PEiAR Laboratorium 6 **08.04.2024**

WIELOPRÓBKI, FILTROWANIE, WYKRESY PUDEŁKOWE

Zad. 1
Ściągnij do katalogu Dokumenty plik Egz_1(csv). Zaimportuj pod nazwą *egzamin_* ramkę danych z jego danymi.
Pobierz z kolumny `punkty` trzydzieści próbek (każda 10-elementowa, pozwalamy na powtórzenia) i przechowaj pod zmienną *multiSample_*.

// W celu jednoczesnego wygenerowania wielu próbek, użyj funkcji *replicate(...)*, która jako I argument przyjmuje pożądaną liczbę próbek natomiast jako II argument – charakterystykę próbki tj. *sample*(..dane z których pobieramy próbkę.., ..liczebność próbki.., *replace*=T).//

Dla każdej z trzydziestu uzyskanych próbek policz średnią, odchylenie standardowe, medianę, wartość najmniejszą i największą. W tym celu, dla dogodniejszego odczytu danych w uzyskanej tablicy *multiSample_*, przetransponuj ją i zapisz pod nazwą *transposed_*.
Wygeneruj wektor *mean_* przechowujący średnie z wierszy tablicy *transposed_*. Analogicznie wygeneruj wektory *sd_* , *median_*, *min_* oraz *max_* i dołącz 5 kolumn ‘Mean’, ‘Sd’, ‘Median’, ‘Min’, ‘Max’ do ramki danych *transposed_*.
Narysuj pięć wykresów: 30 średnich, 30 odchyleń, 30 uzyskanych median, 30 wartości minimalnych i 30 wartości maksymalnych wg schematu:

1	3	4
2		5

Kolejne wykresy twórz np. przy użyciu polecenia *plot*(1:30, ...dane..., type=’l’ / ‘h’ / ‘o’).

Zad. 2
a) Dla danych z ramki *egzamin_* z poprzedniego zadania, oblicz średnią z uzyskanych punktów dla: kobiet, mężczyzn, potem dla każdej z grup A, B, C, D, E.
Użyj do tego funkcji filtrującej elementy ramki
subset(..nazwa ramki.., ..narzucony filtr.., *select* = ..nazwa kol. której wartości chcemy zwrócić..).
Zatem
> *womenScores* = *subset*(egzamin_, plec == ‘K’, select = ‘punkty’)
> *menScores* = …
> *groupAScores* = …
…
> *groupEScores* = …

Przedstaw na jednym wykresie słupkowym średnie dla analizowanych podgrup danych podpisując słupki na poziomej osi: ‘women’, ‘men’, ‘A’, ‘B’, ‘C’, ‘D’, ‘E’. // atrybut *names.arg* //

b) Porównamy wyniki egzaminu dla zestawienia: kobiety / mężczyźni oraz dla zestawienia grup: A / B / C / D / E przy użyciu wykresów pudełkowych (boxplots).
Zainstaluj pakiet ‘ggplot2’ (grammar of graphics package) i użyj funkcji *qplot()*.
Jako pierwszy jej parametr (dane na osi poziomej) wstaw nazwę kolumny: płeć ;
jako drugi (dane na osi pionowej) - nazwę kolumny: punkty;
jako trzeci: data = ..nazwa ramki z której bierzemy kolumny..;
jako czwarty: geom = c(‘boxplot’) // określamy typ wykresu//;

i jako piąty: fill = płeć //wypełnimy wykresy kolorami rozróżniającymi od siebie obie płcie//

Zinterpretuj uzyskane wykresy pudełkowe - odpowiedz na pytania:
- co oznaczają dziwne kropki na dole lub górze wykresów?
- co oznacza pozioma kreska przecinająca pudełko?
- co oznacza dolna krawędź pudełka?
- co oznacza górna krawędź pudełka?

TESTY ZGODNOŚCI

Zad. 3 (Test $\chi 2$ Pearsona – na czym polega? Dla jakich hipotez go wykorzystujemy? Jaki jest zbiór krytyczny?)
Rzucono 90 razy kostką do gry i otrzymano ‘jedynkę’ 19 razy, ‘dwójkę’ 13 razy, ‘trójkę’ 21 razy, ‘czwórkę’ 12 razy, ‘piątkę’ 12 razy oraz ‘szóstkę’ 13 razy.
Stawiamy hipotezę H_0: kostka jest uczciwa.
Używając gotowego w pakiecie R testu $\chi 2$ zweryfikuj postawioną hipotezę na poziomie istotności $\alpha = 0,05$ wobec hipotezy alternatywnej H_1, że używana do gry kostka jest asymetryczna.
>*chisq.test*(..wektor wyników doświadczenia.., *p* = ..wektor pr-stw w podejrzewanym rozkładzie..)

Uwaga: jeżeli w wyniku testu *p-value > 0,05* to wnioskujemy, że nie ma podstaw do odrzucenia hipotezy H_0 (czyli uzyskane doświadczalnie dane nie odbiegają zbyt daleko od ‘uczciwych’ tj. teoretycznych).

Zad. 4
Niech X będzie zmienną losową zwracającą pierwszą cyfrę liczby mieszkańców miast w Polsce w 2017r.
Zweryfikuj hipotezę H_0, że X ma faktycznie rozkład Benforda. Użyj testu $\chi 2$.
Najpierw pod zmienną *population_* przechowaj ponownie dane z pliku ‘2017_lundosc.csv’.
Potrzebujemy z niej wektora liczebności wystąpień cyfr 1, 2, ..., 9 na pierwszej pozycji.
Możemy tym razem użyć np. funkcji filtrującej i wyznaczyć liczbę wierszy w ramce, którą zwróci.
Następnie generujemy wektor prawdopodobieństw w rozkładzie Benforda. Pozostaje nam już tylko użyć testu chi kwadrat z dwoma wygenerowanymi wcześniej argumentami.

Zad. 5
W ciągu 100 dni notowano liczby awarii sieci wodociągowej w Świdniku i uzyskano dane: liczba k awarii: 0, 1, 2, 3, 4, 5 - i odpowiednio - liczba dni n_k : 10, 27, 29, 16, 11, 7.
Czyli np. dni bez awarii było 10, dni z 1 awarią było 27 itd…
Zweryfikuj hipotezę H_0, że rozkład liczby dni awarii jest rozkładem Poissona przy standardowym poziomie istotności $\alpha = 0,05$. Użyj testu $\chi 2$.
Parametr rozkładu λ oszacuj na podstawie próbki. Najpierw wygenerujmy wektor liczebności
> *quantities_* = c(10, 27, ..., 7)
Następnie, ponieważ w rozkładzie Poissona parametr λ jest równy wartości oczekiwanej, w celu oszacowania go - policzmy średnią z naszego zestawu danych (tj. liczby awarii). Jako że dane się powtarzają (mają swoje wagi) to policzmy średnią ważoną.
Generujemy wektor prawdopodobieństw Poissona z oszacowanym parametrem λ dla argumentów 0, 1, 2, 3, 4, 5 i stosujemy test.