**Rachunek prawdopodobieństw i statystyka.**

1. **Wstęp – zdarzenia losowe, prawdopodobieństwo,** 
   1. Pojęcia pierwotne: doświadczenie losowe, zdarzenie elementarne e , Przestrzeń zdarzeń elementarnych Ω.
   2. Zdarzenie losowe definicja: Zdarzeniem losowym nazywamy każdy element przeliczalnie addytywnego ciała Z przestrzeni zdarzeń elementarnych (inaczej sigma ciało zdarzeń).

Uwaga: Gdy Ω jest przeliczalne to każdy podzbiór Ω jest zdarzeniem losowym.

Przeliczalnie addytywne ciało zbioru Ω to niepusta klasa Z podzbiorów zbioru Ω taka, że

1. 
2. Jeśli  to  - stąd wynika, że  - (dopełnienie należy do Z.)
3. Suma co najwyżej przeliczalnej liczby zbiorów Ai należących do Z również należy do Z

Jeśli  to  -

stąd i z ii wynika też, że każda przeliczalna różnica należy do Z, oraz że każdy iloczyn również należy do Z. Z nazywane jest też sigma-algebrą.

Maksymalna ilość elementów w Z to  , gdzie m to ilość zdarzeń elementarnych

* 1. Operacje na zdarzeniach losowych – algebra zbiorów.
     1. Alternatywa (suma):  - zachodzi zdarzenie A lub B
     2. Koniunkcja (iloczyn)  - zachodzi zdarzenie A i B (niekoniecznie jednocześnie w czasie)
     3. Różnica  - zachodzi zdarzenie A i nie zachodzi zdarzenie B
     4.  - zdarzenie przeciwne do A (lub dopełnienie A) stąd mamy:  i 
     5.  - zdarzenie A pociąga za sobą (implikuje) zdarzenie B
     6.  i  to  - zdarzenia są równe
     7.  zdarzenia rozłączne
     8. Podziałem przestrzeni zdarzeń elementarnych jest rodzina  taka, że  dla  i 
  2. Własności działań na zdarzeniach
     1. Przemienność  
     2. Łączność  
     3. Rozdzielność koniunkcji wzg. alternatywy i alternatywy wzg. koniunkcji  
     4. Prawa de Morgana:  
  3. Definicja prawdopodobieństwa – aksjomatyczna (Kolmogorow):

Prawdopodobieństwo to funkcja określona na zbiorze Z i przyporządkowująca każdemu elementowi tego zbioru liczbę P(A) zgodnie z następującymi warunkami:

1.  dla każdego 
2. 
3. Dla dowolnego ciągu parami rozłącznych zdarzeń Ai ( dla  ) należących do zbioru Z zachodzi:  - stąd i z poprzednich dwóch punktów wynika, że 
   1. Własności prawdopodobieństwa
      1. 
      2. Jeśli  to 
      3. 
      4. Jeśli  to 
      5. 
      6. 
      7. Jeśli przestrzeń zdarzeń elementarnych jest co najwyżej przeliczalna i określone są prawdopodobieństwa wszystkich zdarzeń elementarnych to  jeśli zdarzenia elementarne  sprzyjają zdarzeniu A
      8. Klasyczna definicja prawdopodobieństwa (Laplace)

Jeśli przestrzeń składa się z n zdarzeń elementarnych i wszystkie zdarzenia elementarne są jednakowo prawdopodobne to  , gdzie n(A) to liczba zdarzeń elementarnych sprzyjających zdarzeniu A. 🡪 przykład 3 teleturniej 3 drzwi.

* 1. Przestrzeń probabilistyczna: 
  2. Prawdopodobieństwo warunkowe: 
  3. Niezależność zdarzeń: 
  4. Prawdopodobieństwo zupełne

