Zadania ze statystyki cz.3 I rok socjologii ekonomicznej

Zadanie 1

Zdaniem wielu komentatorów, kobiety częściej niż mężczyźni głosują na partię rządzącą. Wyniki badań przedstawia tabela poniżej. Należy sprawdzić czy istnieje związek między płcią a głosowaniem w ostatnich wyborach na PO lub PIS.

|  |  |  |
| --- | --- | --- |
|   | kobieta | mężczyzna |
| PO | 250 | 260 |
| PIS | 380 | 150 |

Zadanie 2

Czy kolor oczu ma znaczenie na wybór partnera? Ten problem został zbadany a wyniki rozkładu cech (kolor oczu badanego i kolor oczu partnera) prezentuje poniższa tabela. Należy ustalić, czy istnieje związek między kolorem oczu osób w związkach oraz siłę tego związku, wykorzystując współczynniki oparte na chi-kwadrat.

|  |  |
| --- | --- |
|   | badany |
| zielone | niebieskie | brązowe |
| partner | zielone  | 15 | 12 | 22 |
| niebieskie | 12 | 28 | 10 |
| brązowe | 29 | 14 | 11 |

Zadanie 3

Proszę wskazać, jaką część pola pod krzywą normalną wyznaczają wartości Z rozkładu dystrybuanty rozkładu normalnego:

- Z > 1,25

- Z < -1,23

- Z < - 1,96 i Z> 1,96

- Z < -2,25 i Z > 1,90

Proszę wskazać, od jakich wartości Z rozkładu normalnego mieści się:

- 25% wszystkich wartości większych niż wartość Z

- 95% pola dla wartości większych niż wartość Z

- 5% skrajnych ( w lewej i prawej strony rozkładu)

Zadanie 4

Średnia rozkładu cechy, która ma rozkład normalny, wynosi 16, a odchylenie standardowe wynosi 4. Wskaż:

- jaki odsetek jednostek mieści się w przedziale między średnią a wartością odległą o 2 odchylenia standardowe od średniej,

-jaki odsetek jednostek mieści się w przedziale między wartościami 12 a 20

- jaki odsetek jednostek mieści się w przedziale od 9 do 21.

- jaką wartość minimalną przyjmuje cecha dla 10 % jednostek o najwyższych wartościach tej cechy

- jaką wartość maksymalną przyjmuje 12% jednostek i najmniejszych wartościach tej cechy

Zadanie 5

Z populacji wyłoniono próbę wielkości 64 jednostek. Średnia arytmetyczna wartość cechy wyniosła 110, zaś odchylenie standardowe 16. Należy wyznaczyć przedział ufności parametru populacji μ, przy poziomie istotności α≤0,05. Jaka będzie wielkość tego przedziału, gdy zmniejszymy poziom istotności do α≤0,01.

Zadanie 6

Przeprowadzono badania dwóch prób pochodzących z tej samej populacji, różniących się wielkością. Pierwsza próba liczyła sobie 100 a druga 400 jednostek. Uzyskano następujący rozkład proporcji badanej cechy w pierwszej próbie p=0,4, a w drugiej p=0,45. Jakie będą przedziały ufności dla uzyskanych frakcji, przy poziomie istotności α≤0,05.

Zadanie 7

Należy określić rodzaj testu (jedno czy dwustronny) oraz wartości krytyczne **z** dla określonych hipotez i ich poziomów istotności:

1. $H\_{0}: μ\_{1}=μ\_{2}; H\_{1}: μ\_{1}\ne μ\_{2}; poziom istotności α\leq 0,05 $
2. $H\_{0}: μ\_{1}>μ\_{2}; H\_{1}: μ\_{1}\leq μ\_{2}; poziom istotności α\leq 0,1 $
3. $H\_{0}: μ\_{1}\leq μ\_{2}; H\_{1}: μ\_{1}>μ\_{2}; poziom istotności α\leq 0,01 $
4. $H\_{0}: μ\_{1}-μ\_{2}=0; H\_{1}: μ\_{1}-μ\_{2}>0; poziom istotności α\leq 0,001 $
5. $H\_{0}: μ\_{1}-μ\_{2}>0; H\_{1}: μ\_{1}-μ\_{2}\leq 0; poziom istotności α\leq 0,02 $

Zadanie 8

Przeprowadzono pewne badanie nad ocenami uczniów. Średnia ocena wszystkich uczniów w szkole wyniosła 4,1, zaś wylosowanej grupy 3,90. Odchylenie standardowe średniej próby wyniosło 0,2. Widząc, że dla 100 osobowej próby różnica między średnimi jest istotna statystycznie przy poziomie istotności α ≤ 0, 05, należy ustalić:

1. czy jest też istotna na poziomie istotności α ≤ 0, 01.
2. Ile powinna wynosić minimalna różnica między średnimi, aby była istotna na poziomie α ≤ 0,001.
3. Jaka powinna być minimalna liczebność próby, aby różnica 0,1 była istotna statystycznie przy poziomie istotności α ≤ 0, 01 (odchylenie standardowe 0,2).

