STATYSTKA MATEMATYCZNA W R

Opracowanie: dr Paweł Drozda

SPIS TREŚCI

1. Podstawowe informacje 3
2. Statystyki opisowe 3
3. Zmienna losowa, rozkłady zmiennej losowej 4
   3.1. Podstawowe rozkłady dyskretne 4
   3.2. Podstawowe rozkłady ciągłe 5
4. Testowanie Hipotez Statystycznych – testy parametryczne 6
   4.1. Test T studenta dla jednej średniej 8
   4.2. Test T studenta dla dwóch średnich – zmienne niezależne 8
   4.3. Test T studenta dla dwóch średnich – zmienne zależne 9
   4.4. Badanie normalności i równości wariancji 9
5. Podłączenie R do Bazy Danych MySQL 10
6. Analiza ANOVA 11
   6.1. prosta ANOVA parametryczna 11
   6.2. prosta ANOVA nieparametryczna 12
   6.3. Analiza post – hoc 12
7. Korelacja i Regresja 13
   7.1. Korelacja 13
   7.2. Regresja 13

1. PODSTAWOWE INFORMACJE

Głównym celem statystyki jest pozyskiwanie, prezentacja i analiza danych. Statystyka pozwala na wnioskowanie z surowych danych. Np. mając próbkę danych pewnej populacji możemy podjąć próbę wyznaczenia wartość oczekiwaną dla średniej. Następnie, można zweryfikować hipotezę, czy obliczona wartość jest zgodna z założeniami. Istnieje wiele dziedzin statystki, które poddają analizie dane w bardzo różny sposób. Możemy wyróżnić:

Statystyka opisowa – głównym celem jest opis danych statystycznych uzyskanych podczas badania statystycznego za pomocą wybranych parametrów, dzięki czemu można dokonać pewnych uogólnień na temat danych, jak również wyciągnąć podstawowe wnioski.

Testowanie hipotez – głównym celem testowania hipotez jest weryfikacja ogólnie postawionych założeń na temat rozkładu, bądź określonych parametrów (takich jak wariancja, czy średnia).

Analiza korelacji i regresji – głównym celem jest określenie współzależności zjawisk za pomocą korelacji oraz wyznaczenie funkcji zależności pomiędzy zmiennymi za pomocą regresji.

Analiza szeregów czasowych – głównym celem jest badanie zależności obserwowanych danych w stosunku do określonego stałego przedziału czasowego (np. godzina, dzień, miesiąc).

2. STATYSTYKI OPISOWE

Wśród podstawowych elementów opisujących zbiorowość, zaliczanych do statystyki opisowej możemy wyróżnić:

- wartość minimalną, wartość maksymalną

- średnią, odchylenie standardowe

- medianę, średnie odchylenie bezwzględne

- kurtozę, skośność

Przykład wyznaczania poszczególnych parametrów w R:

```
> install.packages("e1071")
> library(e1071)
> X <-c(3,5,7,5,3,2,6,8,5,6,9)
> mean(X)
[1] 5.363636
> sd(X)
[1] 2.15744
> mad(X)
[1] 2.9652
> var(X)
[1] 4.654545
> min(X)
[1] 2
> max(X)
[1] 9
> kurtosis(X)
[1] -1.232818
> skewness(X)
[1] 0.06060257
```

kurtosis() – zwraca miarę koncentracji – Kurtozę. Informuje ona o poziomie spłaszczenia rozkładu. Wartość 0 informuje o spłaszczeniu zbliżonym do normalnego, natomiast liczba dodatnia mówi o większej koncentracji wokół średniej, a liczba ujemna o większym spłaszczeniu rozkładu.

skewness() – zwraca skośność rozkładu. Gdy wartość jest bliska zero, oznacza, że rozkład jest symetryczny. Gdy wartość jest mniejsza od zera – rozkład lewostronnie skośny, gdy większa od zera – rozkład prawostronnie skośny.

summary*()* – zwraca podstawowe statystyki opisowe rozkładu

Ćwiczenia

1. Ściągnij plik napoje_po_reklamie.csv. Pozbądź się polskich znaków z nagłówków. Zaimportuj dane do R (read.csv). Sprawdź jaka struktura danych powstała. Oblicz podstawowe statystyki dla zaimportowanych danych dla poszczególnych rodzajów napoi. (**getwd()** – pokazuje domyślny katalog)
2. Dane dla pepsi i fanty zapisz do oddzielnych wektorów. Wyświetl dla nowo stworzonych wektorów podsumowanie podstawowych statystyk.
3. Wgraj plik Wzrost.csv. Użyj dla niego funkcji statystyki opisowej.

