Statystyka wykład

Populacja:

* zbiorowość wszystkich elementów stanowiących podmiot badania (populacja przedmiotowa)
* zbiór wszystkich możliwych do zaobserwowania wartości cechy opisującej badane zjawisko (populacja zdarzeniowa)

Próba – podzbiór populacji dostępny badaczowi i stanowiący podstawę jego wnioskowania o populacji statystycznej

Dane jakościowe – cechy niemierzalne, np. kolor oczu, kształt liścia, ocena bólu, poziom zarobków

Dane ilościowe – cechy mierzalne:

* dyskretne (skokowe) – gdy zbiór wartości jest skończony lub przeliczalny (pomiaru takich cech dokonujemy na ogół poprzez „zliczanie”)
* ciągłe – gdy zbiór wartości jest nieprzeliczalny (pomiaru takich cech dokonujemy na ogół poprzez „mierzenie”)

Zasady konstrukcji szeregu rozdzielczego:

* klasy obejmują wszystkie jednostki badanej zbiorowości
* klasy są rozłączne
* klasy są niepuste

Miary położenia – wskazują wokół jakich wartości skupia się rozkład analizowanych zmiennych. Dzieli się je na klasyczne i pozycyjne. Miary klasyczne to średnie: arytmetyczna, harmoniczna i geometryczna. Do miar pozycyjnych należą dominanta (modalna, wartość najczęstsza) oraz kwantyle. Wśród kwantyli najczęściej stosowane są kwartyle dzielące zbiorowość na cztery części pod względem liczebności.

|  |  |  |
| --- | --- | --- |
| Miara położenia | Szereg | Wartość |
| Średnia  mean(dane) | szereg szczegółowy |  |
| szereg rozdzielczy punktowy |  |
| szereg rozdzielczy przedziałowy |  |
| Dominanta (moda)  -brak- | szereg szczegółowy | Wartość występująca najczęściej |
| szereg pozycyjny |
| szereg rozdzielczy punktowy |
| Kwantyl  quantile(dane, probs = p) | szereg pozycyjny |  |
| szereg rozdzielczy punktowy | - wskazujemy klasę, w której po raz pierwszy liczebność skumulowana osiągnie lub przekroczy *pn*  *-* jest równy wartości cechy we wskazanej klasie |

Kwartyle:

* Q1 – pierwszy kwartyl to 25-kwartyl – dzieli zbiorowość na dwie części w ten sposób, że 25% jednostek zbiorowości ma wartości cechy niższe bądź równe kwartylowi pierwszemu Q1, a 75% równe bądź wyższe od tego kwartyla
* Q2 – drugi kwartyl to mediana
* Q3 – trzeci kwartyl to 75-kwartyl – dzieli zbiorowość na dwie części w ten sposób, że 75% jednostek zbiorowości ma wartości cechy niższe bądź równe kwartylowi trzeciemu Q3, a
* 25% równe bądź wyższe od tego kwartyla

Wariancja – pozwala określić wielkość rozproszenia danych w porównaniu z ogólną wartością średnią - var(dane)

Odchylenie standardowe – miara statystyczna, która mierzy rozproszenie zbioru danych względem jego średniej i jest obliczana jako pierwiastek kwadratowy wariancji - sd(dane)

Współczynnik zmienności

Interpretacja współczynnika zmienności:

* 0 -20% - słabe zróżnicowanie danych
* 20% - 40% - umiarkowane zróżnicowanie danych
* 40% - 60% - silne zróżnicowanie danych
* Powyżej 60% - bardzo silne zróżnicowanie danych

Wykres pudełkowy – pozwala ująć na jednym rysunku miary położenia, rozproszenia i kształtu rozkładu empirycznego badanej cechy. - boxplot(dane)

Interpretacja: Nad osią umieszczony jest prostokąt, którego lewy bok jest wyznaczony przez Q1, a prawy przez Q3. Szerokość pudełka odpowiada wówczas wartości rozstępu ćwiartkowego Rq. Wewnątrz prostokąta znajduje się pionowa linia określająca wartość mediany. Rysunek pudełka uzupełniany jest po prawej i lewej stronie odcinkami zwanymi wąsami, przy czym końce odcinków oznaczają odpowiednio Xmin i Xmax.

Zmienną losową X nazywamy funkcję X = X(ω) określoną na przestrzeni zdarzeń elementarnych Ω o wartościach w zbiorze liczb rzeczywistych *R*.