Zadanie 9

Badania dotyczące palenia wykazały, że nałóg ten dotyczy 72% studentów. Należy wykazać czy:

1. Problem palenia dotyczy większości studentów, wiedząc, że próba do badań wyniosła 150, przy poziomie istotności α ≤ 0, 05.
2. Ilość palących wśród studentów jest większa niż w reszcie społeczeństwa, wiedząc, że palenie dotyczy 60% członków społeczeństwa, przyjmując poziom istotności α ≤ 0, 01.

PODSUMOWANIE WYKŁADÓW

Test χ² chi-kwadrat i mierniki oparte na χ² chi-kwadrat

* Test **χ²** (chi – kwadrat), inaczej nazywany testem niezależności, bada odstępstwa rozkładu empirycznego zmiennych z rozkładem teoretycznym (oczekiwanym), zakładającym pełną niezależność zmiennych.
* Idea **χ²** opiera się na poszukiwaniu takiego rozkładu liczebności , który wskazywałby na niezależność rozkładów dwóch zmiennych względem siebie (rozkład oczekiwany).





* Rozkład empiryczny dwóch zmiennych to łączny rozkład współzmienności dwóch zmiennych, pochodzący z pomiarów
* Rozkład teoretyczny (oczekiwany) to rozkład warunkowy pochodzący z analizy rozkładów brzegowych zmiennych.



$$n\_{i.}- liczebność brzegowa i-tej kategorii pierwszej zmiennej \left(np.w kolumnie\right)$$

$$n\_{.j}-liczebność brzegowa j-tej kategorii drugiej zmiennej \left(np.w wierszu\right) $$

Przykład:

Oto mamy rozkład empiryczny, z rozkładem warunkowym a, b, c, i d:



Tworzymy dla tego rozkładu rozkład teoretyczny (oczekiwany)

  



Następnie porównujemy oba rozkłady, sumujemy ostatnią kolumnę i otrzymujemy χ² chi-kwadrat

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| a  | ta  | a – ta  | (a – ta)²  |  |
| b  | tb  | b – tb  | (b – tb)²  |  |
| c  | tc  | c – tc  | (c – tc)²  |  |
| d  | td  | d - td  | (d – td)²  |  |
|  |  |  |  |  |

* Związek między zmiennymi występuje wtedy, gdy wartość testu χ² (chi – kwadrat) jest większa od zera.
* Brak związku (pełna niezależność dwóch zmiennych) między zmiennymi wyznaczona jest wartością zero testu χ² (chi – kwadrat).
* Wadą testu χ² jest uzależnienie wysokości wyniku od wielkości badanej grupy (od liczebności rozkładów brzegowych i warunkowych). Im większa grupa, a tym samym większa liczebność rozkładów brzegowych i warunkowych, tym większa wartość testu χ² dla tego samego związku między zmiennymi.

Mierniki oparta na chi – kwadrat













Rozkład normalny

* Średnia, mediana i dominanta (modalna) leżą w tym samym punkcie.
* Krzywa jest doskonale symetryczna, gładka i opiera się na nieskończonej liczbie przypadków, więc może być doskonałym przybliżeniem rozkładu empirycznego.
* Charakteryzują ją tylko dwa mierniki: średnia i odchylenie standardowe.
* Punkt maksymalny Y osiągnięty jest gdy wartość xi równa się średniej.
* Im mniejsze odchylenie standardowe tym bardziej spiczasta jest krzywa.
* Pole pod krzywą jest równe 1. Punkt średniej (mediany) dzieli rozkład na dwie równe części.



Umożliwia znalezienie proporcji przypadków zawartych w określonym przedziale.

Niezależnie od konkretnych wartości średniej i odchylenia standardowego , powierzchnia pod krzywą (czyli proporcja) w przedziale od średniej do dowolnego punktu zależy tylko od odległości tego punktu od średniej, **jeśli tę odległość wyrazimy w jednostkach odchylenia standardowego**

Standaryzacja rozkładu i rozkład normalny standaryzowany

* Charakterystyki liczbowe każdej krzywej normalnej (rozkładu o normalnym lub zbliżonym do normalnego charakterze) można przekształcić tak, by przy pomocy **tych samych tablic** (tablic rozkładu normalnego)odnaleźć pole pod krzywą dla dowolnego przedziału.
* Przekształcenie to polega na zamianie rozkładu normalnego ze średnią i odchyleniem standardowym **s** , na rozkład również normalny o średniej równiej **0** i odchyleniu standardowym równym **1, co oznaczamy**
* Rozkład normalny o średniej równej 0 i odchyleniu standardowym równym 1 nazywany ***rozkładem standaryzowanym***, zaś wynik procedury ***standaryzacją***. W efekcie otrzymujemy nowy pomiar, nazywany ***pomiarem standaryzowanym***.
* Procedura polegająca na odjęciu od dowolnego pomiaru (lub dowolnej wartości) wartości średniej i podzieleniu tej różnicy przez odchylenie standardowe.