3. ZMIENNA LOSOWA, ROZKŁADY ZMIENNEJ LOSOWEJ

Zmienna losowa jest to funkcja określona na przestrzeni zdarzeń elementarnych, która przyporządkowuje każdemu zdarzeniu elementarnemu liczbę rzeczywistą z określonym prawdopodobieństwem. Na przykład: każdemu rzutowi kostką przyporządkowujemy wyrzuconą liczbę oczek; przy rzucie monetą przyporządkowujemy 1 orłowi oraz 2 reszce.

Dla pierwszego przykładu – rozkład zmiennej losowej wygląda następująco:

Wartość	1	2	3	4	5	6
Prawdopodobieństwo	1/6	1/6	1/6	1/6	1/6	1/6

Dla pierwszego przykładu – rozkład zmiennej losowej wygląda następująco:

Wartość	1	2
Prawdopodobieństwo	1/2	½

Istnieje wiele różnych rozkładów prawdopodobieństwa, które mają swoje specyficzne cechy i są szeroko wykorzystywane w badaniach statystycznych.

3.1. PODSTAWOWE ROZKŁADY DYSKRETNE

Rozkłady zmiennej losowej dyskretnej charakteryzują się tym, że ilość wartości, które może przyjąć zmienna losowa jest skończona, bądź przynajmniej przeliczalna. Wśród rozkładów dyskretnych możemy wyróżnić:

Rozkład Bernoulliego

Jest to rozkład prawdopodobieństwa, który przyjmuje wartość 1 (sukces) z zadanym prawdopodobieństwem p, oraz wartość 0 z prawdopodobieństwem q = 1 – p.

Do wygenerowania zbioru danych rozkładu Bernoulliego może posłużyć następujący kod:

```
rbinom(N,1,p)
```

gdzie:

N – wielkość próby

P – prawdopodobieństwo sukcesu

Rozkład Dwumianowy

Jest to rozkład prawdopodobieństwa opisujący liczbę sukcesów (na ogół oznaczanych jako k) przy wykonaniu N niezależnych eksperymentów. Każdy z eksperymentów ma założone to samo prawdopodobieństwo sukcesu.

Funkcja prawdopodobieństwa określona jest jako:

`dbinom(k,n,p)` i zwraca prawdopodobieństwo k sukcesów w n próbach z prawdopodobieństwem sukcesu równym p.

Aby wykreślić wykres prawdopodobieństwa sukcesów dla k=0,…,n należy zastosować kod:

```
> x <- seq(0,n,by =1)
> y<- dbinom(x,n,p)
> plot(x,y)
```

Rozkład Poissona

Jest to rozkład, który wyraża prawdopodobieństwo zdarzeń następujących po sobie z daną częstotliwością. Zdarzenia te zachodzą niezależnie. Np. ilość nieobecności w ciągu miesiąca w pewniej grupie można opisać rozkładem Poissona z intensywnością λ = 2 (może się zdarzyć że w miesiącu były 3 nieobecności, a w innym 1. Średnia wyjdzie 2).

Do wygenerowania próby rozkładu Poissona można użyć kodu:

```
rpois(n,2)
```

Do narysowania gęstości rozkładu Poissona można posłużyć się kodem:

```
x <- seq(0,15,by =1)
x_pois <- dpois(x,lambda,p)
plot(x, x_pois, ylab = ‘Rozkład Poissona)
```

Aby obliczyć średnią rozkładu Poissona należy użyć kodu korzystając z rozkładu gęstości:

```
Mean_vec <- x* x_pois

M <- sum(Mean_vec)
```

3.2. PODSTAWOWE ROZKŁADY CIĄGŁE

Rozkłady ciągłe charakteryzują się tym, że pomiędzy dowolnymi dwoma wartościami istnieje możliwość wstawienia dowolnej ilości innych wartości. W konsekwencji, prawdopodobieństwo tego, że zmienna losowa przyjmie dokładną wartość w punkcie wynosi 0. Wśród rozkładów zmiennych losowych ciągłych możemy wyróżnić:

Rozkład normalny

Najbardziej popularny rozkład ciągły i najszerzej stosowany w statystyce. Gęstość prawdopodobieństwa standardowego rozkładu normalnego jest dana wzorem:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2}$$

Do wygenerowania danych rozkładu normalnego możemy posłużyć się następującym kodem:

```
rnorm(n, mean = 0, sd = 1)
```

gdzie:

n – liczba obserwacji, mean – średnia, sd – odchylenie standardowe

Aby narysować wykres gęstości rozkładu normalnego można użyć następującego kodu:

```
> x_norm <- seq(-2,2,by = 0.01)
> x_norm_dist <- dnorm(x_norm, sd = 2)
> plot(x_norm, x_norm_dist)
```

Aby znaleźć wartość dystrybuanty należy użyć funkcji:

`pnorm(q)` – gdzie q określa wartość dla której ma być wyznaczona dystrybuanta

Aby narysować histogram, należy użyć funkcji:

```
hist(X, xlim = c(a,b), breaks = n)
```

gdzie:

X – dane

xlim – przedział X

n – liczba słupków

Ćwiczenia

1. Dla zmiennych losowych przedstawionych w tabelach obliczyć podstawowe statystyki

2. Wygeneruj próby dla n=100 dla następujących rozkładów: Bernoulliego, Dwumianowego, Poissona Policz dla nich statystyki podstawowe (średnią, wariancję, kurtozę i skośność).

3. Dla rozkładów Dwumianowego, Poissona narysuj wykresy rozkładu prawdopodobieństwa (dobierz dowolnie parametry).

4. Dla rozkładów Poissona i Dwumianowego wygeneruj rozkład prawdopodobieństwa dla n = 20, k = 0, …, 20 oraz p = 0.4. Sprawdź, czy suma prawdopodobieństw wygenerowana dla tych rozkładów jest równa 1.

5. Wygeneruj 30 danych dla rozkładu normalnego o średniej równej 0 i odchyleniu standardowym równym 2. Wyznacz statystyki podstawowe – czy są one równe z wartościami teoretycznymi? Sprawdź, czy zwiększenie liczby danych zwiększy dokładność wyliczeń statystyk opisowych.

6. Narysować histogram dla rozkładu normalnego o parametrach: średnia = 1, odchylenie =2, wykres dla rozkładu standardowego, oraz wykres gęstości dla średniej równej -1 oraz odchylenia równego 0.5.

4. TESTOWANIE HIPOTEZ STATYSTYCZNYCH – TESTY PARAMETRYCZNE

https://data-flair.training/blogs/hypothesis-testing-in-r/

Jednym z zadań statystyki jest weryfikacja założeń, które stawia badacz. Przebiega to na ogół jako badanie statystyczne, które składa się z przygotowania badania, gdzie przygotowywane są dane do badania oraz czynione założenia i cel badania.

Następnie należy sformułować najbardziej prawdopodobne ogólne rozwiązanie, które nazywamy hipotezą badawczą. Hipoteza ta powinna być sformułowana w prosty sposób, żeby nie było wątpliwości co chcemy zweryfikować i żeby nie stanowiło problemu dobranie testów.

Jednym z podstawowych badań jest założenie o wartości badanych zmiennych – dla przykładu: Średni wiek chorych wynosi 55 lat. Jako drugi przykład można przytoczyć porównanie parametrów pomiędzy grupami. Na przykład, weryfikacja hipotezy, że średnia wzrostu dzieci w dwóch szkołach jest równa. Kolejnym problemem, który może się pojawić przy weryfikacji hipotez to różnica pomiędzy cechami opisującymi badaną grupę – np. lek A skuteczniej działa niż lek B, kobiety ważą mniej niż mężczyźni. Za pomocą weryfikacji hipotez można również badać zależności między cechami np. zależność (korelacja) pomiędzy wzrostem a wagą, paleniem papierosów a zachorowaniami na raka płuc itd. Dodatkowo można badać kształt zależności – np. zależność logarytmiczna (wzrost a wiek). Testowanie hipotez pozwala również na weryfikację czy dany rozkład jest zbliżony do jakiegoś rozkładu teoretycznego np. zmienna ma rozkład normalny.