Niech zmienna losowa dyskretna (skokowa) X przyjmuje wartości x1, x2… odpowiednio z prawdopodobieństwami p1, p2..;

Zmienna losowa ciągła – zmienna przyjmująca wszystkie wartości z pewnego przedziału liczbowego

Rozkładem prawdopodobieństwa zmiennej losowej dyskretnej X, nazywamy funkcję przyporządkowującą wartościom zmiennej xi (i = 1, 2…) prawdopodobieństwa ich przyjęcia:



W pewnym eksperymencie wykorzystano trzy automatyczne aparaty fotograficzne w celu dokumentowania jego przebiegu. W danych warunkach prawdopodobieństwo wykonania poprawnej fotografii dla każdego aparatu jest takie samo i wynosi p = 0.6. Oblicz prawdopodobieństwo:

* a) nieudokumentowania eksperymentu
* b) zarejestrowania eksperymentu przez co najmniej dwa aparaty

Rozkład prawdopodobieństwa:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 |
|  | 0,064 | 0,288 | 0,423 | 0,216 |

* Dla pojedynczego aparatu możliwe są tylko dwa zdarzenia: zrobienie zdjęcia lub awaria – na każdym poziomie drzewa probabilistycznego rysowane są dwie gałęzie
* Prawdopodobieństwo, że aparat zadziała nie zależy od zadziałania pozostałych aparatów – brak pamięci
* Najmniejszą wartością jaką przyjmuje zmienna losowa X zliczająca aparaty, które zadziałały, jest 0, natomiast największą 3 – liczba wszystkich aparatów (X = 0, 1, 2, 3)

Jest to rozkład dwumianowy: X ~ bin(n, p)

Dystrybuantą zmiennej losowej X nazywamy funkcję F(X) określoną na zbiorze liczb rzeczywistych taką, że:

Dla dowolnych liczb rzeczywistych a i b zachodzi:

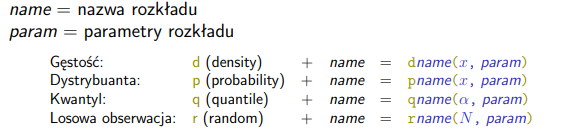
Wartością oczekiwaną zmiennej losowej X nazywamy liczbę oznaczoną symbolem E(x) i określoną wzorem:

Wariancją zmiennej losowej X nazywamy liczbę oznaczoną symbolem D2(X) i określoną wzorem:

Odchyleniem standardowym zmiennej losowej X nazywamy liczbę oznaczoną symbolem D(X) i określoną wzorem:

Do innych rozkładów dyskretnych należą:

* Rozkład równomierny
* Rozkład zero-jedynkowy (Bernoulliego)
* Rozkład Poissona
* Rozkład geometryczny
* Rozkład Pascala
* Rozkład hipergeometryczny

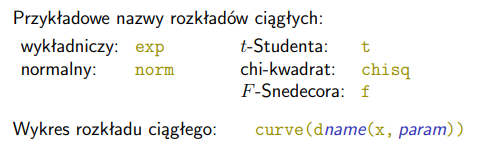


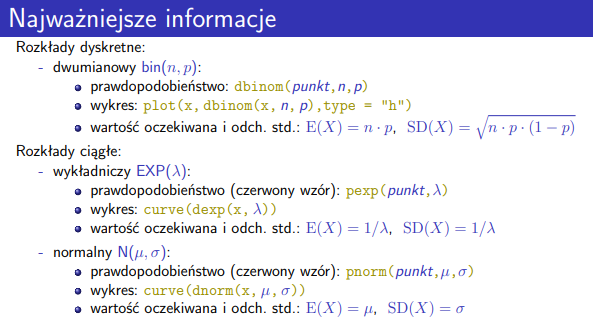
Funkcja gęstości prawdopodobieństwa f(x) – funkcja opisująca rozkład prawdopodobieństwa zmiennej losowej ciągłej.

Wartością oczekiwaną ciągłej zmiennej losowej X nazywamy:

Do innych rozkładów ciągłych należą:

* Rozkład jednostajny (prostokątny)
* Rozkład trójkątny
* Rozkład t-Studenta
* Rozkład Beta
* Rozkład Gamma
* Rozkład chi-kwadrat





Cel statystyki: wyciągnięcie wniosków o populacji na podstawie zbioru obserwowanych danych (próby).

Próba – zbiór znanych, mierzalnych jednostek reprezentujących populację posiadającą badaną cechę.

Statystyka – dowolna funkcja zmiennych losowych X1, X2, X3… stanowiących próbę, nie zawierają nieznanych parametrów, np:

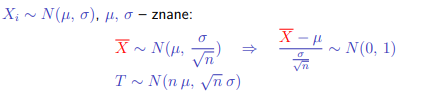
* Średnia próby
* Średnie odchylenie kwadratowe (wariancja) próby
* Odchylenie standardowe
* Wskaźnik struktury (proporcja, prawdopodobieństwo sukcesu)

Statystyka JEST ZMIENNĄ LOSOWĄ, więc posiada swój rozkład.

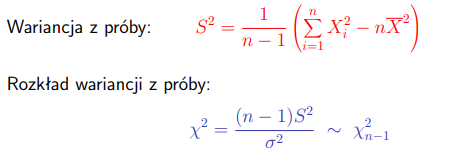
**Centralne Twierdzenie Graniczne**:

W losowym próbkowaniu z dowolnej populacji o wartości oczekiwanej µ i odchyleniu standardowym σ rozkład przy dużym jest w przybliżeniu rozkładem normalnym z wartością oczekiwaną µ i odchyleniem standardowym :

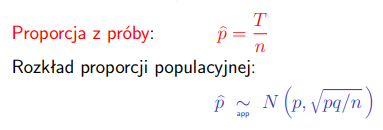
Rozkład średniej z próby:



Rozkład wariancji z próby:



Proporcja populacyjna (próba duża > 100):



Obserwację przed jej pobraniem modelujemy jako zmienną losową X o rozkładzie f(x) – rozkładzie populacji.