**standaryzacja**

Tablica standaryzowanego rozkładu normalnego

* Standaryzowany rozkład normalny jest stablicowany.
* Tablica standaryzowanego rozkładu normalnego opisuje gęstość prawdopodobieństwa w połówce rozkładu (od średniej). Jest również **dystrybuantą rozkładu normalnego f(x)**
* Symetryczny kształt rozkładu sprawia, że miary są identyczne po obu stronach średniej (powyżej i poniżej średniej)
* Wartości **Z** oznaczają punkt na osi X, oznaczający odległość od średniej w jednostkach odchylenia standardowego, równego 1 i wyznaczają granicę jakiegoś przedziału. Z stanowi również wartość krytyczną testu (o tym za dwa tygodnie).

Używając tablicy można określić, jak duża część wszystkich jednostek mieści się w wybranym polu.

Przykłady:

* Pytanie: jaki proporcja rozkładu mieści się w przedziale powyżej 1,20 wartości rozkładu normalnego (jaka proporcja rozkładu mieści się powyżej z = 1,20, czyli jakie jest p dla z ≥ 1,20)?



* Pytanie: ile wynosi pole dla przedziału mieszczącego się od średniej do z = 1,20?

* Pytanie: średnia wieku wynosi 45 lat, odchylenie standardowe 10 lat. Jak wielka będzie proporcja osób w wieku od 40 do 50 lat?

p = ?

Właściwości rozkładu normalnego

* Krzywa jest symetryczna
* Średnia, mediana i modalna leżą w jednym punkcie
* Najwyższa rzędna leży w punkcie średniej, czyli równej 0
* Krzywa jest asymptotyczna. Zbliża się do osi poziomej, lecz nigdy do niej nie dochodzi.
* Punkty zagięcia krzywej znajdują się w miejscach plus minus odchylenie standardowe powyżej i poniżej średniej.
* Między minus i plus jedno odchylenie standardowe od średniej znajduje się 68,26% wszystkich przypadków, plus minus 2 odchylenia 95% wszystkich przypadków, a plus minus 3 odchylenia 99,9% wszystkich przypadków.

**Rozkład t Studenta**

* Twórcą rozkładu *t* był William Gosset. Pisał on swoje prace pod pseudonimem „Student”, stąd nazwa rozkładu.
* Rozkład, a w zasadzie rozkłady *t*, stanowią całą rodzinę rozkładów teoretycznych. Dla każdej liczby stopni swobody istnieje inna wartość *t*.
* Rozkład *t* jest rozkładem symetrycznym, ze średnią równą 0.
* Odbiega on od normalności w przypadku małych wielkości prób.
* Zbiega się on do nieskończoności na obu krańcach, których grubość jest różna niż w przypadku rozkładu normalnego oraz zależna od liczby stopni swobody.
* Stosuje się go w przypadku małych prób (N ≤ 30). Dla wielkich prób, różnice między rozkładem *z* a rozkładem *t* są nieznaczne.
* Dla każdej próby otrzymujemy jedno oszacowanie odchylenia standardowego rozkładu (różne w każdej próbie) oraz błąd standardowy
* Średnią z próby można wyrazić stosując przekształcenie

 

* Przekształcenie to zawiera dwie wartości oszacowania: które są różne przy wielokrotnym pobieraniu prób.
* Wartości rozkładu *t* zależą od liczebności próby (od liczby stopni swobody). Są więc zmienne dla różnych wielkości prób.
* Tabele rozkładu *t* prezentują obszary pod krzywą przy określonej liczbie stopni swobody (df) i poziomie istotności.
* Stopnie swobody związane z odchyleniem standardowym wynoszą N-1, to przy innych statystykach mogą wynosić N-2, N-3, itd.
* Dlatego też tabele sporządzone są wg stopni liczby swobody, a nie zamiast N



**Estymacja i przedziały ufności**

* Oszacowanie lub estymacja to dział wnioskowania statystycznego, będący zbiorem procedur statystycznych, pozwalających określić wielkość nieznanego parametru lub postać rozkładu zmiennej losowej na podstawie badania próby losowej.