Po postawieniu hipotezy następuje jej weryfikacja (hipoteza to przypuszczenie dotyczące próby generalnej – weryfikacja na podstawie próby losowej). Weryfikacja hipotezy przebiega w następujących krokach:

- Postawienie hipotezy zerowej, która będzie podlegała sprawdzeniu (**H0**)
- Postawienie hipotezy alternatywnej (**H1**) – gdy odrzucamy hipotezę **H0** – to przyjmujemy **H1**
- Dobór odpowiedniego testu – obliczenie jego wartości pochodzącej z próby
- Przyjęcie odpowiedniego poziomu istotności (wielkość błędu na jaki wyrażamy zgodę bardzo często α=0,05 – oznacza to że na 100 badań możemy popełnić 5 błędów
- Przy podanym poziomie istotności znajdujemy obszary krytyczne i podejmujemy decyzję

W trakcie weryfikacji możemy popełnić błędy dwóch rodzajów:

- I rodzaju – gdy odrzucamy hipotezę zerową fałszywą – prawdopodobieństwo tego określa poziom istotności
- II rodzaju – przyjęcie fałszywej hipotezy alternatywnej

Aby zminimalizować błąd drugiego rodzaju przeprowadza się testy istotności przy przyjętym z góry poziomie błędu pierwszego rodzaju (poziom istotności).

Hipoteza Zerowa	**DECYZJE**	
	Przyjąć H0	Odrzucić H0

Hipoteza zerowa prawdziwa	Decyzja prawidłowa	Błąd I rodzaju
Hipoteza zerowa fałszywa	Błąd II rodzaju	Decyzja prawidłowa

Podczas testowania hipotez wyznacza się obszary krytyczne w zależności od poziomu istotności oraz postawionej hipotezy. Obszary krytyczne tworzą wartości, dla których hipoteza H0 jest bardzo mało prawdopodobna (<=α). W przypadku jeśli wartość testu trafia do obszaru krytycznego – hipotezę H0 odrzucamy na korzyść hipotezy H1. Wśród rodzajów obszarów krytycznych możemy wyróżnić:

Obszar dwustronny gdy H0: X=A, H1: X ≠A, wtedy P(X<A) = α/2 i P(X>A) = α/2

Obszar lewostronny gdy H0: X=A, H1: X <A, wtedy P(X<A) = α

Obszar prawostronny gdy H0: X=A, H1: X >A, wtedy P(X>A) = α

Bardzo istotną kwestią przy wyborze testu jest założenie o normalności rozkładu zmiennej testowanej oraz o równości wariancji, które będą objęte testem.

Do zbadania normalności rozkładu służą następujące testy:

Shapiro – Wilka, Kołomogorowa – Smirnowa, Lillieforsa

Natomiast do badania równości wariancji w próbach służą testy:

F, Levenea, Bartletta

4.1. TEST T STUDENTA DLA JEDNEJ ŚREDNIEJ

Najprostszym testem do weryfikacji hipotezy mówiącej o równości średniej i pewnej stałej C jest test T studenta. Do wykonania testu weryfikującego hipotezę służy funkcja **`t.test(X, mu = a)`**. Jako przykład wykorzystania może posłużyć następujący kod:

```
> dane = read.csv("Wzrost.csv", header = FALSE)
> t.test(dane1[['V1']], mu = 173)
```

Funkcja zwraca dwie istotne wartości: wartość statystyki t oraz wartość prawdopodobieństwa p określającego poziom, od jakiego nie ma podstaw do odrzucenia hipotezy zerowej. Przy wysokiej wartości prawdopodobieństwa p, hipoteza zerowa nie jest odrzucana. Na ogół przyjmuje się poziom graniczny na p = 0.05 lub p = 0.01.

4.2. TEST T STUDENTA DLA DWÓCH ŚREDNICH – ZMIENNE NIEZALEŻNE

Bardzo podobny test (również T studenta) do poprzedniego weryfikuje hipotezę o równości średnich dwóch populacji. Test ten przyjmuje dwa zbiory danych jako zbiory wejściowe i weryfikuje hipotezę o równości średnich. W R służy do tego ta sama funkcja, co przy jednej średniej. Jako przykład wykorzystania może posłużyć poniższy kod:

```
V1 = c(1,2,4,5,7,6,5,4,3)
V2 = c(2,3,1,3,4,5,6,7,8,1,1)
```

```
t.test(V1,V2)
```

Podobnie jak w poprzednim przypadku, funkcja zwraca dwie istotne wartości: wartość statystyki t oraz wartość prawdopodobieństwa p określającego poziom, od jakiego nie ma podstaw do odrzucenia hipotezy zerowej. Przy wysokiej wartości prawdopodobieństwa p, hipoteza zerowa nie jest odrzucana. Na ogół przyjmuje się poziom graniczny na p = 0.05 lub p = 0.01.