Próba (losowa) prosta o liczebności n – zbiór n niezależnych zmiennych losowych X1, X2, X3… o takim samym rozkładzie f(x) jak interesująca zmienna losowa X w populacji.

X1, X2, X3… - zmienne losowe reprezentujące nieznane pomiary, które w procesie losowania próby zamieniają się w pierwszą, drugą.. ntą obserwację.

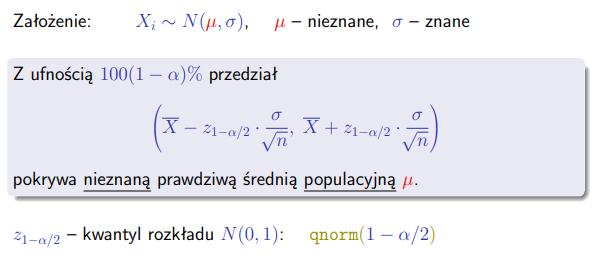
Estymator θ – statystyka podająca sposób obliczania oceny parametru θ.

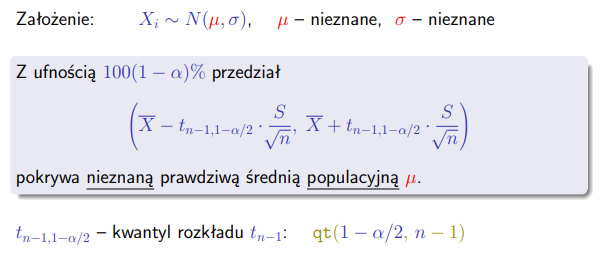
Estymator jest zmienną losową, zatem posiada swój rozkład.

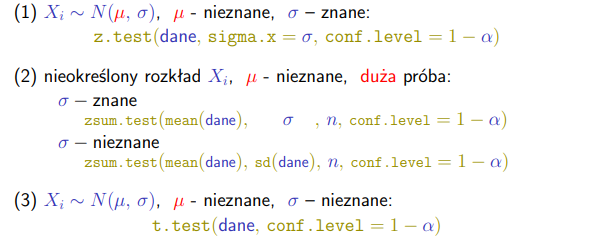
**Estymacja przedziałowa**

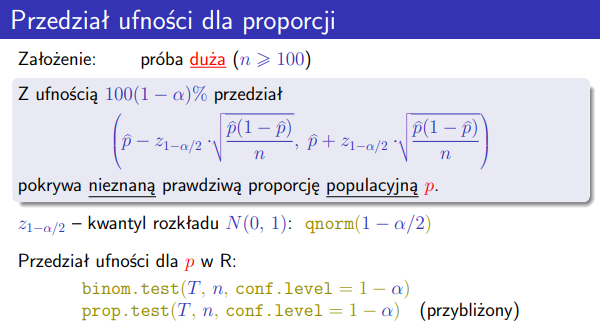
Niech () będzie określonym „dużym” prawdopodobieństwem i niech statystyki (zmienne losowe) L oraz U będą funkcjami próby X1, X2, X3… takimi, że:

Wówczas przedział losowy (L, U) nazywamy przedziałem ufności dla parametru θ, a wartość ) nazywamy współczynnikiem ufności przedziału.



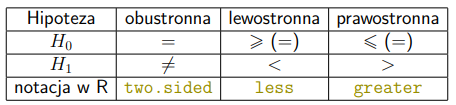






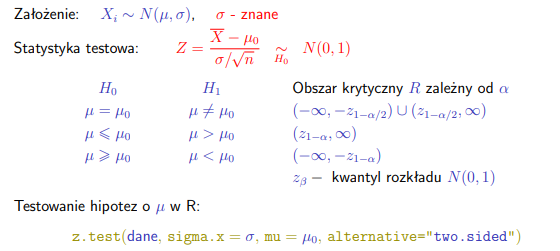
Hipoteza statystyczna (parametryczna) – przypuszczenie dotyczące parametrów badanej populacji.

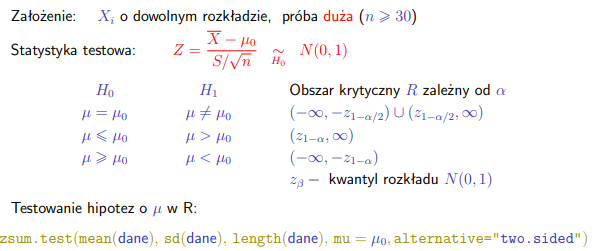
Testowanie hipotez – wnioskowanie statystyczne polegające na podjęciu decyzji na podstawie próby czy hipotezę należy odrzucić czy też nie:

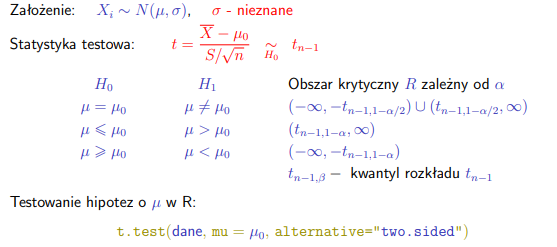


Statystyka testowa – statystyka, której wartości pozwalają zdecydować o odrzuceniu H0

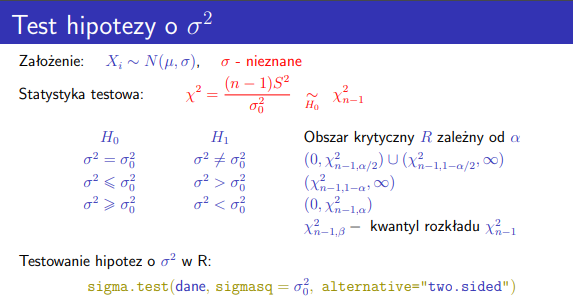
Obszar krytyczny (obszar odrzucenia) testu – zbiór tych wartości statystyki testowej, dla których H0 jest odrzucana.

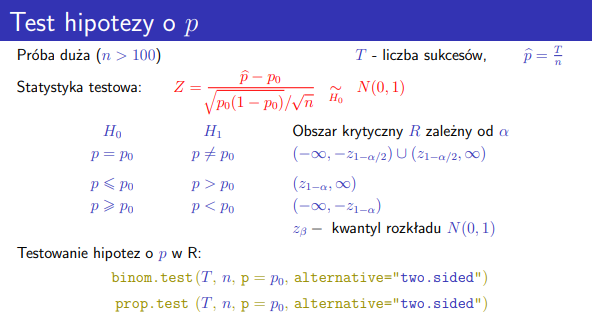




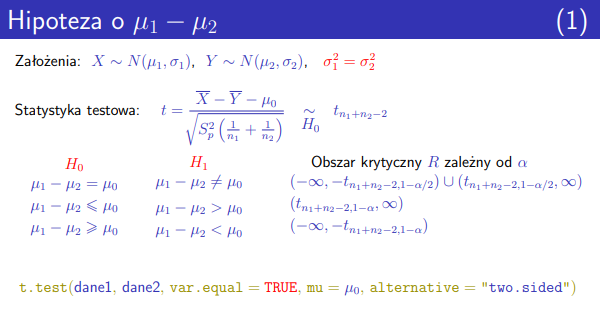


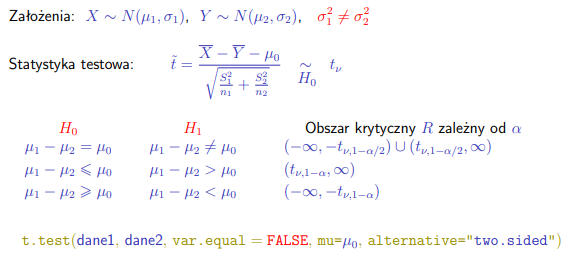
P –value – zmienna losowa oznaczająca wartość „graniczną” decydującą o nieodrzuceniu / odrzuceniu hipotezy H0

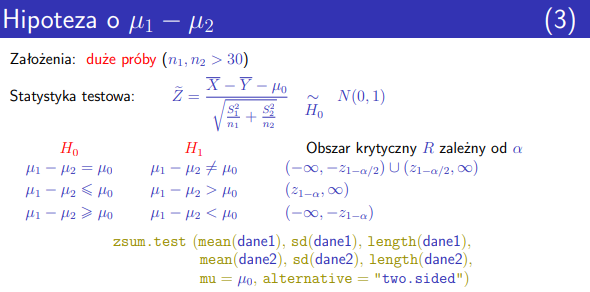




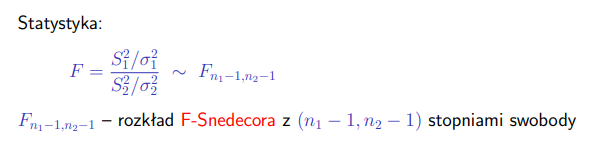
Dwie populacje:

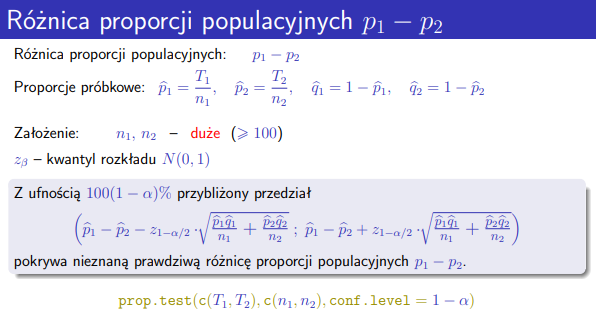


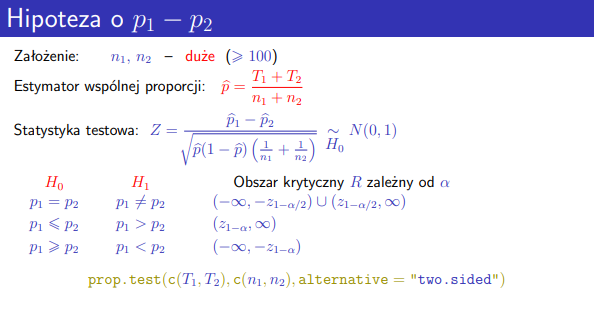




Dla dwóch prób populacyjnych, które są **NIEZALEŻNE** statystyka wygląda następująco:







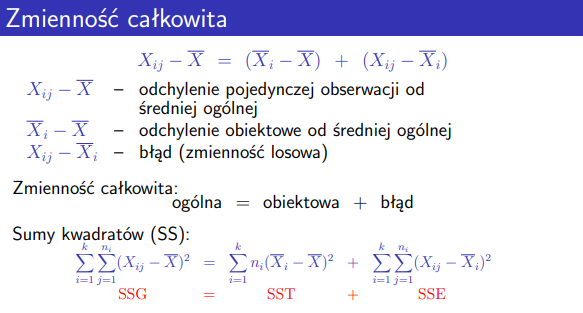
ANOVA – analiza wariancji

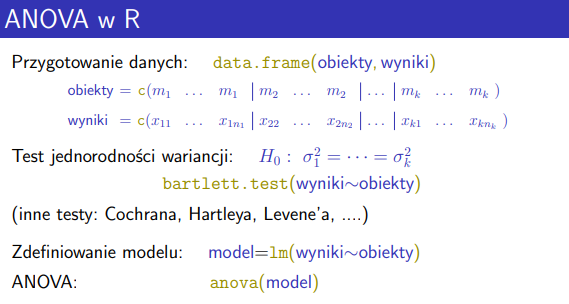
Cele:

* Estymacja punktowa µi oraz σ
* Weryfikacja hipotezy ogólnej o braku różnic między średnimi populacyjnymi, tzn. każdy obiekt ma taki sam wpływ na obserwację:

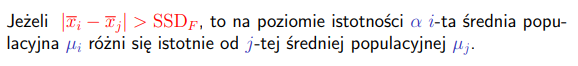


* Wyznaczanie grup jednorodnych (grup obiektów, dla których średnie populacyjne nie różnią się istotnie)





Kiedy jedna średnia populacyjna różni się istotnie od innej średniej populacyjnej:

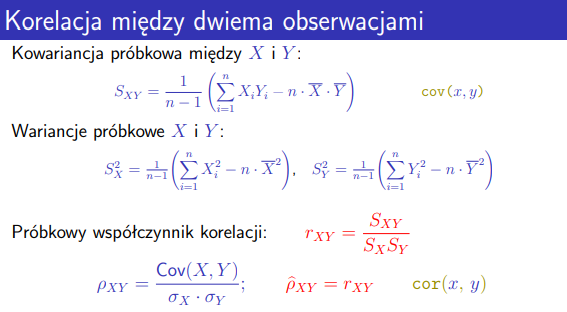


Testy post-hoc:

* Test Scheffiego (najbardziej konserwatywny)
* Test HSD Tukeya - TukeyHSD(aov(wyniki~obiekty))
* Test Newmana Keulsa
* Test Duncana
* Test NIR Fishera (najbardziej liberalny)

Korelacja i analiza regresji:

Próba: pary(Xi, Yi) i = 1,2,3..n



Wartość bezwzględna współczynnika korelacji informuje o sile związku liniowego między cechami X i Y:

* 0 – brak związku liniowego
* 0 - 0.2 – bardzo słaby związek liniowy
* 0.2 – 0.4 – słaby związek liniowy
* 0.4 – 0.6 – umiarkowany związek liniowy
* 0.6 – 0.8 – silny związek liniowy
* 0.8 – 1 – bardzo silny związek liniowy

Analiza regresji – służy do wyrażania zależności między dwiema (lub więcej) zmiennymi losowymi.

Funkcja regresji – matematyczna funkcja opisująca zależność między zmiennymi losowymi:

* Deterministyczna – jeśli każdej wartości zmiennej niezależnej X odpowiada dokładnie jedna wartość zmiennej zależnej Y. Można wówczas zapisać y = f(x)
* Stochastyczna – jeśli każdej wartości zmiennej niezależnej X może odpowiadać kilka wartości zmiennej zależnej Y. Można wówczas zapisać E(Y) = f(x)

Model regresji – funkcja opisująca zależność między X i Y z uwzględnieniem błędu losowego

Regresja prosta:

* Liniowa: Y = br0 + br1x +
* Nieliniowa – jeśli najlepiej dopasowana funkcja jest nieliniowa

Wybór regresji – z wykorzystaniem wykresu korelacyjnego (scatter plot)

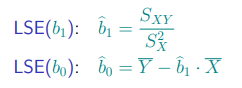
Z modelu liniowego jest jasna interpretacja parametrów. Ponieważ wartości zmiennej obserwowanej są proporcjonalne do wartości zmiennej objaśniającej, wzrost wartości X o 1 pociąga zmianę wartości Y o br1.

