowy.sas7bdat**) W trakcie badań klinicznych nowego leku na ból głowy, zaaplikowano lek 32 pacjentom. Po miesiącu wypełnili oni formularz oceniający jakość życia po zażywaniu leku (quality-of-life, QOL), w którym podawali wyniki w skali 100-punktowej. Wysokie wyniki świadczą o wysokiej jakości życia, niskie wyniki – o niskiej jakości życia. Standaryzowany średni wynik odniesienia to 50 punktów.

a) Wypisać zawartość pliku (**Zadania i programy użytkowe ►Zadania ►Dane ►Listowanie danych**)

b) Narysować histogram (**Zadania i programy użytkowe ►Zadania ►Wykresy ►Histogram**). Na ile przedziałów SAS podzielił zakres zmienności QOL? Do czego służy zakładka lista wyboru **Skala** w części **ROLE** zakładki **USTAWIENIA**?

c) Który przedział ma najmniejszą liczebność? Jaki jest jego punkt środkowy? Jaki jest przedział z największą liczebnością? Jaki jest jego punkt środkowy?

d) Powtórzyć poprzedni podpunkt dla 8, 12 i 3 słupków (w celu zmiany liczby słupków użyć części **KUBEŁKI** w zakładce **WYGLĄD**). Wytłumaczyć zaobserwowane zmiany. Jak dobierać liczbę przedziałów (zob. materiały z wykładu)?

Zadanie 2. Powtórzyć poprzednie zadanie dla poniższych plików:

(a) Plik **recepty.sas7bdat** zawiera dane z badania wydatków amerykańskich emerytów na realizację recept (średnio 1200 USD rocznie). Badanie wykonano na próbie 44 losowo wybranych emerytów. Przeanalizować histogramy, w której wysokość słupków określa częstości w procentach.

(b) Plik **program.sas7bdat** zawiera czasy (w minutach) pojedynczych sesji pewnego programu, które monitorował administrator sieci uniwersyteckiej. Tym razem na histogramie przedstawić proporcje.

Zadanie 3. Plik **groch.sas7bdat.** Celem badań było porównanie plonów dwóch gatunków grochu (A i B). W tym celu każdemu z gatunków przypisano losowo 20 działek, na których go posadzono. Plik zawiera plony w buszelach na akr (takich jednostek używa się w USA).

a) Sprawdzić, że plik zawiera dwie kolumny i zinterpretować ich zawartość.

b) Narysować na jednym wykresie dwa histogramy, każdy dla jednego gatunku (wykorzystać wybór kolumny w **Grupuj analizowane dane** w części **DODATKOWE ROLE** role zakładki **USTAWIENIA**). Co można z nich wywnioskować?

c) Narysować jeden histogram dla 40 wyników niezależnie od gatunku grochu.

Zadanie 4. Plik **hantle.sas7bdat**. Celem badania było określenie wpływu wieku na sprawność fizyczną wśród Amerykanów w wieku od 60 do 89 lat. Jednym z kryteriów była liczba podniesień hantli o masie 3 kg w ciągu 30 sekund (zmienna **arm_culrs**). W badaniu wzięło udział 30 osób. Narysować i porównać histogramy dla trzech grup wiekowych.

Zadanie 5. (plik **bol_glowy.sas7bdat**, zob. Zadanie 1)

a) Określić liczbę obserwacji, średnią, odchylenie standardowe, wartość minimalną i wartość maksymalną (**Zadania i programy użytkowe ►Zadania ►Dane ►Charakterystyka danych**).

b) Obliczyć (np. w Excelu lub za pomocą kalkulatora w Windows) $\bar{x}-s$ oraz $\bar{x}+s$. Jaki procent obserwacji leży między tymi wartościami? Aby to określić, proszę wykorzystać **Zadania i programy użytkowe ►Zadania ►Dane ►Filtrowanie danych**, budując odpowiedni warunek i tworząc w bibliotece WORK nowy plik z wartościami z przedziału $[\bar{x}-s, \bar{x}+s]$. Następnie za pomocą **Zadania i programy użytkowe ►Zadania ►Dane ►Charakterystyka danych** określić liczbę tych wartości i podzielić przez liczbę wartości z oryginalnego pliku.

c) Obliczyć $\bar{x}-2s$ oraz $\bar{x}+2s$. Jaki procent obserwacji leży między tymi wartościami?

d) Obliczyć $\bar{x}-3s$ oraz $\bar{x}+3s$. Jaki procent obserwacji leży między tymi wartościami?

Zadanie 6. Powtórzyć poprzednie zadanie dla plików **recepty.sas7bdat** oraz **program.sas7bdat**, dla każdego gatunku grochu z pliku **groch.sas7bdat**, a także dla każdej z grup wiekowych z pliku **hantle.sas7bdat**.

Zadanie 7. Dla wszystkich dotychczas użytych plików obliczyć następujące parametry: wartość minimalna, P_1, P_5, P_{10}, Q_1, mediana, Q_3, P_{90}, P_{95}, P_{99}, wartość maksymalna (*wskazówka*: przypomnieć sobie, co to są kwartyle i percentyle). W tym celu proszę użyć **Zadania i programy użytkowe ►Zadania ►Statystyka ►Statystyki agregujące** zwracając szczególną uwagę na zakładkę OPCJE.

Zadanie 8. (pliki **groch.sas7bdat** oraz **hantle.sas7bdat**)

a) Zapisać pięcioliczbowe podsumowanie dla tych danych (wartość minimalna, Q_1, mediana, Q_3, wartość maksymalna).

b) Narysować wykres pudełkowy (inaczej: ramkowy) identyfikując ewentualne obserwacje odstające (zob. **Zadania i programy użytkowe ►Zadania ►Wykresy ►Wykres pudełkowy**). Romb blisko mediany oznacza oszacowanie średniej populacji, z której pochodzą dane, a jego szerokość – przedział ufności dla tej średniej. Będzie to szerzej omówiona na jednych z kolejnych zajęć.

Zadanie 9. (pliki **tv.sas7bdat**) Plik zawiera oceny programów nadawanych w TV w środy wieczorem. TVG oznacza, że program nadaje się dla wszystkich (*ang.* general audience). TVPG oznacza, że wskazany jest nadzór rodziców (*ang.* parental guidance). TV14 przestrzega przed oglądaniem przez widzów poniżej 14 roku życia bez nadzoru rodziców.

a) Utworzyć tabelkę i histogramy reprezentujące szereg rozdzielczy punktowy (zob. **Zadania i programy użytkowe ►Zadania ►Statystyka ►Jednoczynnikowe liczebności**). Omówić rezultaty.

b) Narysować wykres kołowy (zob. **Zadania i programy użytkowe ►Zadania ►Wykresy ►Wykres kołowy**).

Zadanie 10. (pliki **wczesna_edukacja.sas7bdat**) Plik zawiera dane z badania umiejętności poznawczych 30 dzieci uczestniczących w sześciu różnych programach edukacyjnych.

a) Płeć jest zakodowana jako **gender** (0 – dziewczynka, 1 – chłopiec). Utworzyć szereg rozdzielczy punktowy oraz narysować wykres kołowy dla tej zmiennej.

b) Zastosowane kodowanie płci jest mało czytelne. Lepiej byłoby posługiwać się nazwami: '**chłopiec'** lub '**dziewczynka'** (w SAS dostępne są tylko dwa typy danych: liczbowy i napisowy). Można to zrobić w dwóch krokach. Najpierw za pomocą **Zadania i programy użytkowe ►Zadania ►Dane ►Przekształcanie danych** zamienia się zmienną o wartościach liczbowych 0 i 1 na zmienną o wartościach znakowych '0' i '1' (w tym celu w polu **przekształcenie niestandardowe** należy wpisać **put(gender, 1.)** co oznacza użycie funkcji zamieniającej wartości gender na napis o długości jednego znaku z zerową liczbą miejsc po przecinku; drugi parametr **put** oznacza format stosowany podczas zamiany). W drugim kroku wykorzystuje się **Zadania i programy użytkowe ►Zadania ►Dane ►Kodowanie wartości** aby zakodować zmienną znakową utworzoną przed chwilą na napisy **chłopiec** lub **dziewczynka** (zakładka **WARTOŚCI**). Proszę zwracać uwagę na to, gdzie te operacje zapisują zbiory wynikowe, bo nieuwaga jest dobrą okazją do wystąpienia błedów.

c) Powtórzyć punkt (a) dla zmiennej **ed_level** oznaczającej stopień wykształcenia opiekuna.