W teście dodatkowo można ustawić założenie o równości wariancji. Należy dodać **`var.equal = TRUE`** jako parametr. Np.:

```
t.test(V1,V2, var.equal = TRUE)
```

W przypadku braku tego parametru, wartość jest ustawiona na FALSE i wykonywany jest test Welcha.

Ponadto, do testu można dodać wersję hipotezy alternatywnej za pomocą parametru **alternative**. Standardowo parametr jest ustawiony na **two.sided**. Parametr można dodatkowo ustawić na **greater** lub **less.**

Kolejnym elementem, który możemy ustawić dla testu jest poziom istotności (**conf.level**).

4.3. TEST T STUDENTA DLA DWÓCH ŚREDNICH – ZMIENNE ZALEŻNE

Kolejnym przypadkiem, jaki należy rozpatrzeć przy testowaniu hipotez statystycznych jest sytuacja, w której porównujemy dwa parametry struktury dla prób, które powstały w sposób zależny. Dobrym przykładem jest np. badanie temperatury pacjentów przed i po zażyciu leku przeciwgorączkowego. Najczęściej zmienne zależne dotyczą tej samej próbki przed i po zaistnieniu jakiegoś zjawiska. Do przeprowadzenia testu dla zmiennych zależnych ponownie służy funkcja `t.test` z tym, że parametr **paired** musi być ustawiony na **TRUE**. Test przyjmuje jako parametry dwa zbiory danych (zbiory powinny być równoliczne). Test ten można zobrazować za pomocą następującego kodu:

```
> V1 = c(1,2,2,2,3,2,1,2,3)
> V2 = c(2,3,9,9,9,9,9,9,8)
> t.test(V1,V2, var.equal = TRUE, paired = TRUE)
```

4.4. BADANIE NORMALNOŚCI I RÓWNOŚCI WARIANCJI

Aby można było zastosować testy parametryczne przy weryfikacji hipotez dane wejściowe muszą spełniać założenia o normalności rozkładu danych. W Pythonie dostępne są następujące testy oceniające normalność rozkładu:

a) **Anderson Darling** – test określa prawdopodobieństwo, tego, że próbka danych pochodzi z rozkładu normalnego. Przykład kodu przedstawia się następująco:

```
> install.packages("nortest")
> library(nortest)
> ad.test(V1)
```

b) test Shapiro – Wilka. Test ten uważany jest za mocniejszy niż test Kołomogorowa – Smirnowa. Przykład w R:

```
> shapiro.test(V1)
```

c) test Kołomogorowa – Smirnowa – testuje, czy rozkład zmiennej jest zbliżony do zadanego rozkładu teoretycznego (w przypadku testowania normalności – do rozkładu normalnego).
Przykład w Pythonie:

```
> ks.test(V1, "dnorm")
```

Drugim kluczowym warunkiem do możliwości zastosowania testów parametrycznych jest założenie o równości wariancji w próbach, które testujemy. Do zbadania równości wariancji w R możemy wykorzystać następujące testy:

a) Test Levene – przykład w R:

```
> dane = read.csv("napoje_test_norm.csv", header = TRUE, sep = ';')
> leveneTest(sprzedaz ~ napoj, data = dane)
```

b) Test Bartlett - przykład w R:

```
> dane = read.csv("napoje_test_norm.csv", header = TRUE, sep = ';')
> bartlett.test(sprzedaz ~ napoj, data = dat)
```