*Na podstawie badań (obserwacji) próby szacuje się (estymuje się) wiele wartości w populacji, np.:*

 *- wielkość poparcia dla partii X lub Y;*

 *- odsetek osób kupujących dany produkt;*

 *- liczbę osób dotkniętych ubóstwem.*

*Problem: dlaczego pojawiają się różne wyniki badań?*

*- W wyniku prowadzenia pomiarów na próbach zawsze istnieje pewna zmienność wyników, będąca efektem zmienności cechy w populacji. Zmienność wyników związana jest z błędem próby.*

* Ze względu na poszukiwaną cechę rozkładu metody estymacji można podzielić na:

 1. **estymację parametryczną** – szacuje się parametr populacji

 2. **estymację nieparametryczną** – szacuje się proporcje rozkładu cechy w populacji

Ze względu na sposób szacowania poszukiwanej cechy rozkładu, można wyróżnić:

 1. **estymację punktową** – szacowanie konkretnej wartości parametru lub proporcji rozkładu cechy w populacji.

 2. **estymację przedziałową** – szacowanie przedziału ufności w jakim mieści się parametr lub proporcja rozkładu cechy w populacji.

Estymacja przedziałowa – przedziały ufności dla średnich z dużych prób

* Estymacja przedziałowa polega na wyznaczaniu pewnego przedziału liczbowego, wewnątrz którego z określonym prawdopodobieństwem znajduje się estymowany parametr.
* Przedział ten nazywamy **przedziałem ufności**, a prawdopodobieństwo tego, że znajdzie się w nim estymowany parametr wyznacza **współczynnik ufności**.
* Rozkład ze średnich z próby zbliża sie coraz bardziej do normalnego wraz ze wzrostem liczebności próby.
* Dlatego też dla dużych prób stosuje sie rozkład normalny w celu oszacowania wyniku.
* Do obliczenia błędu standardowego stosuje sie odchylenie rozkładu z próby średnich, czyli:

 

Przedział ufności dla średnich z dużych prób.

* Stosowany gdy N>30







Przykład

Dla wylosowanych w sposób losowy 100 studentów średnia z ocen z semestru zimowego wyniosła 4,1, odchylenie standardowe zaś 1,2. Jaki będzie przedział ufności tej średniej, przyjmując, że chcemy określić jego wielkość na poziomie istotności 0,05.









* Gdy N <=30





Przykład

* Dla wylosowanych w sposób losowy 20 studentów średnia z ocen z semestru zimowego wyniosła 4,1, odchylenie standardowe zaś 1,2. Jaki będzie przedział ufności tej średniej, przyjmując, że chcemy określić jego wielkość na poziomie istotności 0,05.







Przedziały ufności proporcji dla dużych prób

* Metoda estymacji θ przez p





* Metoda „ostrożna”





Przykład

Przed kampanią wyborczą przeprowadzono sondaż, którego celem było oszacowanie odsetka osób popierających kandydata partii XYZ na prezydenta. Wybrano 1000 osobową próbę. Wyniki sondażu wskazywały na poparcie 55% wyborców. Ile wynosić będzie przedział ufności, przy poziomie istotności α ≤ 0,05?





 





Testy statystyczne

**Test *z*  i  *t* dla jednej próby**

* Badają , czy różnica między średnimi z próby a populacji jest wystarczająco duża, by wyeliminować wahania wynikające z losowego doboru próby.
* Wymaga podjęcia decyzji o charakterze testu (jedno czy dwustronny).
* Wymaga przyjęcia określonego poziomu istotności α (prawdopodobieństwa popełnienia błędu I rodzaju), wyznaczającego **wartość krytyczną** testu (np. α≤ 0,05, wartość krytyczna z = =+-1,96 przy teście dwustronnym lub 1,65 przy teście jednostronnym)
1. **Testy parametryczne - test *z* dla jednej próbyprzy znanym σ**
* Test *z* stosujemy gdy:

 1. próba jest duża (N>30)

 2. znane jest odchylenie standardowe populacji – **σ**

 3. populacja ma rozkład normalny











1. **Testy parametryczne - test *z* dla jednej próbyprzy nieznanym σ**
* Test *z* stosujemy gdy:

 1. próba jest duża (N>30)

 2. nieznane jest odchylenie standardowe populacji – **σ**

 3. rozkład cechy w populacji jest normalny





1. **Testy parametryczne - test *t* dla jednej próby przy znanym σ**
* Test *t* stosujemy gdy:

 1. próba jest mała (N ≤ 30)

 2. gdy znane jest odchylenie standardowe populacji σ

 3. rozkład cechy jest nieznany



1. **Testy parametryczne - test *t* dla jednej próbyprzy nieznanym σ**
* Test *t* stosujemy gdy:

 1. próba jest mała (N ≤ 30)

 2. gdy nieznane jest odchylenie standardowe populacji σ

 3. rozkład cechy jest nieznany



**Test *z* dla frakcji w populacji**

* **Test z dla frakcji stosujemy gdy:**
1. **próba jest duża, co oznacza, że musi zostać spełniony warunek:**

****

 **2. rozkład cechy może być nieznany w populacji**