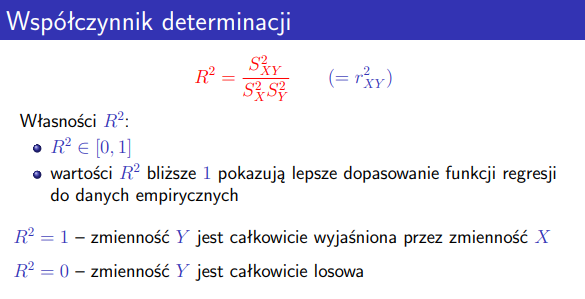
Br0 – wyraz wolny regresji

Br1 – współczynnik regresji

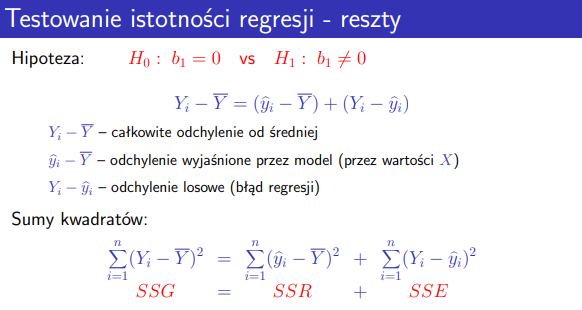
Metoda najmniejszych kwadratów (Least Squares Estimation - LSE) – w celu wyznaczania estymatorów br1, br0 należy zminimalizować wyrażenie:

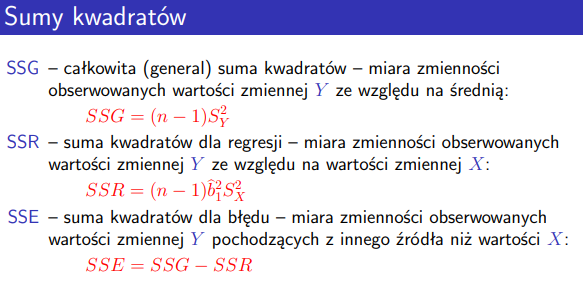






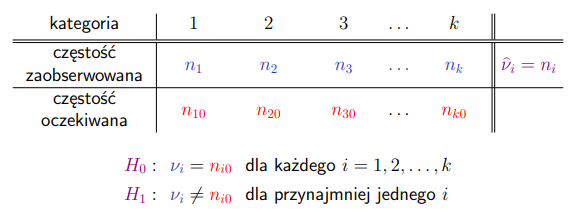
Funkcja do rysowania prostej – abline(prosta)

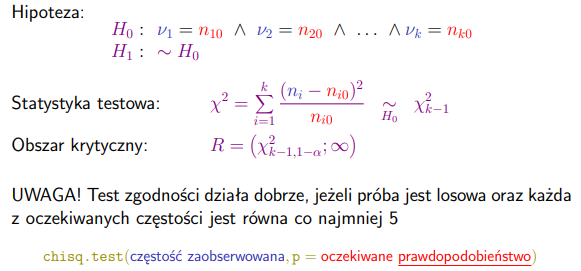




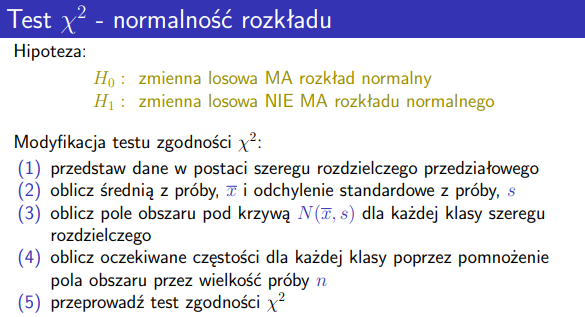
Test zgodności z rozkładem:

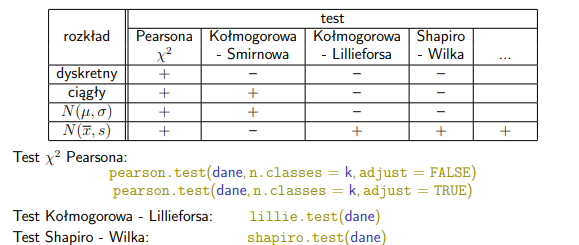
Cel: weryfikacja hipotezy zerowej, że prawdziwy rozkład częstości vi ma określony wzorzec:

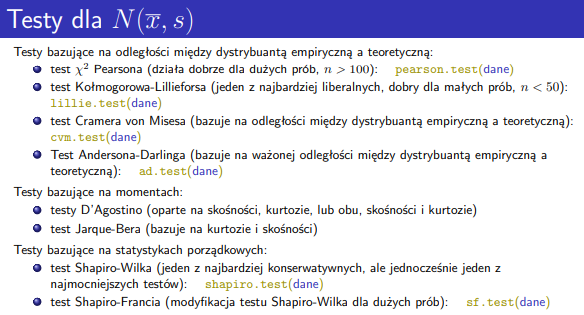




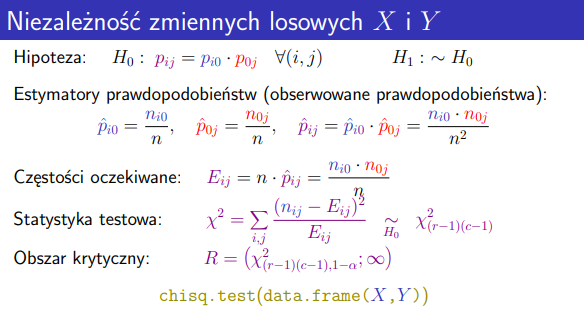
Chi-kwadrat jest testem normalności rozkładu.







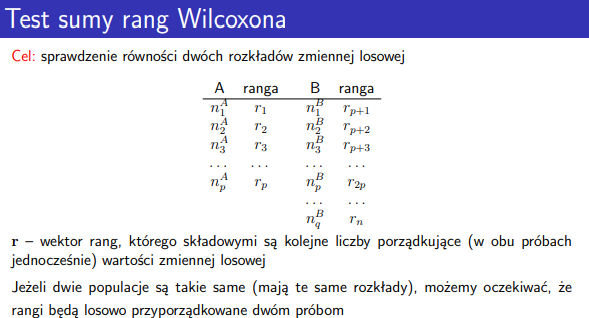
Test niezależności

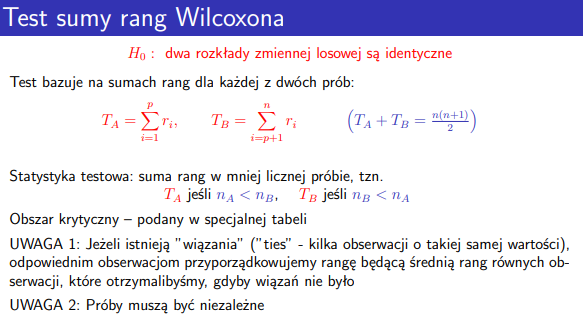


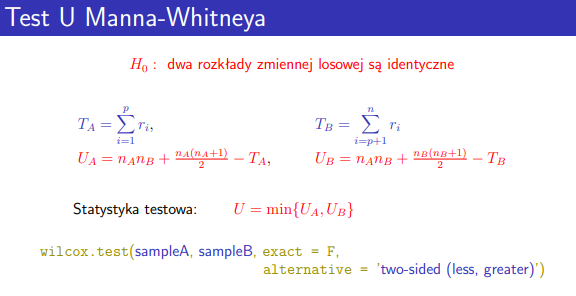
Testy nieparametryczne:

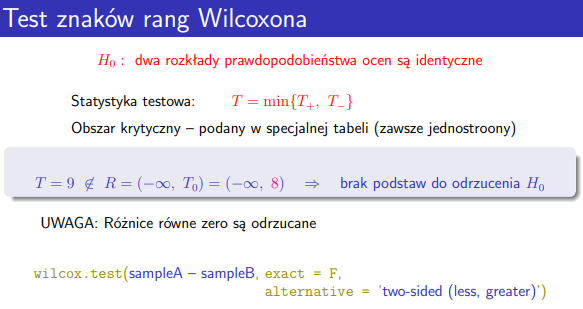
Jeżeli rozkład jest częściowo asymetryczny, wówczas aby porównać średnią, nie można wykorzystać testu *t*, ponieważ założenie o normalności rozkładu może nie być spełnione.

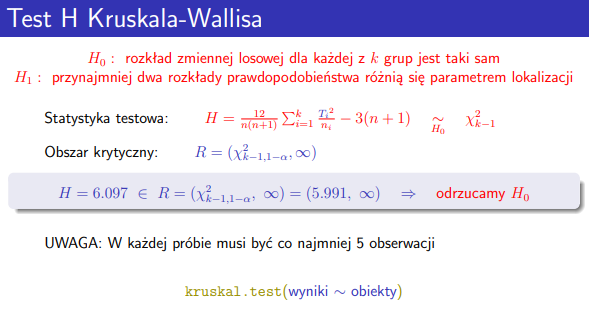
Celem jest sprawdzenie równości dwóch rozkładów zmiennej losowej.