Ćwiczenia

1. Wygeneruj próbę losową dla rozkładu normalnego dla średniej = 2, odchylenia = 30 i liczby elementów = 200. Zbadaj hipotezę mówiącą o tym, że średnia tego rozkładu jest równa 2,5.
2. Wczytaj plik napoje.csv. Zweryfikuj hipotezę że średnie spożycie piwa lech wynosi 60500, coli wynosi 222000, piw regionalne wynosi 43500.
3. Sprawdzić która zmienna w pliku napoje.csv wykazuje normalność
4. Zbadaj równość średnich dla następujących par: okocim – lech, fanta – regionalne oraz cola – pepsi.
5. Zbadaj równość wariancji pomiędzy okocim – lech, żywiec – fanta oraz regionalne – cola.
6. Zbadaj równość średnich pomiędzy latami 2001 i 2015 dla piw regionalnych (ograniczenie liczby wyników – funkcja **subset**).
7. Zbadaj równość średnich dla wartości z roku 2016 oraz dla wartości z pliku napoje_po_reklamie.csv oddzielnie dla coli, fanty i pepsi. Zakładamy, że zmienne te są zależne. (ograniczenie liczby wyników – funkcja **subset**)

5. PODŁĄCZENIE R DO BAZY DANYCH MYSQL

Jednym z najczęściej stosowanych źródeł danych są bazy danych. W wielu przypadkach dane pobierane są z bazy danych do środowiska programistycznego (np. język R), tam są przetwarzane i z powrotem zwracane do bazy danych.

Tworzenie tabel w bazach danych:

Do tworzenia tabeli w bazie danych służy polecenie CREATE TABLE. Przykład:

CREATE TABLE osoba (is_osoby int PRIMARY KEY, nazwisko VARCHAR(40), imie VARCHAR(40), data_ur DATE, zarobki DOUBLE(10,2));

Do wstawiania danych do tabeli służy polecenie INSERT INTO:

INSERT INTO osoba VALUES(1, ‘Kowalski’, ‘Jan’, ‘1995-02-23’, 5000);

Żeby podłączyć się do bazy danych z poziomu języka R należy wykonać ciąg poleceń:

```
> library(RMySQL)
> db_user <- 'nazwaUzytkownika'
> db_password <- 'haslo'
> db_name <- 'nazwa bazy danych'
> db_host <- 'serwer'
> db_port <- 3306
> mydb <- dbConnect(MySQL(), user = db_user, password = db_password, dbname
 = db_name, host = db_host, port = db_port)
> s <- paste0("select * from nazwa_tabeli")
> rs <- dbSendQuery(mydb, s)
> df <-fetch(rs, n = -1)
> print(df)
> on.exit(dbDisconnect(mydb))
```

Ćwiczenia

1. Zaloguj się na serwer MySQL (bad.uwm.edu.pl/phpmyadmin) l: nazwisko p: imie
2. Stwórz tabelę pracownik(id, imie, nazwisko, data_zatrodnienia, stanowisko, zarobki)
3. Wpisz 3 rekordy do stworzonej tabeli
4. Podłącz się do stworzonej bazy danych
5. Dodaj z poziomu R 2 rekordy
6. Odczytaj wszystkie rekordy ze stworzonej tabeli

6. ANALIZA ANOVA

Analiza stosowana jest w przypadku, gdy chcemy porównać średnie więcej niż dwóch grup. Dla przykładu możemy chcieć porównać średnie zarobki w poszczególnych województwach na przestrzeni lat. Nie można porównać każdej średniej z każdą za pomocą testu t – Studenta, ponieważ błąd pierwszego rodzaju wzrasta drastycznie po dodaniu kolejnej grupy. Dla przykładu, mając 4 grupy, przyjmując poziom istotności 0,05 musimy wykonać 6 porównań i wtedy błąd I rodzaju wynosi 1- (0,95)^6 = 0,265.

Dlatego został wprowadzony zestaw metod statystycznych pozwalających na porównania wielu średnich – analiza wariancji. W analizie wariancji ocenę istotności różnic pomiędzy średnimi ocenia się na podstawie porównania wariancji.

Jest to porównanie wartości cech w zależności od czynników klasyfikujących, takich jak np. płeć, wykształcenie, przedział wiekowy itd. Najprostsza analiza bierze pod uwagę jeden czynnik grupujący. Bada zależność, czy średnie w poszczególnych grupach są różne. Dla czynnika klasyfikującego mającego k różnych grup mamy postać:

H0: m1=m2=...=mk

wobec hipotezy alternatywnej:

H1: co najmniej dwie średnie są różne

6.1. PROSTA ANOVA PARAMETRYCZNA

W przypadku spełnienia podstawowych założeń, wykonuje się jednoczynnikową analizę ANOVA. W przypadku odrzucenia hipotezy zerowej należy przeprowadzić analizę post – hoc, która dokładnie wskazuje elementy różnią

się istotnie między sobą. Do prostej parametrycznej analizy ANOVA (jednoczynnikowej) w R służy funkcja **aov**. Przykładowy kod przedstawiający prostą ANOVA został umieszczony poniżej:

```
> my_data <-PlantGrowth
> levels(my_data$group)
> res.aov <- aov(weight ~ group, data = my_data)
> summary(res.aov)
```

