Instytut Sterowania i Systemów Informatycznych
Uniwersytet Zielonogórski

Podstawy analityki biznesowej

Ćwiczenie 2: Statystyka opisowa i proste wnioskowanie

Niniejszy zestaw ćwiczeń oparto o środowisko SAS Studio. Dostęp do niego można uzyskać na trzy sposoby:

1. wykorzystanie wersji programu zainstalowanej na serwerze uczelnianym (zasady dostępu w laboratorium i spoza sieci uczelnianej omówi prowadzący zajęcia);
2. ściągnięcie i zainstalowanie SAS University Edition (http://www.sas.com/pl_pl/software/university-edition.html);
3. dostęp poprzez przeglądarkę do zasobów *SAS OnDemand for Academics* (informacje na stronie http://support.sas.com/ondemand/, rejestracja na stronie https://odamid.oda.sas.com/SASODARegistration/)

Zadanie 1. W eksperymencie przeprowadzonym w Australii w latach siedemdziesiątych wkrótce po wprowadzeniu tam systemu metrycznego, grupa 44 studentów szacowała szerokość sali wykładowej w metrach. Druga grupa, złożona z 69 studentów, szacowała tę samą odległość, ale w stopach. Rzeczywista odległość wynosiła 13.1 m (43 stopy). W zadaniu chcemy określić czy oszacowania w metrach różnią się od oszacowań w stopach, oraz które oszacowania są dokładniejsze.

(a) Wyniki eksperymentu zawiera plik **widths.sas7bdat**. Zaimportować go do folderu **sasue** utworzonego na poprzednich zajęciach i wyświetlić jego zawartość. Zmienna **guess** oznacza oszacowanie odległości, a odpowiednia wartość zmiennej **units** – jednostkę.

(b) Niedogodnością zbioru danych są różne jednostki wyników, co utrudnia ich porównywanie. Utwórzmy więc nową zmienną, w której wszystkie 44 + 96 odległości będzie wyrażonych w stopach. W tym celu, w panelu nawigacyjnym wybierzmy **Pliki i foldery na serwerze**, a następnie przycisk **Nowy** i dalej **Program SAS-owy** (lub **F4**). W oknie należy wpisać poniższy kod (prowadzący zajęcia powinien objaśnić jego działanie):

```
data work.widths;
   set sasue.widths;
   if units = 'metres' then feet = guess * 3.28;
   else feet = guess;
run;
```

Proszę go wykonać wybierając przycisk z biegnącym ludzikiem. W rezultacie, w folderze roboczym **work** pojawi się nowy plik, będący kopią **widths**, ale z dodatkową zmienną **feet**.

(c) Otworzyć **Zadania ►Statystyka ►Statystyki agregujące**. W **Dane ►Dane** wybrać **works.widths**. W **Dane ►Role** wybrać **feet** jako zmienną analizowaną oraz **units** jako zmienną klasyfikującą, po czym kliknąć **Uruchom**.

(d) Jak oceny studentów w metrach i stopach mają się do rzeczywistej odległości? Jaki wniosek można stąd wyciągnąć?

(e) Wnioski wyciągane tylko na podstawie średniej i odchylenia standardowego mogą być mylące w sytuacji występowania obserwacji odstających, które zniekształcają te parametry. Bardziej miarodajne mogą być wykresy pudełkowe. Ich narysowanie umożliwiają **Statystyki agregujące** (poprzez wybór **Opcje ►Wykresy ►Porównawczy wykres pudełkowy**), ale proszę również spróbować otworzyć **Zadania ►Wykresy ►Wykres pudełkowy,** po czym w **Dane ►Dane** wybrać **work.widths**, w **Dane ►Role** wybrać **feet** jako zmienną analizowaną oraz **units** jako zmienną klasyfikującą, a na końcu kliknąć **Uruchom**. Czy występują obserwacje odstające? Które oceny są bardziej asymetryczne? Które oceny odległości charakteryzują się większym rozproszeniem?

(f) Wykresy pudełkowe świetnie sprawdzają się w porównywaniu własności rozkładów pomiarów w różnych grupach. Alternatywą może być porównanie *histogramów*. Proszę otworzyć **Zadania ►Statystyka ►Analiza rozkładu**. W **Dane ►Dane** wybrać **works.widths**. W **Dane ►Role** wybrać **feet** jako zmienną analizowaną. W **Opcje ►Eksploracja danych ►Histogram ►Zmienne klasyfikujące** dodać **units** i kliknąć **Uruchom**. Który rozkład jest bardziej skośny?