Jeśli Ai są parami rozłączne i  oraz  wówczas 

* 1. Twierdzenie Bayesa:  - prawdopodobieństwo a priori i a posteriori.
     1. czułość **testu diag**. (objawu): prawdopodobieństwo że test wypadnie dodatnio zakładając, że pacjent jest rzeczywiście chory P(A|B) (A-jest objaw, pozytywny wynik testu, B – pacjent jest chory na daną chorobę).
     2. swoistość **testu diag**. (objawu): prawdopodobieństwo że test wypadnie ujemnie zakładając, że pacjent nie jest chory. P(A’|B’)
  2. Elementy kombinatoryki:
     1. Ilość permutacji zbioru n-elementowego – n!
     2. Ilość kombinacji k elementowych zbioru n elementowego (bez powtórzeń) 
     3. Ilość kombinacji k elementowych zbioru n elementowego (z powtórzeniami) 
     4. Ilość wariacji bez powtórzeń 
     5. Ilość wariacji k-elementowych z powtórzeniami 
  3. Prawdopodobieństwo w binarnych eksperymentach powtarzalnych

 gdzie p – prawdopodobieństwo sukcesu, q=1-p – prawdopodobieństwo porażki, Pn,m – prawdopodobieństwo, że w n próbach będzie m sukcesów.

1. **Zmienna losowa – dyskretna i ciągła, rozkład prawdopodobieństwa zmiennej losowej.**
   1. Definicja zmiennej losowej:

Niech  będzie dowolną przestrzenią probabilistyczną. Zmienna losowa to dowolna funkcja X określona na zbiorze zdarzeń elementarnych o wartościach ze zbioru liczb rzeczywistych, taka że dla dowolnego ustalonego  zbiór zdarzeń elementarnych, dla których przyjmuje ona wartość mniejszą niż x jest zdarzeniem losowym.



* + 1. Powyżej zdefiniowany podzbiór jest zdarzeniem losowym.
    2. Zdarzeniem losowym są też zbiory:  gdzie R to dowolny zbiór borelowski na prostej.
  1. Definicja dystrybuanty.

Niech  będzie dowolną przestrzeń probabilistyczną. Funkcję FX określoną na zbiorze  wzorem:

 dla 

Nazywamy dystrybuantą zmiennej losowej X.

* 1. Własności dystrybuanty
     1.  dla każdego 
     2.  
     3. F jest funkcją niemalejącą
     4. F jest co najmniej lewostronnie ciągła 
     5. 
     6.  stąd wynika, że gdy F jest ciągła 
     7. Dowolna funkcja G o wartościach rzeczywistych spełniająca warunki b, c, d jest dystrybuantą.
  2. Zmienna losowa typu skokowego (dyskretnego)
     1. Zmienna losowa jest typu dyskretnego, jeśli zbiór jej wartości  jest przeliczalny lub skończony, oraz  dla każdego i, takie, że  ( to punkty skokowe zmiennej X)
     2. Funkcję  nazywamy rozkładem prawdopodobieństwa zmiennej losowej X.
     3. Dystrybuanta zmiennej skokowej 
     4. Znając dystrybuantę można łatwo obliczyć rozkład prawdopodobieństwa korzystając z własności 3f.
  3. Zmienna losowa typu ciągłego
     1. Zmienną losową nazywamy ciągłą jeśli przyjmuje wszystkie wartości z pewnego przedziału i istnieje taka nieujemna funkcja f, że dystrybuantę zmiennej losowej X możemy przedstawić jako  . f nazywa się gęstością prawdopodobieństwa, lub gęstością.
     2. Jeżeli x jest punktem ciągłości f to 
     3. 
     4. 
     5. 
  4. **Funkcje zmiennych losowych**
     1. Jeśli X jest zmienną losową, a g jest dowolną funkcją o wartościach rzeczywistych określoną na zbiorze wartości zmiennej X to wtedy 

( ) jest zmienną losową o rozkładzie:

* + - 1. Dla zmiennej dyskretnej: 
      2. Dla zmiennej ciągłej.

**Twierdzenie:** Jeśli g(X) jest funkcją ściśle monotoniczną, oraz  dla  wówczas gęstość zmiennej U wyraża się wzorem:  , gdzie  oznacza funkcję odwrotną do .

**Dowód** jest prosty. Załóżmy, że wybieramy odcinek (x1, x2), wówczas zmiana zmiennych w całce poniżej pozwala szybko odnaleźć gęstość zmiennej losowej u (przyjmujemy, że u1=g(x1) i u2=g(x2):



Gdy g jest rosnąca (u2>u1) wyrażenie w ostatniej całce to gęstość zmiennej losowej u. Jeśli g jest malejąca to należy zmienić kolejność granic w ostatniej całce. Ponieważ monotoniczność funkcji odwrotnej jest taka sama jak funkcji wyjściowej, więc ostatecznie otrzymujemy: 

**Twierdzenie**: Jeśli u=g(x) jest funkcją przedziałami ściśle monotoniczną,  dla i=0,1,2,…,n to granice odpowiednich przedziałów, a  w tych przedziałach to gęstość zmiennej losowej U wyraża się wzorem:



gdzie  dla i=1,2,…,n oraz  i 

Dowód jest bezpośrednim wnioskiem z poprzedniego twierdzenia.