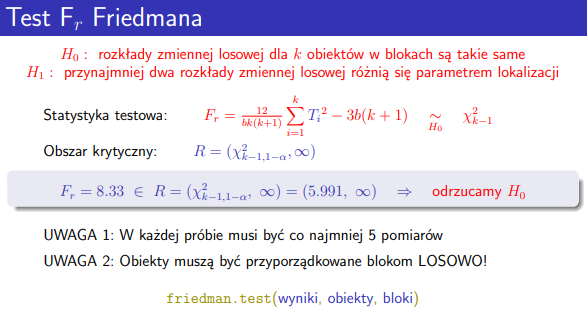


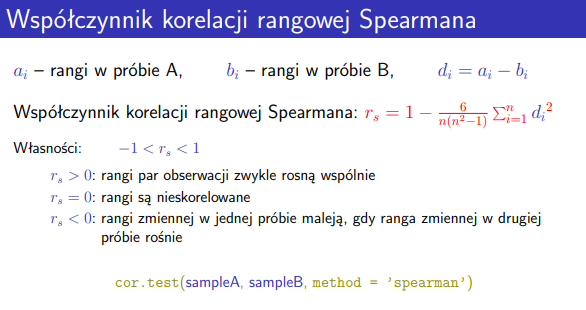


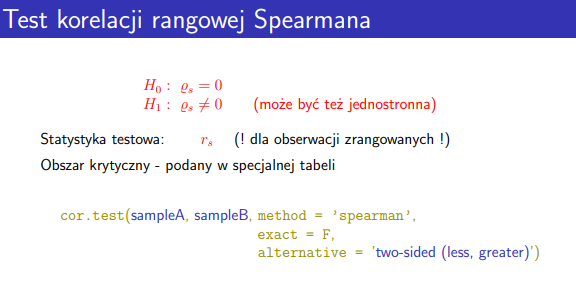












1. Mediana (drugi kwartyl)

a) jest „środkową” wartością w zestawie obserwacji

b) oddziela 75% wyższych obserwacji od 25% niższych obserwacji

c) pojawia się najczęściej wśród wszystkich obserwacji

Wytłumaczenie: Mediana to wartość środkowa, logiczne, patrz strona 2.

2. Rozkład prawdopodobieństwa zmiennej losowej dyskretnej może być przedstawiony:

a) w formie tabeli, z wartościami zmiennej losowej w pierwszym wierszu i odpowiednimi prawdopodobieństwami w drugim wierszu

b) w formie tabeli z wartościami zmiennej losowej w pierwszym wierszu i odpowiednimi częstościami w drugim wierszu

c) jako

Wytłumaczenie: W przypadku zmiennej DYSKRETNEJ można przedstawić wartości w tabeli z poszczególnymi wartościami. W tym przypadku interesuje nas prawdopodobieństwo, więc odp. A

3. Które z poniższych reprezentuje statystyki?

a) proporcja populacji

b) suma zmiennych losowych tworzących próbę

c) średnia populacji

Wytłumaczenie: Statystyki to funkcje danych próbkowych. Proporcja populacji i średnia populacji to parametry!!

4. Niech L i U będą funkcjami zmiennych losowych tworzących próbę spełniającą P(L < sigm^2 < U) = 0.9. Wtedy różnica U –L:

a) jest długością przedziału ufności 90% dla wariancji populacji

b) z prawdopodobieństwem 90% obejmuje prawdziwą wartość wariancji populacji

c) jest przedziałem ufności 90% dla wariancji populacji

Wytłumaczenie: Długość przedziału ufności jest różnicą między górną, a dolną granicą przedziału ufności, który jest dostarczany przez zakres wartości

5. Hipoteza w formie: H0: = 0.9, H1: != 0.9 dotyczy testowania:

a) wariancji populacji

b) poziomu istotności

c) średniej próby

Wytłumaczenie: to wariancja.

6. Załóżmy, że dwie próby są losowo wybrane z populacji o rozkładzie normalnym. Przed skonstruowaniem przedziału ufności dla różnicy średnich populacji, powinniśmy sprawdzić czy:

a) średnie próby są równe

b) wariancje prób są równe

c) wariancje populacji nie różnią się istotnie

Wytłumaczenie: za pomocą bartlett.test musimy sprawdzić jednorodność wariancji. Jednym z założeń testu t jest homogeniczność wariancji próbek.

7. Załóżmy, że przeprowadzono analizę wariancji i hipoteza zerowa o równości kilku średnich została odrzucona. Następnie można przeprowadzić testy post hoc (np. TukeyaHSD). Co możemy wywnioskować z takiego testu?

a) które ze średnich populacji są istotnie podobne / różne

b) które z próbek są jednorodne / niejednorodne

c) ze populacje są / nie są normalnie rozłożone

Wytłumaczenie: test TukeyaHSD pozwala wyznaczyć grupy jednorodne, czyli wartości istotnie podobnych. Jednorodność/niejednorodność – bartlett.test, normalność rozkładu – chi-kwadrat

8. Załóżmy, że wartość kowariancji próbki między dwiema zmiennymi losowymi wynosi (-0.9). Oznacza to, że:

a) wzrost wartości jednej zmiennej implikuje spadek wartości drugiej zmiennej

b) istnieje bardzo silna, ujemna zależność między dwiema zmiennymi losowymi

c) kowariancja nie może być ujemna

Wytłumaczenie: kowariancja może być ujemna. Wskazuje ona, czy istnieje zależność między zmiennymi i czy występuje regresja

9. Która z poniższych funkcji jest przeznaczona do testowania hipotezy o zgodności rozkładu częstości z określonym wzorcem?

a) chisq.test

b) var.test

c) t.test

Wytłumaczenie: test chi-kwadrat jest używany do testowania zgodności rozkładu częstotliwości z określonym wzorcem lub rozkładem, np. normalnym

10. Testy nieparametryczne opierają się na:

a) statystykach konstruowanych jako funkcje miar o rozkładzie normalnym

b) rangach obserwacji

c) graficznych ocenach obserwacji

Wytłumaczenie: współczynnik korelacji rangowej Spearmana:

