

PROPONOWANE TEMATY BADAŃ STATYSTYCZNYCH

Rok akademicki 2015/2016, semestr letni

1. Badanie liczby pojazdów wjeżdżających w ciągu minuty w ulicę Berdychowo przez skrzyżowanie z al. Jana Pawła II. Ważne elementy tego zadania: sposób pobrania próby (sensowny, próba wystarczająco duża); próba oszacowania całkowitej liczby pojazdów przejeżdżających w ciągu doby i osobno w ciągu godzin porannych i popołudniowych. Przetestuj czy w godzinach porannych i popołudniowych ruch jest taki sam. Sugestia: Zadane skrzyżowanie można bardzo wygodnie obserwować z korytarza na 8. piętrze budynku WE PP. Nie ma konieczności obserwacji z narażaniem się na zmienne warunki pogodowe.
2. Badanie wypełnienia (liczby pasażerów w stosunku do liczby miejsc) samochodu osobowego jadącego w Poznaniu. Zastanowić się nad sensownym pobraniem próby (czy konieczne jest rozproszenie geograficzne badania?). Przetestować, czy średnie wypełnienie rano i popołudniu są takie same.
3. Oszacowanie średniej ceny książek obcojęzycznych w Księgarni Powszechnej (Stary Rynek w Poznaniu). Ważne elementy tego zadania: sposób pobrania próby (sensowny, dopasowany do struktury populacji; wystarczającego rozmiaru). Test, czy wartość książek obcojęzycznych jest różna od wartości książek polskojęzycznych.
4. Analiza liczby listów elektronicznych przychodzących na własną skrzynkę pocztową w ciągu dnia. Należy porównać dwie populacje: listy pożądane i spam.
5. Badanie liczby klientów wchodzących do wybranego supermarketu w ciągu minuty w zależności od pory dnia. Ważny element tego zadania: sposób pobrania próby, jej liczebność i reprezentatywność, podział populacji ze względu na płeć. Czy do supermarketu więcej wchodzi kobiet od mężczyzn?
6. Badanie płci osób wchodzących do pewnej drogerii. Ważną sprawą jest sensowne, reprezentatywne pobranie próby (pokrycie całego tygodnia pracy i całego dnia pracy). Odpowiedzieć na pytanie, czy procent mężczyzn wchodzących do drogerii jest taki sam, jak procent mężczyzn w populacji Polski.
7. Badanie polega na pomiarze czasu obsługi pojedynczego klienta przy kasie w hipermarkecie. Interesujący w tym badaniu jest średni czas trwania obsługi: czy jest dłuższy w weekendy niż w ciągu tygodnia? Konieczne byłoby dokładne zdefiniowanie tego, co jest mierzone (od kiedy do kiedy, np. od rozpoczęcia kasowania produktów aż do odejścia od kasy lub zakończenia ewentualnej rozmowy z kasjerem?) oraz co jest weekendem, a co dniem pracy.
8. Analiza średniego kosztu wykonania operacji na strukturach listowych w trzech różnych implementacjach. Należy zaprojektować program, który będzie wykonywał serię operacji wstawienia (add), usunięcia (remove) i pobrania (get) elementów listy. Jako typ elementu należy wybrać liczbę całkowitą (int). Następnie należy ów program wykonać na trzech różnych implementacjach list: biblioteki PCJ (<http://pcj.sourceforge.net/>, klasa IntList), biblioteki commons-collections (<http://jakarta.apache.org/commons/collections/>, klasa TreeList) oraz biblioteki standardowej języka Java (klasa java.util.ArrayList<Integer>). Czy średnie czasy wykonania programu dla tych trzech bibliotek różnią się? Która z bibliotek jest najszybsza?

9. Analiza średniego kosztu wykonania operacji na strukturach słownikowych (Map) w trzech różnych implementacjach. Należy zaprojektować program, który będzie wykonywał serię operacji wstawienia (put), usunięcia (remove) i pobrania (get) elementów słownika (Map). Jako typ elementu należy wybrać liczbę całkowitą (int), jako typ wartości dowolnie jaki obiekt (Object). Następnie należy ów program wykonać na trzech różnych implementacjach list: biblioteki PCJ (<http://pcj.sourceforge.net/>, klasa IntKeyMap), biblioteki commons-collections (<http://jakarta.apache.org/commons/collections/>, klasa HashedMap) oraz biblioteki standardowej języka Java (klasa java.util.HashMap<Integer>). Czy średnie czasy wykonania programu dla tych trzech bibliotek różnią się? Która z bibliotek jest najszybsza?
10. Porównanie dwóch dowolnych wyszukiwarek (np. Yahoo kontra Google, Haka kontra Altavista, Onet kontra Interia) ze względu na średnią pozycję pierwszej wybieranej strony przez użytkownika. Należy przygotować zestaw około 100 słów/hasel kluczowych. Każde słowo/hasło kluczowe jest wpisywane do dwóch badanych wyszukiwarek. Za każdym razem zapisywana jest pozycja pierwszej interesującej użytkownika strony.
11. Zaimplementuj w dwóch różnych językach programowania tablicę haszową, wykorzystując wszystkie cechy wybranych języków pozwalające na jak najbardziej wydajne działanie tablicy haszowej. Przygotuj około 50 instancji problemu polegających na wstawieniu, dodawaniu i usuwaniu elementów z tablicy haszowej. Która z implementacji jest bardziej wydajna?
12. Analiza średniego czasu tworzenia modelu przez wybrane metody z pakietu uczenia maszynowego WEKA (www.cs.waikato.ac.nz/ml/weka/). Należy zbadać w jakim stopniu wybrane metody są wrażliwe na liczbę atrybutów i przykładów opisujących analizowany zbiór danych.
13. Zbadać wiek artykułów (w godzinach lub dniach od daty publikacji - co bardziej sensowne) na stronie www.gazeta.pl w powiązaniu z liczbą wiadomości użytkowników (postów) dołączonych do niego. Pobrać próbę tylko z następujących kategorii: wiadomości, kraj, świat, sport, gospodarka, nauka, kultura - co najmniej 12 z każdej kategorii. Oszacować także średni wiek artykułu i średnią liczbę postów. Testowanie hipotez nie jest konieczne w tym zadaniu.
14. Zbadać strony informacyjne w bibliotece MSDN on-line (<http://msdn.microsoft.com/library>) pod kątem ich oceny dokonywanej przez użytkowników i liczby użytkowników, którzy oceniali stronę. Uwaga! Nie wszystkie strony mają tę cechę dostępną - pobranie próby powinno dotyczyć całej biblioteki, a te strony, które nie mają tych cech podanych, powinny być po wylosowaniu odrzucone. Oszacować także średnią ocenę strony i średnią liczbę oceniających osób. Testowanie hipotez jest niekonieczne w tym zadaniu.
15. Zbadać frakcję (procent występowania) stron w bibliotece MSDN on-line (<http://msdn.microsoft.com/library>), które mają dostępną informację o ocenie zawartości przez użytkowników i liczbie użytkowników, którzy dokonali oceny. Sprawdzić, czy stron takich jest więcej niż 50%.
16. Zbadać, czy w zestawie stron serwisu BBC (www.bbc.co.uk/a-z/) zawiera listę alfabetyczną (stron) jest co najmniej 30% stron poświęconych radiu?
17. Zbadać, czy w zestawie stron serwisu BBC (www.bbc.co.uk/a-z/) zawiera listę alfabetyczną (stron) jest co najwyżej 50% stron poświęconych telewizji?

18. Z badać, jaki procent opisów słów ze słownika angielskiego online Mirriam-Webster (www.m-w.com) dostępnych jest w wersji free tego słownika. Uwaga! Słownik ten posiada alfabetyczny indeks. Sprawdzić, czy dostępnych słów jest więcej niż 50%.
19. Z badać cenę drukarki laserowej w serwisie www.allegro.pl (interesują nas tylko obiekty niepromowane z dostępną ceną „Kup teraz”). Czy konieczne jest podzielenie całej populacji na podpopulacje (warstwy)? Ze względu na jaką pomocniczą cechę należy dokonać rozwarstwienia? Porównać średnie ceny w warstwach między sobą. Czy średnia cena drukarki jest większa niż 400 zł?
20. Z badać cenę książek w serwisie www.allegro.pl (interesują nas tylko obiekty niepromowane z dostępną ceną „Kup teraz”) w dwu kategoriach: literatura piękna i kryminały/thrillery. Czy konieczne jest podzielenie całej populacji na podpopulacje (warstwy)? Ze względu na jaką pomocniczą cechę należy dokonać rozwarstwienia? Porównać średnie ceny w warstwach między sobą. Czy średnia cena kryminału jest inna niż tomu literatury pięknej?
21. Z badać rozkład języków programowania używanych w implementacji projektów gier przechowywanych na www.sf.net. Każdy projekt gry (jest ich kilka tysięcy) ma przypisane języki programowania, w których jest realizowany; badaną cechą jest pierwszy z listy tych języków, jeśli jest ich więcej niż jeden. Interesujące w zadaniu pytania: czy rozkład języków dla gier jest taki sam, jak rozkład języków dla wszystkich projektów na www.sf.net, czy może znacząco inny? Rozmiar próby: co najmniej 60. Sposób losowania: losowanie proste ze zwracaniem lub bez (do wyboru).
22. Wśród studentów Informatyki na Politechnice Poznańskiej panuje powszechna opinia, że pierwszy i trzeci rok studiów jest najtrudniejszy, a drugi rok studiów jest dużo łatwiejszy. Na podstawie próbki przynajmniej 30 studentów, zbadaj czy średnia ocen z drugiego roku studiów jest większa od średniej ocen z pierwszego roku studiów.
23. Sprawdzić, czy tramwaje w Poznaniu się spóźniają. Zgodnie z wytycznymi MPK, tolerancja w stosunku do rozkładu jazdy wynosi 3 minuty opóźnienia. Oszacować średnie opóźnienie wraz z przedziałem ufności oraz sprawdzić za pomocą testu, czy średnie opóźnienie przekracza 3 minuty. Uwaga: przy zbieraniu danych można ograniczyć się do określonej pory dnia i/lub określonego przystanku.
24. Badanie zajętości sal wykładowych na Politechnice Poznańskiej w zależności od pory dnia i dnia tygodnia. Należy zbadać średnią zajętość sali rozumianą jako procent zajętych miejsc na sali. Ważne elementy tego zadania: sposób pobrania próby (sensowny, próba wystarczająco duża). Warto zwrócić uwagę na możliwość podziału populacji na warstwy ze względu na kierunek i rok studiów grup zajmujących sale.
25. Rynek gotowych dań (tzw. "słoiczków") dla niemowląt (dzieci do 1 roku życia) jest w Polsce opanowany przez 3 firmy: Bobo Vita, Gerber oraz Hipp. Z pobieżnego porównania cen wynika, że produkty pierwszej z wymienionych firm są najtańsze, a ostatniej -- najdroższe. Dokonaj weryfikacji takiej obserwacji na podstawie cen produktów tych firm zaobserwowanych w wybranym sklepie (np. Tesco lub Carrefour). Zwróć uwagę na podział populacji na warstwy, np. odpowiadające różnym typom dań (zupki, deserki, obiady) lub pojemnościom opakowań.