* 1. **Niezależność zmiennych losowych** – Zmienne losowe X i Y określone na tej samej przestrzeni zdarzeń losowych są niezależne jeśli zdarzenia  i  są niezależne, tzn. gdy zachodzi: 
     1. Dla zmiennych losowych dyskretnych: 
     2. Dla zmiennych losowych ciągłych 

1. **Zmienna losowa – charakterystyki liczbowe i przykładowe rozkłady.**
   1. Miary położenia
      1. Wartość przeciętna, średnia, oczekiwana -  (zakładamy, że całka i szereg są zbieżne) i jej własności (c to stała):
         1. 
         2. 
         3. 
         4. 
         5. 
         6.  gdy X i Y są niezależne
      2. Mediana – każda liczba , że 
      3. Kwantyl rzędu p – każda liczba , że 
      4. Moda –wartość najbardziej prawdopodobna
   2. Miary rozrzutu
      1. Wariancja  , właściwości
         1. 
         2. 
         3. 
         4.  gdy X i Y są niezależne
         5. 
      2. Odchylenie standardowe 
      3. Współczynnik zmienności 
   3. Momenty zwykłe 
   4. Moment centralny 
   5. Współczynnik skośności 
   6. Współczynnik skupienia – kurtoza  (K>0 – leptokurtyczny, K<0 – platokurtyczny)
   7. Rozkłady skokowe:
      1. Równomierny  , 
      2. Jednopunktowy  , 
      3. Zero-jedynkowy, dwupunktowy, Bernoulliego , 
      4. Rozkład dwumianowy zmiennej K=0,1,2,…,n – (k sukcesów w n etapowym eksperymencie binarnym z rozkładem zero-jedynkowym)



* + - 1. Dla n=1 🡪 rozkład zero-jedynkowy,
      2. dla n>1 🡪 K to suma zmiennych niezależnych o rozkładzie zero-jedynkowym
      3. Dla n🡪∞ (p stałe) rozkład dwumianowy dąży do rozkładu Gaussa.
      4. Rozkład dwumianowy dąży do rozkładu Poissona w granicy n🡪∞ i p🡪0, tak, że np=λ

 gdy 

* + 1. Rozkład Poissona – (ilość zdarzeń w jednostce czasu – czas pojawienia się zdarzenia określony jest rozkładem wykładniczym) zmienna losowa K=0,1,2,3,… ma rozkład Poissona gdy:





* + - 1. Rozkład dwumianowy dąży do rozkładu Poissona w granicy n🡪∞ i p🡪0, tak, że np.=λ

 gdy 

* + - 1. Dla dużej wartości λ i dużych wartości k rozkład Poissona może być przybliżony rozkładem Gaussa o średniej λ i wariancji λ.
  1. Rozkłady ciągłe
     1. Rozkład równomierny – skoncentrowany na przedziale [a,b]





* + 1. Rozkład wykładniczy – np. rozkład czasu bezawaryjnej pracy danego elementu, lub czas w którym pojawi się samochód na jakiejś ulicy, lub czas między jednym a drugim rozpadem jądra, prawdopodobieństwo przetrwania jądra niestabilnego przez czas x, okres między trzęsieniami ziemi, itp.

Można go wyprowadzić wychodząc z założenia, że czas oczekiwania na zdarzenie można modelować jako eksperyment Bernoulliego, w którym czas oczekiwania jest podzielony na bardzo dużą ilość prób Bernoulliego n🡪∞, a prawd. porażki w każdej z tych prób jest liczbą stałą równą 1-t/(λn), gdzie t to czas czekania. Wówczas prawdopodobieństwo, że czas oczekiwania jest większy lub równy t wynosi:  - z rozwinięcia Taylora wzg. T=0 i w granicy n🡪 ∞. – Stąd prawd. że nastąpił sukces w czasie krótszym niż t to 1-exp(-t/λ)



 λ to czas życia

* + - 1. Brak pamięci -  Prawdopodobieństwo, że czas oczekiwania na zjawisko jest dłuższy niż a+b pod warunkiem że minął już czas a jest takie samo jak prawdopodobieństwo, że czas oczekiwania jest dłuższy niż b.
    1. Rozkład normalny, Gaussa