Zadanie 2. Ćwiczenie wprowadza w zagadnienie testowania hipotez w środowisku SAS Studio.

(a) Proszę otworzyć **Zadania ►Dane ►Filtrowanie danych**. W **Dane ►Dane** dodać **works.widths**. W **Dane ►Filtr 1** przypisać **units** jako zmienną 1, ustawić **Porównanie** jako **Równe**, oraz wpisać **feet** w polu **Wartość**. W **Dane ►Zbiór wynikowy** wpisać **feet** i kliknąć **Uruchom**. W przypadku starszej wersji SAS Studio, nie mającej możliwości filtrowania danych, zamiast tego należy wpisać kod

```
data work.feet;
   set work.widths;
   where units = 'feet';
run;
```

Rezultatem jest tymczasowy zbiór danych **feet** w bibliotece **work**.

(b) Powtórzyć to samo dla wyników w metrach. Rezultatem powinien być zbiór danych **metres** w bibliotece **work**.

(c) Przetestujmy hipotezę zerową $H_0: \mu_f = 43.0$. Dotyczy ona ocen odległości w stopach. W tym celu zastosujemy jednopróbkowy test *t* Studenta. Proszę otworzyć **Zadania ►Statystyka ►Testy T**. W **Dane ►Dane** dodać **works.feet**. W **Dane ►Role ►Zmienna analizowana** dodać **guess**. Zwrócić uwagę na to, że **Test jednopróbkowy** jest wartością domyślną. W oknie **Opcje ►Testy ►Hipoteza alternatywna: mi^=** wpisać **43** i odznaczyć **Testy normalności**. Kliknąć **Uruchom**. Jak potraktować otrzymaną *p*-wartość (liczbę wyświetlaną pod nagłówkiem **Pr. > |t|**) w odniesieniu do hipotezy zerowej? Co na ten sam temat mówi 95% przedział ufności dla średniej?

(d) Powtórzyć poprzedni punkt dla obserwacji wyrażonych w metrach, testując hipotezę zerową $H_0: \mu_m = 13.1$.

(e) Teraz chcemy przetestować hipotezę, że średnia ocen w stopach jest równa średniej ocen w metrach (po jej konwersji do stóp), czyli $H_0: \mu_m \times 3.28 = \mu_f$. Można byłoby sądzić, że skoro średnia obserwacji w stopach nie różni się wiele od prawdziwej odległości, a średnia obserwacji w metrach istotnie różni się od prawdziwej wartości, to średnie dwóch grup

mocno się różnią. Niekoniecznie tak jednak musi być. Poprawna ocena prawdziwości rozważanej hipotezy wymaga zastosowania testu *t* dla dwóch prób niezależnych. Proszę otworzyć **Zadania ►Statystyka ►Testy T**. W **Dane ►Dane** dodać **works.widths**. W **Dane ►Role ►Test t** wybrać **Test dwupróbkowy**.W **Dane ►Role ►Zmienna analizowana** dodać **feet**, a w **Dane ►Role ►Zmienne grup** dodać **units**. Kliknąć **Uruchom**. Jak potraktować otrzymaną *p*-wartość (liczbę wyświetlaną pod nagłówkiem **Pr. > |t|** w metodzie wariancji sumarycznej, odpowiadającej założeniu równych wariancji w obu populacjach) w odniesieniu do hipotezy zerowej? Co na ten sam temat mówi 95% przedział ufności dla różnicy średnich populacyjnych?

Zadanie 3. Zajmijmy się teraz sprawdzeniem spełnienia założeń wymaganych przez test *t* oraz alternatywami dla tego testu. Wspomniane założenia są następujace:

- obserwacje są próbkowane z populacji charakteryzującej się rozkładem normalnym;
- w przypadku dwupróbkowego testu *t* populacje mają identyczne wariancje;
- obserwacje wykonuje są niezależnie od siebie.

W praktyce, test *t* jest odporny na lekkie odstępstwa od tych założeń, jednak zawsze warto sprawdzić, na ile są one spełnione.

(a) Jeśli przed wybraniem **Uruchom** w ostatnim podpunkcie poprzedniego ćwiczenia w **Opcje ►Testy** zaznaczymy **Testy normalności**, wśród rezultatów zobaczymy dwie tabele (oddzielnie dla wyników w stopach i w metrach) z wynikami czterech różnych testów normalności danych. Zauważyć, że wszystkie *p*-wartości są bardzo małe. O czym to świadczy?

(b) Wyniki testów z punktu (a) są często mylące, stąd te testy stosuje się raczej rzadko. O wiele częściej podstawę do oceny dają nam odpowiednie wykresy. Proszę przeanalizować histogramy poniżej tabel, o których była mowa w punkcie (a). Na ich tle naniesione są m.in. najlepiej dopasowane krzywe gęstości rozkładu normalnego. Na ile histogramy wpasowują się w kształt gęstości prawdopodobieństwa rozkładu normalnego? Druga z krzywych określa ocenę funkcji gęstości prawdopodobieństwa rozkładu w oparciu o tzw. estymator jądrowy, który nie zakłada żadnej konkretnej postaci rozkładu.

(c) Poniżej histogramów proszę odszukać wykresy kwantylowe. Obrazują one jak zaobserwowane kwantyle odpowiadają kwantylom idealnego rozkładu normalnego. Dla rozkładu normalnego punkty wykresu powinny leżeć mniej więcej na linii prostej. Czy jest tak w przypadku rozważanych danych?

(d) Proszę odszukać tabelę z nagłówkiem **Równość wariancji**. Co o tej równości można powiedzieć na podstawie zamieszczonej poniżej *p*-wartości?

(e) Z dotychczasowej analizy wynika, że w pewnym stopnie naruszone jest założenie o normalności rozkładów, a wariancje populacji się różnią. Test *t* jest w dużej mierze na to odporny. Trudno jednak stwierdzić, do jakiego stopnia sytuację pogarszają obserwacje odstające.

(f) Nierówne wariancje wskazywane w punkcie (d) sugerują inny test służący sprawdzeniu, czy średnie obu populacji są równe: zamiast testu *t* skonstruowanego w oparciu o metodę wariancji sumarycznej, lepiej zastosować *test Satterthwaite'a*. Proszę odszukać odpowiadającą mu *p*-wartość. Jaki wniosek odnośnie hipotezy zerowej można dzięki temu sformułować?

(g) Alternatywą dla testu *t* Studenta, która nie zależy od założenia o normalności rozkładu, jest *test Wilcoxona-Manna-Whitneya*, opierający się o tzw. rangi obserwacji. Proszę go wykonać zaznaczając w **Opcje ►Testy** kwadrat **Test sumy rang Wilcoxona**, a następnie wybrać **Uruchom**. Proszę odnaleźć tabelę z nagłówkiem **Test Wilcoxona dwupróbkowy**, a w niej *p*-wartość dla hipotezy o równości wartości oczekiwanych (**Dwustronne pr. > |Z|**). O czym ona świadczy?

(h) Proszę spróbować powtórzyć zadanie po usunięciu najbardziej odstających obserwacji w oby grupach (*wskazówka:* wykorzystać filtrowanie danych z odpowiednim warunkiem). Czasem taki zabieg poprawia spełnialność założeń testów, jednak powinien być stosowany z dużą rozwagą. Zawsze należy sprawdzić pochodzenie usuwanej obserwacji odstającej, ponieważ może to być wartość całkowicie prawidłowa (wtedy jej nie usuwamy).

Zadanie 4. Ćwiczenie dotyczy danych zebranych podczas badań projektowych urządzenia do generowania energii elektrycznej z energii fal morskich. Eksperymenty wykonywano na modelu w zmniejszonej skali, w zbiorniku ze sztucznie wytwarzanymi falami. Celem był wybór metody cumowania modelu systemu, na który oddziałuje naprężenie zginające powstające w pewnej części urządzenia. Zbiornik może symulować szeroki wachlarz stanów morza (wzburzone, spokojne, umiarkowane, itd.). Model systemu badano dla różnych symulowanych stanów morza, dla każdego stanu testując dwie metody cumowania, z których jedna jest istotnie tańsza niż druga. Dane zawarte w pliku **waves.sas7bdat** opisują moment gnący (w niutonometrach, Nm). Celem naszej analizy będzie określenie, czy naprężenie zginające dla dwóch rozważanych metod cumowania się różni.

(a) Umieścić plik w bibliotece **sasue**, a następnie wypisać jego zawartość. Co oznaczają cztery zmienne w tej tabeli?

(b) Narysować wykresy pudełkowe momentów gnących dla każdej z metod z osobna, a następnie dla ich różnic. Wykorzystać w tym celu **Zadania ►Wykresy ►Wykres pudełkowy**. Co można powiedzieć o symetryczności rozkładów? Ile jest obserwacji? Czy przy takiej niewielkiej ich liczbie można wyciągać konkluzje co do postaci rozkładów? Czy w którymś z przypadków występują obserwacje odstające?

(c) Zwróćmy szczególną uwagę na to, że mamy jedną fundamentalną różnicę względem dotyczas analizowanych danych: obserwacje wystepują w parach, a każda para składa się z wyników dla dwóch metod cumowania dla *identycznego* stanu morza. W ten sposób istnieje duża szansa, że wyniki w obrębie jednej pary są ze sobą skorelowane. Dzięki temu, że wyniki w jednej parze odpowiadają temu samemu stanowi morza, w dużej mierze eliminujemy wpływ innych czynników na rezultaty, a porównaniu podlega jedynie efekt metody cumowania. Do testowania, że średnia różnica wyników jest zerowa, służy parzysty test *t*. Aby go wykonać, proszę otworzyć **Zadania ►Statystyka ►Testy T**. W **Dane ►Dane** dodać **works.waves**. W **Dane ►Role ►Test t** wybrać **Test parzysty**, w **Dane ►Role ►Zmienna grupy 1** dodać **method 1**, a w **Dane ►Role ►Zmienna grupy 2** dodać **method 2**. W **Opcje ►Testy** odznaczyć **Testy normalności**. W **Opcje ►Wykresy** wybrać **Wybrane wykresy**, a w nich **Histogram i wykres pudełkowy**. Wybrać **Uruchom**.Czy otrzymana *p*-wartość pozwala na stwierdzenie, że istnieje istotna różnica w naprężeniu gnącym dla obu metod cumowania? Co na ten temat mówi 95% przedział ufności dla różnicy wyników? Co na podstawie histogramu i wpisanej w niego gęstości prawdopodobieństwa rozkładu normalnego można powiedzieć o spełnieniu wymogu normalności danych?

(d) Przy tak niewielkiej liczbie obserwacji nie ma zbytniego sensu kreślenie wykresu

kwantylowego. Skoro więc trudno zweryfikować założenie o rozkładzie normalnym populacji, wykorzystajmy test sumy rang Wilcoxona. W tym celu powtórzmy wszystkie czynności z punktu (c), przed uruchomieniem wybierając w **Opcje ►Testy** kwadrat **Test znaków i test rangowanych znaków Wilcoxona**. Proszę odnaleźć tabelę z nagłówkiem **Testy położenia: mi0=0 Wilcoxona dwupróbkowy**, a w niej *p*-wartości dla hipotezy o równości wartości oczekiwanych. O czym one świadczą? Zauważyć, że w tej tabeli znajdują się również wyniki dla testu *t* Studenta.

Zadanie 5. Zbiór danych **babies** zawiera wagi 50 noworodków, które zapadły na syndrom samoistnej niewydolności oddechowej (ang. severe idiopathic respiratory distress syndrome, SIRDS). Jest on poważną chorobą mogącą prowadzić do śmierci, co też się stało w przypadku 27 dzieci. Proszę przeanalizować te dane w celu sprawdzenia, czy waga dzieci, które umierają, różni się od wagi dzieci, które przeżywają. Wstępną analizę proszę przeprowadzić w oparciu o techniki graficzne. Prosze również określić 95% przedział ufności dla różnicy średnich wag dzieci obu grup.

Zadanie 6. Zbiór danych **choles** zawiera wyniki badania wykonanego wśród 38 mężczyzn w średnim wieku i z nadwagą (powyżej 102 kg). Jednym z celów badania było określenie związku między typem zachowania a ryzykiem choroby wieńcowej. Zbiór danych zawiera poziom cholesterolu we krwi (w mg/100ml) dla dwóch typów zachowań, oznaczonych jako A i B. Typ A charakteryzuje pośpiech, agresja i duże ambicje, a typ B – zrelaksowanie, brak wygórowanych ambicji i mniej pośpiechu. Proszę przeanalizować, czy poziom cholesterolu zależy w jakiś sposób od wzorca zachowania. Zwrócić szczególną uwagę na sprawdzenie założeń wymaganych przez stosowane testy i obserwacje, które ewentualnie mogłyby zniekształcić konkluzje.

Zadanie 7. Dane zawarte w zbiorze danych **diet** pochodzą z badania diety Stillmana, która składa się przede wszystkim z protein i tłuszczów zwierzęcych, ograniczając przy tym spożycie węglowodanów. Dane dotyczą zawartości trójglicerydów we krwi (w mg/100ml) dla 16 uczestników badania przed rozpoczęciem diety (zmienna **baseline**) oraz po jej zakończeniu (zmienna **final**). Proszę zbadać, czy dieta Stillmana wpływa na poziom trójglicerydów.