6.2. PROSTA ANOVA NIEPARAMETRYCZNA

W przypadku braku spełnienia założenia o normalności powinniśmy użyć nieparametryczny odpowiednik analizy wariancji, czyli test Kruskala -Wallisa. Test sprawdza hipotezę o równości median. W przypadku, gdy odrzucimy hipotezę o równości median, nie oznacza to, że wszystkie mediany różnią się istotnie statystycznie. Aby to zbadać należy wykonać analizę post – hoc. Do wykonania testu w R służy funkcja **kruskal.test**. Poniższy przykład ilustruje zastosowanie tej funkcji:

```
> kruskal.test(weight ~ group, data = my_data)
```

6.3. ANALIZA POST – HOC

W przypadku, gdy któraś z analiz ANOVA odrzuci hipotezę zerową, należy wykonać analizę post – hoc. Przyjęcie hipotezy alternatywnej informuje nas jedynie o tym, że w pomiędzy średnimi prób istnieją różnice istotnie statystycznie, jednak brak jest informacji o dokładniejszym miejscu występowania tych różnic.

W R służy do tego funkcja TukeyHSD:

```
> TukeyHSD(res.aov)
```

W wynikach poszczególne elementy mają następujące znaczenie:

diff: różnica pomiędzy średnimi dwóch grup

lwr, upr: dolna i górna granica przedziału ufności dla poziomu 95%

p adj: Wartość prawdopodobieństwa od którego przyjmujemy brak różnic między średnimi (przyjmujemy poniżej 0.05, że różnice są istotne, powyżej, że są nieistotne)

Ćwiczenia

1. Dla pliku Pracownicy.csv przeprowadź analizę Wariancji ze zmienną grupującą wykształcenie i dla zmiennej zależnej rocznie oraz wiek w dniach wraz z analizą post-hoc. Sprawdź strukturę pliku i dostosuj dane do odpowiednich funkcji.
2. Sprawdź równość wariancji i normalność rozkładów poszczególnych zmiennych. Który test powinien być zastosowany do analizy wariancji.

3. Wczytaj plik Efektywnosc.sta. Przeprowadź analizę wariancji, sprawdź jednorodność wariancji pomiędzy pomiarami. Przygotuj odpowiednio dane. Dla których pomiarów średnie różnią się statystycznie. Dla tych elementów przeprowadź analizę post hoc.

7. KORELACJA I REGRESJA

7.1. KORELACJA

Głównym celem korelacji jest dokładne określenie stopnia powiązania zmiennych. W szczególności korelacja powinna nam określić, czy istnieje związek pomiędzy badanymi próbami oraz, jeśli ten związek istnieje, to powinna zostać określona siła, kierunek i kształt powiązania. Najczęściej bada się zależność pod kątem liniowej zależności, natomiast istnieją przypadki, w których zależność jest krzywoliniowa. Podstawowym współczynnikiem korelacji liniowej jest współczynnik Spearmana określony wzorem:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

Interpretacja tego współczynnika jest następująca:

$r_{xy}=0$ zmienne nieskorelowane

$0<r_{xy}<0,5$ korelacja słaba

$0,5<r_{xy}<1$ korelacja silna

Do obliczania współczynnika korelacji w R służy polecenie **cor**, a do sprawdzenia istotności korelacji **cor.test**. Poniżej znajduje się przykładowy kod obliczający współczynnik korelacji:

```
>  count <- c(9,25,15,2,14,25,24,47)
> speed <- c(2,3,5,9,14,24,29,34)
> cor(count,speed)
> cor.test(count,speed)
```

Alternatywą do parametrycznego współczynnika korelacji jest współczynnik rang Spearmana. W R wystarczy dodać parametr method i ustawić go na ‘spearman’. Przykład kodu obliczającego współczynnik korelacji jest pokazany poniżej:

```
>  count <- c(9,25,15,2,14,25,24,47)
> speed <- c(2,3,5,9,14,24,29,34)
> cor(count,speed, method = ‘spearman’)
```

Korelację więcej niż dwóch zmiennych można obliczyć z macierzy/data.frame:

```
> cor(macierz)
```

7.2. REGRESJA

Głównym celem regresji jest badanie wpływu jednej lub kilku zmiennych **tzw. objaśniających** na zmienną na zmienną **objaśnianą**. Metody regresji służą do opisu kształtowania się poziomu pewnego zjawiska w czasie na podstawie pobieranych z populacji generalnej prób losowych. Na ogół rozpatruje się zależności pomiędzy

wieloma zmiennymi. Mamy wtedy do czynienia z **regresją wieloraką** lub **wielowymiarową**. Głównym celem **regresji wielorakiej** jest ilościowe ujęcie związków pomiędzy wieloma zmiennymi niezależnymi (objaśniającymi) a zmienną zależną (objaśnianą). Do wyznaczenia równania regresji przeprowadza się **analizę regresji,** która polega na estymacji parametrów równania teoretycznego, które w sposób jak najbardziej dokładny odwzorowuje zależność. Podstawowe modele regresji zakładają występowanie zależności **liniowych** istniejących pomiędzy zmienną objaśnianą, a zmiennymi ją objaśniającymi. Zmienne objaśniane i objaśniające powinny cechować się pewnymi właściwościami, aby analiza regresji była przeprowadzona w sposób prawidłowy.

Wśród tych właściwości można wymienić:

- Relacje pomiędzy zmiennymi powinny mieć charakter liniowy
- Rozkłady zmiennych mają kształt zbliżony do normalnego
- Zmienne objaśniające muszą wykazywać związek ze zmienną, którą będą objaśniały.
- Zmienne objaśniające powinny cechować się odpowiednim wskaźnikiem własnej zmienności.
- Zmienne objaśniające nie mogą być współzależne. Znaczy to tyle, że ich wzajemne wskaźniki korelacji muszą wykazywać wartości mniejsze niż korelacji ze zmienną objaśnioną.
- Ilość szacowanych parametrów ($k + 1$) równania nie może przekroczyć liczby okresów (T). Czyli zmiennych w modelu winno być mniej (lub taka sama ilość), co okresów, jakie uwzględniają - jeśli chodzi o modelowanie na podstawie szeregów czasowych.

Równanie regresji liniowej przyjmuje postać:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + e_t$$
$$t = 1,2, \ldots T$$

W R można określić prosty model regresji, gdzie jest tylko jedna zmienna objaśniająca za pomocą:

`lm()`

Jako kod można przedstawić (wykorzystując zmienne count, speed):

```
> dane <-data.frame(count,speed)
> linearModel <- lm(count ~ speed, data = dane)
> print(linearModel)
> modelSummary <- summary(linearModel)
> modelSummary$coefficients
```

Do predykcji możemy zastosować następujący kod:

```
> new.speeds <-data.frame(speed=c(60,90,120))
> predict(linearModel,newdata = new.speeds)
```

Regresję wieloraką przedstawia następujący kod:

```
> cena <-c(200, 250, 280, 300, 320,325, 180, 260, 400, 450)
> powierzchnia <-c(40,45,46,58,60,61,35,50,100, 130)
> pokoje<-c(2,2,2,3,3,3,1,3,4,5)
> dane <- data.frame(cena,powierzchnia,pokoje)
> LinModel <- lm(cena~powierzchnia+pokoje, data = dane)
> summary(LinModel)
```

Ćwiczenia

1. W pliku pracownicy.csv. Oblicz współczynniki korelacji liniowej między zarobkami we wszystkich miesiącach (jedna tabela zmiennych). Oblicz współczynniki korelacji między stażem pracy i zarobkami w poszczególnych miesiącach a także rocznymi zarobkami. (dwie listy zmiennych).
2. Wczytaj plik napoje.csv. Zbadaj czy istnieje zależność między spożyciem poszczególnych napojów.
3. Dla pliku pracownicy.csv zbadaj, czy istnieje zależność przy pomocy współczynników korelacji Spearmana między zarobkami w poszczególnych miesiącach, między zarobkami rocznymi a stażem pracy.
4. Wczytaj plik regresja.xls (arkusz – zadanie2). Wyznacz zależność liczby dzieci od liczby małżeństw. (jeden predyktor). Wyznacz zależność liczby dzieci od pozostałych predyktorów.