Rozkład standaryzowany  🡪 N(0,1)

* + 1. Inne rozkłady ważne w statystyce: gamma, chi2, F-Snedecora, t- Studenta.
  1. Rozkłady ucięte 🡪 prawdopodobieństwo warunkowe.
  2. **Centralne Twierdzenie Graniczne Lindeberga-Levy’ego**:

Jeśli Xn jest ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie, o skończonej wartości oczekiwanej µ i wariancji σ2>0, oraz  to ciąg standaryzowanych średnich arytmetycznych  wówczas ciąg dystrybuant  jest zbieżny do dystrybuanty rozkładu normalnego standaryzowanego N(0,1).



Jeżeli przyjmiemy, że średnia  to średnia z próbki wówczas możemy powiedzieć, że bez względu na rozkład statystyczny cechy w danej populacji rozkład średnich z próbek zbliża się do rozkładu normalnego wraz ze wzrostem rozmiaru próbki.

**Dowód** wymaga odwołania się do funkcji charakterystycznej rozkładu.

* 1. **Mocne prawo wielkich liczb Kołmogorowa:**

Jeśli Xn jest ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie, o skończonej wartości oczekiwanej µ, to wówczas zachodzi mocne prawo wielkich liczb, tzn. że dla

 zachodzi



Mówiąc inaczej średnia z próbki jest dobrym przybliżeniem średniej z populacji jeśli próbka jest dostatecznie duża. Oznacza to, że średnia z próbki jest zgodnym estymatorem.

1. **Zmienna losowa dwuwymiarowa – korelacja i regresja.**
   1. **Dwuwymiarowa zmienna losowa** – Parę zmiennych (X,Y) zmiennych losowych X i Y określonych nie koniecznie na tej samej przestrzeni probabilistycznej nazywamy dwuwymiarową zmienną losową.
      1. ***Dystrybuanta*** – Funkcja  taka, że , własności:
         1. 
         2. 
         3. Dla dowolnych punktów (x1,y1) i (x2,y2) takich, że x1<x2 i y1<y2 zachodzi: 

* Interpretacja geometryczna
  + - 1. F jest niemalejąca i przynajmniej lewostronnie ciągła wzg. Każdego argumentu.
    1. ***Rozkłady brzegowe*** -  dystrybuanta X, gdy Y przyjmuje dowolne wielkości  podobnie dla Y
    2. ***Dwuwymiarowa zmienna losowa typu skokowego -***  i zachodzi 
       1. Rozkłady brzegowe:  i analogicznie dla P(Y=y) 
       2. Rozkłady warunkowe – Jeśli wszystkie p.k są dodatnie to możemy zdefiniować prawdopodobieństwo warunkowe jako:  analogicznie dla P(Y=yk|X=xi)
    3. ***Dwuwymiarowa zmienna losowa (X,Y) jest typu ciągłego*** jeśli istnieje taka nieujemna funkcja f taka, że dystrybuantę tej zmiennej można przedstawić w postaci:  f to gęstość rozkładu prawdopodobieństwa, a F(x,y) to odpowiednia objętość bryły pod wykresem f.
       1. Własności f
          - 
          - W punktach ciągłości (x,y) mamy 
          - Dla obszaru regularnego  mamy , w szczególnym przypadku, gdy B jest prostokątem o skrajnych współrzędnych a,b,c,d, mamy: 
       2. Rozkłady brzegowe -  i analogicznie dla F2(y)
       3. Rozkład warunkowy - 
  1. **Niezależność zmiennych losowych** – Zmienne X i Y zdefiniowane na tej samej przestrzeni zdarzeń elementarnych są niezależne jeśli dla dowolnych borelowskich zbiorów A i B zdarzenia  i są niezależne, czyli  Stąd warunkiem koniecznym i wystarczającym, aby zmienne losowe X i Y były niezależne jest: 
  2. Charakterystyki liczbowe dwuwymiarowej zmiennej losowej
     1. Momenty zwykłe mieszane rzędu r+s 
     2. Momenty centralne mieszane rzędu r+s 

Macierz momentów centralnych rzędu 2 

 - dla niezależnych zmiennych losowych X i Y cov(X,Y)=0

* + 1. Współczynnik korelacji liniowej – Jeśli V(X)>0 i V(Y)>0 to współczynnik  jest miarą zależności liniowej między zmiennymi X i Y. Można wykazać, że jeśli istnieje liniowa zależność między zmiennymi X i Y to jest równe 1 lub -1 to z prawdopodobieństwem 1. Dowód łatwy – z własności V i E. 
  1. Funkcje zmiennych losowych – przypadek szczególny: suma, iloczyn
  2. Linie regresji.
     1. Linie regresji pierwszego rodzaju – określmy warunkowe wartości przeciętne dla rozkładu skokowego:  i dla rozkładu ciągłego:  . Niech  wówczas linią regresji pierwszego rodzaju zmiennej losowej Y wzg. X nazywamy zbiór punktów (x,y) spełniających równanie: .
        1. **Własność**: Średnie odchylenie kwadratowe  zmiennej losowej Y od pewnej funkcji g(X) jest najmniejsze gdy funkcja ta z prawdopodobieństwem 1 jest równa . 

**Dowód:** Powyższa własność wynika z następującej nierówności: ,którą najpierw udowodnimy: 

Stąd wynika, że E((X-c)2) jest minimalne dla c=E(X). Dla ciągłych zmiennych losowych: 

Wyrażenie w nawiasie jest analogiczne do E((X-c)2) ponieważ dla ustalonego x g(x) jest wartością stałą. Wykorzystując udowodnioną powyżej nierówność otrzymujemy oczekiwaną własność. Dla skokowych zmiennych losowych dowód jest analogiczny.

* + 1. Prostą regresji drugiego rodzaju zmiennej losowej Y wzg. zmiennej losowej X nazywamy prostą o równaniu y=ax+b, której współczynniki minimalizują średnią odległość kwadratową między zmiennymi losowymi Y i aX+b. 

Łatwo wykazać, że dla dowolnej dwuwymiarowej zmiennej losowej (X,Y) współczynniki a i b określone są wzorami: 

**Dowód**: Łatwy – wystarczy rozwinąć wyrażenie E((Y-aX-b)2) i zminimalizować przez przyrównanie pierwszych pochodnych po a i b do zera.

* + - 1. **Własność**: Jeśli zmienne losowe X i Y mają rozkład normalny to linia regresji pierwszego rodzaju jest równa prostej regresji drugiego rodzaju.

1. **Statystyka wstęp: statystyka opisowa, skale pomiarowe, estymatory punktowe, przedział ufności.**
   1. Co to jest statystyka i czym się zajmuje? 🡪 opis i estymacja, czyli przewidywanie parametrów dla całej populacji na podstawie badań na próbkach. Pojęcia podstawowe: populacja, próbka, estymatory.
   2. mierzone wielkości i skala pomiarowa: jakościowa i ilościowa. jakościowa: nominalna, porządkowa. ilościowa: interwałowa (równomierna) 🡪 ciągła i dyskretna, ilorazowa
      1. **nominalna** - wynikiem pomiaru jest rozłączna kategoria, np.: kolor oczu, płeć, grupa krwi,
      2. **porządkowa** - podobnie jak nominalna, tylko że wyniki można jednoznacznie uporządkować, np.: stopień znajomości języka: podstawowy, średnio zaawansowany, zaawansowany, biegły, lub masa ciała: niedowaga, norma, nadwaga, otyłość. Skala ta może być wyrażana przy pomocy cyfr, np. tak i nie to 1 i 0, lub skala Apgar (0-10)
      3. **przedziałowa** (interwałowa, równomierna) - tak jak porządkowa, tylko że można obliczyć odległość między wynikami, większość pomiarów należy do tej skali, np.: ciśnienie krwi, masa ciała, temperatura
      4. **ilorazowa** - to samo co skala przedziałowa z tym że iloraz ma sens (istnieje bezwzględne zero), np. wiek,
   3. Sposoby przedstawiania surowych danych (szeregi statystyczne: szeregi szczegółowe, rozdzielcze i czasowe ):
      1. histogramy, zwykłe i skumulowane - skala przedziałowa/ilorazowa - zmienne ciągłe
      2. wykresy słupkowe - zmienne dyskretne - realizowane w statistica przez histogram
      3. wykresy kołowe - wszystkie skale
      4. łodyga i liście - skala przedziałowa /ilorazowa (diagram łodyga i liście - stat. podstawowe)
      5. wykresy rozrzutu - skala przedziałowa/ilorazowa
   4. Statystyka opisowa 🡪 tak jak charakterystyki liczbowe zmiennej losowej, tylko, że dla próbki. Zakładamy prawdopodobieństwo jednostajne, czyli pi=1/n
   5. Estymatory, statystyki
      1. Załóżmy że badamy cechę X pewnej populacji. Wówczas n-elementową próbę prostą możemy przedstawić jako ciąg Xn niezależnych zmiennych losowych o takim samym rozkładzie jak X.
      2. Statystyka to dowolna funkcja g zdefiniowana na próbie 
      3. Estymatorem nazywamy każdą statystykę , której wartości przyjmujemy jako oceny szukanego parametru rozkładu zmiennej X. Dla danego parametru można skonstruować wiele estymatorów. Estymator jest zmienną losową (ponieważ jest funkcją zmiennych losowych) i ma swój rozkład. Optymalny estymator jest:
         1. nieobciążony, czyli 
         2. zgodny 
         3. efektywny – najmniejsza wariancja 
      4. **Przykład 1:** średnia z próbki  jest zgodnym (na mocy twierdzenia wielkich liczb) nieobciążonym () estymatorem średniej z populacji o wariancji  - dla rozkładu normalnego jest to także estymator efektywny.
      5. **Przykład 2:** wariancja  jest estymatorem obciążonym:   

ponieważ

****

Jest to estymator asymptotycznie nieobciążony, ponieważ 

Estymatorem nieobciążonym wariancji z populacji jest  , który jest także asymptotycznie efektywny. Oba estymatory są zgodne.

* + 1. Uwaga jeśli dany estymator  ma pewne pożądane własności, to inny estymator będący funkcją tego estymatora wcale nie musi ich mieć. Np.  jest obciążonym estymatorem odchylenia standardowego, mimo że  jest nieobciążonym estymatorem wariancji.
  1. Sposoby znajdowania estymatorów.
     1. Metoda największej wiarogodności – Jeśli chcemy oszacować k parametrów  na podstawie n-elementowej próby wówczas możemy użyć funkcji wiarogodności L



Gdzie f to gęstość prawdopodobieństwa (zmienne ciągłe), lub funkcja rozkładu prawdopodobieństwa (zmienne dyskretne). Estymatory otrzymuje się przez maksymalizację L, a dokładniej ln(L) co jest wygodniejsze.

 , oraz macierz  jest negatywnie określona, czyli wszystkie jej wartości własne są ujemne.

Estymatory otrzymane tą metodą nie zawsze są optymalne, ale przy dostatecznie regularnych funkcjach rozkładu są zgodne.

* + 1. **Metoda momentów** – obliczamy momenty (zwykle kilka pierwszych) dla próby i przyrównujemy je do momentów otrzymanych z modelowanego rozkładu. W ten sposób otrzymujemy równania, z których wyliczamy parametry rozkładu. Ilość użytych momentów musi być oczywiście co najmniej równa ilości estymowanych parametrów.
    2. Metoda najmniejszych kwadratów 🡪 przy okazji regresji.
  1. Inne estymatory:
     1. Wskaźnik struktury p –  - zgodny, nieobciążony, efektywny - tylko dla rozkładu Bernoulliego.
     2. Współczynnik zmienności - 
  2. Estymacja przedziałowa średniej
     1. Na mocy centralnego twierdzenia granicznego dla dostatecznie dużych próbek średnia z próbki ma rozkład normalny .
     2. Zakładamy, że w realnym eksperymencie (pomiarze) nie otrzymamy wartości zmiennej losowej X, które są bardzo mało prawdopodobne. Wówczas możemy przyjąć, że zredukowana zmienna losowa  znajdzie się w przedziale , gdzie  to kwantyl rzędu  standaryzowanego rozkładu normalnego. Stąd otrzymujemy, że średnia z populacji powinna być w przedziale:  , który nazywamy przedziałem ufności średniej. Przedział ten zależy od konkretnej wartości elementów w próbie i jest różny dla różnych prób. Jeśli skonstruujemy wiele takich przedziałów, dla różnych prób, to w  procentach takich przedziałów powinna znajdować się prawdziwa wartość średniej z populacji, która jest wartością stałą. α jest wartości małą (zwykle 0.05, lub 0.01) i nazywane jest poziomem istotności. 1- α to poziom ufności.
     3. Jeśli nie znamy wariancji z całej populacji (co ma zwykle miejsce), to wówczas zamiast rozkładu normalnego używamy rozkładu t-Studenta, a zamiast wariancji dla całej populacji używamy estymatora dla próby S\*2. Wówczas otrzymujemy następujący przedział ufności dla średniej: , gdzie  to kwantyl rzędu rozkładu t-Studenta z n-1 stopniami swobody. Dla dużych n statystyka t-Studenta dąży do rozkładu normalnego standaryzowanego.
  3. Estymacja przedziałowa odchylenia standardowego
     1. Dla zmiennej X o rozkładzie normalnym korzystamy ze statystykiktóra ma rozkład chi-kwadrat. Wówczas odchylenie standardowe z populacji należy do przedziału:



Gdzie  i  to odpowiednie kwantyle rozkładu chi-kwadrat z n-1 stopniami swobody.

Dla dużych próbek (n>50) można stosować przybliżenie rozkładem normalnym.

1. **Statystyka – testowanie hipotez statystycznych, parametryczne testy istotności.**
   1. *Hipoteza statystyczna* to każde przypuszczenie dotyczące rozkładu badanej cechy w populacji. Przyjmujemy zwykle dwie hipotezy, tzw. Zerową H0 i alternatywną H1 (HA), które się wzajemnie wykluczają.
      1. Hipotezy dwustronne – np.:  
      2. Hipotezy jednostronne – np.:  
   2. *Weryfikacja hipotez* dokonuje się przez obliczenie statystyki testowej na próbie. W zależności od wartości statystyki wybieramy jedną z hipotez. Rodzaj użytej statystyki zależy od rodzaju problemu. Wybór hipotezy nie oznacza, że jest ona prawdziwa. W każdym wypadku możemy popełnić błąd.

|  |  |  |
| --- | --- | --- |
| prawdopodobieństwo | H0 prawdziwa | H1 prawdziwa |
| Nie odrzucamy H0 | ok - 1-α | β – błąd 2 rodzaju |
| akceptacja H1 | α - błąd 1 rodzaju | ok - 1-β |

Moc testu to prawdopodobieństwo 1-β, że jeśli hipoteza H1 jest prawdziwa to H1 zostanie zaakceptowana.

* 1. Istnieje relacja między błędem pierwszego i drugiego rodzaju. Im większy błąd I rodzaju tym mniejszy błąd II rodzaju i odwrotnie. Wartość  powinna być niewielka (zwykle 0.05 lub 0.01), ale taka aby moc testu była duża. 🡪 rys.
  2. Moc testu – Im test jest mocniejszy tym łatwiej wykazać prawdziwość H1. Moc testu zależy od wielkości próby, sformułowania hipotez oraz rozrzutu wyników w próbie. Im większa próba i mniejszy rozrzut wyników tym test jest mocniejszy.
  3. Testy parametryczne – zakładają pewien rozkład cechy w populacji – najczęściej jest to rozkład normalny.
  4. Przykładowe testy parametryczne:
     1. Test dla jednej próby – porównanie z wartością tablicową:
        1. Założenia – Próbka ma rozkład normalny, lub jest dostatecznie duża, tak, ze można skorzystać z Centralnego Twierdzenia Granicznego.
        2. Hipotezy:
        3. Statystyka: 
        4. Obszar krytyczny: Jeśli  to wybieramy hipotezę H1. W przeciwnym przypadku zostajemy przy H0 i mówimy że nie mamy podstaw do odrzucenia hipotezy zerowej.
        5. Uwaga: Ten test jest równoważny sprawdzeniu, czy  należy do przedziału ufności dla średniej z populacji: . Jeśli tak to nie mamy podstaw do odrzucenia hipotezy zerowej zakładającej, że  jest średnią z populacji. W przeciwnym wypadku wybieramy hipotezę alternatywną.
        6. Rys. wyjaśniający test oraz błędy I i II rodzaju. Wartość P.
     2. Test dla porównania średnich z dwóch prób zależnych (związanych) – test t-Studenta dla różnic
        1. Założenia: Różnica z próbek ma rozkład normalny, lub próbki są dostatecznie duże, aby można było zastosować Centralne Twierdzenie Graniczne.
        2. Hipotezy:
        3. Statystyka:  gdzie  to średnia z różnic związanych cech.
        4. Obszar krytyczny: Jeśli  to wybieramy hipotezę H1. W przeciwnym przypadku zostajemy przy H0 i mówimy że nie mamy podstaw do odrzucenia hipotezy zerowej. Dla wartości P: Jeśli  🡪 H0 .Jeśli  🡪 H1.
        5. Uwaga: Ten test jest równoważny testowi dla jednej próbki jeśli za elementy naszej próbki weźmiemy różnice elementów związanych: 
     3. Test dla porównania średnich z dwóch prób niezależnych (niezwiązanych) – test t-Studenta
        1. Założenia: Obie próbki mają rozkład normalny, lub są dostatecznie duże, aby można było zastosować Centralne Twierdzenie Graniczne.
        2. Hipotezy:
        3. Statystyka:  , gdzie  i ilość stopni swobody df=2n-2 – dla próbek równolicznych.
        4. Obszar krytyczny: Jeśli  to wybieramy hipotezę H1. W przeciwnym przypadku zostajemy przy H0 i mówimy że nie mamy podstaw do odrzucenia hipotezy zerowej. Dla wartości P: Jeśli  🡪 H0 .Jeśli  🡪 H1.
        5. Uwaga: Istnieją również wersje tego testu dla próbek z różną wariancją, oraz różnolicznych. Wtedy używa się bardziej skomplikowanych statystyk, ale reguły wyboru hipotez są takie same.
     4. Testy pomocnicze – Często prócz testów głównych potrzebne są inne testy np. w celu zweryfikowania założeń testów głównych.
        1. Test Shapiro-Wilka – sprawdzanie normalności rozkładu próbki. Hipotezy: H0: Rozkład próbki jest zgony z rozkładem normalnym H1: Rozkład próbki różni się od rozkładu normalnego
        2. Test Levena – sprawdzanie jednorodności wariancji. H0: wariancje są jednorodne H1: wariancje są różne.

1. **Zależność między zmiennymi – wsp. Korelacji liniowej i regresja.**
2. Wstępne określenie rodzaju zależności na podstawie wykresu rozrzutu.
3. Jeśli jednocześnie zachodzą (relacja liniowa, nie ma wyników odstających, ani podgrup, normalny rozkład obu zmiennych) wtedy stosujemy **współczynnik korelacji liniowej Pearsona r.** Tutaj r szacowane jest dla próby. Jest to zgodny, ale obciążony estymator wartości tego współczynnika (ρ) dla populacji.



* 1. r2 – **współczynnik determinacji** jest miarą (ułamkową) zmienności y, która może być wyjaśniona jej liniową zależnością od x --> rysunek przy regresji
  2. Testowanie hipotez:
     1. H0: ρ=0, H1: ρ≠0 🡪 zmienna testowa test t-studenta z n-2 stopniami swobody
     2. H0: ρ=ρ0, H1: ρ≠ρ0 🡪 zmienna testowa Gaussian test - transformacja odwrotna 

Przedział ufności dla z --> 

1. **Regresja liniowa** – regresja drugiego rodzaju – obliczana wtedy gdy zachodzą: relacja liniowa, nie ma wyników odstających, ani podgrup, normalny rozkład obu zmiennych. Dla pary zmiennych (X,Y) możemy obliczyć regresję Y wzg. X, X wzg. Y lub, regresję ortogonalną. Załóżmy, że interesuje nas regresja Y wzg. X. Zakładamy, że w populacji istnieje następująca zależność: Y=aX+b. Wówczas estymatorami parametrów a i b obliczonymi dla próby o wielkości n są:



Krzywa Y=AX+B minimalizuje odległość między yi i AX+B. Współczynniki A i B zostały obliczone metodą najmniejszy kwadratów.

1. Błąd standardowy estymacji



1. Test hipotezy H0:a=0 jest taki sam jak dla testowanie istotności współczynnika korelacji liniowej.
2. Przedział ufności dla współczynnika kierunkowego a:



1. Przedział ufności dla wyrazu wolnego b



1. Obszar ufności dla prostej regresji. Jeśli mamy próbkę dwuwymiarową (xj,yj) j=1,…,n to możemy utworzyć dwuwymiarowe klasy o środkach (xi,yk), i=1,…,l k=1,…,m. Wówczas przedział ufności dla wartości oczekiwanej Y dla danego Xi (Y|Xi) wyraża się wzorem:

